Observation of Human Eye Movements to Simulate Visual Exploration of Complex Scenes

Ahmed A. Faisal, Markus Fislage, Marc Pomplun,

Robert Rae[*] and Helge Ritter

Neuroinformatics Group, Faculty of Technology,

University of Bielefeld, P.O. Box 100 131,

D-33501 Bielefeld, Germany

{impomplu|robrae}@techfak.uni-bielefeld.de

14th April 1999

[*] To whom correspondence should be send.

Abstract

The human eyes are always in action, they explore the environment every second we are awake. But what attracts our visual attention? In this paper we examine the eye movements of human subjects observing a breakfast scenario using an eye-tracking system. Moreover, we developed a hierarchical model consisting of three modules. This model is applied to the same visual stimuli and generates saccades depending on an empirically based set of image features extracted from the input image. The architecture of our model is motivated by the anatomy of the human visual pathway and the results from the eye-tracking experiment. The model uses low-level image features, extracted by filters similar to those found in the receptive fields of cortical neurons in mammals. The filter responses are combined in the adaptive multi-layered *cortical map* module. The activity in the *saliency map* of this first module roughly estimates the next fixation point. The second module refines this point using a neural network which shifts the model's attention to locally interesting image regions. The third module finally determines the fixation point and recognizes the object found in the focused image region. To obtain a quantitative comparison of the common features between gaze trajectories of the human subjects and our model, we introduce the *Markov-path distance* measure. Based on this measure, we analyze common eye movement strategies between different subjects. Furthermore, we can show that our model generates "virtual eye movements" similar to those found in the eye-tracking experiment.

Acknowledgements

This work was largely done within the "Projektseminar Modellierung menschlicher Augenbewegungen". Participants: Peter Biermann, Ahmed Faisal, Markus Fislage, Robert Haschke, Benjamin Hell, Alexander Kazik, Thorsten Köpsel, Andre Krause, Sandra Plahl and Martin Schmitz. Organizers: Marc Pomplun and Robert Rae.

Contents

1 Introduction

An individual's attention is usually attracted by sensory information of several types, e.g., sound, pain, or visual perception. Especially our eyes are always in action and usually we are unaware of these sometimes rapid gaze movements. The visual focus shifts depending on the task we perform (e.g., reading a text), reflexes (e.g., eye-protection) or in response to salient scene features (e.g., bright colors). Since the human eye has only a small field of high resolution in its center—the *fovea*—, gaze shifts are important for a detailed visual analysis of a scene. Researchers distinguish between two basic types of eye movements: (i) saccades and (ii) smooth pursuit. While saccades are rapid adjustments used to scrutinize interesting objects in the actual scene, smooth pursuit is used to track moving objects. Both actions allow us to center the fovea at an interesting visual target. This focusing of attention enables a very efficient processing of visual information in our brain since only interesting image regions become processed at foveal resolution. In addition, evolution has created mechanisms and strategies to accomplish different tasks, like visual search or object recognition. Today, many researchers are interested in these basic eye movement behaviors and the underlying processes. Experiments using modern eye-tracking devices provide an insight into the cognitive mechanisms of eye movements [Pom98],[Yar67].

Results from this interesting research area can also be used to design technical vision systems and there is an increasing number of publications in the field of active vision [RB95a],[SS93]. Since real-time processing of visual images is computationally expensive, it is important to reduce the amount of visual data without losing its inherent information. In this paper, we present a hierarchical model which uses a multi-layered low-level feature integration scheme to simulate visual attention. This model selects interesting and relevant parts of the input image for a closer inspection by an object recognition module. The architecture of our model is based on the concept of nested region of interest, using decreasing image sizes at increasing resolution as input for the successively working modules. A typical problem in such an attentive active vision system is the "where-to-look-next" task. Below, we will concentrate primarily on approaches to strategies of camera movements designed for efficient static scene exploration and object recognition ("what" task).

Niebur and Koch developed a model based on new physiological data on the primate visual pathway [NK96]. They use a saliency map to control the focus of attention. The map is computed by summation of the output of several low-level image feature filters. Afterwards, the most salient image region is selected by a winner-takes-all network. The systems mimics the control of selective visual attention and identifies the most salient points in a visual scene which is demonstrated on many different input images.

The VISIT model of Ahmad is based on physiological experiments on human attention [Ahm91]. VISIT consists of a highly parallel connectionist network performing two visual tasks: (i) computation of spatial relations and (ii) object recognition based on image features. Using a second model (SWIFT) to minimize the number of fixations, the resulting system is very efficient and flexible and works on high resolution images.

The approach of Fellenz [Fel97] concentrates on the problem of feature binding and employs a multi-layered hierarchical processing model. The first layer is a pre-attentive synchronization mechanism for figure-ground separation. It is followed by an attentive process which extracts the segmented object in a low resolution image focus. This knowledge is used on a higher level to control the data acquisition. The feedback to the camera closes the action-perception cycle of the system, which allows the analysis of complex images.

Kattner implemented a model of attention to link low- and high-level vision using "active reconstruction" [Kat94]. In this model, the first cues for fixation are calculated using standard low-level methods. The focused region is classified and looped back into an integration scheme on a higher level. The model works in two directions: Attention is guided both by simple image features *and* by high-level knowledge. This results in a very efficient fixation scheme working on various types of real-world images and recognition problems.

In the present contribution, we propose a model which imitates eye movements of human subjects exploring images of a real-world "breakfast scenario" (see Fig. 2). The image database of a typical breakfast table showing cups, plates, cutlery, etc. is suitable for eye-tracking research since we can easily change position, color and form of the objects (see also [RB95b]). Moreover, it provides a convenient basis for modelling low- and high-level domain knowledge since it includes only a limited number of different items. Although our system was mainly designed for this special scenario, we will show that it also works well in new, previously unseen environments with new objects and lighting conditions. The system architecture employs the following adaptive modules:

1. The *cortical map module* (COM) analyzes the whole field of view at a very coarse resolution modelling the peripheral perception. This module determines the target of the next fixation and simulates a biologically motivated *saliency map* (sometimes also referred to as *motor map*).

2. The *re-center and hypothesis module* (RHM) uses a smaller viewing angle with a more detailed resolution and re-adjusts the fixation point generated by the COM. At the same processing step, a hypothesis about the identity of the fixated object is made.

3. The *object recognition module* (ORM) simulates the fovea and has only a very small field of view with a high resolution. This module is the object recognition stage; it uses a multi-dimensional feature-vector, derived from the foveated part of the scene, together with the object hypothesis from the RHM to determine the fixated object's identity.

The results of the object recognition are finally used to “fade out” the already visited image position in the COM. This feedback loop generates a dynamic system behavior with saccades similar to those employed by human subjects exploring a static scene.

To compare the saccades of our model with the behavior of the subjects we introduce a novel similarity measure, the *Markov-path distance*. This quantitative measure allows us to calculate the similarity between the different saccade paths created on the images by both the model and the human subjects.

In the next section, we describe the eye-tracking experiment in the breakfast scenario. The recorded data are further examined to explore the influence of low-level image features—like color or structure of the objects—on the eye movement behaviors of the human observer. The third section is dedicated to the definition of the Markov-path distance, a measure to compare different trajectories connecting discrete fixation points. Subsequently, we introduce our hierarchical model whose architecture is motivated by physiological anatomy, and discuss results obtained by applying our novel Markov-path distance measure. Issues of the qualitative behavior of the hierarchical model are discussed as well. To demonstrate the flexibility of the approach, we show some examples of view paths obtained with our model "viewing" images from other domains. We close with a summary and an outlook on future work.

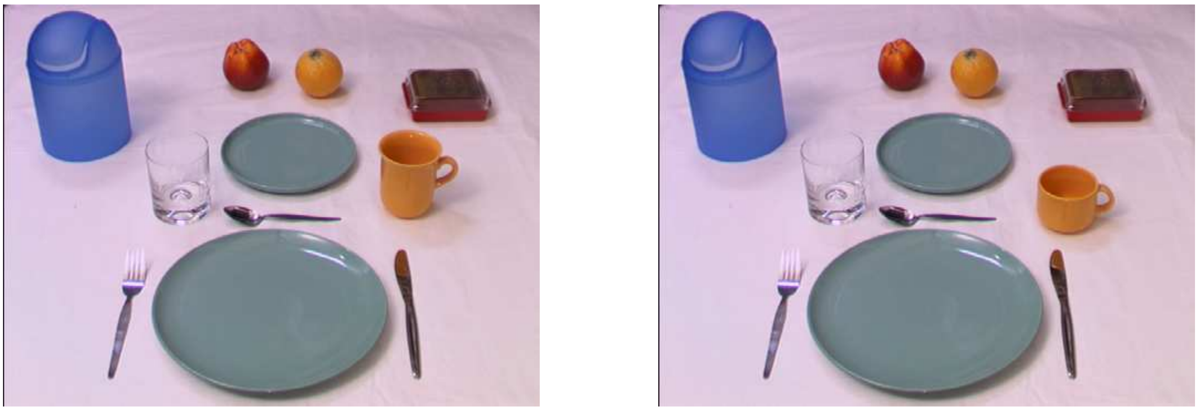
2 Eye-Tracking Experiment

The eye-tracking experiment was conducted in order to obtain empirical eye-movement data about visual exploration of realistic scenes. In order to induce the visual exploration of the displayed scenes, subjects were told to detect a difference between two successively presented similar scenes. Fig. 1 shows one of the image pairs with some typical objects used in the experiment.

