Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Computer Vision and Image Understanding 100 (2005) 64–106

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision

Fred H. Hamker [*]

Allgemeine Psychologie, Psychologisches Institut II, Westf. Wilhelms-Universität, Fliednerstrasse 21, 48149 Münster, Germany

Received 8 January 2004; accepted 29 September 2004
Available online 21 July 2005

Abstract

Technologies such as video surveillance and vision guided robotics require flexible vision systems that interpret the scene according to the current task at hand. Attention has been suggested to play an important role in the process of scene understanding by prioritizing relevant information. However, the underlying processes that allow cognition to guide vision have not been fully explored. Our procedure has its origin in current findings of research in attention. We suggest an approach in which high-level cognitive processes are top-down directed and modulate stimulus signals such that vision is a constructive process in time. Prior knowledge is combined with the observation taken from the image by a population-based inference in order to dynamically update the conspicuity of each feature. Any decision, such as object detection, is based on these distributed conspicuities. We demonstrate this concept on a goal-directed object detection task in natural scenes.

Keywords: Attention; Natural scenes, Object detection; Object recognition; Cognitive control; Top-down inference; Computational neuroscience

[*] Fax: +49 251 8334180.
E-mail address: fhamker@uni-muenster.de.

1077-3142/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2004.09.005

1. Introduction

Emerging new technologies require vision systems to flexibly focus on the relevant information in a visual scene. The processing of a full scene in parallel up to a high level description has turned out to be problematic and not fully successful in computer vision [1–3]. Attention might be the solution. The idea is to control the information flow and thus to improve vision by focusing the resources merely on aspects of the whole visual scene. Early, preattentive, and parallel vision modules compute a set of basic features from the scene, which are then attentively integrated and further processed. Such attentive processing has been described as a 'spotlight of attention' [4] that highlights an area of interest by routing that information into higher areas for further processing [5–7]. The guidance of an attentional focus can be implemented by a winner-takes-all process within a saliency map which indicates potentially relevant locations [6]. Based on this paradigm, sophisticated models of information control have been developed, in which the complex problem of scene understanding is transferred into a sequential analysis of image parts. Such spatial selection is computationally efficient [1,2]. However, we have to consider other crucial issues of efficiency as well. First of all, selection should be effectively guided by the task at hand. It would be problematic if we had to scan several salient items before focusing the relevant item. Thus, we have to elaborate mechanisms that integrate high level knowledge into the selection process. Second, the mechanisms of attention must result in a representation that facilitates further processing. For example, object recognition in natural scenes would not benefit a lot if we simply determined a point in space by some competitive mechanism. Even a region of interest can be problematic if it does not sufficiently cover the object of interest. Thus, we need forms of selection that enhance the features of an object in space.

The most crucial issue of attention deals with the integration of information from different modules (or brain areas). This could be implemented as a central process that collects the information from different modules and then controls those by a single attentional signal. For example in the Guided Search framework [8] a bottom-up map is combined with a top-down map in order to determine the activity in a master map of locations. The location of the highest activity could then be used to control a single attentional focus. An alternative has been outlined by the integrated competition hypothesis [9], in which different specialized modules (or brain areas) have to coordinate themselves to let a distributed system operate on the same event. We present an approach that follows this idea. At its core is a population code that encodes in a dual coding principle a feature and its respective conspicuity. The term conspicuity here reflects stimulus-driven saliency as well as task relevance and relates to the probability that a feature is present in the scene. We developed a population-based inference approach to continuously update the conspicuity 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 region in space enhances the probability of a feature being detected. In this respect, attention emerges in the vision process in order to serve in a flexible manner the needs of the task at hand.

Before we explain our approach in more detail the most influential concepts in modeling attention with relevance to computer vision are discussed.

2. Approaches to modeling attention

2.1. The spotlight of attention

The most often used analogy of attention is the spotlight metaphor. Ahmad [10] literally implemented a spotlight as a circular focus which gates processing to the next stage. The Selective Tuning model [11,12] and SCAN [13] offer an illustrative explanation of the spotlight within a hierarchy of processing layers. In each of those processing layers a set of gating nodes determines which information is allowed to project to the next higher layer. In the Selective Tuning model, competition among the gating nodes starts from the top of the hierarchy and is sent downwards such that a beam emerges which covers the area of the selected feature surrounded by inhibition. Mozer and Sitton [14] proposed an elastic spotlight model which essentially emerges by a competition among populations.

The Shifter Circuit model [15] and related approaches [16,17] preserve the spatial relationship of features within a window of attention for invariant object recognition by routing a retinal input via a connection matrix or copying procedures into a focus of attention.

In addition to a spotlight in space, a selection of the level of spatial resolution has been proposed [18].

2.2. The saliency map

An essential aspect of all models of attention is the origin of selection. The Selective Tuning model [11] starts with looking for a feature on a very broad scale and then tracking it to the image resolution, whereas most other models use a spatially organized saliency map [6]. They try to find conspicuous scene sections in each feature map and integrate them into a single saliency map [19,20]. One of the most influential among these approaches uses center-surround operations to determine the conspicuity of each feature [21,22]. In this model attention is purely stimulus-driven. Other models try to include top-down, task-driven knowledge into the saliency map [23,24] or determine probability values of feature–target pairs [25]. Other approaches have suggested to dynamically adapt the saliency map according to the task at hand by controlling the preattentive flow of information by a neural network [26]. Despite the fact that such a top-down influence is computationally efficient, the guidance in visual search has been psychophysically verified in numerous experiments and illustrated in the Guided Search model [27,8].

2.3. Segmentation and gating

Many connectionist models of attention follow the spotlight metaphor. The idea is that the area of a spotlight is highlighted and its content is gated into higher levels for further processing. Thus, computation at high levels typically requires prior selection. In computer vision which deals with natural scenes a simple spotlight does hardly serve other processes such as object recognition. Thus, the dominant paradigm of modeling attention in computer vision is a parallel computation of key features followed by parsing the image into constituent components which define a region of interest [19,28–32]. The idea is to provide an object-related focus of attention, similar as suggested by the Selective Tuning model [11].

2.4. Feature maps

The Feature Integration Theory [5] suggested that basic features 'pop-out' and can easily be detected without extensive serial search. Although several features such as color, motion, and depth can provide very good cues for selection, a universal set of features has not been identified. Since serial search is expensive, research in computer vision has been directed to find the feature maps which best provide reliable cues for a given task (e.g. [33]).

2.5. Attentional selection

To implement the selective behavior of attention a winner-takes-all process has been suggested [6]. As an alternative to this most common approach, dynamic neural fields [34] have been used [35,30,36,20]. As compared to a simple winner-takes-all process, which detects the highest entry in a saliency map, dynamic neural fields detect an area of highly salient entries by forming an activity cluster.

The decision where to attend is usually determined after the selection process is settled. Selection does not necessarily has to occur at a single place (e.g., in a saliency map). Recently, models have been proposed in which selection operates on a global scale, although competition is defined locally [37–42]. These distributed approaches are able to accumulate evidence over time and over different areas (or modules), and estimate the consequences of a planned but not finalized decision.

2.6. Feature-based attention

In applied computer vision usually only the task-relevant features are computed and weighted according to their correspondence with the goal. A more task-independent approach could dynamically enhance the relevant target features for the task at hand. Indeed there is evidence for a global feature-specific feedback signal in the brain [43–45]. Only recently models of vision have started to incorporate aspects of feature-based attention [46,11,47,37,48–50,41,42].

