Invited contribution. In: H.-M. Groß et al. (eds.), SOAVE 2004, 3rd Workshop on Self Organization of Adaptive Behavior. Fortschritt-Berichte VDI, Reihe 10, Nr. 743. Düsseldorf: VDI Verlag, 79-93, 2004.

# Vision as an anticipatory process

Fred H. Hamker
Allgemeine Psychologie, Psychologisches Institut II
Westf. Wilhelms-Universität
Fliednerstrasse 21, 48149 Münster
fhamker@uni-muenster.de

**Abstract**

*Vision is a crucial sensor. It provides a very rich collection of information about our environment. The difficulty in vision arises, since this information is not obvious in the image, it has to be constructed. Wheres earlier approaches have favored a bottom-up approach, which maps the image onto an internal representation of the world, more recent approaches search for alternatives and develop frameworks which make use of top down connections. In these approaches vision is inherently a constructive process which makes use of a priory information. Following this line of research, a theory of anticipatory vision is outlined and demonstrated by a computational model of primate vision.*

## 1 Introduction

The brain accomplishes the myriad of visual tasks seemingly without effort, yet we do not understand its machinery. Among others our enormous ability to recognize objects has puzzled researchers for decades. The robustness of our visual system in object recognition has still not been achieved by current models. Object recognition is approached from two perspectives: the one tries to build models for invariant recognition (location invariance, view-point invariance, size invariance) in scenes which contain just a single object [9], [56], [40], [58] whereas another research direction focuses on the recognition problem in cluttered scenes [17], [49], [36] – the ones we typically encounter. Imagine you come late at night home with an awful headache. You know you have an aspirin bottle in the bathroom. So you enter the bathroom and directly look onto the sink (Fig. 1). How might the brain accomplish this task? For example, a visual system could construct an internal world model like a labeled image where each name stands for a 3-D model of the object. Such a full reconstruction of a scene [29] has been criticized in the last years [15] [11]. Findings in Change Blindness experiments indicate that we are even not able to fully memorize the scene in one shot. The brain could be overloaded with all the items in the scene. Object recognition, generally implemented in a hierarchical bottom-up process [13], [33], [56], [40] in which the complexity of representation along with the receptive field size increases, leads to a strong overlapping of populations encoding features belonging to different objects. These ambiguities in cell populations encoding features within the same receptive field limit the use of these approaches for non-segmented object scenes like natural images.

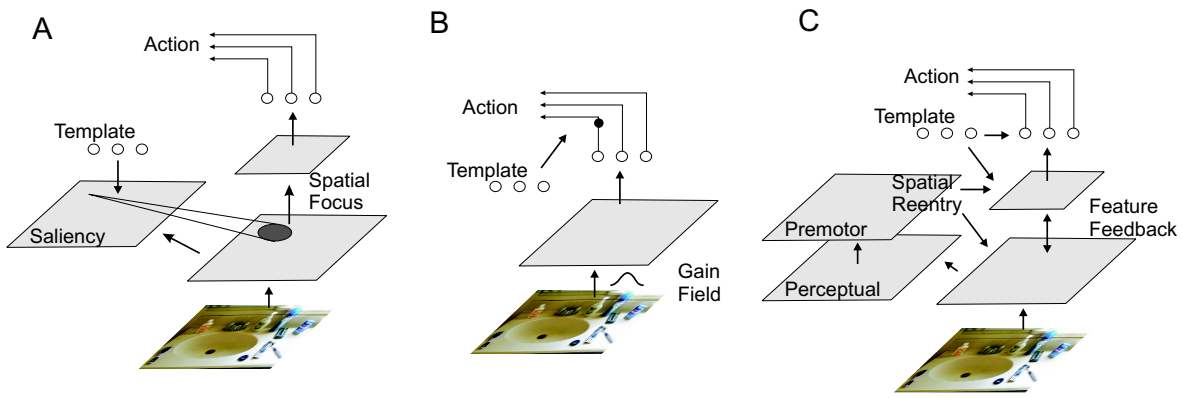


*Figure 1. Three models of perception and object recognition. A) The classical approach of visual attention. An object can be recognized and used for an action, if it is first selected by spatial attention and then routed into higher areas. Templates can be used the modify the saliency map on basis of simple pre-attentive features. B) Space-feature representation. A scene is represented in parallel and then read out by gain fields. Templates can only route the relevant object to action networks. C) Distributed processing with spatial and feature reentry. A template enhances the sensitivity of specific populations encoding the features of interest. Top-down connections send new templates downwards to early areas. Areas responsible for action preparation send a spatially organized signal back to enhance populations encoding stimuli at a specific location.*

The closely linked paradigms of active vision, purposive vision and animate vision [1], [33] have accepted that bottom-up directed vision is an ill-posed-problem and suggested each task requires its own specific algorithm. In this regard, an universal, general vision is not possible. According to these paradigms, the fundamental problem of vision is the selection of the relevant information within the scene and the computation of an appropriate representation. An ”active” vision system is able to acquire the necessary information on demand by focusing on the relevant areas within the visual scene and taking different views from the same object. Regarding visual agents, such as robots, vision provides the necessary information to act in the environment. Within the concept of the action-perception-cycle vision becomes interactive vision.

The approach of ”Deictic Codes for the Embodiment of Cognition” aims to provide a framework for describing the phenomena that appear at about one third of a second in the perception-action process [3]. Deictic primitives dynamically refer to points in the world with respect to their crucial describing features (e.g., color or shape). The outcome of the processing after one-third second, which is the natural sequentiality of body movements can be matched to the natural computational economies of sequential decision systems through a system of implicit reference (called deictic) in which pointing movements are used to bind objects in the world to cognitive programs. Ballard et al. [3] suggested visual routines [25], [54], [23] to divide one complex task into subtasks, such as selection and identification.

Research in attention has focused on how the selection of objects occurs. How could then attention facilitate object recognition? The basic idea is that once an object is selected by a focus of attention it can easily be recognized, because the object is processed preferably and it can be placed onto an internal canvas. This view has its origin in the classical approach of perception that separates between a pre-attentive and attentive stage [51], [24], [59] (Fig. 1A). Computer implementations of these types of models use a saliency map to indicate a location of interest [22] and compute a focus of attention that selects an object [32]. This focus could be guided by some rough knowledge about an object, such as the color of the aspirin bottle. A combination of these attention models with hierarchical

models of object recognition would solve the translation invariance problem, similar as suggested by gain-field [46], [10] and gating models [53], [57] (Fig. 1B).