The scenes consisted of everyday breakfast items like plates, cups, and cutlery. The general arrangement was identical for all images in our database. Between the first and the second image of an image pair, one of the following four types of local differences occurred:

- one of the objects disappeared
- an additional object appeared
- one of the objects changed its color
- one of the objects changed its form

In order to detect the difference, subjects had to scan the first scene thoroughly. The comparison task did not only ensure visual exploration, but also—to some extent—it eliminated higher



(a) (b)

Figure 1: *Typical "breakfast scenario": (a) First image presented to the subject and (b) second image presented with a slightly different setup—the cup has changed.*

cognitive processes from the subjects' scene inspection, because the demand to memorize all objects within a limited duration did conceivably not allow the subjects to apply high-level (semantic) knowledge about the stimuli. As a consequence, it should be possible to simulate the subjects' eye-movement behavior in an adequate way.

2.1 Method

The 16 subjects were students of various fields at the University of Bielefeld. All subjects had normal or corrected-to-normal vision, none had pupil anomalies and all were able to distinguish between colors. The stimulus pictures were presented on a computer screen with a spatial resolution of 640×480 pixels. The pictures showed standardized "breakfast scenes" (see Fig. 2). The size of the breakfast items ranged from about 2.5 to 15 degrees of visual angle in diameter.

Eye movements during stimulus presentation were measured using the OMNITRACK 1 system [Sta93]. OMNITRACK 1 is a non-invasive imaging eye tracking system consisting of a 486DX2-66 type computer equipped with image processing hard- and software plus a headset. Subjects were seated about 60 cm away from a 17" color monitor. They were wearing a head-set equipped with two miniature infrared video cameras yielding on-line information about the position of both the pupil of the subjects' right eye and the subjects' head, so that any head movements during viewing cannot impair the accuracy of eye movement measurement. From the camera data, actual fixation points on the screen were calculated at a frame rate of 60 Hz. Only fixations that lasted for at least five frames (i.e. 83 ms) were measured. The measurement of fixations included their absolute time of occurrence, their duration, and their screen coordinates. Additionally, the subject's actual pupil size was registered for each fixation. Prior to experimentation, a calibration procedure was performed by making the subject fixate nine specified points on the screen. The absolute precision of the calibrated system lay within 0.7 to 1.0 degrees of visual angle, corresponding to about 7 to 10 mm on the screen.

Subjects were tested individually. In each trial, they were sequentially shown three different images: First, a breakfast scene A (15 seconds), second, a distractor image (5 seconds), and third, a breakfast scene B (maximal 10 seconds). While viewing the last on of these images, the subjects' task was to find the only difference between the scenes A and B. Subjects were to press a mouse key placed in front of them as soon as they had detected the mismatch. Pressing the mouse key finished the trial immediately. Each subject viewed 20 sequences of images. Before the presentation of each sequence, the eye tracker was recalibrated using a single target in the center of the screen in order to compensate for any sliding of the head-set due to head movements by the subjects. The order of stimulus presentation was randomly generated for every subject.

2.2 Results

The analysis of eye movements was restricted to the presentation of the "memorization" scenes. Fig. 2 shows an example scanpath, where the small squares mark the fixation points, and the brightness at the fixation point visualize the fixation duration at this point. Thus, the highlighting of fixation points signifies that the glass and the cup attracted most of the subject's attention.

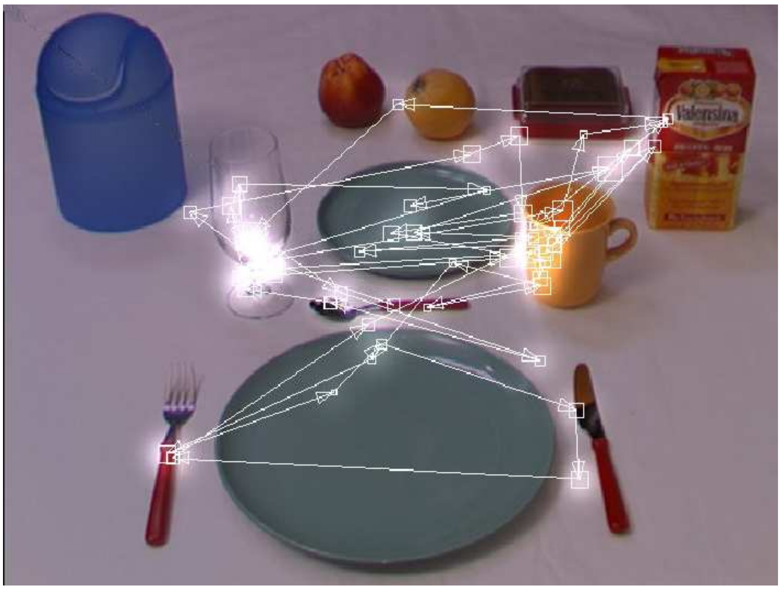
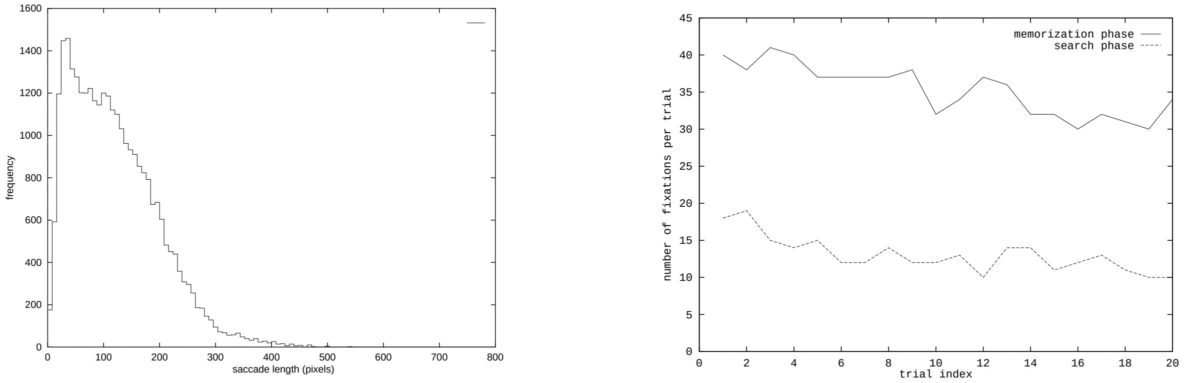


Figure 2: *Example of a gaze trajectory generated by one of the subjects. The size of a square indicates the duration of the corresponding fixation; surroundings of fixations are highlighted, indicating salient areas.*

Fig. 3(b) shows the number of fixations per trial as a function of the trial's position within a session. The number of fixations seems to decrease slightly with an increasing number of trials completed before. In fact, an analysis of variance (ANOVA) revealed that the mean number of fixations per trial is significantly higher for trials 1 to $10 (36.15)$ than for trials 11 to $20 (33.24)$ $(F(1, 159) = 27.56; MSe = 68.27; p < 0.0001)$. This small but significant difference suggests that subjects are able to benefit from the constant general arrangement of the stimuli. As soon as they have learned which objects can appear in which locations, they do not need to memorize the whole information given in the scene anymore, which results in fewer fixations.

In order to analyze the empirical gaze trajectories in a quantitative way, they were converted into object-by-object scanpaths. Each object was assigned an elliptical "area of attention" containing the object itself plus an appropriate error margin compensating for errors in gaze-position measurement. All fixations located in such an area were attributed to the respective object; fixations hitting none of the areas constituted an additional group named



(a) (b)

Figure 3: *(a) Histogram of saccade length for all subjects and all memorization stimuli; (b) Number of fixations during one experiment averaged over all persons.*

“orientational”, since they were likely to be employed for the subjects’ orientation within the stimulus.