2.7. Grouping and object-based attention

Although attention is typically linked to space there is evidence that objects are considered as an entity rather than a collection of features in space [51]. Object-based attention is far from being understood and has computationally only been addressed on the level of principles [52–55]. Sun and Fisher [56] presented an algorithm that translates the idea of salient locations [21] to grouped units and demonstrated its performance on synthetic and natural scenes. However, the difficult grouping process has been assumed to occur prior to their algorithm.

2.8. Feature tuning

A number of single unit recordings have shown that attention correlates with an increase of visual salience. More systematic studies observed the effect of attention on the tuning curve. For example, a cell's orientation tuning curve is determined by systematically varying the orientation of a stimulus presented within the receptive field of a cell and observing the cells response. These studies have revealed that the tuning curve increases by a gain factor when attention is directed to the stimulus [45,57]. Such effects have only been modeled within a computational neuroscience framework [58,59]. However, this finding seems to be relevant for computer vision. If we consider neural cells as feature detectors indicating the probability that the encoded feature is present in the scene, this finding provides a concept of how to increase the conspicuity of a feature.

2.9. Biased competition

The Biased Competition framework [60] is also routed in electrophysiology. It has been observed that neuronal populations compete with each other when more than a single stimulus is presented within a receptive field. Such competition can be biased by top-down signals. As a result, the irrelevant stimulus is suppressed as if only the attended one had been presented. The idea of Biased Competition has been explored with several computational models [61–63,37,55,38,39,41,64], but so far it has not been applied in computer vision. However, competition among feature representations could be a useful mechanism to filter out irrelevant stimuli for object recognition. A spatial focus of attention can reduce the influence of features outside the focus, whereas a competition among features has the potential to select objects without the need of a segmentation on the image level. We have recently shown that such a competition among features allows to detect objects in natural scenes [65].

3. Approach

We propose a model for generic computer vision, which, in the long run, could be used for object recognition and tracking (e.g., in autonomous robots). As discussed in Section 1, there seems to be no generic, data driven solution to computer vision, but rather the model itself (by a cognitive control structure) has to provide appropriate top-down knowledge for each task. Top-down connections modulate the processing along the levels of the hierarchy by inference. Attention is a natural consequence of using top-down knowledge in solving a task. It arises in the vision system and serves to emphasize the task-relevant information.

The present version is in many regards simplified. We use a simple feature set which only allows for object detection rather than object recognition. We also do not address the issue of invariance in object recognition. Thus, the emphasis of the present approach lies on the demonstration of the attentional mechanisms for goal-directed visual search in real world scenes. We show how a cognitive goal can penetrate vision to direct the processing resources towards the relevant aspects in a visual scene.

3.1. Population code

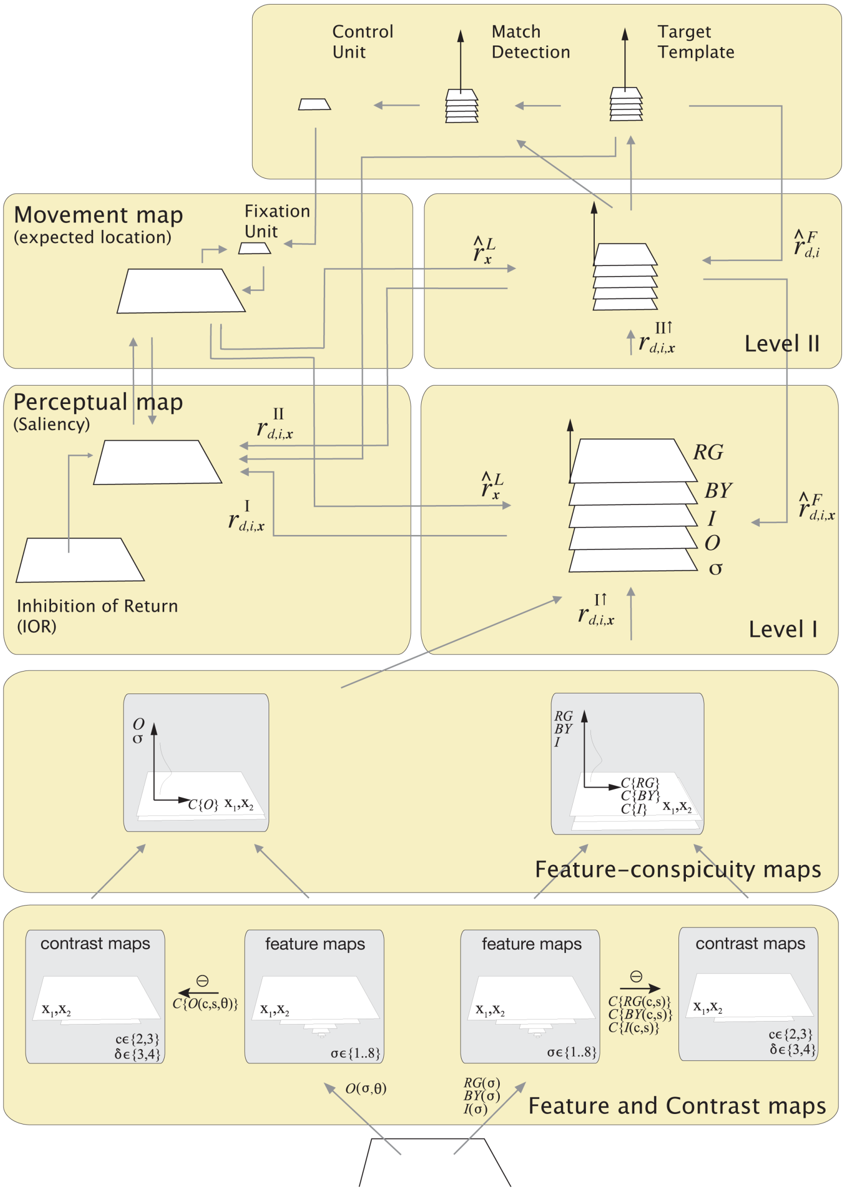
Decision making requires information being adequately represented. If we want to make a decision whether an object is in the scene, we have to accumulate evidence from several sources. Each process, however, will accumulate its own evidence necessary to make a decision, so that the information is distributed across processes. For example, if we want to find a vertical bar, the fit of contours in an image with an appropriate filter influences our decision. Moreover, the surrounding can make the bar more visible. Prior knowledge about the exact shape of this bar can improve its detectability as well. Provided we have an initial guess about its location we can use space information for inference as well.

Decision making also involves uncertainty arising from noise in sensation and the ill-posed nature of perception. Thus, we have to represent alternatives until a decision is found. Such constraints can be well handled by a population code. Population coding has been used as theoretical basis for describing the computation in the brain [66,67,47,68]. 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—its firing rate. The conspicuity represents the accumulated evidence. Our algorithm describes the local rules that will affect the conspicuity of each feature.

3.2. Overview

Fig. 1 illustrates our approach. For the purpose of a clear notation we distinguish between feature space and physical space. The preferred feature value of a cell is determined by the counter i and $\mathbf{x}$ is the center of an area (the receptive field) from which a cell receives input. We also introduce the variable d which refers to different channels such as orientation (O), intensity (I) or red-green (RG), blue-yellow (BY), or spatial resolution (σ). From the image we compute d sets of features $\mathscr{F}_d$ at each location $\mathbf{x} = (x_1, x_2)$. Each feature set is modeled as a continuous space with $i \in N$ sampling units by assigning each unit a conspicuity $r_{d,i,\mathbf{x}}$. The initial conspicuity is determined by center-surround operations in contrast maps given the scene [21] and then continuously updated to reflect the task-relevance.

