Grasping of Extrafoveal Targets: A Robotic Model

Wolfram Schenck[1]*, Heiko Hoffmann[2], Ralf Möller[1]

[1] Computer Engineering Group, Faculty of Technology, Bielefeld University,
Bielefeld, Germany

[2] University of Southern California, Los Angeles, USA

Email: wschenck@ti.uni-bielefeld.de

February 19, 2008

Abstract

We present a computational model of grasping of non-fixated (extrafoveal) target objects which is implemented on a robot setup, consisting of a robot arm with cameras and gripper. This model is based on the premotor theory of attention (Rizzolatti et al., 1994) which states that spatial attention is a consequence of the preparation of goal-directed, spatially coded movements (especially saccadic eye movements). In our model, we add the hypothesis that saccade planning is accompanied by the prediction of the retinal images after the saccade. The foveal region of these predicted images can be used to determine the orientation and shape of objects at the target location of the attention shift. This information is necessary for precise grasping. Our model consists of a saccade controller for target fixation, a visual forward model for the prediction of retinal images, and an arm controller which generates arm postures for grasping. We compare the precision of the robotic model in different task conditions, among them grasping (1) towards fixated target objects using the actual retinal images, (2) towards non-fixated target objects using visual prediction, and (3) towards non-fixated target objects without visual prediction. The first and second setting result in good grasping performance, while the third setting causes considerable errors of the gripper orientation, demonstrating that visual prediction might be an important component of eye-hand coordination. Finally, based on the present study we argue that the use of robots is a valuable research methodology within psychology.

KEYWORDS: Perceptual Motor Processes, Robotics, Eye Fixation,
Visual Perception, Grasping, Neural Networks

PsycINFO classification: 2330, 4140, 2323

* corresponding author

1 Introduction

In everyday life, reaching and grasping movements are mainly carried out under visual control. The most important information about the position and shape of target objects is obtained from accompanying eye movements and retinal activation. A considerable amount of research adresses the question how eye and arm movements are coordinated and which information is used at which stage of motor planning and execution (e.g., Bekkering and Sailer, 2002; Frens and Erkelens, 1991; Horstmann and Hoffmann, 2005; Mather and Fisk, 1985; Neggers and Bekkering, 2000; Prablanc et al., 1979). Experimental studies show that saccades for target fixation usually precede arm movements (Abrams et al., 1990; Neggers and Bekkering, 1999; Vercher et al., 1994). Even when the onset time of eye and arm movements is the same, eye movements are finished more rapidly, providing the eye orientation as input for the completion of the arm movement. Nevertheless, as everyday experience shows, it is possible for humans to reach for and grasp objects while the saccade to the target is suppressed. But this ability comes at a price: Several studies have shown that the accuracy of limb movements suffers in such a setting (Abrams et al., 1990; Mather and Fisk, 1985; Prablanc et al., 1979; Vercher et al., 1994). In conclusion, grasping of and reaching to both fixated and to non-fixated target objects is possible, although the former allows for more precise arm and hand movements.

In a previous study (Hoffmann et al., 2005), we explored the necessary coordinate transforms for both settings, and presented a computational model for grasping movements with a robot arm. In the present work, we concentrate on the sensorimotor processing for grasping of non-fixated target objects which are projected on the extrafoveal region of the retina. Our starting point is the premotor theory of attention (Rizzolatti et al., 1994) which states that spatial attention is a consequence of the preparation of goal-directed, spatially coded movements. Because the neural mechanisms for foveal vision in primates and humans appear to be highly developed, oculomotor maps coding space for eye movements play a central role in selective attention according to this theory. Experimental evidence for the close coupling of saccade preparation and visual attention has been found in several studies, for example in the work by Deubel and Schneider (1996) and Irwin and Gordon (1998). Moreover, there is a considerable overlap between frontoparietal control structures which are activated during covert shifts of visual attention and during saccade preparation, as functional imaging studies have shown (Beauchamp et al., 2001; Nobre et al., 2000; Perry and Zeki, 2000). Muggleton et al. (2003) were able to modulate attentionally guided performance in visual search tasks by transcranial magnetic stimulation over the frontal eye fields. In summary, there is strong experimental evidence for the link between visual attention and saccade preparation. The link between manual response preparation and shifts of spatial attention has been less convincing, but several studies (Baldauf et al., 2006; Deubel et al., 1998; Eimer et al., 2005, 2006; Schiegg et al., 2003) provide support for the claim that covert preparation of manual responses is linked to shifts of spatial attention as well.

We propose a computational model of grasping of extrafoveal targets which is implemented on a robot setup. This model is based on the premotor theory of attention and adds one specific hypothesis: Attention shifts caused by saccade planning imply a prediction of the retinal images after the saccade. The foveal region of these predicted retinal images is required to determine movement parameters for the manual interaction with objects at the target location of the attention shift.

Without visual prediction, grasping towards extrafoveal target objects is difficult because of the heavy distortions found in retinal images (with the term "retinal image" we refer here to the activation pattern of receptors in the retina). These distortions have at least three distinct sources. First, the retina has approximately the shape of a half-sphere (Atchinson and Smith, 2000). This brings about that the projection of one and the same object on the retina has a different shape, depending on its position relative to the optical axis of the eye. Second, the lens system of the eye suffers from chromatic and monochromatic aberrations in various forms, causing varying image quality (focus, shape of point spread function) throughout the retina (Atchinson and Smith, 2000). And third, the distribution of light receptors (rods and cones) on the retina is non-uniform. Cones are used for color vision under strong light, rods for monochromatic vision under low light levels. The cones are densely packed in the fovea (around 0 degrees eccentricity) with rapidly decreasing density towards the periphery of the eye. The density of rods decreases much slower towards the periphery but they are completely absent from the fovea (for illustration, see for example Fig. 6.12 in McIlwain, 1996, p. 93). Because of the non-uniform sensor distribution, the pattern of rod and cone activation caused by the projection of a certain object on the retina varies considerably with the retinal position of this projection. This non-uniformity is also found in the retinotopic maps in the visual cortex (Mallot, 1985).

Considering this background information, a grasping task in which the eyes do not fixate the target object poses a special difficulty because the object-related retinal activation differs depending on the object's position relative to the eyes. Any mechanism which extracts the necessary information for proper grasping (e.g., object orientation) from this activation pattern has be tuned to the exact position on the retina onto which the object is projected. This would cause considerable computational overhead and the need to learn complex input-output relationships between retinal activation and grasping parameters. To avoid this overhead, the system could predict what the foveal representation of the target object would look like after a successful saccade, and use a much simpler sensorimotor model which takes just the predicted foveal activation as input to generate the grasping parameters as output. This model could be the same sensorimotor model as the one which is applied to fixated target objects. Thus, visual prediction would allow us to apply one and the same model for the sensorimotor processing for both grasping of foveal and extrafoveal target objects. We hypothesize that such a prediction actually takes place when humans and other primates grasp towards extrafoveal targets. In accordance with the premotor theory of attention, the first step would be that spatial attention is shifted towards the object by preparing the motor command for making a saccade towards this object (but this saccade is never carried out). The second step is to use this saccadic motor command as input for a visual forward model to generate the predicted foveal representation of the target object. In the third step, the planned new eye position and the predicted foveal representation of the target object are provided as input for the sensorimotor model which associates this information with an appropriate motor command for grasping.

For the visual prediction, we use a forward model (FM). We do not make strong assumptions about the visual representation underlying the prediction. In our robot implementation, the prediction takes place on the level of artificial "retinal images" (see Sect. 4.1) which mimic roughly the cone distribution on the human retina (see the right of Fig. 5 for a resolution plot). The important analogy to biological retinal activation patterns is the fact that a target object appears in a different shape depending on its location in the retinal image.

The anticipation of sensory consequences in the nervous system of biological organisms is supposed to be involved in several sensorimotor processes: First, many motor actions rely on feedback control, but sensory feedback is generally too slow. Here, the output of FMs can replace sensory feedback (Miall et al., 1993). Second, FMs may be used in the planning process for complex motor actions (Tani, 1996). Third, FMs can help to distinguish self-induced sensory effects (which are predicted) from externally induced sensory effects (which stand out from the predicted background) (Blakemore et al., 2000). Fourth, it has been suggested that perception might rely on the anticipation of the consequences of motor actions which could be applied in the current situation; the anticipation would be accomplished by FMs (Hoffmann and Möller, 2004; Hoffmann, 2007; Möller, 1999). In our model of grasping of extrafoveal targets, the prediction of visual data serves as a replacement for sensory feedback and is used in the planning process for motor control (although it is only a one-step "planning" for the generation of a single movement). Therefore, the visual FM in this study contributes to the first and second of the above-listed four applications of FMs.

The learning of adaptive visual FMs is a rather new field. It is difficult because of the high dimensionality of visual data and because part of the output may be non-predictable. In fields like robotics or artificial life, studies using FMs for motor control focus mainly on navigation or obstacle avoidance tasks with mobile robots. The sensory input to the FMs are rather low-dimensional data from distance sensors or laser range finders (e.g.: Tani, 1996; Ziemke et al., 2005), optical flow fields (Gross et al., 1999), or preprocessed visual data with only a few remaining dimensions (Hoffmann and Möller, 2004). Only in a recent study by Hoffmann, a visual FM is implemented which predicts images with a size of 40×40 pixels. It is used for distance estimation and deadend recognition to demonstrate that perception by anticipation actually works (the fourth function of FMs mentioned above) (Hoffmann, 2007). The visual FM in the present study is an adaptation of the work by Schenck and Möller (2007) where we proposed a learning algorithm for visual FMs which overcomes the problems of high dimensionality and non-predictability.

In the following, the components of the overall model are explained in detail, and it is described by which learning procedures they are acquired. Afterwards, the final experiments and their results are described. The purpose of these experiments is to show that a robot implementation of our model is actually capable of grasping of extrafoveal targets. Moreover, we hypothesize that grasping of fixated targets results in slightly better performance than grasping of extrafoveal targets, and that grasping of extrafoveal targets without visual prediction results in low grasping success, illustrating the need for a visual FM. These hypotheses are tested in our experiments. In the discussion section, we will relate our robotic model to general methodological issues regarding the use of robots in psychological research. We will argue that our robotic approach demonstrates in multiple ways that robots are a valuable research instrument within psychology.

2 Overall System Architecture

2.1 Overview

Our model consists of three parts (see. Fig. 1). First, a saccade controller acquired through an iterative learning procedure (Schenck and Möller, 2006); second, a visual FM predicting retinal images (with decreasing image resolution towards the corners in analogy to the sensor distribution on the human retina; Schenck and Möller, 2007); and third, an arm controller for grasping movements which receives the output of the saccade controller and the orientation of the target object as inputs (similar to the controller presented by Hoffmann et al., 2005).

When the model is used for grasping of extrafoveal targets, a single trial starts with the presentation of the grasping target, a red wooden block, at a random location within the working space on a table surface. The cameras are in a random posture. The saccade controller generates the necessary motor command for proper fixation with the cameras, but this movement is not carried out, only the suggested motor command is recorded as input for the visual FM and the arm controller. Afterwards, the visual FM predicts the retinal images after the (hypothetical) saccade. From these predicted images, the orientation of the block is determined. Finally, the arm controller uses both the saccadic motor command and the block orientation in the predicted images as inputs to generate the grasping movement.

In the final experiments, the grasping performance of four different versions of the robotic model is compared: (1) for grasping towards target objects which are precisely fixated by a series of saccades, using the actual retinal images instead of the predicted ones; (2) for grasping towards target objects which are fixated just by one saccade, also using the actual retinal images instead of the predicted ones; (3) for grasping towards non-fixated target objects using visual prediction; (4) for grasping towards non-fixated target objects without visual prediction.

2.2 Setup

The experimental setup (see Fig. 9) consists of a robot arm with six rotational degrees of freedom and two-finger gripper (PowerCube, Amtec Robotics). Except in singularities, the inverse kinematics of the robot arm allows for eight different solutions for every gripper position and orientation within the working range. In practice, considering collisions of the robot arm with its environment or with itself, usually only two or four different solutions are applicable. A table in front of the robot arm is used to place target objects for grasping.