Moreover, the objects were divided into groups of identical colors and identical shapes. This enabled us to investigate whether the subjects preferred transitions between specific combinations of object colors or shapes. Table 1 presents a list of all breakfast items and their attributes.

		color	shape
a	dustbin	light blue	complex
b	apple	red	circular
c	orange	orange	circular
d	butter-dish	red	angular
e	small plate	green	circular
f	cup	orange/blue	complex
g	glass	transparent	complex
h	carton	orange	angular
i	spoon	silver	longish
j	knife	silver	longish
k	fork	silver	longish
l	large plate	green	circular

Table 1: *The breakfast items and their color and shape attributes.*

	l.blue	red	orange	green	blue	transp.	silver	orient.
light blue	0	195	63	95	1	248	45	88
red	211	383	1000	440	85	218	85	252
orange	100	915	474	925	10	463	375	170
green	119	446	779	229	130	729	974	213
blue	7	73	7	147	0	67	53	13
transparent	177	248	443	839	51	0	547	125
silver	26	72	304	940	51	487	632	114
orientational	60	260	214	245	15	120	103	0

Table 2: *Frequency of transitions between color attributes (standardized to a maximum value of 1000).*

	circular	angular	complex	longish	orient.
circular	596	598	974	475	181
angular	618	364	795	246	178
complex	1000	835	858	616	194
longish	463	214	532	407	74
orientational	220	182	186	66	0

Table 3: *Frequency of transitions between shape attributes (standardized to a maximum value of 1000).*

3 Markov Path Distance

In this section, we describe a method to compare different gaze strategies. Generally this is accomplished by interpreting each saccade as a two dimensional movement vector.

However, to characterize a complete gaze trajectory we are not primarily interested in its precise geometric shape, but instead in the order in which it visits the different objects. Therefore, interpretation of spatial information is not sufficient and large-scale quantitative analysis of eye-tracking data, especially the comparison of saccade-trajectories on non-static images, requires a new kind of trajectory comparison measure.

For this purpose, we developed a new kind of distance measure, the *Markov-Path-Distance* (MPD). Its aim is to measure the similarity between gaze strategies in such a way, that the geometric details of the individual trajectories play only a minor role. The serial occurrence of semantically important features, such as object identity, form and color, are used to characterize a gaze strategy. The basics for the comparison of the MPD is a more abstract representation of the set of gaze trajectories. The representation is found by representing each trajectory as a Markov sequence of discrete states. This is an stochastic approximation of the unknown biological process.

Thus, the entire gaze strategy is characterized by the statistics of the Markov process. Mathematically we need to compute the Markov transition matrix M whose elements M_{ij} describe the probability of fixating feature i after having fixated feature j. The probability of visiting a particular feature j after a larger number, say, n successive fixations, is given by the matrix M^n. A straightforward way to compute two different gaze strategies A and B might be to measure the similarity of the associated Markov matrix M_A and M_B. However, M describes only the statistic short term structure of a trajectory. On the other extreme, the matrix $M^{\inf} = lim_{n \to \inf}$ (if it exists) describes essentially the proportion of the stationary state $p^{\inf}$ of the assumed Markov process[1]. Therefore, by choosing an intermediate value of n, we can strike a compromise between both extremes. For our investigation it has turned out that a suitable value is $n = 3$. We analyzed the trajectories by counting how often each of the possible object combination of n elements, i.e. paths of length n, were found. We varied n from 2 to 5 and found only for $n = 3$ that the distribution was non-uniform. Hence, we selected this path length, as it suggested more inherent information.

Now we can define the non-similarity *distance* between two trajectories A and B, which is the difference of the two Markov process matrices $P_A := M_A^n$ and $P_B := M_B^n$. The distance $||D||$ between two trajectories A and B is

$$||D|| := ||P^A - P^B|| = \sum_{ij} (D_{i,j})^2 \qquad (1)$$

We use $\sum (D_{i,j})^2$ as a matrix measure to describe the similarity between the two matrices. Large differences in the processed matrices M contribute to a higher proportion in the dissimilarity between two trajectories. Thus, the *Markov Path Distance* $||D||$ is an effective method to describe similarities between any two trajectories.

[1] $p^{\inf}$ denotes the probability distribution that is invariant under M: $Mp^{\inf} = p^{\inf}$

With the MPD we are now able to describe fixation trajectories in a translation-invariant way. In Section 5 we have several results on the quality of our MPD measure.

4 System Architecture

Our aim is to design a model which imitates some aspects of human visual behavior. Based on a quantitative analysis of empirical human eye-tracking data we build a hierarchical model which generates saccades. The model receives the visual input at different resolutions and sizes, simulating an image sensor similar to the human retina, where the resolution decreases logarithmically with growing distance from the center. As the human visual system controls visual attention by controlling the gaze direction [KSJ95], our aim is to construct a model based on this principle.

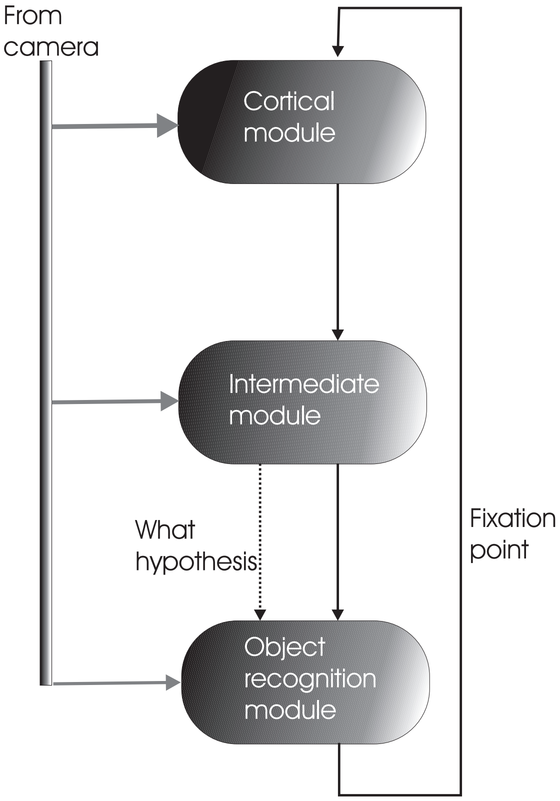


Figure 4: *Hierarchical module concept.*

We distinguish two (roughly) separate processing levels, a pre-cognitive and a cognitive level. At the cognitive level, sub-conscious high-level knowledge and context information is used to direct selective visual attention. At the pre-cognitive level, low-level information is analyzed. This level uses features extracted from the visual input, e.g., color and form, to determine *where-to-look-next*. This is achieved by assigning a "saliency" value to each point in the visual field. Thus, a complete *saliency map* of the field of view is constructed [TG80],[NK96]. Peak values in this saliency map correspond to areas of high visual interest and visual attention is preferably directed to these locations. Indeed, there are several hints for the existence and location of a saliency map in biological visual systems [PRM87],[KMKI88],[LB90],[DWTS91],[RP92].

In our approach, we concentrate on low-level visual attention and design a hierarchical model consisting of three modules. Each module processes visual information at a finer and smaller region of interest (RoI) than its preceding module (see Fig. 4). Therefore, the number of operations required to analyze a whole scene is significantly smaller than the number of operations required to analyze the scene at high resolution.

The central saccade-mechanism is implemented in the first module, the "cortical map module"—COM. In contrast to the remaining modules, this module has a multi-layered architecture. It receives inputs from different filters which extract various low-level features, e.g., vertical and horizontal edges. The responses of all these filters are represented in *feature layers*. All layers are individually weighted and accumulated to calculate the *cortical map*. To determine suitable weight values for each feature layer, we use an adaptation algorithm which dynamically changes the influence of the different feature layers on the resulting cortical map (this algorithm will be explained in Section 4.1.1). The values in the cortical map represent the saliency of each image position, and the highest peak in this map is a rough estimation of the next visual target.

This first estimation of the fixation point is further refined in the remaining two modules. The second module analyzes the image region around the fixation point of the COM and shifts the fixation position towards adjacent objects. This module also generates an object hypothesis from the image region around the re-centered fixation position ("re-center and hypothesis module"—RHM). The last module performs an object recognition ("object recognition module"—ORM), which uses the fixation position determined by the RHM. It analyzes the image region at a higher resolution and classifies the objects located in the extracted RoI. The output of the ORM is compared with the object hypothesis of the RHM and the result determines the final object class.

Thus, the first two modules COM and RHM select *where-to-look-next* for an interesting object and focus the attention on it. Saccades generated by these modules can be compared with real human eye movements. With the ORM we have a further possibility to check the efficiency and stability of the whole attentional system by analyzing the recognition performance of this module. In the following section we will discuss each of the three modules in detail.

4.1 Cortical Map Module (COM)

The cortical map module (COM) is an abstract model of the layered feature processing structures in the mammalian visual pathway [KSJ95]. The COM generates a *cortical map* using a low-resolution feature analysis scheme. This cortical map is our simulation of a saliency map.