The relevance of each feature is determined by the search template (target). For simplicity we define the target $\mathscr{T}_d$ by the same sets of features $\mathscr{F}_d$. Thus, a target object is defined by the expected features $\hat{r}^F_{d,i}$, independent of their location. For visual search we 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 position **x** in parallel. If the observation is similar to the expectation we increase the conspicuity. As we will explain, we apply a population-based inference approach in which the expectation enhances the gain of the observation. In this typical visual search situation, the search space is initially focused in the feature dimension and invariant in location.



To detect an object in space we combine in parallel the conspicuity across all d feature sets as well as all i sampling units and generate an expectation in space $\hat{r}^L_{\mathbf{x}}$. 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. Thus, a location with high conspicuity in different channels d tends to have a high expectation in space $\hat{r}^L_{\mathbf{x}}$. Analogous to the inference in feature space we iteratively compare the expected location $\hat{r}^L_{\mathbf{x}}$ with the observations $r^{\uparrow}_{d,i,\mathbf{x}}$ in $\mathbf{x}$ and enhance the conspicuity of all features with a similarity of expectation and observation. The conspicuity is normalized across each map. Such iterative mechanisms finally lead to a preferred encoding of the features and space of interest. Thus, attention emerges by the dynamics of vision.

We perform this iterative procedure not only on a single level but in a hierarchy of processing levels, to which we refer as level I and level II. It is well known that the receptive field size and the complexity of features increases along the ventral pathway [69]. In the present version we consider only an increase of the receptive field size.

3.3. Determination of initial conspicuity values

It is well known that the arrangement of stimuli determines perception [70]. For example, if we present a red bar surrounded by green bars, it is more easily detected than a red bar within a more heterogeneous collection of bars. Center-surround operations have been shown to provide a good estimate of such a stimulus-driven saliency [22,71]. Thus, in order to determine the initial conspicuity values we: (i) create multi-resolution feature maps, (ii) compute multi-resolution contrast maps using center-surround operations and (iii) combine both in feature conspicuity maps. For computing the first two steps we largely follow Itti et al. [21]. Details are given in Appendix A.

Fig. 1. Model of attentive vision. From the image we obtain $d = 5$ feature maps ($d \in \{RG, BY, I, O, \sigma\}$). For each feature at each location $\mathbf{x}$ we compute its conspicuity in 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. This initial, stimulus-driven conspicuity is now dynamically updated within a hierarchy of levels. From level I to level II we pool across space 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 $\hat{r}^F_{d,i}$ which are used to compute the (posterior) conspicuity at level II. Similarly level II represents the expectation for level I. As a result, the conspicuity of all features of interest is enhanced regardless of their location. In order to identify candidate objects we integrate across all five channels to determine the saliency. The saliency is then used in the eye movement map to compute the expected region of an object $\hat{r}^L_{\mathbf{x}}$, which in turn enhances the conspicuity of all features at levels I and II within the expected region. Thus, objects at expected locations are preferably represented. By comparing the conspicious features in level II with the target template in the match detection we can continuously track if the object of interest is within the expected region. If we loose a match an inhibition of return is triggered which marks the expected region as being visited. Otherwise the expectation increases until an overt shift occurs.

3.3.1. Feature maps

We currently use color, intensity, and orientation as basic features. Starting from r, g, and b, the color values (red, green, and blue) of the input image, an intensity image $I = (r+g+b)/3$ and the color maps $R = r - (g+b)/2$ for red, $G = g - (r+b)/2$ for green, $B = b - (r+g)/2$ for blue, and $Y = (r+g)/2 - |r-g|/2 - b$ for yellow are obtained. The color values are transferred into color opponency (RG, BY). All features are represented within a Gaussian pyramid, which is constructed by progressively low-pass filtering and sub-sampling the input images of the channels [72].

The detection of local orientation at each point in the image x_1, x_2 is achieved using overcomplete steerable filters $O(\sigma, \theta)$ [21,73] with varying resolution (or frequency) σ and 20 different orientations θ.

3.3.2. Contrast maps

Contrast maps represent the conspicuity of each feature. In analogy to the known influence of lateral excitation and surround inhibition, center-surround operations '$\ominus$' calculate the difference of maps with a fine scale σ and a coarse scale $s = \sigma + \delta$. This operation across spatial scales is done by interpolation to the fine scale and then point-by-point subtraction. The variation of the distance δ between resolutions results in a multi-scale feature extraction [21]. For each pixel of the resolution σ we create intensity contrast maps $\mathscr{I}(c,s) = |I(c) \ominus I(s)|$ by subtracting the map with the coarse scale s from the one with center scale c. A similar mechanism is applied in the color channels, which leads to the known double opponent system $\mathscr{RG}(c,s)$, $\mathscr{BY}(c,s)$. In the center, the cells are exited by one color (e.g., red) and inhibited by its opponent (green), while in the surround the opposite takes place. Double opponency determines the conspicuity of a stimulus, but does not alter the stimulus feature.

We average the maps obtained by a different course scale $s = \sigma + \delta$ to receive one contrast value per channel and center scale $\mathscr{I}(c)$, $\mathscr{RG}(c)$, and $\mathscr{BY}(c)$. To obtain orientation contrast maps $\mathscr{O}(c,s,\theta)$ we apply for each orientation θ the center surround operation with a fine scale c and a course scale $s = \sigma + \delta$.

The orientation channel is not averaged across scale s, since we use this information to determine another channel termed 'spatial frequency' as described in the next section.

3.3.3. Feature conspicuity maps

We now compute the initial conspicuity r_i by combining the feature value $\mathbf{V}$, as determined in the feature maps, with its gain P into a population code. We construct a space, whose axes are defined by the represented features and by the conspicuity (Fig. 2). As we have explained earlier, conspicuity is related to contrast. Thus, although we encode absolute values like the intensity, there will be no response on a black scene, since there is no contrast. The population is defined by $i \in N$ units sampling the space, with each unit tuned around its preferred value $\mathbf{u}_i$. The preferred value is the feature value for which the response of the unit is maximal. For each unit i we obtain an initial conspicuity value:

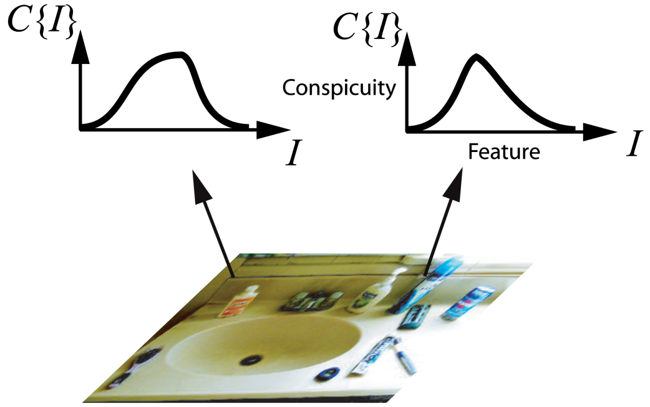


Fig. 2. Feature conspicuity for the intensity channel, illustrated at two positions. At each location x_1, x_2 we obtain a population of the conspicuity $C(I) = \mathscr{I}$ over the feature values $\mathbf{V} = I$. Thus, at each location the feature values and their related conspicuity is represented.

$$r_i = P \cdot g(\mathbf{u}_i - \mathbf{V}), \tag{1}$$

using a Gaussian tuning curve g. We perform this at each location x_1, x_2 and for each feature value $\mathbf{V} \in \{\theta, I, RG, BY\}$ with the associated conspicuity $P \in \{\mathscr{O}, \mathscr{I}, \mathscr{RG}, \mathscr{BY}\}$. After normalizing the conspicuity value to fit into a range between 0 and 1, we obtain the populations for intensity ($r_{I,i}(c, \mathbf{x})$), red–green ($r_{RG,i}(c, \mathbf{x})$), and blue–yellow ($r_{BY,i}(c, \mathbf{x})$) with scale c. The orientation information is transferred into two channels, one for scale or spatial frequency $r_{\sigma,i}(\theta, \mathbf{x})$ and one for orientation $r_{\theta,i}c, \mathbf{x})$.