The discussed paradigms more closely describe human vision and they provide new concepts to solve some ill-posed problems. However, the paradigms of purposive vision and animate vision did not revolutionize models of vision, since one fundamental problem of visual processing within one-third second has not been solved. The crucial issue at this point is that we have no successful algorithms to implement visual routines for difficult vision problems such as object recognition. Most approaches separate between a network that computes attention for control and one that performs object recognition. Object recognition within the focus does not benefit from attention any more. Clutter and different backgrounds typically strongly influence the results. A given scene at a particular point of time is selectively processed but the mechanisms used to acquire selection and recognition only try to match the outcome after one-third second – we are missing implementations of the processes that act within the first 300 ms.

I propose an approach of anticipatory vision to solve the recognition problem inherent to feed-forward solutions. Anticipatory vision is a theory of vision for intentional or goal-directed systems, such as robots. It acts within one third of a second and solves the problem of binding objects in the world to cognitive programs. In this approach vision is an active, dynamic, constructive process. Object recognition is a search according to the task at hand. The task I have to look for the aspirin bottle produces top-down expectations, which meet the bottom-up processed stimulus features in the ventral pathway (Fig. 1C). Recognition is not just a rigid comparison of a target template with an incoming pattern but a flexible process that enhances the features of interest. In parallel, areas responsible for action selection in the fronto-parietal network start to plan appropriate responses. By means of reentry the different specific modules coordinate themselves and tend to work on the same problem. Such a network might dynamically connect the external world to our intention. As a result, the necessary computations to approach the aspirin bottle are processed in the responsible brain areas. Strictly speaking, a representation of the scene in the brain does not exist. The only representation is the scene itself and the brain just connects to it by developing an interpretation of the scene according to the task at hand.

# 2 Anticipatory vision – a formal approach

## 2.1 Computation in one-third second

The proposed concept relies on top-down connections in vision, which have been discussed and its usefulness has been demonstrated for several times [14], [31], [55], [50], [53], [34], [35], [16], [12], [17], [6], [21], [48], [37], [38], [19]. I am now going to outline a formal approach of anticipatory vision with emphasis on object recognition.

The detection problem in one-third second can be formulated as follows: A high level goal description, such as a given category has to be used by prefrontal areas to determine an appropriate target template. Now all feature combinations which are consistent with this template have to be supported. At higher areas these are typically complex patterns. At lower levels these complex patterns have to be decomposed into more simple patterns. Thus, the inference mechanism has to rely on the reverse weights to decompose a pattern into the parts that build the pattern. This powerful mechanism would allow to flexibly filter out the information which is inconsistent with the high level goal description. However, the sensory evidence of the encoded items does not always allow to rule out all objects but one. Typically this inference only strengthens the expected features, which are not necessarily the to be reported ones. Thus, the graded response resulting from this adaptive filter will guide further inspection, for example by focusing on one object. The target location of the planed eye movement is

used for a location specific inference operation which now filters out irrelevant locations. This spatial attention effect could be interpreted as a shortcut of the actual planned eye movement. It facilitates planning processes to evaluate the consequences of the planned action. As a result of both inference operations, the high level goal description is bound to an object in the visual world.

## 2.2 Population-based Inference

On the theoretical level I suggest a population-based inference approach to implement anticipatory vision. Decision making involves uncertainty arising from noise in sensation and the ill-posed nature of perception. Thus, alternative interpretations should be represented until a decision is found. Such constraints can be well handled by a population code. It offers a dual coding principle. A feature is represented by the location of a cell $i$ within the population, and the conspicuity of this feature is represented by a value $r_i$. The conspicuity represents the accumulated evidence and reflects stimulus-driven saliency as well as task relevance. I developed a population-based inference approach to continuously update the conpicuity using prior knowledge in form of generated expectations. The idea is that all mechanisms act directly on the processed variables and modify their conspicuity. Attending a certain feature or a location in space enhances the probability of a feature being detected.

The relevance of each feature is influenced by the search template (target) in each dimension $d$. For simplicity I define the target $\mathcal{T}_d$ by the same sets of features $\mathcal{F}_d$. Thus, a target object is defined by the expected features $\hat{r}^F_{d,i}$. For visual search I infer the conspicuity $r_{d,i,\mathbf{x}}$ by comparing the expected features $\hat{r}^F_{d,i}$ with the observation $r^{\uparrow}_{d,i,\mathbf{x}}$ at each location $\mathbf{x}$ in parallel. If the observation is similar to the expectation the conspicuity is increased. Such a mechanism enhances in parallel the conspicuity of all features which are similar to the target template (Fig. 2).

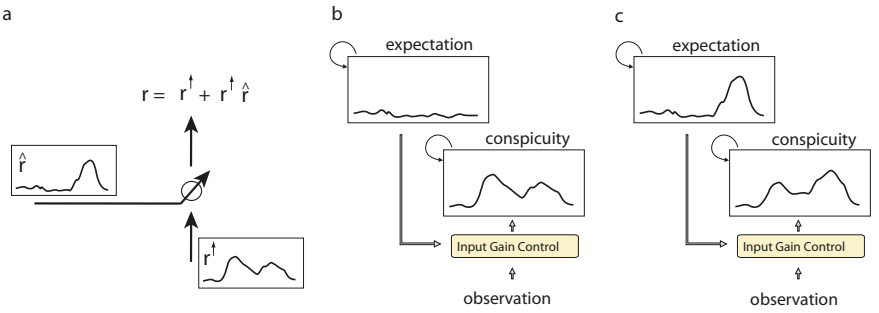


*Figure 2. Concept of the population based inference approach. (a) The expectation $\hat{r}$ acts on the input $r^{\uparrow}$ and increases the gain as depicted by the arrow through the circle. The y-axis encodes the firing rate of cells and the x-axis the feature space, e.g. orientation, color, or location. For simplicity, the feature space of the involved areas is identical. To give an example, the expectation could originate from a population of cells in IT and modulate the conspicuity in V4. (b) Population activity without a significant top-down influence. In this case the content is simply processed in a bottom-up manner. (c) Population activity after top-down expectation multiplicatively increases the gain of the cells and therefore emphasizes a specific pattern (or location). Due to competitive interactions the population response for the non-supported stimuli decreases resulting in a dynamic attention effect. The figure is reproduced from [20] with permission from Cerebral Cortex: Oxford University Press.*

In order to detect an object in space the conspicuity across all $d$ dimensions as well as all $i$ sampling units is combined and an expectation in space $\hat{r}^L_{\mathbf{x}}$ is generated. The higher the individual conspicuity $r_{d,i,\mathbf{x}}$ across $d$ at one location relative to all other locations the higher is the expectation in space $\hat{r}^L_{\mathbf{x}}$

at this location. Analogous to the inference in feature space the expected location $\hat{r}^L_{\mathbf{x}}$ is iteratively compared with the observations $r^{\uparrow}_{d,i,\mathbf{x}}$ in $\mathbf{x}$ and the conspicuity of all features with a similarity of expectation and observation is enhanced. The conspicuity is normalized across each map. Such interative mechanisms finally lead to a preferred encoding of the features and space of interest. Thus, attention emerges by the dynamics of vision.