The COM is structured in *cortical layers* (see Fig. 5). Each cortical layer contains the responses of biologically motivated feature detectors (see Section 4.2). Thus, each layer is a mapping of a distinct feature over the whole visual field at a low resolution[2]. Information from the same area in the visual field is located at the same position on each layer. This organization of the layers is conform to the principle of *retinotopical* processing in biological visual systems [KSJ95]. According to this principle, information neighboring in the visual input will preserve neighborhood relationship in representations at different processing stages. A rectan-

[2] The camera's visual field was $29.8^o \times 22.6^o$, which corresponds to 485×365 pixels sub-sampled to 100×74 pixels.

gular vertical section through the cortical layers reflects all features found at the corresponding area of the visual field. This section can be interpreted as a “cortical” column (see Fig. 5). This principle is found in neurobiology results from the hyper-columns of mammalian visual cortex [KSJ95].

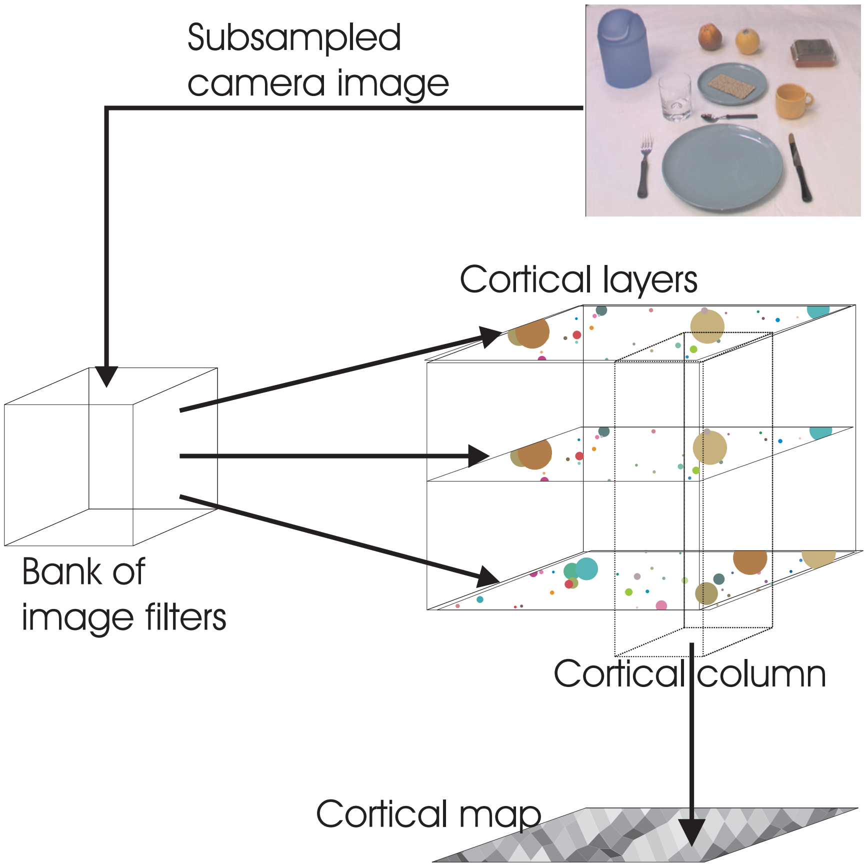


Figure 5: *The multi-layered cortical map module.*

In our system, these columns contain the responses of different cortical layers, i.e., feature detectors. As a measure for the presence of a relevant feature in the correlated visual area we take the sum of all values within a column. To respect the influence of all features, occuring with different intensities over the whole visual input, each feature layer is weighted differently. Hence, the cortical map module combines two different kinds of information. Columns represent local information, because each column processes its own corresponding local visual input. The different weights of the layers represent a form of scene (or global) information.

Since the weight within one layer is equal, the layer values provide information of the effect of the corresponding feature on the whole scene.

We use the weighted sum of a column as the resulting saliency value of the corresponding area in the visual input. The whole visual input (processed in the cortical map module) is covered by these columns. Therefore, all values together represent the *cortical map* of the visual input, assigning a saliency value to each image position.

A generalizing mapping of the saliency must be able to adapt the weights dynamically in order to perform robust on changes in a scene or even on changing scenes. We determine the set of weights by adapting them with a special “training”-algorithm. The aim of this algorithm is to modify the weights in such a way that the resulting cortical map will be most similar to an ideal “teacher” map. This map has a peak in its center and decreases logarithmically from the center—in the same way as resolution in the human eye is distributed. The teacher map is based on the data of our eye-tracking experiment, where we found that objects closer to the current focus are more interesting and are preferably inspected next. We assume that an object closer to the fovea generally has higher saliency values, due to the higher resolution—thus, more “interesting” details can be seen. This heuristical method is used to select one target of attention among the possible ones.

To select a new focus of attention, the three highest peaks in the saliency map are analyzed. One peak is selected according to an empirically found distance constraint and passed as a saccade coordinate to the subsequent modules of the hierarchical model.

The object position from the recognition module (ORM) is used to generate an additional layer, the inhibition layer. In this layer the already inspected areas are accumulated and it can be used as a “memory” for the next cycle through the model. This additional layer provides inhibitory feedback to the cortical map. The values in this layer typically decay exponentially in each cycle (exponent$= 0.03$). Thus, the inhibition layer represents the history of recently visited areas as negative values at these positions. Hence, the addition of this layer to the feature layers inhibits already inspected areas in the cortical map. This can be interpreted as a “*fade out*” of visited areas. Inspected locations become unlikely to be revisited, enabling the system to explore the whole scene (“Inhibition of return”). Thus, “interesting” objects are visited first and less interesting ones will be visited later.

4.1.1 Mathematical Description

In this section, we present the mathematical model describing our adaptive cortical map and its properties. Its description is general as it can be applied to multiple teacher maps and different scenes simultaneously. Our current model uses only one teacher map. The training of the cortical module’s weights is achieved by minimization of an error function E. This function is defined as the average square error between the teacher maps x^α and the result of the corresponding cortical map. Note that in our model all x^α are equal and correspond to a two-dimensional logarithmic peak. The different layers j are high dimensional vectors $\vec{v}_j$. Thus, each component of this vector corresponds to the activity of the input layer at a certain location. We express the cortical map as a linear combination of all “layer vectors” $\vec{v}_j$ weighted with a scalar w_j. Hence, an optimal set of weights minimizes the difference E between the teacher map and the actual cortical map.

$$E = \frac{1}{2}\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}[(\sum_{j=1}^{L} w_j\vec{v}_j^\beta) - \vec{x}^\alpha]^2 \tag{2}$$

We calculate the optimal weight set by an iterative gradient descent on the error function E. At a minimum of E we must have:

$$\begin{aligned} 0 &= \frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{2}\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}[(\sum_{j=1}^{L} w_j\vec{v}_j^\beta) - \vec{x}^\alpha]^2 \\ &= \sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}[(\sum_{j=1}^{L} w_j\vec{v}_j^\beta) - \vec{x}^\alpha]^T\vec{v}_i^\beta \end{aligned} \tag{3}$$

Therefore,

$$\begin{aligned} w_i &= \frac{1}{NM}\frac{\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}[(\vec{x}^\alpha)^T\vec{v}_i^\beta]}{\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}[||\vec{v}_i^\beta||^2]} - \\ &\quad \frac{1}{NM}\sum_{j\neq i}^{L}\frac{\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}[w_j(\vec{v}_j^\beta)^T\vec{v}_i^\beta]}{\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}[||\vec{v}_i^\beta||^2]} \end{aligned} \tag{4}$$

Eq. 4 contains a term for w_i which is independent from all other $w_j, j \neq i$. Because the algebraic solution of this term is computationally expensive (approx. $O(L^3)$) we use an iterative approximation algorithm instead. It performs the task in approximately $O(MNL)$ (note that in our model M and N are 1), proof not provided here.

Algorithm 1 Training algorithm for the cortical layer weights.

```
t=0
**for all** cortical layers $i$ **do**
    initialize all $w_j(0)$ with a normal distributed random variable with mean $0$ and variance $1$
**end for**
**repeat**
    $t = t + 1$
    **for all** cortical layers $i$ **do**
        compute the new weight $w_i(t+1)$ according to Eq. 4 using the old weights $w_j(t)$
    **end for**
**until** both $\Delta w_i$ and $\Delta E$ are below the thresholds $\epsilon_{\Delta E}$ and $\epsilon_{\Delta w_i}$ during the last $k$ iterations
$(k \approx 3..10)$
```

Note that Eq. 4 consists of two independent terms. The first one $C = \frac{1}{NM} \frac{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N} [(\vec{x}^{\alpha})^T \vec{v}_i^{\beta}]}{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N} [||\vec{v}_i^{\beta}||^2]}$ is constant and can be computed prior to learning. The second term $D = \frac{1}{NM} \sum_{j \neq i}^{L} \frac{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N} [w_j (\vec{v}_j^{\beta})^T \vec{v}_i^{\beta}]}{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{N} [||\vec{v}_i^{\beta}||^2]}$ depends on the changing weights w_j. Therefore, the efficiency of the algorithm can be improved by randomly selecting a weight w_a to adapt, while all other weights $w_c, c \neq a$ are kept constant. This corresponds to an one-dimensional line optimization in the weight space.