We now have $\#c$ maps, where $\#c$ is the number of center scales, each with a population at every position $\mathbf{x}$. To combine these maps across different levels of spatial resolution into a single map with the lowest resolution (highest c) we introduce the notion of a receptive field (RF). We have to consider that variables $\mathbf{V}(\mathbf{x})$ are encoded at each location $\mathbf{x}$ within a RF and that the encoding population can get input from different locations within the receptive field. We use a convergent mapping function $\mathscr{R}$ of the projection from areas $S \in RF(\mathbf{x})$ to the target population T:

$$\mathscr{R}\colon S \mapsto T, \quad r^{\mathrm{T}}_{i,\mathbf{x}} = \max_{\mathbf{x}' \in RF(\mathbf{x})} r^{S}_{i,\mathbf{x}'}. \tag{2}$$

In a pyramidal structure the RF can be defined as the number of units necessary in each resolution encoding an area of equal size, e.g., the RF at the level of $c = 3$ is determined by one unit in $\mathbf{x}$ and at $c = 2$ by 4 units and so on. By means of this operation we achieve the final feature conspicuity maps $r_{\theta,i}(\mathbf{x})$, $r_{I,i}(\mathbf{x})$, $r_{RG,i}(\mathbf{x})$, $r_{BY,i}(\mathbf{x})$, and $r_{\sigma,i}(\mathbf{x})$.

3.4. Modification and transformation of conspicuity values

We now address how we represent conspicuity, the rules that allow us to modify it and to transform the conspicuity of features into other maps. We will explain the population-based inference approach and determine how we can cope with multiplicity in a hierarchical architecture. Please refer to Appendix A for a more detailed mathematical explanation of the model.

3.4.1. Population-based inference

We now explain how prior information iteratively shapes the conspicuity in each channel $d \in \{\theta, I, RG, BY, \sigma\}$. The system is given as a set of difference equations suited for computer implementation. We define the conspicuity $r_{d,i,\mathbf{x}}$ at time step $t+h$ as

$$\begin{aligned} r_{d,i,\mathbf{x}}(t+h) &= r_{d,i,\mathbf{x}}(t) + \frac{h}{\tau}\Delta r_{d,i,\mathbf{x}}(t), \\ \Delta r_{d,i,\mathbf{x}}(t) &= G\left(r^{\uparrow}_{d,i,\mathbf{x}}, \hat{r}^{F}_{d,i,\mathbf{x}}, \hat{r}^{L}_{\mathbf{x}}\right) - H\left(r_{d,i,\mathbf{x}}, \sum_{i} r_{d,i,\mathbf{x}}, \sum_{x} \max_{i} r_{d,i,\mathbf{x}}\right), \end{aligned} \tag{3}$$

where $G()$ is an activation term that determines the match of the actual observation $r^{\uparrow}_{d,i,\mathbf{x}}$ with the expected feature $\hat{r}^{F}_{d,i,\mathbf{x}}$ and with the expected location $\hat{r}^{L}_{\mathbf{x}}$, respectively. The idea of Bayesian probability theory is to use prior knowledge about scenes which is combined with image features to infer the most probable interpretation of a scene [74]. However, a true Bayesian inference would require to determine probability density functions and to ensure the independence of the prior from the observation. Based on earlier work [75], we developed a related but simpler population-based inference approach in which we combine the observation with the prior on the population level in order to compute the posterior conspicuity (Fig. 3).

$$\begin{aligned} G() = {} & r^{\uparrow}_{d,i,\mathbf{x}}(t) + r^{\uparrow}_{d,i,\mathbf{x}}(t) \cdot \left(\sum_{j} w^{d}_{ij} r_{d,j,\mathbf{x}} + \sum_{\mathbf{x}'} w^{d}_{\mathbf{x},\mathbf{x}'} r_{d,i,\mathbf{x}'}\right) \\ & + \Gamma\left(A - \max_{i}(r_{d,i,\mathbf{x}})\right) \cdot \left(w^{L} r^{\uparrow}_{d,i,\mathbf{x}} \cdot \hat{r}^{L}_{\mathbf{x}} + w^{F} \max_{\mathbf{x}'}(r^{\uparrow}_{d,i,\mathbf{x}} \cdot \hat{r}^{F}_{d,i,\mathbf{x}'})\right); \\ \Gamma(a) = {} & \max(a, 0). \end{aligned} \tag{4}$$

The activation term increases if the expected feature matches the actual observation. This is consistent with the idea that a global feature-based feedback signal (the prior) enhances the gain of a cell (Section 2.6). In addition, there has been evidence for a spatial gain control of V4 cells by the frontal eye field [76]. The effectiveness of inference is reduced with the strength of the maximal conspicuity which relates to the contrast dependence of attention [77]. On the population level the inference approach tunes the expected feature by increasing the gain (Section 2.8).

$H()$ induces competition and normalizes the activity. It is explained in more detail in Appendix A. From Eq. (3) we see that the conspicuity is constant if $G()$ and $H()$ are balanced. It is important to note at this point that the conspicuity indicates the evidence and relevance of a certain feature. This approach may remind the reader to relaxation approaches in the sense that we aim to reduce iteratively the ambiguity between the representation in the system and the search template. Given the template as a constraint and the connections in the network as coefficients we have defined a

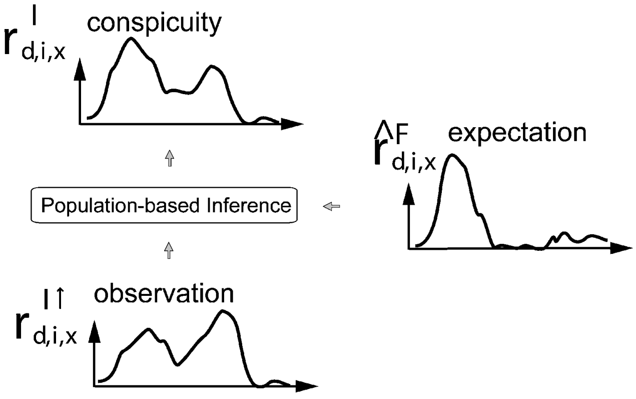


Fig. 3. Illustration of the population-based inference approach. As an example we illustrate the effect of a feature specific inference at level I with the expectation $\hat{r}^{F}_{d,i,\mathbf{x}}$. In order to compute the conspicuity at each time step the observation $r^{I\uparrow}_{d,i,\mathbf{x}}$ is modulated by the expectation. As a result, the conspicuity of the features that match the expectation is enhanced, whereas the conspicuity of other features is slightly reduced.

cooperative/competitive algorithm. A crucial difference to relaxation labeling, however, is that we do not assign each pixel a label nor do we wait until the relaxation process settles down to a final value. Our algorithm uses slight advantages to enforce a quick decision. This avoids an unpleasant slow convergence, as predicted by extensive simulations. The result of the population-based inference is consistent with the Biased Competition framework (Section 2.9), since the network shifts into a state in which it filters out the irrelevant stimuli.