In the following I will describe how this concept of inference in vision might be implemented in the visual system.

## 2.3 Feedforward pathway

The brain has developed specific functional areas in the visual cortex, which can be divided into two major streams. Form and color travel from V1 through V2, V4 of the occipital lobe into TEO and TE of the inferior temporal lobe [60], [27]. The ventral pathway is known to encode object identity. It is generally accepted that the complexity of encoded features increases along the ventral pathway (Fig. 3). V1 neurons can be driven by simple properties of a stimulus, such as the orientation of a bar. TE neurons, however, encode highly sophisticated shape properties. These ”experts” have probably evolved to meet the statistics of stimuli we typically encounter. The receptive field size has also been suggested to increase along the ventral pathway as well. Most of the receptive field size mappings have been done with anestesized monkeys. The idea is that the increasing receptive field size supports location invariance, since a cell in higher areas is less sensitive to the position of a stimulus. The ventral pathway has been often proposed of having a limited capacity. Referring to a bottleneck, the processing in early stages is supposed to operate in parallel, whereas further processing in higher areas has been proposed to require stimulus selection [51], [24], [59]. Recent findings suggest that these conclusions might depend on the artificial stimuli used in attentional experiments. In statistically richer data sets, such as natural scenes, the similarity between target and distractors is probably much lower than in most of the artificial stimuli used. For example, it has been shown that the detection of animals in natural scenes is easy even in dual-task conditions [26], [43], [44]. These findings could be taken as evidence that we can do all processing of natural scenes within the feedforward sweep. The detection of trained and thus familiar objects placed into natural scenes, however, requires monkeys to serially search for the target [47] .

Although visual processing seems to rely on a massively parallel feedforward pathway, at least two factors limit the capacity of feedforward processing: increasing receptive field size and competition for visual short term memory. The latter seems to have a roughly fixed capacity of about four items [28]. I propose that the former capacity limit rather depends on the statistics of the stimuli used and their spatial arrangement than it reflects a general capacity limitation. As long as the overlapping of the neural populations encoding target and clutter is low, detection can be done within the feedforward sweep at least without extensive use of feedback loops. In other cases reentry is required to tune the visual system dynamically towards the relevant information in the visual scene.

## 2.4 Feature-specific reentry

The capacity limitation in the visual system probably arises to force the system to decide between alternative interpretations on the high level scene interpretation and low level receptive field competition. The competition at the level of a high level scene interpretation becomes obvious in bistable figures. The simultaneous perception of both interpretations is impossible, the percept switches from one to the other but they are never perceived that the same time. At the local level competition within the receptive field could led to a competition at the neural level among competing patterns. In order to give priority to one pattern over the other, e.g. by attending to the location of one stimulus, a top-down bias was proposed [7], [39]. Similarly, a feature-based mechanism allows to emphasize one