Moreover, a stereo camera head belongs to the setup. Each camera (Imaging Source DFK 50H13) provides an RGB color image with a resolution of 320×240 pixels. The horizontal and vertical angles of view are 61.9 and 48.5 degrees, respectively. Each camera is mounted on a pan-tilt unit (Directed Perception PTU $46 - 17.5$) with two degrees of freedom. In this study, the valid range for the pan angle is between -60.4 and 23.8 degrees, for the tilt angle between -42.9 and 21.4 degrees. In this range, the camera images always capture at least a small part of the white table shown in Fig. 9 below the cameras and in front of the robot arm.

For the training of the saccade controller (Sect. 3) and the visual FM (Sect. 4), not the real setup was used, but instead "virtual" camera movements were carried

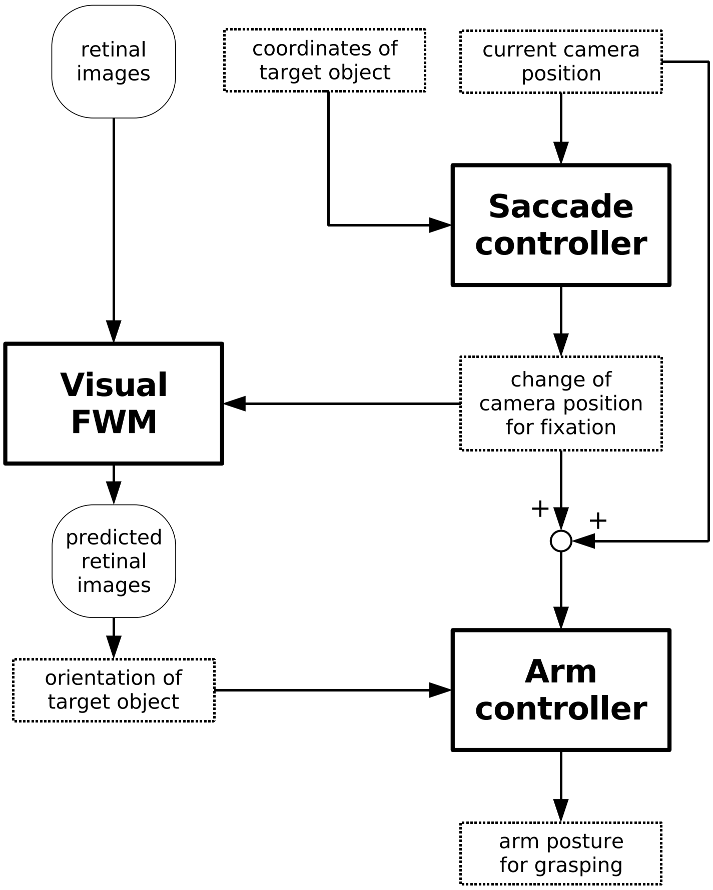

Figure 1: The overall system architecture. The main components of the model are a saccade controller, a visual FM, and an arm controller (for details see text) (adapted from Schenck and Möller, 2007, © Springer).

out using an image database. This image database contains the camera images for more than $120,000$ different camera positions within the above-mentioned pan/tilt range. Instead of using the cameras directly, we retrieved the images from the database. The recorded scene shows the white table with 56 colored wooden blocks on its surface — 14 blocks each from the colors red, green, blue, and yellow (see Fig. 3).

3 Saccade Controller

In primates and humans, saccades are fast eye movements for the fixation of interesting target regions in the visual surroundings. After a successful saccade, the target region is projected on the foveae of both eyes. The physiological and neural mechanisms of saccade control have gained a lot of interest from psychology, biology, and neurophysiology (for a comprehensive overview, see Leigh and Zee, 1999). In computer science, especially in the field called "active vision", research is centered to a large extent on the development of technical solutions for artificial saccades of robot camera heads (e.g., Klarquist and Bovik, 1998). Nevertheless, several studies propose models of saccade generation which are closely related to neurophysiological

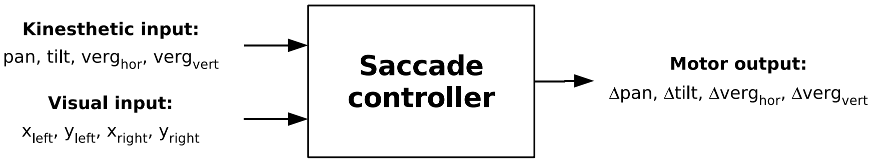
Kinesthetic input:
pan, tilt, verg_hor, verg_vert
Visual input:
x_left, y_left, x_right, y_right
Saccade controller
Motor output:
Δpan, Δtilt, Δverg_hor, Δverg_vert

Figure 2: Input and output of the saccade controller (adapted from Schenck and Möller, 2007, © Springer).

findings (Dean et al., 1994; Gancarz and Grossberg, 1999). In this area, which is related to both robotics and biology, methods of adaptive saccade learning are of special interest (Bruske et al., 1997; Pagel et al., 1998). In a previous study, we compared different learning strategies for saccade control (Schenck and Möller, 2006). In the present study, we use a similar saccade controller on the basis of a multi-layer perceptron (MLP) (Rumelhart et al., 1986) which is trained by a strategy called "continuous learning by averaging" (Schenck and Möller, 2006).

3.1 Controller input and output

The task of the saccade controller is to fixate target objects with both cameras so that the target object is projected onto the center of both camera images. In time step t, the saccade controller receives the current sensory state $\mathbf{s}_{\text{SAC}}^{(t)}$ as input, composed of a kinesthetic and a visual part (see Fig. 2). The kinesthetic input $\mathbf{s}_{\text{KIN}}^{(t)}$ consists of the current position of the cameras, defined by a conjoint pan-tilt direction $(\text{pan}, \text{tilt})$, and a horizontal and vertical vergence value $(\text{verg}_{\text{hor}}, \text{verg}_{\text{vert}})$. The visual part $\mathbf{s}_{\text{VIS}}^{(t)}$ represents the position of the target object in the left and right camera image relative to the image center: x_{left}, y_{left}, x_{right}, y_{right}. The motor output $\mathbf{m}_{\text{SAC}}^{(t)}$ of the saccade controller is defined as change of the motor position. It consists of four values: Δpan, Δtilt, $\Delta\text{verg}_{\text{hor}}$, and $\Delta\text{verg}_{\text{vert}}$. The new position of the cameras is computed as $\mathbf{s}_{\text{KIN}}^{(t+1)} = \mathbf{s}_{\text{KIN}}^{(t)} + \mathbf{m}_{\text{SAC}}^{(t)}$, and the cameras are moved accordingly. All sensory variables are scaled to the range $[-1; 1]$, the motor output variables to the range $[-2; +2]$.

3.2 Image processing

The image processing is restricted to a central area of 213×213 pixels in each camera image. For simplicity, in the following (throughout Sect. 3) the term "camera image" refers to this cropped region. The image processing extracts the position of the target object in the left and right camera image. Before any saccade, an appropriate target object has to be selected, after the saccade, it has to be re-identified to evaluate the success of the saccadic movement. In our setup, target objects are colored wooden blocks (see Fig. 3). During target identification and re-identification, their centers of mass are calculated via a color detection algorithm, finally yielding a list of red, green, blue, and yellow object coordinates.

Before any saccade, one of the detected objects is chosen from either the left or right camera image, depending on the current task. Afterwards, the matching target object in the other camera image is identified by searching for the image

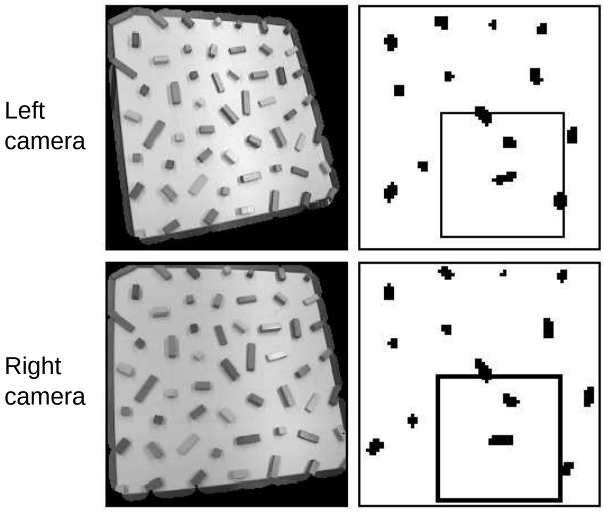

Figure 3: Camera images (left column; the surroundings of the table are already blanked) and salience images for red objects (right column) before a saccade. The selected target is marked with a rectangle in the salience images. It has been identified first in the right camera (bold rectangle); by a correlation approach, the corresponding image region in the left camera has been found.

region with the highest local pixel intensity correlation. In this way, the target object coordinates $\mathbf{s}_{\mathrm{VIS}}^{(t)} = \left(x_{\mathrm{left}}^{(t)}, y_{\mathrm{left}}^{(t)}, x_{\mathrm{right}}^{(t)}, y_{\mathrm{right}}^{(t)}\right)$ are determined. After the saccade, the target object is re-identified in both camera images by the same correlation approach, providing the coordinates $\mathbf{s}_{\mathrm{VIS}}^{(t+1)} = \left(x_{\mathrm{left}}^{(t+t)}, y_{\mathrm{left}}^{(t+t)}, x_{\mathrm{right}}^{(t+t)}, y_{\mathrm{right}}^{(t+t)}\right)$.

3.3 Implementation

The saccade controller is implemented by an MLP. It has 8 inputs and four linear output units (see also Fig. 2). The single hidden layer has four units with hyperbolic tangent as activation function; in addition, the inputs are also directly connected to the output layer ("shortcut connections"). In the beginning, the network weights are initialized to random values, resulting in erratic output. The network is trained by providing proper learning examples for weight adjustment as outlined in the following section.

3.4 Learning by averaging

Like most motor learning tasks, saccade learning suffers from the problem of the "missing teacher signal". Whenever an incorrect motor command is carried out, the resulting error is only measurable in the sensory domain. The motor error and therefore the correct motor output remains unknown. In the literature, several learning strategies are proposed to overcome this problem (e.g., Kuperstein, 1988; Kawato, 1990; Jordan and Rumelhart, 1992). We suggested a new algorithm called "continous learning by averaging" (CLbA) for the saccade learning task (Schenck and Möller, 2004, 2006). Its basic idea is to search at random in the neighborhood of the network output in motor space for saccades which are slightly better than

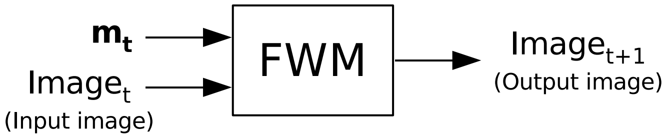

Figure 4: Visual forward model (FM) (adapted from Schenck and Möller, 2007, © Springer).

the saccade produced by the controller network, and which bring the target object closer to the center in both camera images. These improved saccades are used as learning example for network adaptation. In the process of learning, over- and undershoot saccades cancel each other out, resulting in more precise motor output of the network. This "canceling out" only works because the MLP as function approximator adapts to the average of the over- and undershoot saccades. Since saccade learning is not a central aspect of the overall model for extrafoveal grasping, we would like to refer the reader to a previous publication (Schenck and Möller, 2006) for a more detailed description of the CLbA algorithm.

To quantify saccade precision in the following, we define a measure called "radial target distance" as $r = r(\mathbf{s}_{\mathrm{VIS}}) = \frac{1}{2\sqrt{2}} \left(\sqrt{x^2_{\mathrm{left}} + y^2_{\mathrm{left}}} + \sqrt{x^2_{\mathrm{right}} + y^2_{\mathrm{right}}} \right)$, with $r = 0.0$ indicating a perfect saccade after which the center of mass of the target object is projected exactly on the center of both camera images, and with $r = 1.0$ being the worst value (as long as the target is not completely lost which is even worse). The saccade controller network of this study was trained over 450 learning trials, the average radial target distance over 50 test saccades amounted finally to $r < 0.018$. In the course of these 450 learning trials, 4435 saccades were carried out.