3.4.2. Invariance to multiplicity

An efficient processing within a population code requires that the conspicuity of each feature is adequately combined across hierarchy levels. Due to the increase in RF size, the conspicuity of features from different locations converges onto a single location. A weighted sum of the conspicuity of identical features across space shows a multiplicity effect: increasing the number of identical stimuli within a RF enhances the conspicuity [59]. Assume a cell whose preferred stimulus is a vertical bar. If we model the receptive field with a weighted sum over all inputs the input will increase with the number of vertical bars (or with a similar orientation) placed within the receptive field. A better strategy would be to separate between the content and relevance of a stimulus independent of the number of stimuli.

Let us assume we project from hierarchy level I to level II. Thus, we want to determine the observation at position $\mathbf{x}$ on level II given the conspicuity $r^{\mathrm{I}}_{d,i,\mathbf{x}'}$ on level I at the positions $\mathbf{x}' \in RF(\mathbf{x})$ within the receptive field of $\mathbf{x}$. The strength of the feedforward projection depends on the similarity of the encoded feature, i.e., the distance in feature space between the unit at level I $\mathbf{u}^{\mathrm{I}}_{i}$ and the unit at level II $\mathbf{u}^{\mathrm{II}}_{i}$. We define the weighting function $F(r^{\mathrm{I}}_{d,i,\mathbf{x}'}) = r^{\mathrm{I}}_{d,i,\mathbf{x}'} \cdot g(\|\mathbf{u}^{\mathrm{II}}_{i} - \mathbf{u}^{\mathrm{I}}_{i}\|)$ using a Gaussian g. In extension to Eq. (4) we write for the activation term

$$\begin{aligned} G() = {} & \max_{i,\mathbf{x}' \in RF(x)} \left(F\left(r^{\mathrm{I}}_{d,i,\mathbf{x}'}\right)\right) + \max_{i,\mathbf{x}' \in RF(x)} \left(F\left(r^{\mathrm{I}}_{d,i,\mathbf{x}'}\right)\right) \\ & \cdot \left(\sum_j w^d_{ij} r^{\mathrm{II}}_{d,j,\mathbf{x}} + \sum_{\mathbf{x}''} w^d_{\mathbf{x},\mathbf{x}''} r^{\mathrm{II}}_{d,i,\mathbf{x}''}\right) + \Gamma\left(A - \max_i (r^{\mathrm{II}}_{d,i,\mathbf{x}})\right) \\ & \cdot \left(w^L \max_{i,\mathbf{x}' \in RF(x)} \left(F\left(r^{\mathrm{I}}_{d,i,\mathbf{x}'}\right) \cdot \hat{r}^{II_L}_{\mathbf{x}'}\right) + w^F \max_{i,\mathbf{x}' \in RF(x)} \left(F\left(r^{\mathrm{I}}_{d,i,\mathbf{x}'}\right) \cdot \hat{r}^{II_F}_{d,i,\mathbf{x}'}\right)\right), \end{aligned} \tag{5}$$

that now considers a convergent mapping of the conspicuity at locations $\mathbf{x}'$ at level I to the location $\mathbf{x}$ at level II using a maximum operation. The conspicuity from different locations does not add up, but is simultaneously represented in the feature space. Thus, the presentation of two equal objects does not result in an increase of the conspicuity. Two different objects are encoded in parallel by different conspicious features. It has been shown that such a max-pooling allows to reproduce the data of Reynolds et al. [63] showing the influence of attention on the competition within a receptive field [59]. The following maps are implementations of the general Eqs. (3)–(5) above.

3.4.3. Level I

Level I 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. 1). The expectation of features at level I originates in level II $\hat{r}^{\mathrm{I}_F}_{d,i,\mathbf{x}'} = r^{\mathrm{II}}_{d,i,\mathbf{x}}$ and the expected region in the eye movement map $\hat{r}^{\mathrm{I}_L}_{\mathbf{x}'} = w \cdot r^m_{\mathbf{x}'}$. Please note that even level II has a coarse dependency on location.

3.4.4. Level II

The features with their respective conspicuity and location in layer I project to layer II, but only within the same dimension d, so that the conspicuity of features at several locations in level I converges onto one location in level II: $r^{\mathrm{II}\uparrow}_{d,i,\mathbf{x}} = w \max_{i,\mathbf{x}' \in RF(x)} (F(r^{\mathrm{I}}_{d,i,\mathbf{x}'}))$. We simulate a map containing nine populations with overlapping receptive fields. For simplicity, we do not increase the complexity of features from level I to level II. The expected features at level II originate in the target template $r^{\mathrm{II}_F}_{d,i,\mathbf{x}} = w \cdot r^{\mathrm{T}}_{d,i}$ and the expected region in the eye movement map $\hat{r}^{\mathrm{II}_L}_{\mathbf{x}} = w \cdot r^{\mathrm{m}}_{\mathbf{x}}$.

3.4.5. Target template

When we present an object to the model, its "working memory" memorizes a target template, i.e., the most conspicious feature in each channel. The memory units can hold a pattern even when the input is removed. They receive their input from the full visual field, i.e., from all nine locations of level II units. It is important to note that these units are able to encode a stimulus without spatial attention. In a visual scene with more than one object a spatial selection would ensure that the information in each channel belongs to a single object.

3.4.6. Match detection

To determine if an actively encoded pattern at level II fits with the target template we define match detection units (md) that compare in parallel the encoded pattern with the target template. This allows to close the loop from setting a target template, over selection to recognition. If both, the encoded and expected pattern are similar, the activation term G is high.

$$\Delta r^{\mathrm{md}}_{d,i} = G\left(r^{\mathrm{T}}_{d,i}, \max_{\mathbf{x}}(r^{\mathrm{II}}_{d,i,\mathbf{x}})\right) - H\left(r^{\mathrm{md}}_{d,i}, \sum_{j} r^{\mathrm{md}}_{d,j}\right). \tag{6}$$

3.4.7. Perceptual map

The perceptual map (v) indicates salient regions by integrating the conspicuity of level I and II across all channels as defined by the first two terms of $G()$.

$$\begin{aligned}\Delta r^{\mathrm{v}}_{\mathbf{x}} = {} & G\left(\sum_{d} \max_{i} r^{\mathrm{I}}_{d,i,\mathbf{x}}, \sum_{d} \max_{i} r^{\mathrm{II}}_{d,i,\mathbf{x}'\in RF(x)}, r^{\mathrm{m}}_{\mathbf{x}}, r^{\mathrm{I}}_{d,i,\mathbf{x}}, r^{\mathrm{T}}_{d,i}\right) \\ & - H\left(\sum_{x} r^{\mathrm{v}}_{\mathbf{x}}, \sum_{\mathbf{x}''} w^{d}_{\mathbf{x},\mathbf{x}''} r^{\mathrm{v}}_{\mathbf{x}''}\right).\end{aligned} \tag{7}$$

In addition to the conspicuity in level I and II the activation term $G()$ includes the match of the target template with the features encoded in level I at all locations simultaneously by the product $\prod_{d} \max_{i,\mathbf{x}'\in RF(\mathbf{x})} r^{\mathrm{T}}_{d,i} \cdot r^{\mathrm{I}}_{d,i,\mathbf{x}'}$. This implements a bias to regions with a high joint probability of encoding all searched features in a certain area. The variation in $\mathbf{x}'$ ensures that the expected feature in each channel can slightly vary in location.