4.1.2 Statistical Interpretation

In this section, we analyze our cortical model, showing further interesting properties. Eq. 4 can be decomposed into the two terms C and D.

$$\begin{aligned} C &= \frac{1}{NM}\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}\frac{(\vec{x}^{\alpha})^T\vec{v}_i^{\beta}}{||\vec{v}_i^{\beta}||^2} \\ &= \langle\frac{(\vec{x}^{\alpha})^T\vec{v}_i^{\beta}}{||\vec{v}_i^{\beta}||^2}\rangle_{\alpha,\beta} \end{aligned} \tag{5}$$

The mathematical interpretation of term C is a discrete average of $\frac{(\vec{x}^{\alpha})^T\vec{v}_i^{\beta}}{||\vec{v}_i^{\beta}||^2}$. Hence, Eq. 5 can be interpreted as the average length of the projection of the teacher map $\vec{x}^{\alpha}$ onto cortical layer $\vec{v}_i^{\beta}$. It is a similarity measure between layer i and the teacher map. This can be interpreted: if a feature in a layer is similar to the teacher map, then its influence should become higher in the whole cortical map. Term D can be analyzed in a similar way:

$$\begin{aligned} D &= \frac{1}{M}\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N}\sum_{j\neq i}^{L}\frac{w_j(\vec{v}_j^{\beta})^T\vec{v}_i^{\beta}}{||\vec{v}_i^{\beta}||^2} \\ &= \langle\sum_{j\neq i}^{L}\frac{w_j(\vec{v}_j^{\beta})^T\vec{v}_i^{\beta}}{||\vec{v}_i^{\beta}||^2}\rangle_{\alpha,\beta} \end{aligned} \tag{6}$$

D decreases the influence of C on the weight w_i, since

$$w_i = C + D. \tag{7}$$

The term inside the sum of D

$$\frac{w_j(\vec{v}_j^{\beta})^T\vec{v}_i^{\beta}}{||\vec{v}_i^{\beta}||^2} \tag{8}$$

is averaged over the training data. This term describes the weighted length of the projection of all layers j to layer i. Thus, the length describes the overall similarity of layer i to all other layers $j, j \neq i$.The similarity between the two layers i and j is weighted with the influence—w_j—of layer j on the cortical map. If layer i is too similar to a layer j^*, w_i will drop. This ensures that the layer weights compete against each other for their influence in the resulting cortical map.

The properties described above allow the dynamic generation of the cortical map from heterogeneous feature layers. The mixture of different layers ensures the variety of relevant features forming the saliency map. The weights "compete" continuously against each other during adaptation. A change in the inspected scene, e.g., changes in lighting or changes due to motion or switching to a new scene will immediately effect the weight set. The algorithm will adapt the weights to reflect the new scene characteristics in the cortical map and limit the weights to an upper bound.

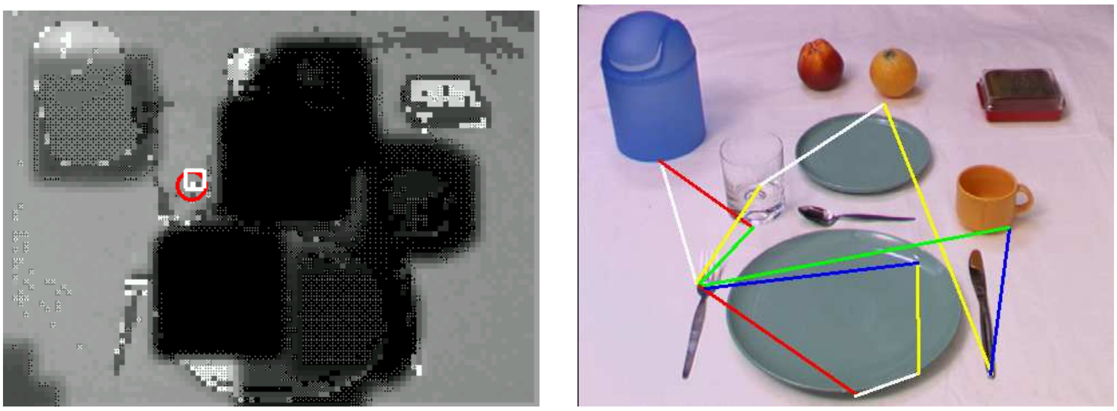


Figure 6: *State of the cortical map module while fixating a glass. Grey scale representation of the cortical map layer, with "fade outs" from the feedback inhibition layer (see text).*

4.2 Inputs to the Cortical Layer

We use 8 cortical layers to calculate the cortical map, each receiving its input from a different biologically motivated image filter. As an approximation of complex and end-stopping sensitive cells in primate visual cortex [KSJ95] and [Eng96] we designed a novel filter mask, which we call the "Comet" filter.

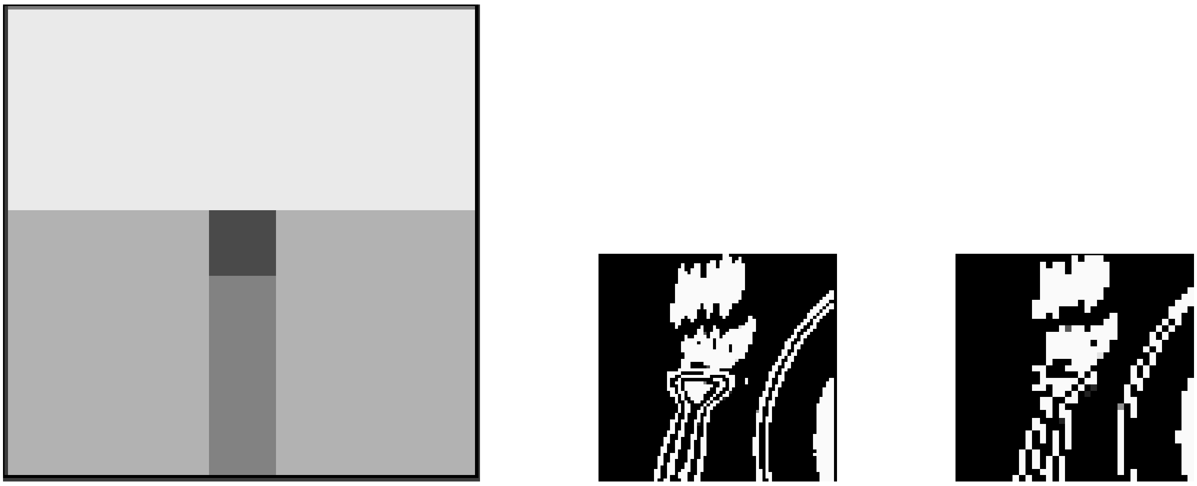


Figure 7: *The left picture shows the shape of the Comet filter in greyscale (white corresponds to high values, black to low ones). The Comet filter shows high sensitivity (white areas) at ending contours and performs a robust orientation sensitive edge detection due to the “comet’s tail” (right pictures).*

The Comet filter mask is constructed using the following definition:

$$m_{\text{Comet}}(x,y) = \begin{cases} -0.5 & \text{if} \quad (y > 0), \\ 0.5 & \text{if} \quad (x = 0) \wedge (y < 0), \\ 0 & \text{else} \quad (x = y = 0). \end{cases} \tag{9}$$

This filter mask is convolved with the sub-sampled camera image. Fig. 7 shows the result of the Comet filter applied to a typical scenario RoI. We use 4 different orientations of this filter mask corresponding to 4 cortical layers. The selection of 4 orthogonal orientations is biologically plausible, since the human visual system is especially sensitive to them (see Fig.7). Furthermore, we use responses of 4 different color segmentation algorithms, each calculated by a color angle and intensity threshold in the HSI-space. We select the 4 color shades which we found most attractive for the subjects in our eye-tracking experiment (red, orange, green and shininess/reflective). Thus, we have 8 feature layers and the negative feedback of the inhibition layer as input for the calculation of the cortical map.

4.3 The Re-center and Hypothesis Module (RHM)

The COM module suggests a coarse fixation point. Typically, this position is at an object's boundary. At an intermediate resolution with a smaller field of view the "re-center and hypothesis module" (RHM) determines a new fixation point (centering) and provides a classification hypothesis. The centering and hypothesis task is solved by two distinct artificial neural networks (Multi Layer Perceptrons).

class:	big plate	small plate	cup	glass	fruit	h. cutlery
color:	black	white	pink	light blue	yellow	dark blue
class:	v. cutlery	dust-bin	butter-dish	toast	cartoon	
color:	green	red	cyan	-	-	

Table 4: *Different object classes of the RHM and the color code for the objects used in Fig. 8.*

The first neural network is trained to refine the position of the fixation and the second to generate an object hypothesis at the newly centered position. The input field of view is a 50×50 grey-scale image, which is used to generate a 512-dimensional feature vector for the networks. We locally convolve the input image at 16×16 equidistant positions with first order 7×7 *Gabor masks* in horizontal and vertical orientations ($2 \times 16 \times 16 = 512$) [Dau85]. With this method, the edge extraction is independent from lighting and contrast.