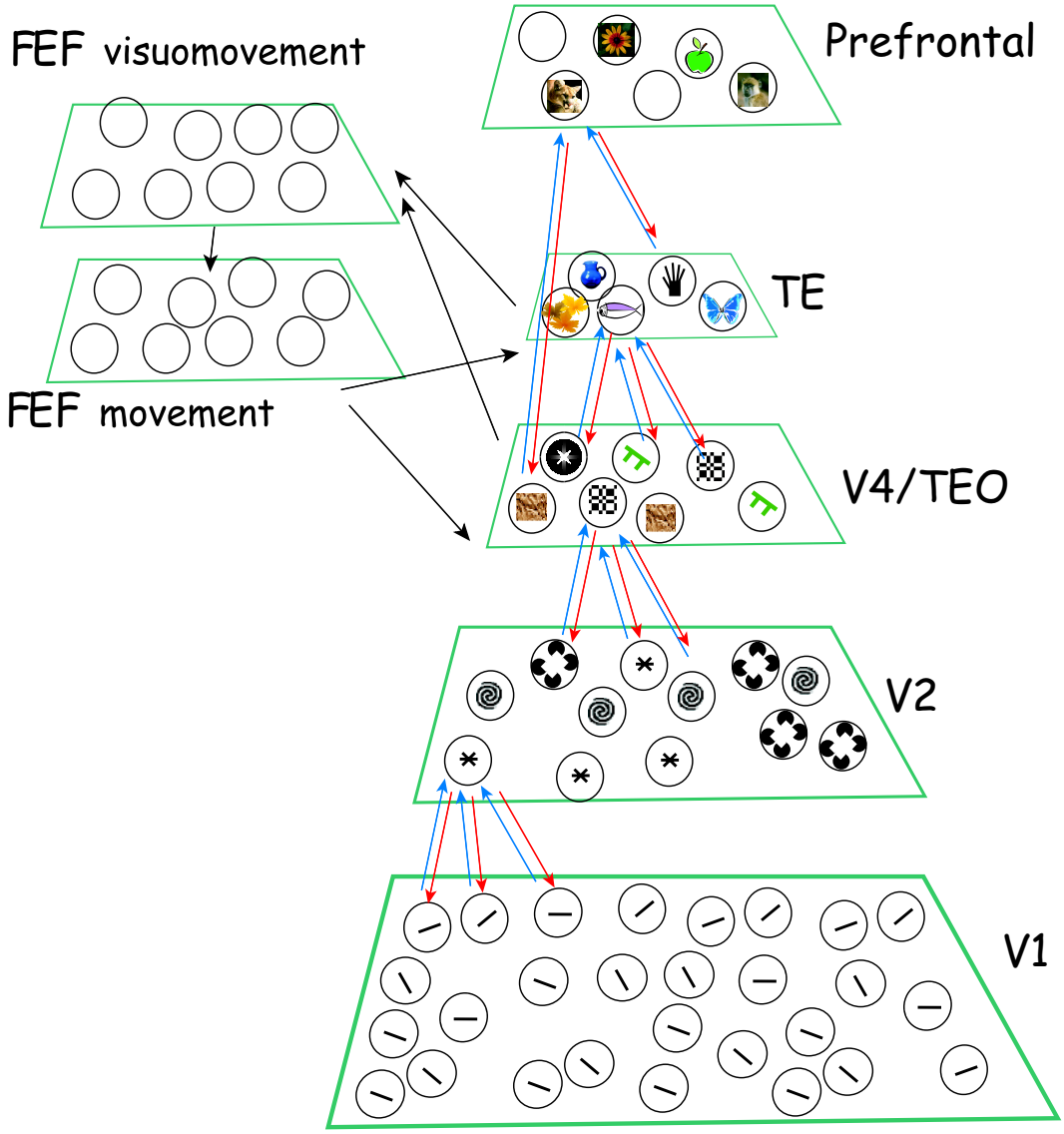


*Figure 3. A formal approach to anticipatory vision.*

pattern over others [5], [52]. Thus, more general, feedback allows to resolve ambiguities and to reveal visual details. Extending the idea of a mere ”attentional” bias, I propose that a target template travels the ventral pathway downwards via massive feedback connections [41], [42] and enhances the firing rate of cells supporting the target template by the proposed inference approach. This mechanism implements a dynamic filter to compute the relevant properties within a general purpose machinery of localized experts.

The target template is generated in prefrontal areas. I hypothesize, it is not limited to simple properties, such as color, it can be highly complex. The prefrontal cortex (PFC) has a vital role in the control of goal-directed behavior. Recent evidence suggests that the prefrontal cortex relies on an adaptive neural coding [8] which could compute and provide a rich target template given the context of the present task to perform. Thus, I suggest that the prefrontal cortex guides visual perception by generating an appropriate target template in time, which is then used for an inference mechanism implemented by the visual system.

## 2.5 Spatial reentry

I have proposed that occulomotor areas responsible for planning an eye movement, such as the frontal eye field, influence perception prior to the eye movement [18]. The activity reflecting the planning of an eye movement reenters the ventral pathway and provides a spatially selective expectation used in the described inference mechanism. In the following, I focus on the FEF as a putative source of this reentry signal [20]. The FEF has connections to occipital, temporal and parietal areas, the thalamus, superior colliculus and prefrontal cortex [45]. It can be subdivided into a lateral and medial part. The lateral FEF, which generates short and precise saccades is connected to the dorsal (LIP, MT, MST, V3) and ventral (TEO, V4, V2) pathways, the ventrolateral prefrontal cortex [45]. The projections from V2 and V3 are weak, while the one from V4 are intermediate. Strong projections from TEO, MT and MST suggest that the FEF uses features after several stages of processing for target selection [45]. The neurons in the FEF can be categorized based on both their responses to visual stimuli and to saccade execution into visual, visuomovement, fixation and movement cells [4], [45].

By means of reentry into extrastriate visual areas from the FEF, neurons in V4 and IT that have their receptive fields at the location of an intended eye movement could increase their sensitivity by the proposed inference mechanism. There is some recent experimental evidence that the source of the reentry signal is indeed the FEF [30]. Simulations specifically support the movement cells in the FEF as an ideal reentry source [20].

# 3 A model

The validity of the outlined framework is now demonstrated by a computational model on an object detection task in a naturalistic image. The areas, their interactions and the temporal activity of the cells are grounded in neuroanatomy and neurophysiology [20]. I demonstrate how a top-down directed expectation, which alters the gain of the feedforward signal, modulates the vision process to allow for detecting an object in a natural scene. The detection problem is simplified in two aspects. Firstly, the high level goal description is not constructed by the model on its own but a target template is presented to the model. Secondly, the target template is defined only in low level feature space. This constraint occurs, since the complexity of the feature space does not increase along the models "what" pathway.

## 3.1. Overview

Each feature set is modeled as a continuous space with $i \in N$ cells at location $\mathbf{x} = (x_1, x_2)$ by assigning each cell a conspicuity $r_{d,i,\mathbf{x}}$. From the feature maps we determine the contrast maps according to a measure of stimulus-driven saliency (Fig. 4). Feature and contrast maps are then combined into feature conspicuity maps which encode the feature and its initial conspicuity by means of a population code. The mechanisms in the model act directly on the processed variables encoded by the population and modify their conspicuity.

The conspicuity of each encoded feature is altered by the target template. A target object in prefrontal cortex is defined by the expected features $\hat{r}^{PFC}_{d,i}$. We infer the conspicuity of each feature in TE $r^{TE}_{d,i,\mathbf{x}}$ by comparing the expected features $\hat{r}^{PFC}_{d,i}$ with the bottom-up signal $r^{TE\uparrow}_{d,i,\mathbf{x}}$. If the bottom-up signal is similar to the expectation we increase the conspicuity. Such a mechanism enhances in parallel the conspicuity of all features in TE which are similar to the target template. The same procedure is performed in V4 where the expected features are those from TE.

In order to detect an object in space the conspicuity across all $d$ channels in the perceptual map is combined and an expectation in space $\hat{r}^{FEFm}_{\mathbf{x}}$ is generated in the movement map. The higher

the individual conspicuity $r_{d,i,\mathbf{x}}$ across $d$ at one location relative to all other locations the higher is the expectation in space $\hat{r}^{FEFm}_{\mathbf{x}}$ at this location. Thus, a location with high conspicuity in different channels $d$ tends to have a high expectation in space $\hat{r}^{FEFm}_{\mathbf{x}}$. Analogous to the inference in feature space the expected location $\hat{r}^{FEFm}_{\mathbf{x}}$ is iteratively compared with the bottom-up signal $r^{\uparrow}_{d,i,\mathbf{x}}$ in $\mathbf{x}$ and the conspicuity of all features with a similarity of expectation and bottom-up signal is enhanced. The conspicuity is normalized across each map by competitive interactions. Such interative mechanisms finally lead to a preferred encoding of the features and space of interest.