3.4.8. Eye movement map

The projection of the perceptual map onto the movement map (m) transforms the salient regions into a few candidate regions which provide the expected region for level I and level II units. We achieve this by subtracting the average saliency from the saliency at each location $w^{\mathrm{v}} r^{\mathrm{v}}_{\mathbf{x}} - w^{\mathrm{v}}_{inh} \sum_{\mathbf{x}} r^{\mathrm{v}}_{\mathbf{x}}$. Simultaneously, the movement units indicate the target location of an eye movement. This is consistent with several findings indicating a strong overlap between spatial attention and eye movements [78–81]. Please refer to [64] for a discussion about the origin of spatial attention and eye movement planning. A shift of the visual scene as a consequence of moving the eye is not explicitly modeled. All movement units can be inhibited by the activity of a fixation unit r^{f}.

$$\Delta r^{\mathrm{m}}_{\mathbf{x}} = G\left(w^{\mathrm{v}} r^{\mathrm{v}}_{\mathbf{x}} - w^{\mathrm{v}}_{inh} \sum_{\mathbf{x}} r^{\mathrm{v}}_{\mathbf{x}}\right) - H\left(r^{\mathrm{f}}, \sum_{x} r^{\mathrm{m}}_{\mathbf{x}}, \sum_{\mathbf{x}''} w^{d}_{\mathbf{x},\mathbf{x}''} r^{m}_{\mathbf{x}''}\right). \tag{8}$$

3.4.9. Inhibition of return

An inhibition of return avoids revisiting a region during covert and overt scanning. We regard a location $\mathbf{x}$ as inspected, dependent on the selection of an eye movement, or when the match detection units indicate no match, given a region shows a high expectation. An eye movement occurs at the time t_0 when the activity of a movement unit exceeds a threshold $(r_{\mathbf{x}}^{\mathrm{m}}(t_0) > \Gamma_0^{\mathrm{m}})$.

In this case the IOR units are charged with the signal $I_{\mathbf{x}}^{m}$ around the location of the strongest unit in the movement map $\mathbf{x}_m$ for a period of time T^{IOR}. This reduces the saliency of the recently attended region. IOR units get slowly discharged by a decay with a low weight w_{inh}.

$$\Delta r_{\mathbf{x}}^{\mathrm{IOR}} = (1 - r_{\mathbf{x}}^{\mathrm{IOR}})(w^{\mathrm{m}} I_{\mathbf{x}}^{\mathrm{m}} - w_{inh} r_{\mathbf{x}}^{\mathrm{IOR}}),$$

$$I_{\mathbf{x}}^{\mathrm{m}} = \begin{cases} \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_{\mathrm{m}})^2}{0.01}\right) & \text{if } t < t_0 + T^{\mathrm{IOR}} \\ 0 & \text{else} \end{cases} ;\ r_{\mathbf{x}_m}^{\mathrm{m}} = \max_{\mathbf{x}}(r_{\mathbf{x}}^{\mathrm{m}}). \qquad (9)$$

4. Results

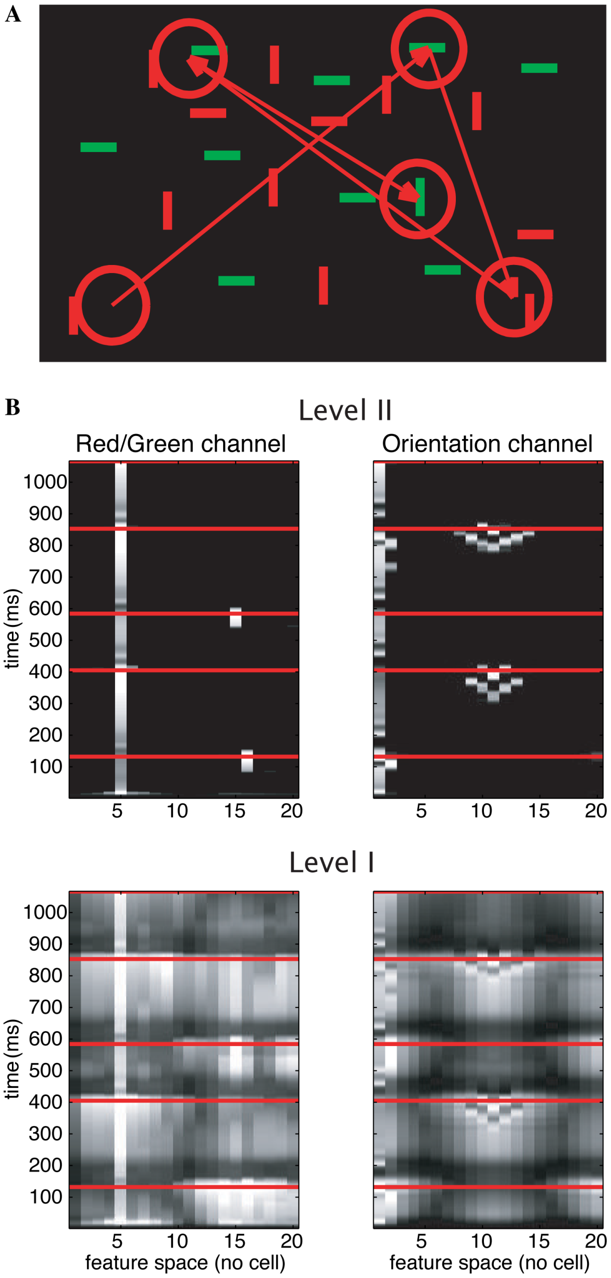
We illustrate our approach with a conjunction visual search task. We then demonstrate the model's properties and the emergence of attention in an object detection task using 16 different target objects in various natural scenes.

4.1. Conjunction search

We simulated a conjunction visual search task in which overt attention (eye movement) is allowed. In conjunction search a target is unambiguously defined by the conjunction of two or more separately processed features. Conjunction search was often associated with serial search—and feature search with parallel search [5]. Parallel search is defined as a visual search condition showing no set size effect. Otherwise, search is assumed to be serial. Under the assumption that each item in a scene is analyzed separately, the performance of search can be expressed by the number of scanned items per second. However, the time to

Fig. 4. Conjunction visual search, while allowing the model to overtly search until it selected the target. The target is the green, vertical bar. (A) Scan path. (B) Conspicuity values of level II and level I units in the red-green and orientation channel over time. The conspicuity is indicated by brightness. Red lines indicate the time of the eye movement. The target color "green" is encoded by units with lower numbers (e.g., unit 5) and the target orientation vertical as well (e.g., unit 1). To illustrate the conspicuity of features over time we removed all spatial information by simply taking the maximum conspicuity of each feature regardless of its location. The time depends on the time constant in our equations and on the chosen connections. Since we start presenting an input to level I which is comparable with V4 one would have to add another 80 ms to compare the amount of time the model requires to make a decision with the time of responses in experiments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

complete search shows a high variability in feature and conjunction search [82] such that knowing the number of items per second does hardly allow to predict the underlying search mode. In addition, the assumption that each item is visited one after the other predicts a very fast covert scan of the scene of about 30–40 ms



per item for which no physiological counterpart has been found yet. There is more evidence for a slow serial component in search (at least larger than 100 ms to shift spatial attention) as found in a rapid visual presentation task [83], difficult visual search [84], and EEG [85].