For the classification task, we distinguish the 11 object classes given in Table 4. The training data for the 11 object classes consist of 10 randomly displaced images for each object. In these images the objects are located at different positions. After some training and test runs we found the following network-architectures to be best suited: A $512 - 80 - 2$ MLP network for the re-centering task and a $512 - 60 - 11$ MLP network for the classification task.

The output of the RHM is illustrated in Fig. 8, it consists of a vector-field, which converges locally to the object centers. The centering vectors are visualized by lines. If the classification output of the hypothesis network is above a threshold, a colored dot at the vector ends is plotted, indicating the identity of the object according to the hypothesis (see Table 4 for the color encoding scheme).

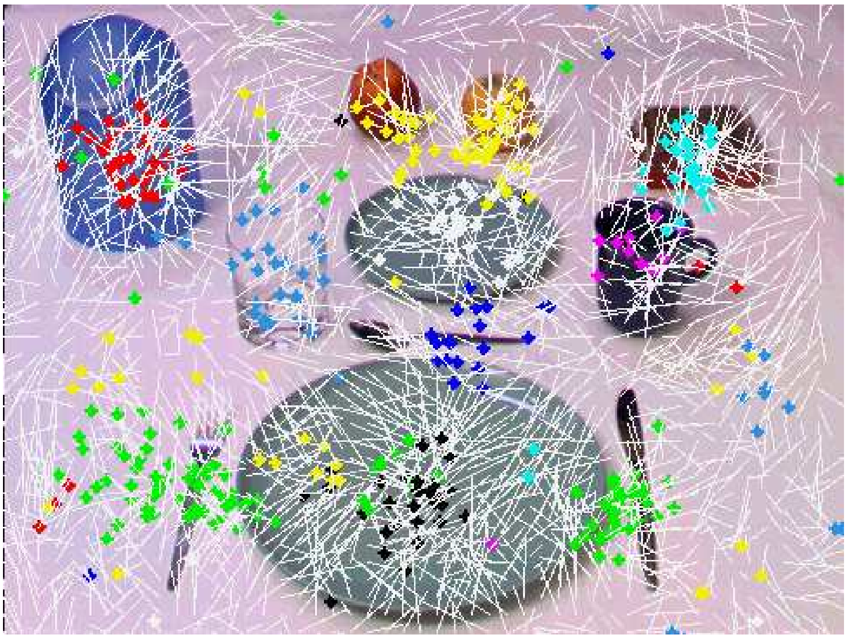


Figure 8: *Position refinement and classification results of the RHM.*

Fig. 8 shows that the vectors are directed towards the objects' centers and the classification output is suitable. To establish a final reliable recognition output for the fixated object, we use a further object recognition module which is applied to the re-centered image location.

4.4 The Object Recognition Module (ORM)

The RHM determines a fixation position at the object center. This position is used by the object recognition module (ORM), which operates on a small image area at high resolution (70×70), containing the object to be classified. First, we employ a cluster-algorithm with a heuristically determined color threshold to remove the background information. Thus, we separate the object from its background. The resulting cluster is used to shift the fixation point nearer to the center of gravity and to finally determine the input area for the object recognition.

To solve the recognition task we apply a MLP network, using a high dimensional feature vector with image structural and color information. We construct the feature vector using Gauss masks of first- and second order derived *Gabor filters* [Dau85]. By convolving these Gauss filter mask with the *hue*-image (angle of color from the HSI model) we extract the feature vector which encodes color information. Structural and orientational information is extracted by convolving the Gabor filter masks at four different orientations $(0, 45, 90, 135$ degrees). To convolve the masks at equidistant positions on the input area, we choose three different mask sizes to cover the high resolution input image with a minimal overlap over a 1×1, 2×2 and 3×3 field. Thus, we obtain a 168-dimensional feature vector $(3\times4\times(1+4+9))$.

To generate the training feature vectors we take 10 randomly placed images near the objects' positions. With these examples we train the MLP-network using the resilient backpropagation algorithm. We obtain satisfying results with a network dimension of $168-50-15$. The 15 objects are: big plate, small plate, cup, glass, orange, apple, knife, spoon, fork, little cup, dust-bin, butter dish, toast, big glass and carton.

To integrate the hypothesis from the second module (RHM), we use a heuristically constructed evaluation matrix. The rows and columns correspond to the separate classification outputs from the two modules (RHM and ORM). The matrix entries represent the joint probability of the row and column indices. This means, e.g., the hypothesis of cutlery supports the recognition of a fork or a knife. We call this the *combination*-matrix C. Thus, the classification output $\vec{o}$ is obtained by:

$$\vec{o} = \left((C^T \cdot \vec{h}) + (\alpha\ \cdots\ \alpha)^T\right) \cdot \left(\vec{e} + (\beta\ \cdots\ \beta)^T\right) \tag{10}$$

Where the output vector from the RHM is represented by $\vec{h}$ and the ORM output by $\vec{e}$. The constants α and β are used to obtain robustness against possible negative values from the RHM or ORM ($\alpha = 2, \beta = 0.5$). Finally, we apply a *winner-takes-all* rule, i.e. selecting the largest component of $\vec{o}$. If, e.g, a vertically oriented cutlery is detected by the RHM, the detection of a knife or a fork is favored.

5 Results

In this section we analyze the similarity of our model-generated saccades to the data observed during the eye-tracking experiment. Of course, it is very difficult to compare our simple simulation model with the sophisticated mechanism of the human brain. Especially the high-level knowledge influences the reflexive behaviors of the human visual system. However, our eye-tracking experiment is designed to suppress this effect by imposing time-pressure upon the subjects and presenting varying stimuli (see Section 2). The results we obtain from our experiment show that some image features appear to be particularly interesting and immediately attract visual attention.

5.1 Similarities between Model and Eye Tracking Experiment

We use the Markov-path distance measure (MPD; described in Section 3) to detect common features in the gaze trajectories. First, we apply the model to the same visual stimuli used for the eye-tracking experiment. Second, the recorded fixation data of the model are processed in the same way as the eye-tracking data (see Section 2.2): The fixation data are translated into trajectories representing fixations located in object regions—approximated by ellipses—or on the background. Next, we compute three matrices—representing the transition probabilities in each of the three feature categories (objects, colors and shapes)—for different lengths of trajectories. For our comparison, we only use results from trajectories of path length $n = 3$ (c.f. Sec. 3). Next, we multiply each matrix three times by itself to compute the Markov-path matrices (MP). These matrices represent stochastical and geometrical information on the trajectories (path length $n = 3$) generated by our model. In the third step, we compute the Markov-path distance between the MP-matrices referring to our model or a human subject observing an image.

Since we conducted an eye-tracking experiment with 16 subjects watching 20 pairs of images, we acquired a huge amount of data consisting of 960 MP-matrices. Consequently, we only show some typical samples of trajectories produced by our model and the subjects. To

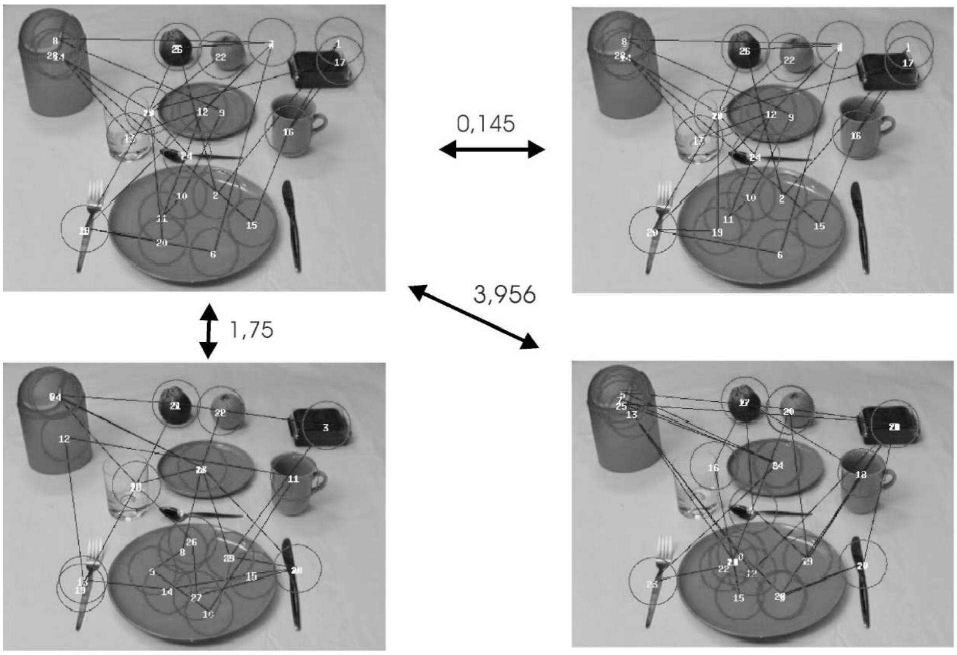


Figure 9: *Example of Markov-path distances between different model trajectories superimposed over the original image.*