**Early visual processing:** We compute feature maps for Red-Green opponency ($RG$), Blue-Yellow opponency ($BY$), Intensity ($I$), Orientation ($O$), and Spatial Resolution ($\sigma$). We determine the initital conspicuity by center-surround operations [22] from the feature maps which gives us the contrast maps. The feature-conspicuity maps combine the feature and conspicuity into a population code, so that at each location we encode each feature and its related conspicuity.

**V4:** V4 has $d$ channels which receive input from the feature conspicuity maps: $r_{\theta,i,\mathbf{x}}$ for orientation, $r_{I,i,\mathbf{x}}$ for intensity, $r_{RG,i,\mathbf{x}}$ for red-green opponency, $r_{BY,i,\mathbf{x}}$ for blue-yellow opponency and $r_{\sigma,i,\mathbf{x}}$ for spatial frequency (Fig. 4). The expectation of features in V4 originates in TE $\hat{r}^{V4_F}_{d,i,\mathbf{x}'} = r^{TE}_{d,i,\mathbf{x}}$ and the expected location in the movement map $\hat{r}^{V4_L}_{\mathbf{x}'} = r^{FEFm}_{\mathbf{x}'}$. Please note that even TE has a coarse dependency on location.

**TE:** The features with their respective conspicuity and location in V4 project to layer TE, but only within the same dimension $d$, so that the conspicuity of features at several locations in V4 converges onto one location in TE. A map containing 9 populations with overlapping receptive fields is simulated. The complexity of features from V4 to TE is not increased. The expected features in TE originate in the target template $r^{TE_F}_{d,i,\mathbf{x}} = w \cdot r^{PFC}_{d,i}$ and the expected location in the movement map $\hat{r}^{TE_L}_{\mathbf{x}} = w \cdot r^{FEFm}_{\mathbf{x}}$

**FEF visuomovement cells:** The FEF visuomovement cells (FEFv) indicate salient locations by integrating the conspicuity of V4 and TE across all channels. In addition to the the conspicuity in V4 and TE the match of the target template with the features encoded in V4 is considered by computing the product $\prod_d \max_i r^{PFC}_{d,i} \cdot r^{V4}_{d,i,\mathbf{x}}$. This implements a bias to locations with a high joint probability of encoding all searched features in a certain area.

**FEF movement cells:** The projection of the visuomovement cells onto the movement cells (FEFm) transforms the salient locations into a few candidate locations which provide the expected location for V4 and TE units. This is achieved by subtracting the average saliency from the saliency at each location $w^{FEFv} r^{FEFv}_{\mathbf{x}} - w^{FEFv}_{inh} \sum_{\mathbf{x}} r^{FEFv}_{\mathbf{x}}$. Simultaneously, the movement units indicate the target location of an eye movement.

# 4 Results

I now demonstrate how the principle of inference facilitates object detection in cluttered scenes (Fig. 5). I present an object to the model for 100 ms and let it memorize some of its features as a target template. I do not give the model any hints which feature to memorize. The model has to identify in parallel the location where the features of the target template sufficiently match the encoded features. The inference operation in the model allows the binding of features encoded in high level areas (TE)