In our example, the model searches for a green, vertical item in a scene composed of other green and vertical items. The model takes a serial search and finds the target after the 5th eye movement (Fig. 4A). Note, that Motter and Belky [86] found a preference for color over orientation in the guidance of visual search. We could account for this finding by decreasing the feedback weight in the orientation channel. However, this does not reveal anything new about the underlying architecture of guidance in human vision. From the perspective of computer vision, the channels are equivalent and if necessary, the weights could be tuned for a given task.

The initial conspicuity of each feature depends on the local arrangement of the stimuli. In the model the difficulty in conjunction search arises due to the high target-distractor similarity. The population-based inference for green and vertical is independent of each other, so that distractors manage to increase their conspicuity as well (Fig. 4B). The advantage of two simultaneous matches (in the red–green and orientation channel) within a localized area does not necessarily overrule the stimulus-driven saliency. Thus, the model performs additional eye movements until the target is detected. This idea of guidance is similar to the Attentional Engagement [70] and Guided Search [8] proposal. However, these models are very abstract. Since our model operates on a much deeper level of the putative underlying mechanisms we now briefly discuss some predictions of the model.

4.1.1. Variability of the time to complete visual search

Although the model takes a serial search mode, the time it takes to make an eye movement is highly variable (as indicated by the red bars in Fig. 4B). First, our model predicts that spatial selection is slow as we have discussed in-depth elsewhere [42]. Second, the time of the decision about the location of the potential target varies. We have shown with an earlier version of the model using artificial input that the variability in time depends on the target-distractor similarity [42] and on the strength of the top-down signal [64]. Thus, we explain the variability in visual search by the number of shifts of spatial attention and by the variable amount of time to make a decision where the potential target is located—a variability in a parallel decision process.

4.1.2. Distributed nature of attention

Initially, the expectation to find a green, vertical target, reinforces the conspicuity of green and vertical in level II so that these features dominate the response as soon as the activity travels upwards from level I to level II. The conspicuity at level II provides the expectation for level I so that the target template travels downwards. The conspicuity of other features at level I decreases due to the normalization but remains at a significant level. The selection of a non-target in the eye movement map involves a spatial expectation which reinforces the conspicuity of all features in the respective regions so that prior to the first eye movement the conspicuity of red increases at level I and II (Fig. 4B). The distribution of conspicuity across the feature red is much broader, since it emerges from increasing the gain of the whole tuning curve of each neuron encoding a stimulus within the expected region. The region of the selected target emerges dynamically by reading out the perceptual map through the eye movement map. The perceptual map can be compared to the saliency map (Section 2.2). The primary difference is, however, that the selection dynamics are more distributed over the network. The model goes beyond of selecting an item in space and passing the features within the selected area to some arbitrary recognition module as proposed by the classical approach of gating and segmentation (Section 2.3). The model integrates the "what" and "where" aspect of vision by adapting its internal representation (updating the conspicuity of feature variables). This is the groundwork of all other processes that operate on these variables—and thus, the mechanisms that implement attention ensure a distributed type of processing and a dynamic binding.

4.1.3. Variable focus of attention

A more close look into the spatial selection process reveals that not just a single region is picked but a bunch of potential ones (Fig. 5). This pattern of selection does not necessarily fit with the spotlight hypothesis (Section 2.1). Although the model shows a spatial organization of the distribution of expectation values in the movement map, it is related to items and much more flexible than a unitary focus. The general idea is that spatial selection is a dynamic process [14]. We call this the reentry hypothesis of spatial attention [41,42]. Indeed there is mounting evidence against a unitary spotlight from experiments showing a split of spatial attention [87–90]. This item related reentry signal in our model predicts that more than a single item could be analyzed in parallel per scan—a hypothesis that was also acknowledged by advocates of serial search [5,8].

4.2. Object detection in natural scenes

We now demonstrate the performance of our approach for object detection in different natural scenes, with emphasis on the flexibility and variability of the model. The model is able to handle a stream of images. We will present a target object on a black background to the model for 100 ms and it will memorize conspicious features of the object. We only apply this procedure to ensure that the features taken into memory are from the target object. We do not give the model any hints which feature to memorize. It is also possible to present the model a natural scene and let it memorize the first object to which it makes an eye movement [91]. We do not use an optimization method (e.g., learning) for target-distractor discrimination. We then present a black scene for 50 ms in order to let the conspicuity decay, which prevents the location of the target from influencing the result. The model works as well without the black scene in between, but then the performance of the model could depend on where we presented the target to be memorized. Finally, the search scene is pre-

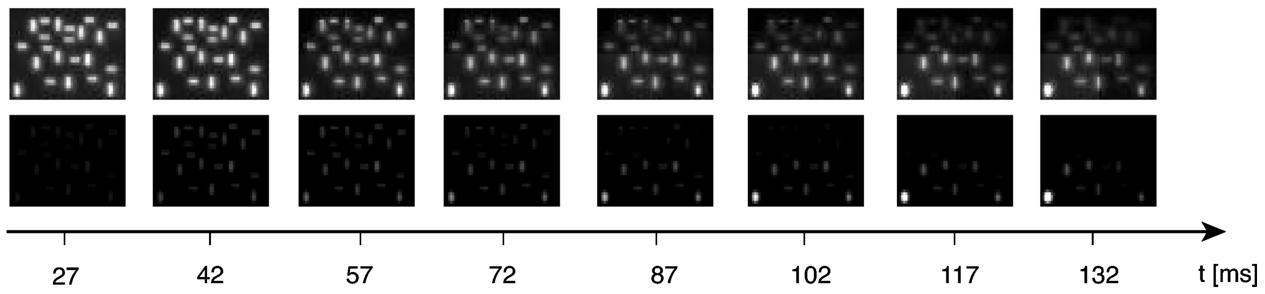


Fig. 5. Saliency of the items in the perceptual map (top) and expectation in the movement map (bottom) until the first eye movement. Initially all items are salient, although some show higher values, but over time, saliency decreases at most locations. Expectations in the movement map are initially low but distributed over several locations, since no specific evidence has accumulated. Prior to the eye movement a high expectation at a single location has built up. However, even at this time some expectation is directed to other locations.

sented, and the model's task is to make an eye movement towards the target, whereas it should avoid making an eye movement to a distractor. We now present a few illustrative examples in detail and then give an overview of all 16 tested trials—none of the tested trials was rejected in order to tune the impression about the models performance. All runs were performed with a single parameter set, the same as in the conjunction search.

4.2.1. Correct eye movement selection

Fig. 6A shows an example in which the first eye movement of the model is correct. The conspicuity in level II immediately follows the target template (Fig. 6B) which in turn guides level I to emphasize the features of the target as well. In the blue–yellow channel the target template is not dominant initially, but the modulation by the expectation from level II overwrites the initial conspicuity. This emphasis of specific features allows for a good discrimination in space so that the model quickly converges to the correct region.

4.2.2. Covert search

In the previous example attention emerged as part of the process of planning an eye movement. Although spatial attention occurred as well, it did not play an important role, since the initial guidance of level I had been already correct. We now show an example in which spatial attention turns out to be crucial for the task (Fig. 7). The model visits three regions before it executes a correct eye movement to the target (Fig. 7B). Here, the feature-based inference does not allow to sufficiently discriminate the target in order to immediately determine its location. The model first expects the target as being at the location of the man in white, as indicated by the high expectation in the eye movement map at 75 ms (Fig. 7D). As we have explained, the model continuously compares the encoded conspicuity in level II with the target template. The high expectation around the man in white increases the conspicuity of features in this region. If those features do not match the target template the model re-fixates (i.e., inhibits the expectation in the movement map) and initiates an inhibition of return in order to mark this region as being visited and analyzed. Fig. 7C reveals that the orientation channel first indicated a poor match with the template. The level II clearly shows a high conspicuity of a non-target feature. Although initially the target features have a high conspicuity, the reinforcement of features at a non-target region begins to dominate the level II and suppresses the conspicuity of target features. In the orientation channel the dominant vertical orientation of the man's shape suppresses the horizontal orientation of the van. In the second and third deployment of spatial attention the intensity channel first indicates a poor match.