4 Visual Forward Model

The task of a visual FM is to predict future visual sensory states. In the framework of our robot setup, this means to predict what the camera images will look like after a movement of the camera head. The input of a visual FM is the current image at time step t and the motor command $\mathbf{m}^{(t)}_{\mathrm{FM}}$, the output is a prediction of the resulting image in time $t+1$ (see Fig. 4). Learning of this input-output relationship is difficult because of the high dimensionality of the image data, and because of the fact that part of the future image may not be predictable at all.

In a previous study (Schenck and Möller, 2007), we suggested an algorithm for the learning of visual FMs which is based on the idea of learning the mapping between corresponding pixel positions in the images of time step t and $t+1$ instead of directly predicting pixel intensities. Moreover, this algorithm is successful in identifying non-predictable regions in the future image. In the following, this algorithm is described. The description is an updated and abbreviated version of the presentation in Schenck and Möller (2007).[1]

[1] The copyright of the original publication is held by Springer. The permission for the publication of this modified version was kindly granted by Springer.

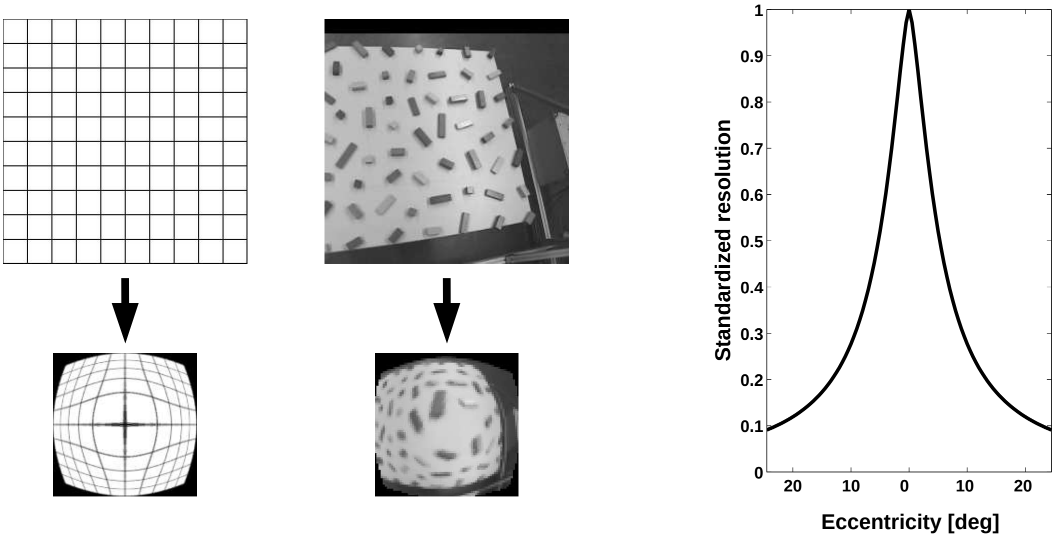

Figure 5: Left: Retinal mapping for an image depicting a regular grid. Center: Retinal mapping for a camera image. Right: Resolution of the artificial retinal images depending on the distance from the image center. Resolution values are standardized to a maximum of 1.0.

4.1 Retinal images

Instead of directly working with the camera images, we first apply a retinal mapping to create the artificial retinal images which are discussed in the introduction. This mapping maintains full image resolution in the image center, and decreases this resolution progressively towards the corners. Fig. 5 shows the effect of this mapping on an image depicting a regular grid and on a camera image (in the following, throughout Sect. 4, the term camera image refers to a 240 × 240 pixel region in the center of the original camera image).

The mapping between camera and retinal images is specified in polar coordinates. The origins of the coordinate systems are located at the image centers. They are scaled in a way that in both images the maximum radius (along the horizontal/vertical direction) amounts to 1.0. r_R is the radius of a point in the retinal image, r_C is the radius of the corresponding point in the camera image, the angle of the polar representation is kept constant. r_C is computed by $r_C = \lambda r_R^\gamma + (1-\lambda) r_R$, $\gamma > 1$, $0 \leq \lambda \leq 1$. Here we use $\gamma = 2.5$ and $\lambda = 0.8$. The resolution of the final retinal image is 69 × 69 pixels.

4.2 Structure of the visual FM

The same visual FM works for both the left or the right camera of our setup. As input, it receives $\mathbf{I}^{(t)}$, the retinal image of time step t (also called input image in the following), and the motor command $\mathbf{m}_{\mathrm{FM}}^{(t)} = (\Delta\mathrm{pan}, \Delta\mathrm{tilt})$ for the repositioning of a single camera.[2] The output of the FM is $\widehat{\mathbf{I}}^{(t+1)}$, the predicted retinal image of time

[2] Because the pan and tilt axes cross in close vicinity to the nodal point of the camera-lens system, the current camera position is not needed as input for the FM; for the same reason, depth

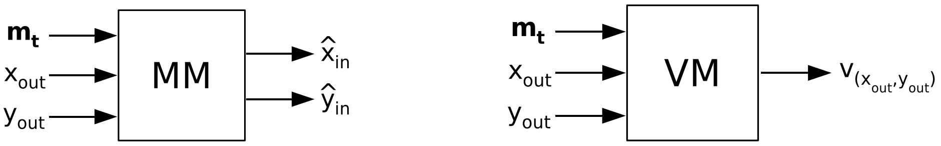

Figure 6: Left: Mapping model (MM). Right: Validator model (VM) (for details see text) (adapted from Schenck and Möller, 2007, © Springer).

step $t+1$ (also called output image in the following). While $\mathbf{I}^{(t)}$ is an unmodified retinal image, $\widehat{\mathbf{I}}^{(t+1)}$ is a center region with a size of 53×53 pixels. This is necessary to clip the black corners of the retinal image without valid information (see Fig. 5) which are just a technical artifact.

The visual FM consists of a so-called "mapping model" (MM) and a "validator model" (VM). Both are used to generate the output image. The mapping model (MM) is depicted in Fig. 6: As input, it receives the motor command $\mathbf{m}_{\mathrm{FM}}^{(t)}$ and the location of a single pixel $(x_{\mathrm{Out}}, y_{\mathrm{Out}})$ of the output image; as output it estimates the previous location $(\widehat{x}_{\mathrm{In}}, \widehat{y}_{\mathrm{In}})$ of the corresponding pixel (or region) in the input image. The overall output image is constructed by iterating through all of its pixels and computing each pixel intensity as $\widehat{\mathrm{I}}^{(t+1)}_{(x_{\mathrm{Out}},y_{\mathrm{Out}})} = \mathrm{I}^{(t)}_{(\widehat{x}_{\mathrm{In}},\widehat{y}_{\mathrm{In}})}$ (using bilinear interpolation). Moreover, the validator model (VM) generates a signal $v_{(x_{\mathrm{Out}},y_{\mathrm{Out}})}$ indicating whether it is possible at all for the MM to generate a valid output for the current input. It predicts which pixels of the output image are at a position that does not correspond to any pixel of the input image. This is necessary because even for small camera movements parts of the output image are not present in the input image. In this way, the overall FM (Fig. 4) is implemented by the combined application of a mapping and a validator model.

4.3 Learning of the MM and VM

The basic idea of the learning algorithm for the MM can be outlined as follows for a specific $\mathbf{m}_{\mathrm{FM}}^{(t)}$ and $(x_{\mathrm{Out}}, y_{\mathrm{Out}})$: During learning, the motor command is carried out in different environmental settings. Each time, both the actual input and output image are known afterwards, thus the intensity $\mathrm{I}^{(t+1)}_{(x_{\mathrm{Out}},y_{\mathrm{Out}})}$ is known as well. It is possible to determine which of the pixels of the input image show a similar intensity. These pixels are candidates for the original position $(x_{\mathrm{In}}, y_{\mathrm{In}})$ of the pixel $(x_{\mathrm{Out}}, y_{\mathrm{Out}})$ before the movement. Over many trials, the pixel in the input image which matches most often is the most likely candidate for $(x_{\mathrm{In}}, y_{\mathrm{In}})$ and therefore chosen as MM output $(\widehat{x}_{\mathrm{In}}, \widehat{y}_{\mathrm{In}})$. When none of the pixels matches often enough, the MM output is marked as non-valid (output of the VM).

4.3.1 Grid of cumulator units

The input space of the MM and VM consists of four dimensions: Δpan, Δtilt, x_{Out}, and y_{Out}. A four-dimensional grid $\mathbf{P}$ of points $\mathbf{p}_{ijkl} = (\Delta\mathrm{pan}^{(i)}, \Delta\mathrm{tilt}^{(j)}, x^{(k)}_{\mathrm{Out}}, y^{(l)}_{\mathrm{Out}})$ is embedded in this space, with $i, j = 1, .., 11$ and $k, l = 1, .., 13$. $\Delta\mathrm{pan}^{(i)}$ and $\Delta\mathrm{tilt}^{(j)}$

information is irrelevant for the prediction task.

cover the range from -29 to $+29$ degrees with constant step size, while $x_{\mathrm{Out}}^{(k)}$ and $y_{\mathrm{Out}}^{(l)}$ form an equally spaced rectangular grid covering the whole output image.

To each point $\mathbf{p}_{ijkl}$, a so-called "cumulator unit" C_{ijkl} is attached. Such a unit is basically a single-band image with the same size as the input image. Thus, the input image and the cumulator units have the same number of pixels in the horizontal and vertical direction. Each "pixel" of a cumulator unit can hold any non-negative integer value. They are used to accumulate and store the number of matches between input and output image at their specific position during the learning process.

4.3.2 Learning process

The goal of the learning process is to accumulate activations in the cumulator units. At the beginning, all pixels of these units are set to zero. In each learning trial, the pan-tilt unit is first moved into a random $(\mathrm{pan}, \mathrm{tilt})$ position. The input image for the FM is recorded and processed. Afterwards, the algorithm iterates through all points of the grid $\mathbf{P}$, the corresponding motor command is executed (relative to the initial random position), and the output image is generated from the camera image after the movement. For each point $\mathbf{p}_{ijkl}$, the intensity of the output image at the coordinates $(x_{\mathrm{Out}}^{(k)}, y_{\mathrm{Out}}^{(l)})$ is compared to the intensities of all pixels $(x_{\mathrm{In}}, y_{\mathrm{In}})$ in the current input image. Whenever the intensity difference (computed as Euclidean distance in RGB color space) is below 3.5% of the overall intensity range, the value of pixel $(x_{\mathrm{In}}, y_{\mathrm{In}})$ in cumulator unit C_{ijkl} is increased by one.

In the present study, 100 trials were carried out, each with $11 \times 11 \times 13 \times 13 = 20449$ iteration steps (size of the grid $\mathbf{P}$). In each trial, the initial camera position was varied, resulting in different input images.

4.3.3 Generating the MM and VM

After the cumulator units have been acquired in the learning process, raw versions of the MM and VM can be created whose output is defined at the grid positions $\mathbf{p}_{ijkl}$ in input space. The output $(\widehat{x}_{\mathrm{In}}, \widehat{y}_{\mathrm{In}})$ of the MM at grid point $\mathbf{p}_{ijkl}$ are the coordinates of the pixel with maximum intensity in the cumulator unit C_{ijkl}. The output $v_{(x_{\mathrm{Out}}, y_{\mathrm{Out}})}$ of the VM at point $\mathbf{p}_{ijkl}$ is set to 1 (indicating a valid output of the MM at this point) whenever the maximum pixel intensity in unit C_{ijkl} is above a certain threshold. Otherwise, $v_{(x_{\mathrm{Out}}, y_{\mathrm{Out}})}$ is set to 0. The threshold is computed as the product of the maximum pixel intensity of all cumulator units and a factor equal to 0.41. This proved to be the value resulting in the most correct separation.