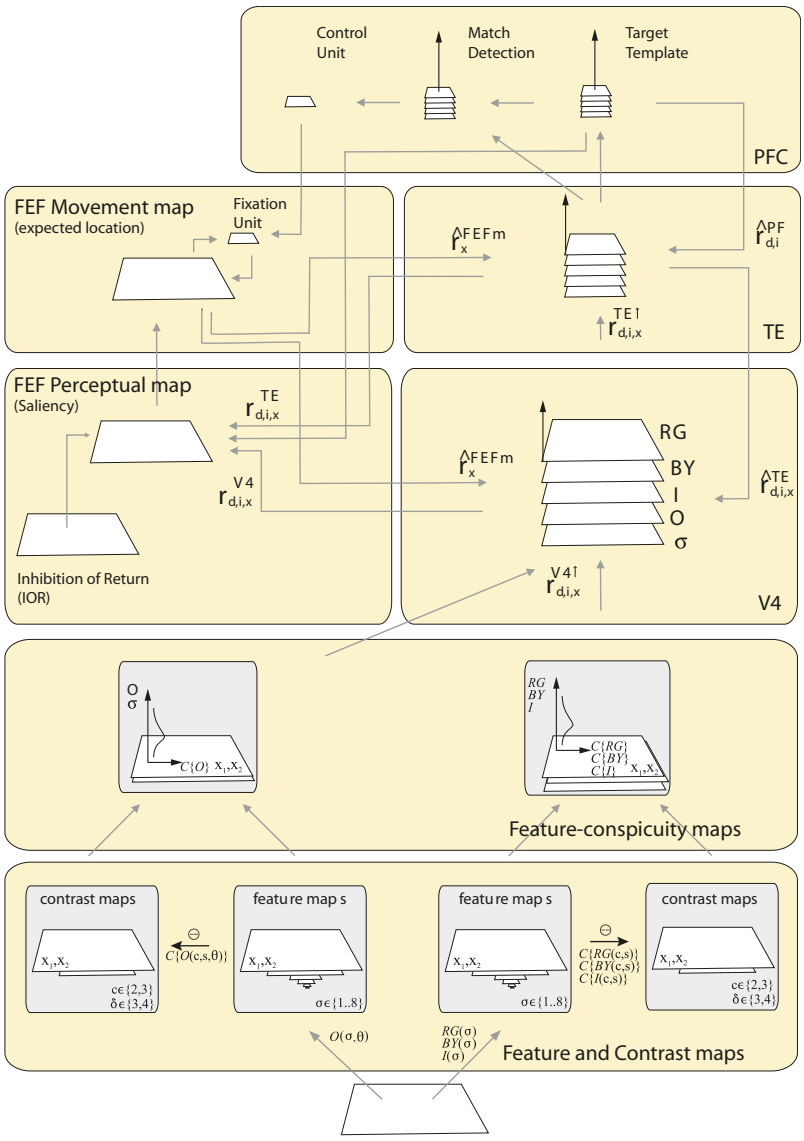


*Figure 4. Model of anticipatory vision. From the image, 5 feature maps (RG, BY , I, O, σ) are obtained. For each feature at each location* $\mathbf{x}$ *the conspicuity in the contrast maps is computed. The feature-conspicuity maps combine the feature and conspicuity into a population code, so that at each location a feature and its related conspicuity is encoded. This initial, stimulus-driven conspicuity is now dynamically updated within a hierarchy of levels. From V4 to TE a pooling across space is performed to gain a representation of features with a coarse coding of location. The target template holds the to be searched pattern regardless of its location. It represents the expected features* $\hat{r}_{d,i}^{PFC}$ *which are used to compute the (posterior) conspicuity in TE. Similarly TE represents the expectation for V4. As a result, the conspicuity of all features of interest is enhanced regardless of their location. In oder to identify candidate objects by their saliency the activity across all 5 channels is integrated. The saliency is then used to compute the expected locations of an object* $\hat{r}_{\mathbf{x}}^{FEFm}$*, which in turn enhances the conspicuity of all features in V4 and TE at these locations. Thus, objects at expected locations are preferably represented. By comparing the conspicious features in TE with the target template in the match detection it is possible to continuously track if the object of interest is at the expected location. If the match is lost an inhibition of return is triggered which marks the expected location as being visited. Otherwise the expectation increases until an overt shift occurs.*

with only crude location information with the same features at a lower level (V4) with intermediate location information. The difficulty of this search task lies in the heavy overlapping of the target with distractors. The template, e.g. the color and orientation of edges provides only little evidence for the target, since many other objects will show at least a partial match with the target template. The reason is that the present model does not use the highly sophisticated feature detectors that humans use. Thus, the difficulty of the search task for the model cannot be directly compared to the difficulty humans have in the same scene. The search task itself, however, describes a similar problem that humans face if they have to search for stimuli in natural scenes.

When presenting the search scene, TE cells that match the target template quickly increase their activity to guide V4 cells. Thus, the feature-specific inference mechanism implements a dynamic filter. It emphasizes all features throughout the ”what” pathway from higher areas to lower areas which are consistent with the target template. As a result the model produces patches of enhanced feature and location specific activity. This information allows to guide the planning of the saccade in the FEF. The process of planning an eye movement provides the expectation for a location specific inference. This, now implements a decision process by enhancing the gain of all cells at the target location of the intended eye movement.

I now specifically demonstrate that the feature-based inference operation can be beneficial for object detection. The orientation filter responses of the lighter in isolation are vertical at the left and right corner and close to horizontal in the middle (Fig. 5 A). The model has obviously no problem in detecting the lighter in a cluttered scene (Fig. 5 B), although it memorized only the slightly tilted orientation (Fig. 5 C). However, the spatial focus (Fig. 5 D) increases the conspicuity of all features within its area, so that the vertical edge of the cigarette box gets dominant as well (Fig. 5 C, E). The V4 conspicuity in the orientation channel initially exhibits a dominance for slightly tilted horizontal edges due to the top-down feature-based inference (Fig. 5 E). The emergence of a spatial focus (Fig. 5 D), however, increases the conspicuity of all features within its area. Thus, a spatial focus of attention does not sufficiently resolve the interference of distractors. In densely cluttered scenes features from distractors are enhanced as well. A purely feedforward approach of object recognition could be impaired by the clutter. The feature-based inference facilitates the detection, since knowing the target features keeps those dominant against the influence of distractors, so that even when distractor features become conspicuous the target features remain represented to allow a match. Thus, feature-based inference serves as a cue to ”segment” the object features in feature-space similar in effect to cue based region segmentation.

# 5 Discussion

I have outlined a theory of anticipatory vision. Anticipatory vision provides a concept of binding objects to cognitive programs. Such binding is necessary to ensure that all parallel modules from action selection to recognition operate on the same event. In this respect, it is essential to recognize that a vision system must have the capability to dynamically change its internal filter. The paradigm of anticipatory vision localizes the recognition problem not solely at a high level of abstractions but at many levels of abstraction.

I have suggested a population based inference approach to implement the theory of anticipatory vision into a model of primate vision. This model has been demonstrated to be able to detect objects in natural scenes. I have specifically shown the influence of feature-based inference. Rather than selecting just the location of an object by some spatial mechanism, features of interest are dynamically enhanced. The definition of a task, e.g., looking for red items, is reflected by an enhancement of cells encoding these properties. An activity landscape is therefore constructed according to internal goals which is not equivalent to a representation of the external world.