visualize the saccades in the corresponding images, the recorded fixation points are numbered (starting with 0) and connected by lines. Fixations are illustrated by circles. Fig. 9 shows an example of four different model trajectories superimposed on the same image and the MPD (calculated on the basis of object identity transition probabilities, see below) between them. The upper left image visualizes the trajectory generated by our model receiving the shown image as input. The other trajectories are created by the same model but with three different images serving as stimuli. We use these trajectories to calculate the object identity transition probabilities for the original image shown in Fig. 9. These artificial examples illustrate the effect of different trajectories on the MPD measure.

While the trajectories of the upper two images are very similar to each other (MPD: 0.145) the difference is larger between the upper left and the lower left image (MPD: 1.75). The butter-dish and the fork, for example, receive different numbers of fixations. An even larger distance is measured between the upper left and the lower right image, since there are many fixations on the bin in the upper left corner (MPD: 3.965).

After analyzing most trajectories with the MPD measure, we find that the principle of similarity of state transitions—corresponding to a certain value of the MPD—can be roughly divided into four classes as shown in Table 5. To obtain a broader basis for our results, we calculate the mean and standard deviation over all trajectories recorded during several experiments.

MPD $\in$	$[0, 1]$	$[1, 2]$	$[2, 4]$	> 4
class	very similar	semi-similar	small (local) similarities	no similarities.

Table 5: *The Markov-path distance heuristically divided into four classes of similarities.*

First, we want to test the quality of the MPD measure. Assuming that during the eye-tracking experiment the subjects explore the images following a common principle, our MPD measure should be in the lower range if it reveals this principle. The MPD is calculated over all 16 subjects watching 20 images from the breakfast scenario. We choose three different kinds of stimuli influencing the eye movement behavior: (i) the objects' identity; the (ii) objects' shape; and (iii) the objects' color (see Section 2). We calculate the MPD with mean and standard deviation for all three kinds of state transitions. The resulting differences between the subjects are shown in the left column of Table 6.

Markov-path distance	between subjects		subjects and model	
	$\varnothing$	σ	$\varnothing$	σ
object identity	1.742	0.36	2.550	0.309
object shape	0.801	0.15	0.618	0.16
object color	0.645	0.21	1.116	0.375

Table 6: *The three different Markov-path distances measured (mean and standard deviation) between the human subjects (left side) and the human subjects and the model (right side).*

This shows that the subjects' trajectories are similar to each other since all three MPD's are below 2 and the standard deviation is small. Especially the transitions on the level of color and shape reveal common strategies of eye movements. In general, certain colors and shapes seem to be particularly attractive to the human eye. Fig. 10 visualizes two trajectories created by different subjects watching the same breakfast image. Since the corresponding object identity MPD measure is 1.61, the trajectories are semi-similar.

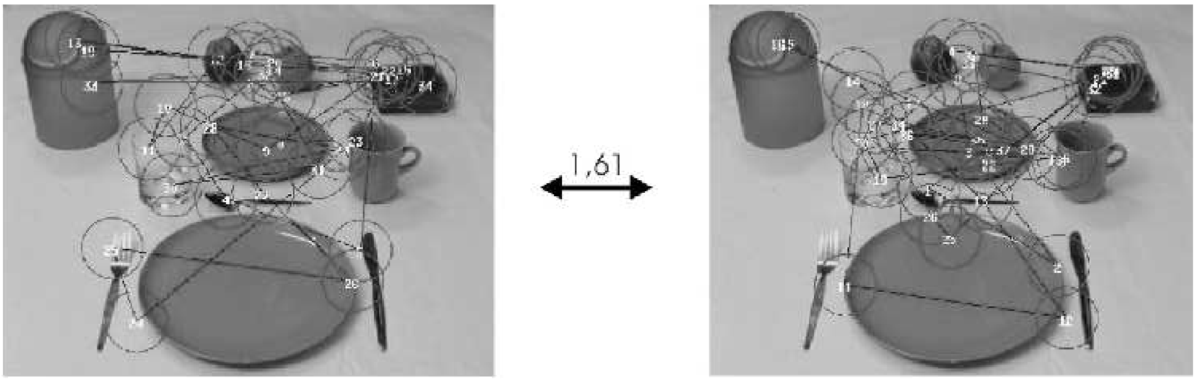


Figure 10: *Typical eye movements of two subjects observing the breakfast scenario. The MPD of 1.61 shows that both subjects exploring the several objects with a similar strategy.*

We now compare the trajectories of the subjects with those of our hierarchical model. For that purpose, we apply the model to the same 20 images. The saccades are recorded and compared with the corresponding trajectories generated by the subjects on the same stimulus. Again, we show the MPD of the three different state transition matrices in Table 6 (right column).

Differences occur mainly at the object identity MPD ($\Delta \approx 0.8$). The MPD's calculated on the level of color and shape show similarities between our model-generated saccades and the subject's eye movements. Fig. 11 shows the difference between the trajectory of one subject (left) and the model (right). In this case, they are semi-similar and the resulting object identity MPD is 1.465. In the following, we explain why there are such substantial differences in the recorded transitions from object to object.

Fig. 11 shows a conspicuous detail: The human subjects sometimes fixate the objects at the peripheral regions while our model always fixates the objects' center. The human eye is able to recognize objects in the peripheral field of view. Our model, however, needs an exact focus point for object recognition. The human observer also employs his knowledge and his memory to avoid saccades to formerly visited image regions. We have already seen this effect of decreasing numbers of saccades during the eye-tracking experiment (see Section 2). In contrast, our model has no knowledge about the objects of the scenario and thus uses only the

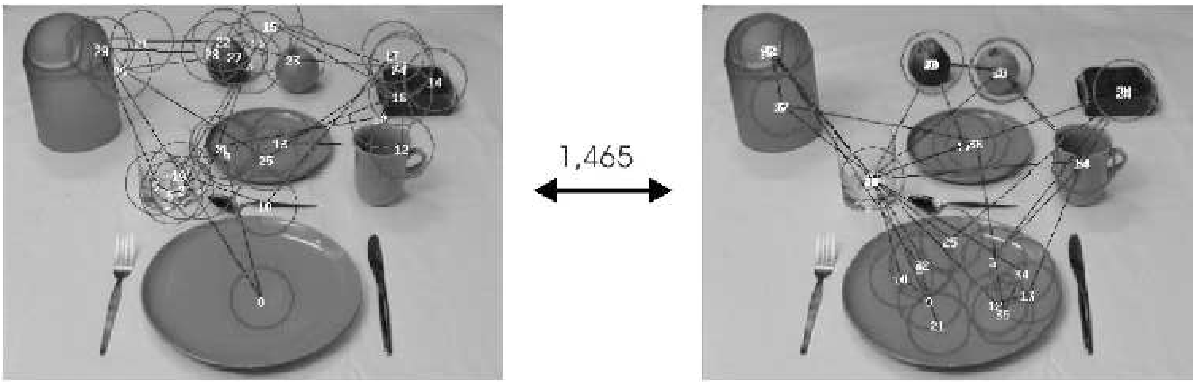


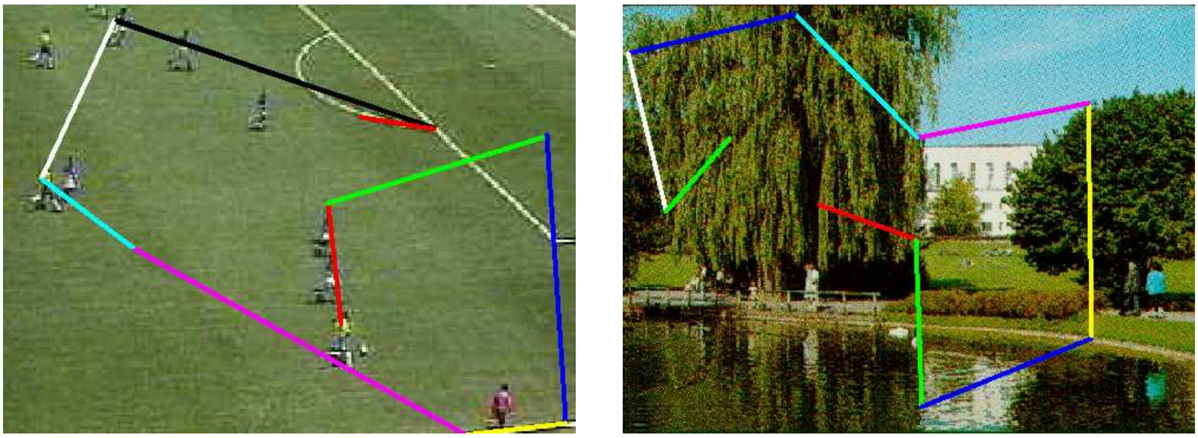
Figure 11: *Saccades of a subject (left) and the model (right). The object identity MPD of 1.45 shows the similarity between the empirical and the simulated eye movements based on the transition probabilities between the objects. The MPD is independent from the frequency of the observation of an object.*

fade-out in the saliency map to simulate a short-term memory. Future developments of our model should integrate higher-level knowledge, e. g. by implementing some rules ("the cup is at the top of the plate") or further learning strategies.