For illustration, Fig. 7 shows the cumulator units for the center pixel of the output image for four different $(\Delta\mathrm{pan}, \Delta\mathrm{tilt})$ positions. The larger the camera movement, the more the intensity maximum in the respective cumulator unit vanishes until no prediction is possible any longer (movement 4).

The output of the raw versions of the MM and the VM is only defined at the grid points $\mathbf{p}_{ijkl}$. To get the output in-between, function interpolation is necessary. For this purpose, the raw versions of the MM and the VM were replaced by radial basis function networks (RBFN) (Moody and Darken, 1989) in the final step of the learning algorithm. These networks have the same input/output structure as the MM and the VM, respectively (see Fig. 6). The training data for both networks was

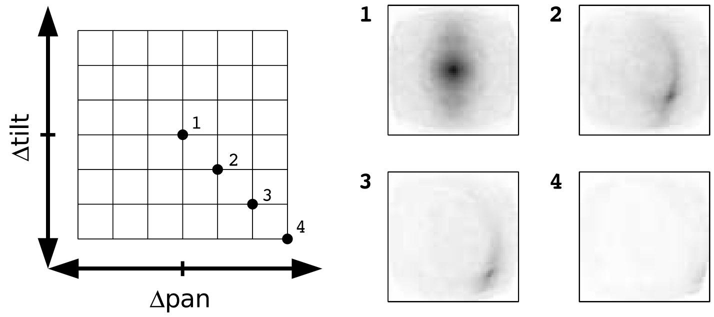

Figure 7: Cumulator units for the center pixel for four different (Δpan, Δtilt) positions. All depicted cumulator units were normalized by the same scaling factor so that a pixel value of zero corresponds to white and the overall maximum pixel value to black (adapted and updated from Schenck and Möller, 2007, © Springer).

generated from the output of the raw versions of the MM and the VM at the grid points $\mathbf{p}_{ijkl}$ (overall, there are $11 \times 11 \times 13 \times 13 = 20449$ grid points). For the MM network, training data was restricted to the 10523 grid points with valid output (as indicated by the raw version of the VM). For more details on the network training, see Schenck and Möller (2007).

4.4 Results

The MM and VM network were used to implement the overall visual FM for predicting the output image. Especially, non-predictable regions of the output image were marked by the VM network. The prediction works rather precise as shown exemplary in Fig. 8. The actual and the predicted output image are compared for four different motor commands (Δpan, Δtilt) (camera movements to the lower right of increasing length as in Fig. 7). Moreover, the region of each output image which is marked as non-predictable by the VM is shown in black color in the third row of images. The input image (the same for all four movements) is displayed as well. Movement 1 is a zero movement. The actual and the predicted output image are very similar and show the center region from the input image. Movements 2 and 3 are of increasing size. The non-predictable regions mask parts of the output images which have no correspondence in the input image. The center of the predicted images is slightly blurred and distorted because the mapping generated by the MM network has to enlarge a region of a few pixels in the input image to a much larger area (especially for movement 3). Movement 4 is so large that the center of the output image is non-predictable. Nevertheless, the small upper left part of the output image which is predicted corresponds closely to the actual output.

5 Arm Controller

The purpose of the arm controller is to generate the motor command for the final grasping movement. As input, it receives the orientation of the target object, a red wooden block on the table surface (see Fig. 9), and the position of the cameras $\mathbf{s}_{\text{KIN}}^{(t+1)}$

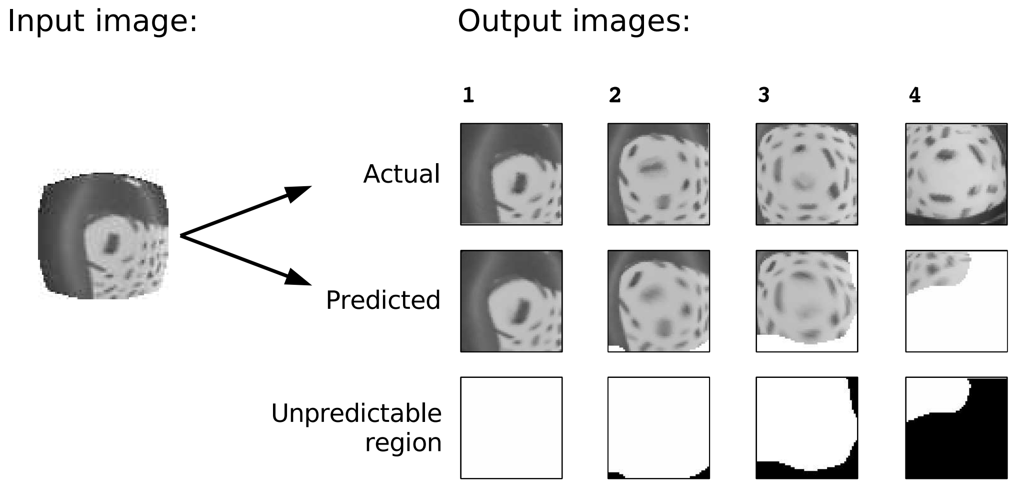

Figure 8: Comparison of actual and predicted output images at four different $(\Delta\text{pan}, \Delta\text{tilt})$ positions (the same as in Fig. 7; for details see text) (adapted and updated from Schenck and Möller, 2007, © Springer).

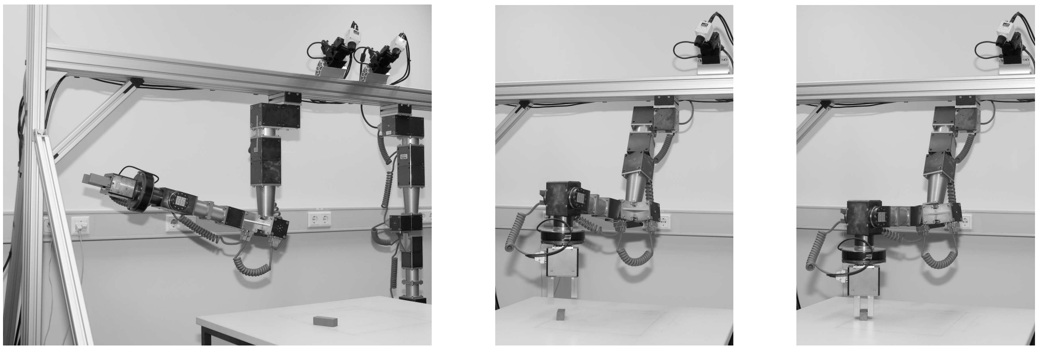

Figure 9: Resting, pre-grasping and grasping posture (from left to right).

after a successful fixation movement towards the target object. The camera position implicitly encodes the position of the red block. As output, the arm controller produces two sets of joint angles, for the pre-grasping and the grasping posture. The pre-grasping posture serves as via point for the robot arm when it moves from its resting position to the final grasping position. This is necessary to avoid collisions with the environment and with the block before it is grasped. Figure 9 shows the resting, the pre-grasping, and the grasping posture for a single grasping trial. In a perfect grasping movement, the approach direction of the gripper is perpendicular to the table surface. Because of the geometry of the robot arm, this movement is only possible over a restricted area of the table. Here, we use a rectangular region of 380 × 250 mm for the placement of the target objects.

5.1 Data preprocessing

The arm controller is implemented by a neural network algorithm called "NG-PCA" (Hoffmann and Möller, 2003; Möller and Hoffmann, 2004) (details follow in

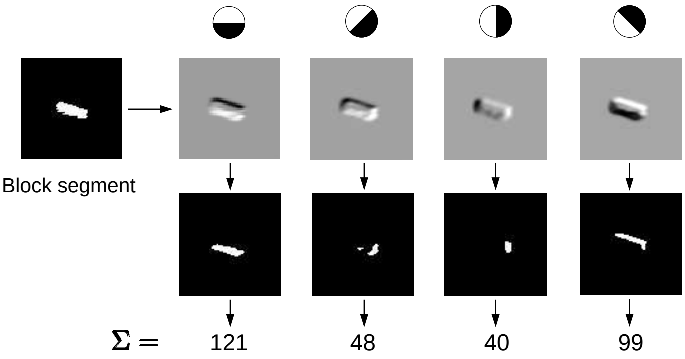

Figure 10: Image processing to encode the block's orientation. On the very left is the block segment. In each column, the preprocessing steps for one compass filter (top) are shown, the edge-image, the threshold image, and the sum of white pixels in the threshold image (adapted from Hoffmann et al., 2005, © Springer).

Sect. 5.3). To achieve maximum learning success with this algorithm, certain preprocessing of the controller input and output is necessary. We use similar methods as in the study by Hoffmann et al. (2005).

The visual input of the arm controller is the orientation of the red block. To determine this orientation, the left retinal image *after* the successful fixation saccade towards the target object is used as default (although this is later varied in the experiments). A color filter is used to generate an image where the block appears as single white segment on a completely black background. In the next step, four compass filters enhance the edges in four different directions (0°, 45°, 90°, and 135°) (see Fig. 10). After thresholding, the remaining pixels in each image are counted to give a value for the distribution of edges in a given direction. The resulting four values are normalized so that there sum yields 1.0. These normalized values form a "compass filter histogram" which uniquely encodes the orientation of the block independent of its size.

All postural variables (camera position, arm joint angles) are encoded by tuning curves: A variable x is represented by the values of four Gaussian functions $f_i(x) = \exp(-(x - c_i)^2/(2\sigma^2))$ whose centers c_i are uniformly distributed within the maximal range of the variable. σ equals the distance between two neighboring centers. Overall, there are 20 input values for the arm controller (4 compass filter values and 4 × 4 values for the camera position), and 48 output values (2 arm postures with 6 × 4 values each).

5.2 Collection of training data

Each single training example for the arm controller network is collected in the following way (similar to the procedure suggested by Hoffmann et al., 2005): First, a random block position and orientation on the table surface are generated. By the analytical solution of the inverse kinematics of the robot arm, a corresponding pre-grasping and grasping posture are determined. If the inverse kinematics yields several applicable solutions, one of them is chosen at random. The robot arm is moved to this position with the red block held by the gripper, and releases the red

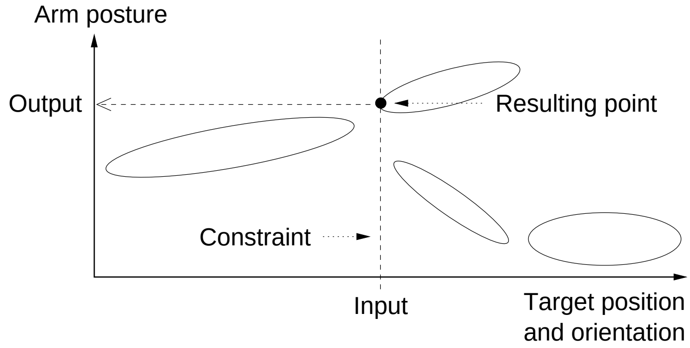

Figure 11: Recall by constrained subspace (for details see text) (adapted from Hoffmann et al., 2005, © Springer).

block on arrival. Afterwards, the arm returns to its resting position. The saccade controller from Sect. 3 is used to fixate the red block; to enhance precision, a second corrective saccade is carried out if the radial target distance r is larger than 0.015 after the first saccade. Afterwards, the image of the left camera is recorded and mapped to the retinal image. From the information which is gathered in this sequence, a full learning example with input (camera position and block orientation) and output (pre-grasping and grasping posture) is constructed. This way of collecting learning examples is a technical solution and not intended for biological modeling.

Altogether, 3213 learning examples were collected. All input and output dimensions were normalized to mean 0.0 and variance 1.0. The postural dimensions were normalized *before* the encoding to tuning curve values.

5.3 Neural network algorithm

Since one of the possible solutions of the inverse kinematics is chosen at random, the training data represents a one-to-many mapping. For this reason, function approximator networks like MLPs are not suitable for the implementation of the controller. This is a general problem found in many motor control tasks. Möller and Hoffmann (2004) suggested so-called “abstract recurrent neural networks” as solution. These networks consist of a set of hyperellipsoids in the sensorimotor space which comprises both the input and output dimensions. The hyperellipsoids describe the training data manifold with considerably fewer parameters than the original training data contains.