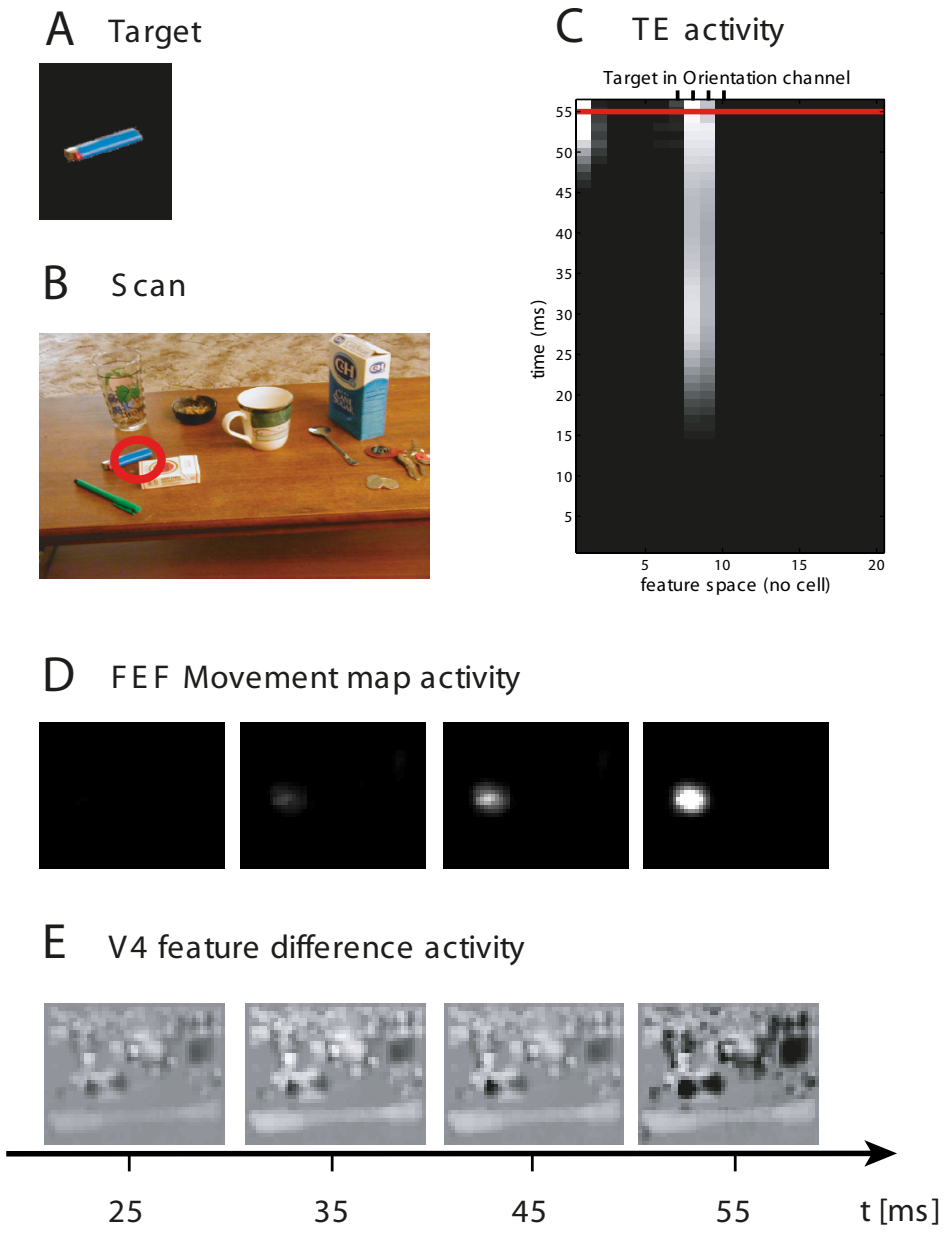


*Figure 5. Illustration of effect of feature-based and spatial inference. A) Target object. B) Target detection in the scene. C) Conspicuity of each cell in the orientation orientation channel in TE over time. The target orientation encoded in PFC is shown by the tick marks at the top. Initially the conspicuity in only high for the target features. Prior to the correct detection a non-target feature (vertical orientation) raises its conspicuity. D) The spatial expectation builds up over time and it is directed towards the location of the target. E) Feature difference of the V4 conspicuity. At each location, the activity for vertical orientation is subtracted from the feature activity of the orientation of the lighter to illustrate the relative strength of the target feature. The strength of the conspicuity is scaled by the intensity. Dark areas represent the activity of the distracting feature – in this case the vertical orientation. Although the vertical edge from the cigarette box becomes dominant by the spatial inference operation, the feature-based inference facilitates the match detection of the lighter such that the lighter can be detected up to 55 ms after scene onset even in the presence of a strong distractor.*

Despite recurrent interactions in the system, the detection of objects is still fast. Processes in the model do not have to wait until a selection process in one area is settled, they can immediately start to plan an action or make a decision on basis of a slight preference in the input. Thus, the concept of reentry does not require extensive recurrent processing to achieve a solution.

The present model argues for two lines of future research. Firstly, a usage of this model in autonomous robots demands that the goals of vision have to be internally constructed. The present model has no knowledge what to search for. Thus, an object is presented as a cue, and the model memorizes its features as a target template. I suggest that the use and generation of templates is a natural way of perception. Learning to see means learning to generate appropriate templates from the context of a scene. Thus, approaches that allow to generate such templates will be part of future research.

Secondly, the present model uses only simple features as detectors of visual properties. However, object recognition requires a much richer set of detectors. If we want to incorporate those into the model we have to ensure that the feedforward and feedback connections are consistent with each other. Learning appropriate feedforward and feedback connections by the statistics of the visual scene would allow to generate consistent complex feature detectors. With such an extension the model would be able to shed more light on the puzzling issues of perception in natural scenes. Furthermore, it could provide an interesting alternative in computer vision tasks which require a high flexibility of vision – any knowledge about an object can be used to increase the conspicuity of object features without the need of an engineer to pre-determine the cue information.

# References


[1] Y. Aloimonos. Introduction: Active Vision Revisited. In: *Acitve Perception*. Lauwrence Erlabaum Associates, 1-18, 1993.

[2] D. Ballard. Annimate Vision. *Artif Intell*, 48:57-86, 1991.

[3] D. Ballard, M. M. Hayhoe, P. K. Pook. Deictic codes for the embodiment of cognition. *Behav Brain Sci*, 20:723-742, 1997.

[4] C. J. Bruce, M. E. Goldberg. Primate frontal eye fields. I. Single neurons discharging before saccades. *J Neurophysiol*, 53:603-635, 1985.

[5] L. Chelazzi, J. Duncan, E. K. Miller, R. Desimone. Responses of neurons in inferior temporal cortex during memory-guided visual search. *J Neurophysiol*, 80:2918-2940, 1998.

[6] S. Corchs, G. Deco. Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data. *Cereb Cortex*, 12:339-348, 2002.

[7] R. Desimone, J. Duncan. Neural mechanisms of selective attention. *Annu Rev Neurosci*, 18:193-222, 1995.

[8] Duncan J. An adaptive coding model of neural function in prefrontal cortex. *Nat Rev Neurosci*, 2:820-829, 2001.

[9] S. Edelman. *Representation and recognition in vision*. Cambridge, Mass. : MIT Press, 1999.

[10] S. Edelman, N. Intrator. A productive, systematic framework for the representation of visual structure. In: *Advances in neural information processing Systems (Vol. 13)*. Leen, TK et al., ed., MIT Press, 10-16, 2001.

[11] S. Edelman. Constraining the neural representation of the visual world. *Trends Cogn Sci*, 6:125-131, 2002.

[12] A. K. Engel, P. Fries, W. Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nat Rev Neurosci*, 2:704-716, 2001.

[13] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biol Cybern*, 36:193-202, 1980.

[14] S. Grossberg. How does the brain build a cognitive code? *Psychol Rev* 87:1-51, 1980.

[15] F. H. Hamker, H. M. Gross. Intentionale Aufmerksamkeit: Ein alternatives Konzept für technische visuo-motorische Systeme. In: *Proceedings des Workshops der GI-Fachgruppe 1.0.4 Bildverstehen "Aktives Sehen in technischen und biologischen Systemen"*, Hamburg, 101-108, 1996.

[16] F. H. Hamker. The role of feedback connections in task-driven visual search. In: D. Heinke, G. W. Humphreys, A. Olson (Eds.), *Connectionist Models in Cognitive Neuroscience*. London: Springer Verlag, 252-261, 1999.