However, the MPD results on the level of color and shape state transitions show a general similarity between the behavior of the human subjects and our model. The weights between the color and shape features in the COM seem to be well adapted. On this level we have a good simulation of the eye movements of a human observer. Thus, our model imitates the common principle of the subjects' visual exploration behavior. This, and the fact that the model-generated saccades are also sufficient for our object recognition scheme (see Section 4.4)—producing an error rate of less than 2%—makes our model suitable for other challenging computer vision tasks.

5.2 Generalization Capabilities of the Model

Here, we show some examples of stimuli different to the breakfast scenario. We applied the same model described above, with the same parameter setting to images with natural scenar-



(a)

(b)

Figure 12: *The model applied to stimuli different to the breakfast scenario.*

ios. First, we choose an image from a tv-screenshot of a soccer game (Fig. 12(a)). Starting from the image center, the model explores the most salient image regions. The other example shows a park picture with many textures and colors (Fig. 12(b)). Again, the system fixates interesting image regions. These examples show the generalization capabilities of our model in different real-world images. Thus, the exploration behavior of the model can be used for other applications such as object recognition or surveillance cameras, since it can reduce the amount of data to be processed.

6 Summary and Outlook

In accordance to biological studies concerning the control of visual attention we have developed an artificial vision system intended to simulate some aspects of the human visual system. For the exploration of complex scenes under real-time constraints, a promising approach is to model saccadic eye movements. Such mechanisms can be found by observing human eye movements. Starting from biological principles, we assume that the saccades are motivated by simple low-level stimuli like edges, contrasts, color thresholds, and movements. A model using only these stimuli as input shows reactions similar to the human saccadic system. From our eye-tracking experiment we obtain empirical data describing some aspects of the human scene exploring behavior. These experiments are based on breakfast scenario images and yields some qualitative aspects supporting the construction of our model and improving its performance. The experimental data leads us to the selection of the biologically motivated stimuli integrated in our attention control model.

To simulate the *where-to-look-next* task we have implemented three hierarchical modules operating on successively smaller fields of view on different resolutions. In the first module (COM) different stimuli are combined in a *cortical map*. The influence of each stimulus depends on a specific weight. These weights are determined by an adaptive algorithm, which is used to simulate the data derived from the empirical results of the eye-tracking experiment. The resulting *cortical map* represents a saliency map, which generates a rough fixation position for the intermediate module (RHM). At a smaller viewing angle but higher resolution the RHM re-centers the fixation point. This shifting mechanism is implemented by a neural network, calculating a “correction saccade” closer to the object’s center. This replacement determines the field of view for a second network which generates a hypothesis for the object’s class. The third module (ORM) calculates the final object recognition output and uses the hypothesis of the second module for further verification of the output. For the main recognition task in the ORM we use again a neural network to classify a feature vector, that is extracted from a region of interest determined by the second module. Using a cluster algorithm, the ORM shifts the input position to the final fixation point near the object’s center. The model memorizes the already visited image positions in an inhibition layer to decrease the values of the *cortical map* at the appropriate positions (“fade out”). The final fixation position marks the point of attention and determines the next simulated saccadic movement.

To compare the trajectories of the model with the empirical data obtained from the eye-tracking experiment, we presented the *Markov-Path-Distance* designed for comparison of trajectories of human subjects. It also indicates similarities between the simulated and the empirical eye movements. The hierarchical model was also tested on several scenarios different from the original breakfast scene, and it shows a robust and reliable behavior. The presented system runs on a common PC workstation with a frame rate of one image (one saccade) per second and was completely developed with NEO/NST[3].

The current recognition tasks depend on a constant background and the combination of the two recognition units is problem-dependent by design. To improve the system, routines for automatic background segmentation and an adaptive system for combining the hypothesis with the object recognition output will be integrated.

The current model operates on static pictures and simulates saccadic human eye movements. By presenting the same selection of pictures to human subjects in the eye-tracking experiment and to the model we measure similarities between the model and the human. Based on this static simulation model we will implement a dynamic system operating on changing scenarios. Our aim is to process a dynamic visual input and to simulate eye movements using an active stereo camera head. A further development is to integrate information on moving objects detected in the scene. This information has a strong influence on the saccades observed in the human visual system and is a part of our current research activities.

In the near future the presented approach will be used for gesture-based interaction. Our aim is to develop an active vision system which automatically explores complex scenes and is sensible for human hand pointing gestures to guide the visual attention of the artificial observer.

[3] NEO/NST is a graphical simulation tool developed by Helge Ritter in the Neuroinformatics Group of the University of Bielefeld.

References

[Ahm91] Subutai Ahmad. *VISIT: An efficient computational model of human visual attention*. PhD thesis, ICSI, Univ. of California, Berkley, 1991.

[Dau85] John G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, July 1985.

[DWTS91] R. Desimone, M. Wessinger, L. Thomas, and W. Schneider. Effects of deactiviation of lateral pulvinar or superior colliculus to selectively attend to a visual stimulus. *Society of Neuroscience Abstracts*, 15:162, 1991.

[Eng96] A.K. Engel. Prinzipien der Wahrnehmung: Das visuelle System. In G. Roth and W. Prinz, editors, *Kopf-Arbeit - Funktionen und kognitive Leistungen*, chapter 5, pages 181–207. Spektrum der Wissenschaften, Berlin, Heidelberg, New York, Tokyo, 1996.

[Fel97] Winfried A. Fellenz. Ein Neuromorphes System für die Datengetriebene Szenenanalyse. *Forschungsberichte VDI, Reihe 10:Informatik/Kommunikationstechnik Nr. 479*, 1997.

[Kat94] H. Kattner. Using attention as a link between low-level and high-level vision. Technical Report TUM-I9439, Mathematisches Institut und Institut für Informatik, TU München, 1994.

[KMKI88] T. Kubota, M. Morimoto, T. Kanaseki, and H. Inomata. Visual pretectal neurons projecting to the dorsal lateral geniculate nucleus and pulvinar nucleus in the cat. *Brain Research Bulletin*, 20:573–579, 1988.

[KSJ95] E.R. Kandel, J.H. Schwartz, and T.M. Jessel. *Essentials of Neural Science and Behaviour*. Appleton and Lange, New York, 1995.

[LB90] D. Laerge and M.S. Buchsbaum. Positron emission tomographic measurements of pulvinar activity during an attention task. *Journal of Neuroscience*, 10:613–619, 1990.

[NK96] E. Niebur and C. Koch. Control of selective visual attention: Modelling the ”where” pathway. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8 NIPS*95*. Bradford MIT Press, 1996.

[Pom98] M. Pomplun. *Analysis and Models of Eye Movements in Comparative Visual Search*. PhD thesis, Universität Bielefeld, Technische Fakultät, 1998.

[PRM87] S.E. Petersen, D.L. Robinson, and J.D. Morris. Contributions of the pulvinar to visual spatial attention. *Neuropsychologia*, 25:97–105, 1987.

[RB95a] R.P.N. Rao and D.H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505, 1995.

[RB95b] R.P.N. Rao and D.H. Ballard. Learning saccadic eye movements using multiscale spatial filters. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, pages 893–900. MIT Press, Cambridge, 1995.

[RP92] D.L. Robinson and S.E. Petersen. The pulvinar and visual salience. *Trends in Neuroscience*, 15:25–42, 1992.

[SS93] M. Swain and M. A. Stricker. Promising directions in active vision. *International Journal of Computer Vision*, 11(2):109–126, 1993.

[Sta93] D.M. Stampe. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods, Instruments, and Computers*, 25:137–142, 1993.

[TG80] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[Yar67] A.L. Yarbus. *Eye movements and vision.* Plenum Press, New York, 1967.