To determine the center and shape of the hyperellipsoids, different algorithms are suggested in the literature (for an overview, see Hoffmann, 2004). We applied the NGPCA method by Möller and Hoffmann (2004). To recall data in such a network after training, certain dimensions are defined as input dimensions. The input data defines a constrained subspace. The hyperellipsoid with the smallest normalized Mahalanobis distance to the constrained subspace is chosen. Then, on this constraint, the point that is closest to the chosen ellipsoid gives the desired output values. Figure 11 illustrates this process.

For the arm controller, we used a network consisting of 100 hyperellipsoids with 4 dimensions. From the collected 3213 learning examples, 2900 randomly selected examples were used for training (and 313 for the test set). $100,000$ learning iterations were applied. After training, the arm controller showed the following performance figures on the test set: The average horizontal distance between the gripper tip and the block on the table surface amounted to 8.7 mm. The average difference between the orientation of the gripper and the block orientation amounted to 3.4 degrees (for a more detailed specification of these performance indicators see Sect. 7).

6 Experiments

In the final experiments, we tested if the overall robotic model shows the hypothesized performance in different task conditions. These conditions vary with respect to the experimental sequence within a single trial. In general, an experimental trials starts by generating a block position and orientation at random. The robot arm is used to place the red block on the table surface at exactly this position and in this orientation. Afterwards, the robot arm returns to its resting posture.

Task condition WW represents grasping of properly fixated targets.[3] In this task condition, the saccade controller is used for a very precise fixation movement towards the red block. A maximum of five correction saccades is allowed to reduce the radial target distance to less than $r = 0.015$. The left camera image after the last saccade is used to compute the retinal image and the compass filter values as input for the arm controller. Moreover, the final camera position is used as input $\mathbf{s}_{\text{KIN}}^{(t+1)}$ for the arm controller.

In task condition OW, grasping of extrafoveal target objects is carried out.[4] After the red block has been placed on the table surface, the cameras are moved to a random position where (1) the block is visible in both camera images as input to the saccade controller, and (2) the full shape of the block is visible in the left retinal image as input to the visual FM. Afterwards, the saccade controller is used to generate one saccade towards the red block, but this saccade is never carried out, only $\mathbf{s}_{\text{KIN}}^{(t+1)}$ is computed. The visual FM predicts the hypothetical retinal image after the saccade, and from this image the compass filter values are determined as input for the arm controller.

Task condition WW1 serves as comparison: It is equal to WW, but only *one* saccade is carried out, accepting a less than optimal target fixation.[5] This allows a more fair comparison with OW, where only a single hypothetical saccade is determined, but no correction saccade. This reduces the quality of the camera position input $\mathbf{s}_{\text{KIN}}^{(t+1)}$ of the arm controller, and the retinal images after the saccade may differ slightly from the images which were used during training (where a corrective saccade was carried out if necessary).

Task condition OO is used as a control experiment to demonstrate that the extraction from orientation information from the retinal images is not trivial and depends actually on the position of the block in the retinal image.[6] Here, the sequence is similar to condition OW, but the visual FM is not used. Instead, the retinal image before the hypothetical fixation saccade is used to compute the compass filter values

[3] WW: With saccade execution, With proper retinal image

[4] OW: withOut saccade execution, With proper retinal image

[5] WW1: With saccade execution, With proper retinal image, only 1 saccade

[6] OO: withOut saccade execution, withOut proper retinal image

as input for the arm controller.

In all conditions, the grasping movement which is finally generated as output from the arm controller was carried out at the end of every sequence, and the grasping success was evaluated. Overall, 100 trials were performed in every task condition.

7 Results

To evaluate the grasping success, the most important indicator is the percentage of successful grasping trials; a trial was rated as success if the gripper of the robot arm was able to grasp the red block firmly and to lift it. This measure tolerates small position and orientation errors since the distance between the gripper jaws amounted to 60 mm when it approached the red block. The red block itself has a horizontal cross section of 74 × 23 mm. Moreover, the following indicators are used to evaluate the grasping precision:

- Block position error: The Euclidean distance between the center of mass of the red block and the center of the open gripper, projected onto the table surface.
- Vertical position error: The difference between the ideal height of the gripper tip above the table surface (held constant for all learning examples) and the actual height.
- Block orientation error: The difference between the block's orientation on the table surface and the orientation of the perpendicular to the line that connects both gripper jaws, projected onto the table surface.
- Vertical orientation error: In all learning trials, the approach direction of the gripper is exactly perpendicular to the table surface. The vertical orientation error is the difference between this ideal approach orientation and the actual approach orientation.

7.1 Grasping success

First of all, the percentage of successful grasping trials clearly shows that our model of extrafoveal grasping actually works as expected (see Table 1): In condition `OW`, the success rate amounts to 85%. As expected, the success rate in grasping towards precisely fixated target objects (condition `WW`) is higher, amounting to 97%. Condition `WW1` (target objects only fixated with one saccade) has a success rate of 89%, which shows that the performance difference between conditions `WW` and `OW` is largely attributable to the less precise camera position information if only one saccade is scheduled. The baseline condition `OO` has only a success rate of 40%, clearly indicating that the prediction of the retinal image is not just a trivial add-on, but instead crucial for successful grasping towards extrafoveal targets. The pairwise differences are statistically significant on the $p < 0.01$ level, expect of the differences `WW` vs. `WW1` (only $p < 0.05$) and `WW1` vs `OW` (not significant) (four cell Chi-square test).

WW	WW1	OW	OO
97 %	89 %	85 %	40 %

Table 1: Success rate (over 100 grasping trials in each experimental condition).

Error	WW	WW1	OW	OO
Block position [mm]	8.9 (7.2)	12.9 (12.2)	15.4 (9.7)	23.6 (27.1)
Vertical position [mm]	2.1 (2.9)	2.4 (4.8)	4.0 (6.4)	6.1 (12.1)
Block orientation [deg]	4.2 (9.4)	6.4 (13.2)	12.5 (13.1)	37.9 (23.0)
Vertical orientation [deg]	0.5 (0.4)	0.7 (1.2)	1.1 (1.5)	2.0 (5.1)

Table 2: Average position and orientation errors for the different task conditions. Standard deviations are given in brackets.

7.2 Grasping precision

The indicators for grasping precision show results which are consistent with the grasping success rate. Table 2 presents the average position and orientation errors for the grasping posture of the robot arm. Regarding the mean value of all trials, condition WW shows always the best precision, closely followed by WW1, and with a certain distance by OW. OO is always the worst performer, especially with regard to the block orientation error. The last result illustrates very clearly the impact of the missing visual FM.

The vertical position and orientation errors are much smaller than the block position and orientation errors in all conditions. This is no surprise since the distance between gripper tip and table surface and the default approach direction of the gripper are constant for all learning examples and thus rather easy to learn by the adaptive arm controller network.

We restricted the statistical tests to a pairwise comparison of the mean values. For each error measure, we computed pairwise t-tests (two-sided) for independent samples between the four task conditions. We corrected the degrees of freedom to compensate for the unequal estimated population variances (Bortz, 1993). All 24 tests yielded significant results at least on the $p < 0.05$ level with the following exceptions: Block position error: WW1 vs. OW; vertical position error: WW vs. WW1, WW1 vs. OW, OW vs. OO; block orientation error: WW vs. WW1; vertical orientation error: WW vs. WW1, WW1 vs. OW, WW1 vs. OO, OW vs. OO.

7.3 Saccadic precision

In condition WW, the average radial target distance after the final saccade amounts to $r = 0.012$, while in condition WW1 with only one saccade it amounts to $r = 0.018$. This shows that the lower saccadic precision in condition WW1 is actually the most plausible source of the larger mean block position error and smaller success rate found in WW1 compared to WW. Furthermore, in condition WW1 the correlation between saccade length (for the left camera) and radial target distance amounts to $r_{Corr} =$

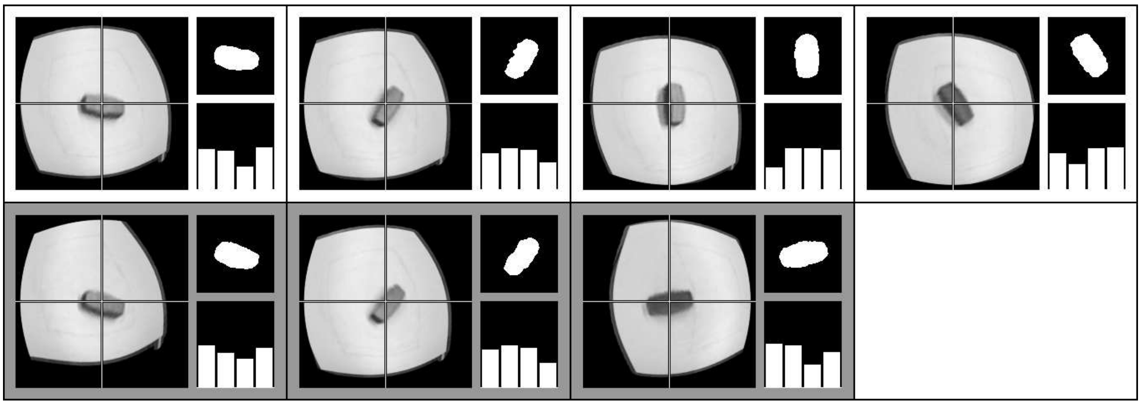

Figure 12: This figure shows the retinal images which are used to extract the block orientation for the WW task condition. A cross marks the center of each retinal image. In addition, the block segment and the corresponding compass filter histogram are shown for each retinal image. The first row with white background depicts successful trials, the second row with gray background failed trials. There are only three failed trials in the WW condition.

0.16. Inspired by this finding, we investigated if there is a direct relationship between saccade length and grasping precision. For conditions WW1, OW, and OO, we computed the correlations between saccade length and the different position and orientation errors. The largest absolute correlation coefficient is found between saccade length and block orientation error in the OO condition ($r_{Corr} = 0.11$). However, even this correlation value is not significantly different from zero ($t = 1.1; df = 98$), thus we cannot draw any firm conclusions from these correlation coefficients.

7.4 Visualization

Figures 12 to 15 show some exemplary retinal images which are used for the computation of the compass filter values for the different task conditions. In the first row of each figure (white background), the images from four successful trials are shown, in the second row (grey background) from three or four failed trials. In addition to the retinal image, the segment which is identified as red block is shown together with the compass filter histogram which is computed from this segment. Moreover, Fig. 14 for task condition OW shows on top of this information the retinal image which is used as input for the visual FM. The predicted retinal image is depicted underneath together with the identified block segment and compass filter histogram.

Figure 12 shows the examples for the WW condition. The block is well centered in the retinal image, indicating good saccadic accuracy. For the first row with successful trials, four different block orientations have been chosen. The compass filter histogram reflects the block orientation by the position of the minimum within the histogram.

Figure 13 is dedicated to the WW1 condition. The retinal images reveal that even in the successful trials the red block is not as well centered as in the WW condition, resulting in slightly banana-shaped segments (top left trial).