[17] F. H. Hamker, J. Worcester. Object detection in natural scenes by feedback. In: H.H. Bülthoff et al. (Eds.), Biologically Motivated Computer Vision. Lecture Notes in Computer Science. Berlin, Heidelberg, New York: Springer Verlag, 398-407, 2002.

[18] F. H. Hamker. The reentry hypothesis: linking eye movements to visual perception. *J Vis*, 11:808-816, 2003.

[19] F. H. Hamker. A dynamic model of how feature cues guide spatial attention. *Vis Res*, 44:501-521, 2004.

[20] F. H. Hamker. The Reentry Hypothesis: The Putative Interaction of the Frontal Eye Field, Ventrolateral Prefrontal Cortex, and Areas V4, IT for Attention and Eye Movement. *Cereb Cortex*, in press.

[21] S. Hochstein, M. Ahissar. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, 36:791-804, 2002.

[22] L. Itti, C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res., 40:1489-1506, 2000.

[23] M. A. Just, P. A. Carpenter. Eye fixations and cognitive processes. *Cog Psychol*, 8:441-480, 1976.

[24] C. Koch, S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Psychol*, 4:219-227, 1985.

[25] S. M. Kosslyn. *Image and Brain. Cambridge*, MA: MIT Press (A Bradford Book), 1994.

[26] F. F. Li, R. VanRullen, C. Koch, P. Perona. Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci USA*, 99:9596-9601, 2002.

[27] M. S. Livingstone, D. H. Hubel. Segregation of form, color, movement and depth. *J Neurosci*, 7:3416-3468, 1988.

[28] S. J. Luck, E. K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390:279-281, 1997.

[29] D. Marr. *Vision*. San Francisco: Freeman, 1980.

[30] T. Moore, K. M. Armstrong. Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, 421:370-373, 2003.

[31] D. Mumford. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern*, 66:241-251, 1992.

[32] B. Olshausen, C. Anderson, D. van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J Neurosci*, 13:4700-4719, 1993.

[33] D. I. Perrett, M. W. Oram, The neurophysiology of shape processing, *Image and Vis Comp*, 11:317-333, 1993.

[34] R. P. Rao, D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*, 2:79-87, 1999.

[35] R. P. Rao. An optimal estimation approach to visual perception and learning. *Vis Res*, 39:1963-1989, 1999.

[36] R. P. Rao, G. J. Zelinsky, M. M. Hayhoe, D. H. Ballard. Eye movements in iconic visual search. *Vis Res*, 42:1447-1463, 2002.

[37] R. P. Rao, T. J. Sejnowski. Self-organizing neural systems based on predictive learning. *Philos Transact Ser A Math Phys Eng Sci*, 361:1149-75, 2003.

[38] R. P. Rao. Bayesian computation in recurrent neural circuits. *Neural Comput*, 16:1-38, 2004.

[39] J. H. Reynolds, L. Chelazzi, R. Desimone. Competetive mechanism subserve attention in macaque areas V2 and V4. *J Neurosci*, 19:1736-1753, 1999.

[40] M. Riesenhuber, T. Poggio. Hierarchical models of object recognition in cortex, *Nat Neurosci*, 2:1019-1025, 1999.

[41] K. S. Rockland, G. W. van Hoesen. Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey. *Cereb Cortex*, 4:300-313, 1994.

[42] K. S. Rockland, K. S. Saleem, K. Tanaka. Divergent feedback connections from areas V4 and TEO in the macaque. *Vis Neurosci*, 11:579-600, 1994.

[43] G. A. Rousselet, M. Fabre-Thorpe, S. J. Thorpe. Parallel processing in high-level categorization of natural images. *Nat Neurosci*, 5:629-630, 2002.

[44] G. A. Rousselet, S. J. Thorpe, M. Fabre-Thorpe. Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vis Res*, 44:877-894, 2004.

[45] J. D. Schall, A. Morel, D. J. King, J. Bullier. Topography of visual cortex connections with frontal eye field in macaque: Convergence and segregation of processing streams. *J Neurosci*, 15:4464-4487, 1995.

[46] E. Salinas, L. F. Abbott. Invariant visual responses from attentional gain fields, *J Neurophysiol*, 77:3267-3272, 1997.

[47] D. L. Sheinberg, N. K. Logothetis. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J Neurosci*, 21:1340-1350, 2001.

[48] V. Stephan, H. M. Gross. Visuomotor anticipation - a powerful approach to behavior-driven perception. *Künstl Intell*, 2:12-17, 2003.

[49] S. Thorpe. Ultra-Rapid Scene Categorisation with a Wave of Spikes. In H.H. Bülthoff et al (eds), *Biologically Motivated Computer Vision*, Lecture Notes in Computer Science, 2525, Springer-Verlag, Berlin, 1-15, 2002.

[50] G. Tononi, O. Sporns, G. Edelman. Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. *Cereb Cortex*, 2:310-335, 1992.

[51] A. Treisman, G. Gelade. A feature integration theory of attention. *Cog Psychol*, 12:97-136, 1980.

[52] S. Treue, J. C. Martínez-Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399:575-579, 1999.

[53] J. K. Tsotsos, S. M. Culhane. Wai, W., Lai, Y., Davis, N., Nuflo, F. Modeling visual attention via selective tuning. *Artif Intell*, 78:507-545, 1995.

[54] S. Ullman. Visual Routines, *Cognition*, 18:97-157, 1984.

[55] S. Ullman. Sequence seeking and counter streams: A computational model for bidirectional flow in the visual cortex. *Cereb Cortex*, 5:1-11, 1995.

[56] G. Wallis, E.T. Rolls. Invariant face and object recognition in the visual system. *Prog Neurobiol*, 51:167-194, 1997.

[57] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, C. Koch. Attentional Selection for Object Recognition - a Gentle Way. In: H.H. Bülthoff et al. (Eds.), *Biologically Motivated Computer Vision*. Lecture Notes in Computer Science. Berlin, Heidelberg, New York: Springer Verlag, 472-479, 2002.

[58] J. Wieghardt, C. von der Malsburg. Pose-Independent Object Representation by 2-D Views. *Biologically Motivated Computer Vision*, 276-285, 2000.

[59] J. M. Wolfe. Guided search 2.0: A revised model of visual search. *Psycho Bull*, 1:202-238, 1994.

[60] S. Zeki. Functional specialization in the visual cortex of the rhesus monkey. *Nature*, 274:423-428, 1978.