Figure 14 displays exemplary trials of the OW condition. The difference between

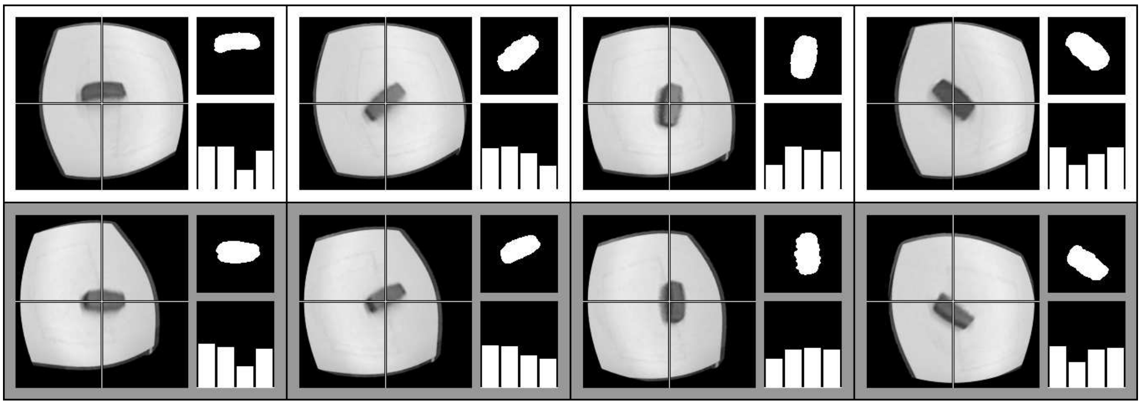

Figure 13: Retinal images of the `WW1` task condition (for further explanation see caption of Fig. 12).

the depicted retinal images which are input and output of the visual FM illustrate its performance. Shape and orientation of the block in the retinal images differs considerably between input and output. But it becomes also clear that the prediction is sometimes not accurate enough and generates atypical block shapes like in the second trial in the first row (nevertheless successful) or in the second trial in the second row (failed). The predicted block shape in the latter is nearly quadratic resulting in a compass filter histogram with equal values. Histograms like this do not occur in the learning examples for the arm controller network, thus this input is outside the learned data manifold and causes erratic extrapolation and failure.

Finally, Fig. 15 shows exemplary trials for the `OO` condition. Here, no saccade and no prediction takes place, and the retinal images which are recorded at the initial camera position are used to generate the compass filter histograms from the block segment. The shapes of the block segment differ strongly from the ideal shapes shown in Fig. 12 in the context of the `WW` condition. Accordingly, the compass filter histograms are sometimes ill-shaped (especially showing too large differences between minima and maxima). Moreover, the correction of the orientation of the block segment, which is accomplished by the visual FM in condition `OW`, is missing. These findings correlate well with the low grasping success rate in the `OO` condition.

8 Discussion

8.1 Evaluation of the results

The most important goal of our robotics study was to show that our model of grasping towards extrafoveal targets actually works as expected. All important components of the model — a saccade controller, a visual FM, and an arm controller — were implemented for the use with a a robotic real-world setup for this grasping task. The results show that the suggested architecture is actually capable of fulfilling this task. This supports the claim that spatial attention shift are accompanied by the preparation of eye movements (as the premotor theory of attention states; Rizzolatti et al., 1994), and corroborates our specific hypothesis that a visual FM predicts how the target object would appear in the fovea, and that this prediction is used to

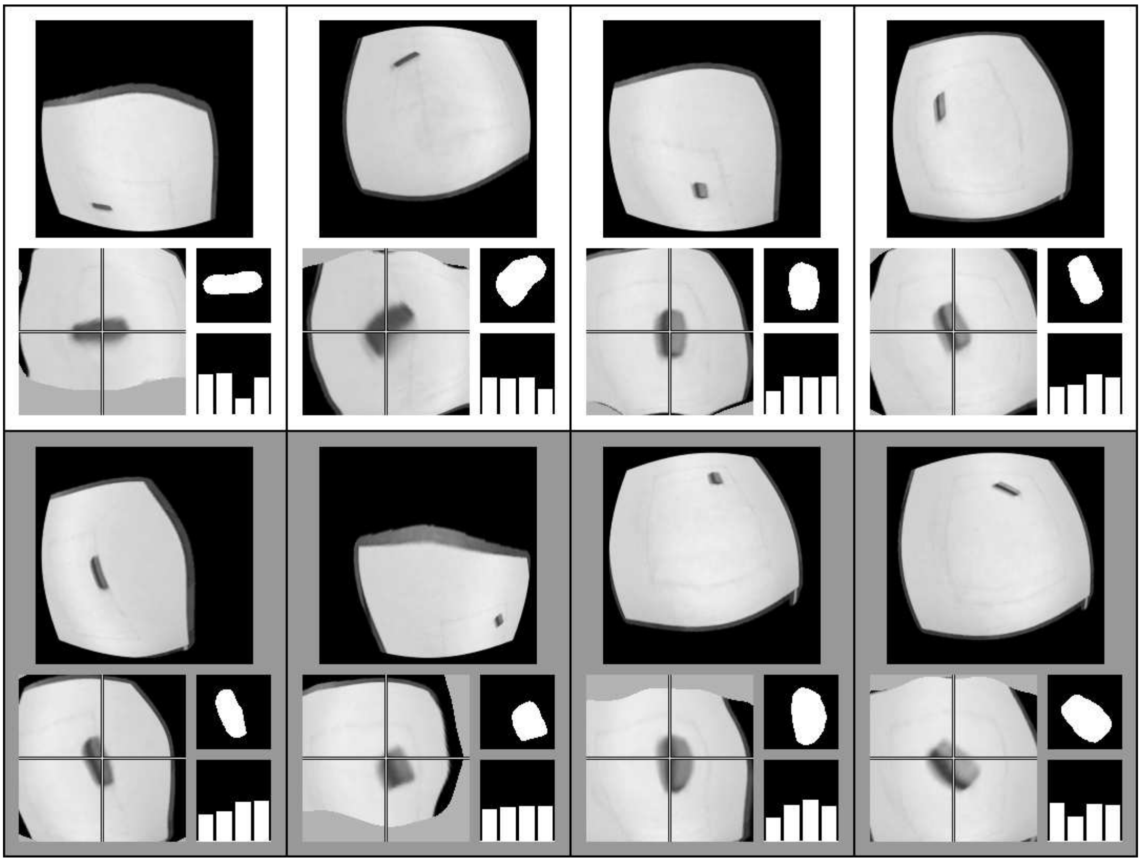

Figure 14: Retinal images of the `OW` task condition (for further explanation see caption of Fig. 12; in addition, the retinal image which is used as input for the visual FM is shown for each trial on top of the other images).

extract precise information about orientation (or in a more general sense, about shape).

Furthermore, we expected that grasping towards precisely fixated target objects results in a better grasping performance than grasping towards extrafoveal target objects (as suggested by the literature on eye-arm coordination; Abrams et al., 1990; Vercher et al., 1994). This expectation was confirmed in a comparison of the respective task conditions with regard to the overall grasping success and with regard to the grasping precision. In an additional task condition, the influence of saccadic accuracy on grasping success was explored. The fixation movement was restricted to one saccade regardless of the resulting accuracy (like in the extrafoveal condition). This revealed that the superior performance of grasping towards precisely fixated targets compared to extrafoveal targets can be attributed to a large extent to the inferior saccadic accuracy. Only part of the performance difference has to be explained by insufficient visual prediction.

The baseline condition without saccade execution and without visual prediction was used to show that the retinal mapping causes non-trivial changes of object shape and orientation depending on the position in the retinal image. As expected, directly extracting orientation information from the non-predicted retinal images and feeding it to the arm controller resulted in low grasping success. Furthermore,

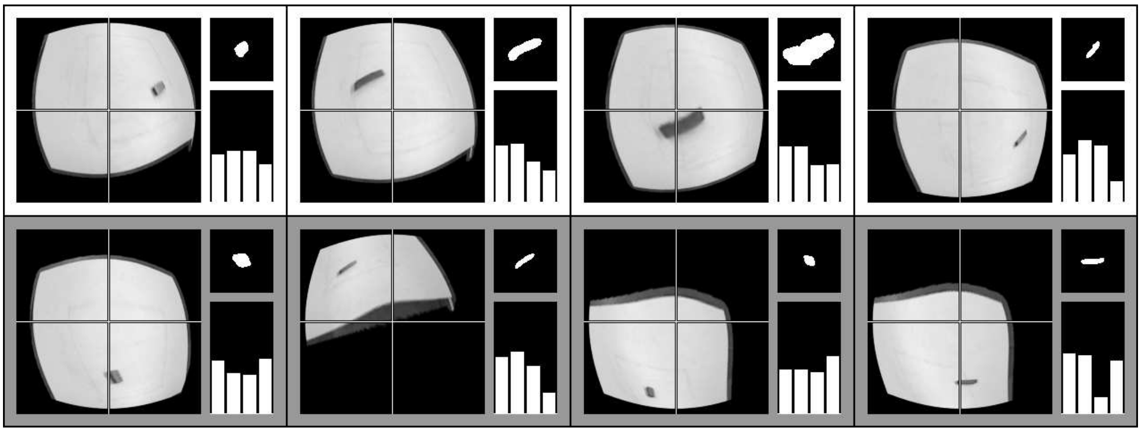

Figure 15: Retinal images of the OO task condition (for further explanation see caption of Fig. 12).

especially the block orientation error in this task condition was very high compared to the other conditions.

8.2 Premotor theory revisited

The premotor theory of attention (Rizzolatti et al., 1994) is supported by many studies which demonstrate a close coupling between attentional "selection for perception" and attentional "selection for action", especially in the context of eye movements (Baldauf and Deubel, 2008; Deubel and Schneider, 1996; Godijn and Theeuwes, 2003; Rolfs et al., 2005). However, these findings are generally compatible with two different theoretical accounts. One of them is the premotor theory which states that attention shifts are accomplished by the preparation of (non-executed) motor commands (thus, motor commands first), the other approach states that both selection mechanisms are driven in parallel by the attentional allocation (thus, motor commands last). For example, the "visual attention model" by Schneider (1995) follows the second route. At the moment, it is not possible to draw clear conclusions in favor of one of these theories based on experimental findings. In a neurophysiological study by Juan et al. (2004) on macaque monkeys, the experimenters were able to dissociate spatial attention from saccade preparation which contradicts the premotor theory. On the other hand, Craighero et al. (2004) showed in a behavioral study with human subjects that it is not possible to shift the attention to peripheral targets on the retina which are outside the reach of saccadic eye movements. This finding strongly supports the premotor theory. Furthermore, Eimer and colleagues (Eimer et al., 2005, 2006) interprete their experimental results on shifts of spatial attention and manual response preparation in favor of the premotor theory.

Our model of extrafoveal grasping relies on the preparation of a saccadic motor command whenever the attention is shifted towards a peripheral target object. This motor command is required as input for the visual FM and the arm controller. In the visual attention model, the generation of the saccadic motor command is not mandatory, while it is in the premotor theory. For this reason, our approach to extrafoveal grasping fits better with the premotor theory.

8.3 Model compared to neurophysiological findings

Overall, our combined model operates at a rather abstract level. The input and output of the different components are intended to model the information flow on a level of abstraction which is feasible to be implemented in a robot setup. Nevertheless, in the following we would like to point out some correspondences between our model and neurophysiological findings.

An important input to the arm controller is the camera position which corresponds to the gaze direction of the eyes in biological systems. In the literature on eye-arm coordination, this information is often referred to as "extraretinal eye position information" (EEPI) (e.g., in Bockisch and Miller, 1999). EEPI plays in important role in the localization of targets which appear as visual stimuli on the retina (e.g., Battaglia-Mayer et al., 2003; Bock, 1986). EEPI allows the transformation from eye-centered to head-centered coordinates; head position information is needed to further localize the target in body-centered coordinates. As simplification, the latter is omitted in the model, assuming a fixed head position. Both kinesthetic eye position information and the efference copy of the eye positioning commands are plausible sources for EEPI (Bridgeman, 1995; Weir, 2006). Here it is supposed that both sources are compatible with each other and can be added up to compute the hypothetical eye position information after a non-executed saccade. Otherwise, an additional internal model would be necessary for this transformation. Despite this simplification, it is important for the plausibility of our model that the "hypothetical EEPI" is actually available to the nervous system. Actually, experimental studies show that EEPI starts to change before saccade onset (Bockisch and Miller, 1999; Dassonville et al., 1992; Matin et al., 1970); thus, the nervous system has the means to update EEPI before the eye movement takes place and therefore before the kinesthetic eye position information can change.

Many studies on eye-arm coordination emphasize the necessary coordinate transforms for the localization of target objects in body- or arm-centered space after they have been perceived as visual input on the retina. For example, Snyder (2000) presents the finding that in some cortical regions the locally represented retinal position is modulated by the population code of the gaze direction (this modulation has been termed "gain fields"). Coordinate transforms like this allow to compute the position of a target object in head- or body-centered coordinates, but they do not explain how the overall object shape is transformed. Here, the visual FM of our model offers a plausible mechanism.

The output of the arm controller in our model are final arm postures, not trajectories. This is consistent with the result of Graziano et al. (2002) in a study on monkeys, in which the stimulation of certain motor cortex neurons lead to hand locations independent of the initial arm posture. Thus, this level of encoding is biologically plausible.

8.4 Remarks on the components of the model

The components of the overall model — the saccade controller, the visual FM, and the arm controller — are not pre-wired. Instead, they are acquired by different learning strategies. These strategies have been presented and discussed in previous publications (Hoffmann et al., 2005; Schenck and Möller, 2006, 2007). For this reason, we restrict the discussion here to a few remarks.

For the saccade controller, we use learning by averaging although many authors favor feedback error learning for saccade control in humans and primates (Dean et al., 1994; Gancarz and Grossberg, 1999). Nevertheless, learning by averaging offers a new way of adaptive motor control which is genuinely simple and low-level in its algorithmic structure and therefore a viable candidate for biological or psychological modeling. Future research has to show for which motor tasks it is suited as plausible mechanism.

The visual FM is learned by matching regions in the retinal images before and after the saccade. Over many learning trials, correspondences emerge during the matching process. From these correspondences, a mapping between pixel positions in the retinal images before and after the saccade is constructed, and non-predictable image regions are detected by the lack of any clear correspondence. This learning process is restricted to low-level visual processing and therefore a plausible candidate for biological modeling. Studies on "predictive remapping" support the claim that visual prediction takes place in the brain. In these studies, neurons which shift their visual receptive fields in anticipation of an upcoming saccade were discovered in various brain areas (Duhamel et al., 1992; Umeno and Goldberg, 1997; Walker et al., 1995).

8.5 Alternative solutions

The retinal mapping in our model is used to change the pattern of sensor activation depending on the position of a visual stimulus in a retinal image (in analogy to biological systems as pointed out in the introduction). These activation differences are still relevant on the next processing level where compass filters are used to detect edge orientations (like the simple cells in the visual cortex; Hubel and Wiesel, 1962). In consequence, the arm controller which has been adapted to orientation information gained from the foveal region of the retina cannot work successfully with extrafoveal target objects. As solution, we offered visual prediction of the foveal region by a FM. This has the advantage that the system can solve the task by a single arm controller which is also used for grasping of precisely fixated targets.

Within our framework, there exist two additional approaches which offer alternative solutions. Both cause considerably more overhead on part of the arm controller. First, instead of visual prediction, one could use a multitude of arm controllers, each adapted to a certain region of the retina. This could work in theory, but would require a lot of storage effort for the parameters of the large number of arm controllers. Moreover, each arm controller would need its own learning examples in which the target object is exactly depicted on the retinal region for which the controller is responsible. This means a lot of additional effort in the collection of learning examples and in the adaptation process. As second alternative solution, one could use one arm controller which also takes the retinal position of the target as input. This seems to be a straightforward solution, but it suffers from the disadvantage that the manifold of training data becomes considerably more complex. The relation between compass filter values and joint angles would be mediated by retinal position in a non-linear fashion, which basically adds two non-redundant dimensions to the data manifold. Thus, the interpolation task of any neural network algorithm (whether biological or artificial) would become more difficult and would require more complex network structures. Moreover, the required amount of training data for such a network would be much larger: Enough learning examples for all retinal regions would be needed to allow for adequate interpolation.

8.6 What does our robotic model offer psychology?

The use of robots within psychological research serves multiple purposes. On the one hand, there is the field of human-machine interaction in which robots serve as the interaction partner of human subjects (e.g., Tanaka et al., 2007). On the other hand, robots are applied to replace formal models by material models (Rosenblueth and Wiener, 1945) or machine instantiations (Tamburrini and Datteri, 2005). For the latter reason, robots are used in the present study.

Many psychological models attempt to explain the mechanisms underlying specific classes of behavior. Instead of just focusing on the sensory input and the corresponding overt behavior as in behaviorism (e.g., Watson, 1994), these models provide an insight into the "black box" which transforms system inputs into system outputs. In the following, we will call these models "functional models" because they offer an account to the inner functioning of the psychological system. A good functional model allows the generation of new hypotheses of how specific changes in the input would result in specific changes in the output.

Functional models can differ with regard to how precisely they are specified. In the one extreme, a formal model is just described by its main structural components and a sketch of the information flow between them as in Fig. 1. In the other extreme, the model is defined such precisely that it can be implemented on a robot setup, including surplus details which are only implementation-specific and not relevant for the model per se. The step from the formal model to the material model or machine instantiation is expensive with regard to development time and equipment costs, thus it has to be justified by the expected scientific gain.

In the view of Rosenblueth and Wiener (1945), there are two main reasons for a material model: It "may enable the carrying out of experiments under more favorable conditions than would be available in the original system" (p. 317), and it has to suggest experiments "whose results could not have been easily anticipated on the basis of the formal model alone" (p. 318). In a more recent review on biorobotics research, Webb (2000) summarizes the studies in this field under the four headlines "testing hypotheses", "characterizing the problem and understanding the environment", "integrating data and enforcing completeness", and "producing new hypotheses". This agrees well with the arguments of Rosenblueth and Wiener (1945), but emphasizes as additional benefit of robotic models that they enforce a deep understanding of the complete sensorimotor loop. In contrast, in a formal "box model" like in Fig. 1 logical flaws may stay unnoticed, and the relationship between the abstract inputs and outputs of the model and the real-world environment remains unclear (see also Datteri and Tamburrini, 2007; Webb, 2001).

However, one cannot expect that robot experiments will replace behavioral and neurophysiological studies. Final evidence in favor of a formal model has still to be obtained from the real psychological (or biological) system. Robotic models are an additional research tool which supplements conventional theoretical and experimental work. Robot implementations can be used to reject a formal model, both during the development of the robot instantiation (because logical flaws of the formal model become obvious) and during the final robot experiments (because it is not possible to generate the predicted effects). Whenever it is difficult to design a good behavioral or neurophysiological study with human subjects, the use of robots can be a valuable approach to further shape the formal model, to generate additional hypotheses, and — in case — to falsify the formal model in an early stage of development. In the following, we will examine which of these benefits apply to our robotic model of extrafoveal grasping and its components, in this way demonstrating how the use of robots can be advantageous for psychology on a theoretical and experimental level.

First of all, in our attempt to implement the complete sensorimotor loop for extrafoveal grasping, we became aware that the premotor theory of attention (Rizzolatti et al., 1994) might not be sufficient to constitute all the necessary information processing steps. Usually, it is assumed that attention coincides with enhanced sensory processing in the attended region (Kanwisher and Wojciulik, 2000), but the "enhanced processing" itself remains rather unspecified. We filled this gap with the visual prediction hypothesis to enable extrafoveal grasping on our robot setup. In this respect, the creation of the robotic model helped to enhance the underlying formal model. Furthermore, without the robot implementation it was not possible to determine definitely if visual prediction is really necessary in this task domain. For this reason, our final experiment incorporated the important comparison to the baseline condition without fixation and without prediction. Through this comparison, clear evidence was obtained that the orientation errors which are related to the distortions in the retinal images have the expected negative impact — or more generally spoken that the uneven sensor distribution on the retina does not allow for a mechanism which extracts target information without reference to the specific target location on the retina. In this way, the use of a robot setup was the right method to "characterize the problem and understand the environment" (Webb, 2000). In this context, it is important to note that we used retinal images instead of straight camera images. Although the artifical retinal images are still an abstraction from the real retinal sensor activation, their usage is not a mere irrelevant implementation detail. Instead, the retinal images are very important for a close structural similarity between the biological and the robotic system in an area which is highly relevant for the validity of the overall robot experiment ("structural accuracy" in the terms of Webb, 2001). In contrast, the learning algorithm for the arm controller is for example merely a technical detail.

Furthermore, our robotic approach allowed for experimental variations which would be difficult to achieve with human subjects. In the robot experiment, it was easy to switch between the task conditions with and without visual prediction, while this would be rather difficult in a behavioral or neurophysiological study. However, one has to admit that the results of the robot study are only a "proof of principle". Further work with human subjects is needed to collect supporting evidence from the real psychological system. The goal of such a behavioral experiment would be to decide if visual prediction is actually required for extrafoveal grasping. It might be possible to design such an experiment by enforcing "new" sensorimotor relationships (e.g., with prism goggles), and by suppressing prediction learning somehow in one task condition without touching the ability to learn motor controllers for grasping (although this dissociation seems very difficult to achieve). In this way, our robot implementation serves as an in advance check of the visual prediction hypothesis and as starting point for the design of corresponding experimental studies with humans subjects.

Similar arguments hold for the learning of visual prediction per se. Here we suggested an algorithm based on accumulating low-level visual feature matches. With the current research methods in neurophysiology and psychology, one cannot expect to collect hard evidence for such a mechanism through experiments with human subjects or other primates. Nevertheless, the use of a robot implementation enabled us to test the feasibility of this learning mechanism in the real world. The methodological value of the robot experiment lies in its ability to falsify the formal model of the learning mechanism in this early stage of development. As result, the proposed learning algorithm was *not* falsified.

In summary, our overall robotic model and its single components allow for experiments which are difficult if not impossible to carry out with human subjects at the moment, they generate results which are not possible to obtain just by the underlying formal model (and which help in characterizing the problem and in understanding the environment), and the successful robot implementation corroborates our theoretical approaches. In this way, our study demonstrates in multiple ways that the use of robots is a valuable research methodology within psychology.

8.7 Final conclusion

Our overall model offers a novel functional framework for grasping of extrafoveal targets based on the premotor theory of attention which has gained a lot of experimental support in the past (e.g., Craighero et al., 2004; Eimer et al., 2005, 2006). It identifies visual prediction as an important putative component of eye-hand coordination in this task domain. Moreover, its applicability to a robotic real-world setup was successfully demonstrated, including novel ways to learn saccade controllers and visual FMs for eye movements.

Acknowledgements

We thank Olaf Gerstung and Josch Pauling for their help in supervising the robot arm setup during the collection of the learning examples for the arm controller.

References

Abrams, R. A., Meyer, D. E., and Kornblum, S. (1990). Eye hand coordination — oculomotor control in rapid aimed limb movements. *Journal of Experimental Psychology-Human Perception and Performance*, 16(2):248–267.

Atchinson, D. A. and Smith, G. (2000). *Optics of the human eye*. Butterworth-Heinemann, Oxford, UK.

Baldauf, D. and Deubel, H. (2008). Properties of attentional selection during the preparation of sequential saccades. *Experimental Brain Research*, 184(3):411–425.

Baldauf, D., Wolf, M., and Deubel, H. (2006). Deployment of visual attention before sequences of goal-directed hand movements. *Vision Research*, 46(26):4355–4374.

Battaglia-Mayer, A., Caminiti, R., Lacquaniti, F., and Zago, M. (2003). Multiple levels of representation of reaching in the parieto-frontal network. *Cerebral Cortex*, 13(10):1009–1022.

Beauchamp, M. S., Petit, L., Ellmore, T. M., Ingeholm, J., and Haxby, J. V. (2001). A parametric fMRI study of overt and covert shifts of visuospatial attention. *Neuroimage*, 14(2):310–321.

Bekkering, H. and Sailer, U. (2002). Commentary: Coordination of eye and hand in time and space. *Progress in Brain Research*, 140:365–373.

Blakemore, S. J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *NeuroReport*, 11(11):R11–R16.

Bock, O. (1986). Contribution of retinal versus extraretinal signals towards visual localization in goal-directed movements. *Experimental Brain Research*, 64(3):476–482.

Bockisch, C. J. and Miller, J. M. (1999). Different motor systems use similar damped extraretinal eye position information. *Vision Research*, 39(5):1025–1038.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler*. Springer-Verlag, Berlin, Heidelberg, New York, fourth edition.

Bridgeman, B. (1995). A review of the role of efference copy in sensory and oculomotor control systems. *Annals of Biomedical Engineering*, 23(4):409–422.

Bruske, J., Hansen, M., Riehn, L., and Sommer, G. (1997). Biologically inspired calibration-free adaptive saccade control of a binocular camera-head. *Biological Cybernetics*, 77(6):433–446.

Craighero, L., Nascimben, M., and Fadiga, L. (2004). Eye position affects orienting of visuospatial attention. *Current Biology*, 14(4):331–333.

Dassonville, P., Schlag, J., and Schlagrey, M. (1992). Oculomotor localization relies on a damped representation of saccadic eye displacement in human and nonhuman-primates. *Visual Neuroscience*, 9(3-4):261–269.

Datteri, E. and Tamburrini, G. (2007). Biorobotic experiments for the discovery of biological mechanisms. *Philosophy of Science*, 74(3):409–430.

Dean, P., Mayhew, J. E. W., and Langdon, P. (1994). Learning and maintaining saccadic accuracy: A model of brainstem-cerebellar interactions. *Journal of Cognitive Neuroscience*, 6(2):117–138.

Deubel, H. and Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36(12):1827–1837.

Deubel, H., Schneider, W. X., and Paprotta, I. (1998). Selective dorsal and ventral processing: Evidence for a common attentional mechanism in reaching and perception. *Visual Cognition*, 5(1-2):81–107.

Duhamel, J. R., Colby, C. L., and Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye-movements. *Science*, 255(5040):90–92.

Eimer, M., Forster, B., Van Velzen, J., and Prabhu, G. (2005). Covert manual response preparation triggers attentional shifts: ERP evidence for the premotor theory of attention. *Neuropsychologia*, 43(6):957–966.

Eimer, M., van Velzen, J., Gherri, E., and Press, C. (2006). Manual response preparation and saccade programming are linked to attention shifts: ERP evidence for covert attentional orienting and spatially specific modulations of visual processing. *Brain Research*, 1105(1):7–19.

Frens, M. A. and Erkelens, C. J. (1991). Coordination of hand movements and saccades — evidence for a common and a separate pathway. *Experimental Brain Research*, 85(3):682–690.

Gancarz, G. and Grossberg, S. (1999). A neural model of saccadic eye movement control explains task-specific adaptation. *Vision Research*, 39(18):3123–3143.

Godijn, R. and Theeuwes, J. (2003). Parallel allocation of attention prior to the execution of saccade sequences. *Journal of Experimental Psychology-Human Perception and Performance*, 29(5):882–896.

Graziano, M. S., Taylor, C. S., and Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. *Neuron*, 34(5):841–851.

Gross, H.-M., Heinze, A., Seiler, T., and Stephan, V. (1999). Generative character of perception: A neural architecture for sensorimotor anticipation. *Neural Networks*, 12(7-8):1101–1129.

Hoffmann, H. (2004). *Unsupervised Learning of Visuomotor Associations.* MPI Series in Biological Cybernetics. Logos Verlag, Berlin.

Hoffmann, H. (2007). Perception through visuomotor anticipation in a mobile robot. *Neural Networks*, 20(1):22–33.

Hoffmann, H. and Möller, R. (2003). Unsupervised learning of a kinematic arm model. In Kaynak, O., Alpaydin, E., Oja, E., and Xu, L., editors, *Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP 2003, LNCS*, volume 2714, pages 463–470. Springer, Berlin.

Hoffmann, H. and Möller, R. (2004). Action selection and mental transformation based on a chain of forward models. In Schaal, S., Ijspeert, A., Billard, A., Vijayakumar, S., Hallam, J., and Meyer, J.-A., editors, *From Animals to Animats 8, Proceedings of the Eighth International Conference on the Simulation of Adaptive Behavior*, pages 213–222, Los Angeles, CA. MIT Press.

Hoffmann, H., Schenck, W., and Möller, R. (2005). Learning visuomotor transformations for gaze-control and grasping. *Biological Cybernetics*, 93(2):119–130.

Horstmann, A. and Hoffmann, K. P. (2005). Target selection in eye-hand coordination: Do we reach to where we look or do we look to where we reach? *Experimental Brain Research*, 167(2):187–195.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154.

Irwin, D. E. and Gordon, R. D. (1998). Eye movements, attention and trans-saccadic memory. *Visual Cognition*, 5(1-2):127–155.

Jordan, M. I. and Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16(3):307–354.

Juan, C. H., Shorter-Jacobi, S. M., and Schall, J. D. (2004). Dissociation of spatial attention and saccade preparation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15541–15544.

Kanwisher, N. and Wojciulik, E. (2000). Visual attention: Insights from brain imaging. *Nature Reviews Neuroscience*, 1(2):91–100.

Kawato, M. (1990). Feedback-error-learning neural network for supervised motor learning. In Eckmiller, R., editor, *Advanced Neural Computers, Elsevier, North-Holland*, pages 365–372.

Klarquist, W. N. and Bovik, A. C. (1998). Fovea: A foveated vergent active stereo vision system for dynamic three-dimensional scene recovery. *IEEE Transactions on Robotics and Automation*, 14(5):755–770.

Kuperstein, M. (1988). Neural model of adaptive hand-eye coordination for single postures. *Science*, 239(4845):1308–1311.

Leigh, R. J. and Zee, D. S. (1999). *The Neurology of Eye Movements.* Oxford University Press, UK.

Mallot, H. A. (1985). An overall description of retinotopic mapping in the cats visual-cortex area-17, area-18, and area-19. *Biological Cybernetics*, 52(1):45–51.

Mather, J. A. and Fisk, J. D. (1985). Orienting to targets by looking and pointing — parallels and interactions in ocular and manual performance. *Quarterly Journal of Experimental Psychology Section A — Human Experimental Psychology*, 37(3):315–338.

Matin, L., Matin, E., and Pola, J. (1970). Visual perception of direction when voluntary saccades occur. 2. Relation of visual direction of a fixation target extinguished before a saccade to a subsequent test flash presented before saccade. *Perception & Psychophysics*, 8(1):9–14.

McIlwain, J. T. (1996). *An introduction to the biology of vision*. Cambridge University Press, Cambridge, UK.

Miall, R. C., Weir, D. J., Wolpert, D. M., and Stein, J. F. (1993). Is the cerebellum a Smith predictor? *Journal of Motor Behavior*, 25(3):203–216.

Möller, R. (1999). Perception through anticipation — a behavior-based approach to visual perception. In Riegler, A., Peschl, M., and von Stein, A., editors, *Understanding Representation in the Cognitive Sciences*, pages 169–176. Plenum Academic / Kluwer Publishers, New York.

Möller, R. and Hoffmann, H. (2004). An extension of neural gas to local PCA. *Neurocomputing*, 62:305–326.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294.

Muggleton, N. G., Juan, C. H., Cowey, A., and Walsh, V. (2003). Human frontal eye fields and visual search. *Journal of Neurophysiology*, 89(6):3340–3343.

Neggers, S. F. W. and Bekkering, H. (1999). Integration of visual and somatosensory target information in goal-directed eye and arm movements. *Experimental Brain Research*, 125(1):97–107.

Neggers, S. F. W. and Bekkering, H. (2000). Ocular gaze is anchored to the target of an ongoing pointing movement. *Journal of Neurophysiology*, 83(2):639–651.

Nobre, A. C., Gitelman, D. R., Dias, E. C., and Mesulam, M. M. (2000). Covert visual spatial orienting and saccades: Overlapping neural systems. *Neuroimage*, 11(3):210–216.

Pagel, M., Maël, E., and von der Malsburg, C. (1998). Self calibration of the fixation movement of a stereo camera head. *Machine Learning*, 31(1-3):169–186.

Perry, R. J. and Zeki, S. (2000). The neurology of saccades and covert shifts in spatial attention — an event-related fMRI study. *Brain*, 123(11):2273–2288.

Prablanc, C., Echallier, J. F., Komilis, E., and Jeannerod, M. (1979). Optimal response of eye and hand motor systems in pointing at a visual target. 1. Spatio-temporal characteristics of eye and hand movements and their relationships when varying the amount of visual information. *Biological Cybernetics*, 35(2):113–124.

Rizzolatti, G., Riggio, L., and Sheliga, B. M. (1994). Space and selective attention. In Umiltà, C. and Moscovitch, M., editors, *Attention and Performance VI: Conscious and Nonconscious Information Processing*, pages 231–265. MIT Press, Cambridge (MA).

Rolfs, M., Engbert, R., and Kliegl, R. (2005). Crossmodal coupling of oculomotor control and spatial attention in vision and audition. *Experimental Brain Research*, 166(3-4):427–439.

Rosenblueth, A. and Wiener, N. (1945). The role of models in science. *Philosophy of Science*, 12(4):316–321.

Rumelhart, D. E., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA.

Schenck, W. and Möller, R. (2004). Staged learning of saccadic eye movements with a robot camera head. In Bowman, H. and Labiouse, C., editors, *Connectionist Models of Cognition and Perception II*, pages 82–91, London, NJ. World Scientific.

Schenck, W. and Möller, R. (2006). Learning strategies for saccade control. *Künstliche Intelligenz*, Iss. 3/06:19–22.

Schenck, W. and Möller, R. (2007). Training and application of a visual forward model for a robot camera head. In Butz, M. V., Sigaud, O., Pezzulo, G., and Baldassarre, G., editors, *Anticipatory Behavior in Adaptive Learning Systems: From Brains to Individual and Social Behavior*, number 4520 in Lecture Notes in Artificial Intelligence, pages 153–169. Springer, Berlin, Heidelberg, New York.

Schiegg, A., Deubel, H., and Schneider, W. X. (2003). Attentional selection during preparation of prehension movements. *Visual Cognition*, 10(4):409–431.

Schneider, W. X. (1995). VAM: A neuro-cognitive model for visual attention control of segmentation, object recognition and space-based motor actions. *Visual Cognition*, 2:331–376.

Snyder, L. H. (2000). Coordinate transformations for eye and arm movements in the brain. *Current Opinion in Neurobiology*, 10(6):747–754.

Tamburrini, G. and Datteri, E. (2005). Machine experiments and theoretical modelling: From cybernetic methodology to neuro-robotics. *Minds and Machines*, 15(3-4):335–358.

Tanaka, F., Cicourel, A., and Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46):17954–17958.

Tani, J. (1996). Model-based learning for mobile robot navigation from the dynamical systems perspective. *IEEE Transactions on Systems, Man, and Cybernetics — Part B*, 26(3):421–436.

Umeno, M. M. and Goldberg, M. E. (1997). Spatial processing in the monkey frontal eye field. 1. Predictive visual responses. *Journal of Neurophysiology*, 78(3):1373–1383.

Vercher, J. L., Magenes, G., Prablanc, C., and Gauthier, G. M. (1994). Eye-head-hand coordination in pointing at visual targets - spatial and temporal analysis. *Experimental Brain Research*, 99(3):507–523.

Walker, M. F., Fitzgibbon, E. J., and Goldberg, M. E. (1995). Neurons in the monkey superior colliculus predict the visual result of impending saccadic eye-movements. *Journal of Neurophysiology*, 73(5):1988–2003.

Watson, J. B. (1994). Psychology as the behaviorist views it (reprinted from psychological review, vol 20, pg 158, 1913). *Psychological Review*, 101(2):248–253.

Webb, B. (2000). What does robotics offer animal behaviour. *Animal behaviour*, 60(5):545–558.

Webb, B. (2001). Can robots make good models of biological behaviour? *Behavioral and Brain Sciences*, 24:1033–1050.

Weir, C. R. (2006). Proprioception in extraocular muscles. *Journal of Neuro-Ophthalmology*, 26(2):123–127.

Ziemke, T., Jirenhed, D.-A., and Hesslow, G. (2005). Internal simulation of perception: A minimal neuro-robotic model. *Neurocomputing*, 68:85–104.