This example demonstrates that covert search can emerge from an overt search process under the additional constraint to make an eye movement only to the target. It is not necessary to implement a separate covert search process. Covert spatial attention emerges in this model by a planned but not executed eye movement. In addition, attention serves a specific purpose, it guides the eye and reduces interference to improve recognition. However, overt search and covert search are not equivalent. Overt search shifts the fovea, the area of the highest resolution in the retina, to

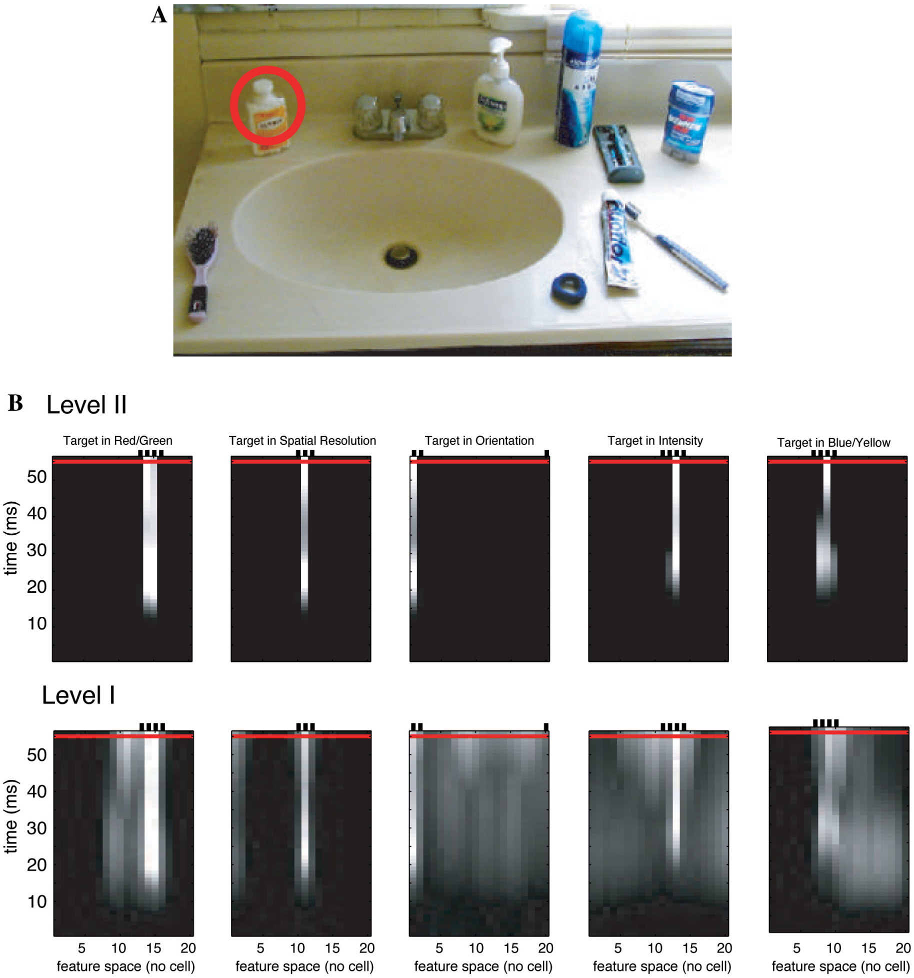


Fig. 6. Visual search in natural scenes. The asprin bottle in the upper left corner was presented to the model before the scene appeared and in each dimension the most conspicious feature was memorized in order to generate a target template. Then the model searched for the target. (A) Indication of the first eye movement, which directly selects the target. (B) Conspicuity values of level II and level I units in all channels over time. The strength of conspicuity is indicated by brightness. The red line indicates the time of the eye movement. The target template is indicated by the bars at the top of each figure. The conspicuity of each feature occurs first in level I and then travels upwards to level II. Level II, however, first follows the target template, which then travels downwards to level I. This top-down inference is clearly visible in the blue-yellow channel, where initially other features than the target feature are conspicious. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

allow for an inspection of an object in high resolution. Outside the fovea, the spatial resolution decreases with increasing eccentricity. These retinal effects are not considered in the model.

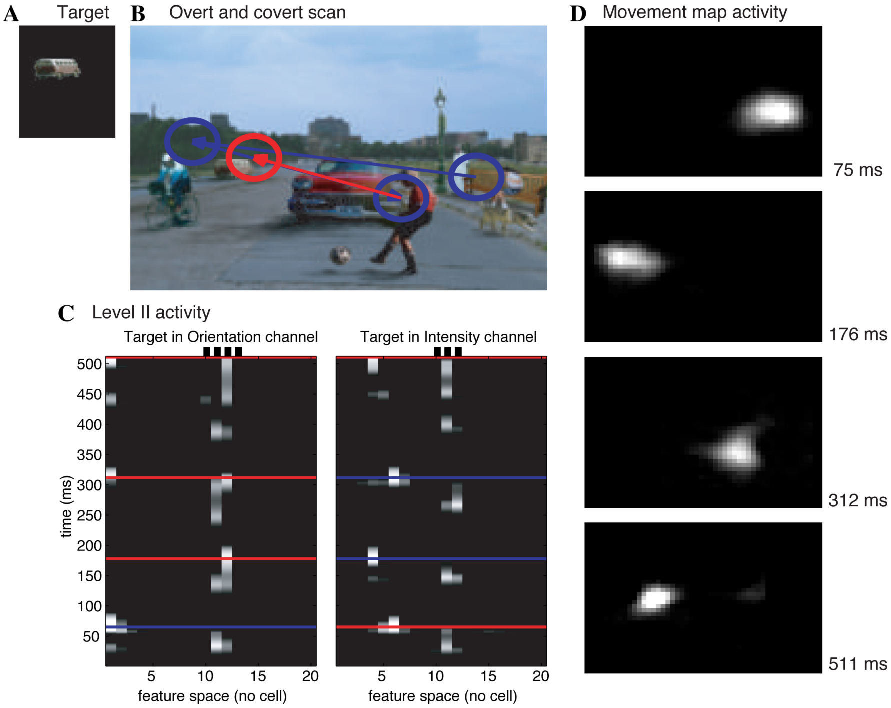


Fig. 7. Mechanisms for overt and covert scan. We initially present the target shown in (A) for 100 ms to the model and allow the model only to make an eye movement to the target. (B) Locations of overt (red) and covert (blue) attention. The model covertly visits three locations until it makes an eye movement to the target. (C) Conspicuity in level II. The target template is indicated by the bars at the top of each channel. The model did not achieve a sufficient match in either the orientation or the intensity channel, as indicated by the blue line. The blue line refers to the channel in which a match was poor and as a result, an inhibition of return was initiated. (D) Expectation in the movement map at each covert or overt shift. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

4.2.3. Split of spatial attention

Early findings in attentional research suggested a unitary spotlight [4] or a zoom lens [92], which allows a variation in size. However, recent experiments have reported a split of spatial attention [87–90]. Our reentry model results in a behavior which is more consistent with the latter findings. In the recent example (Fig. 7) spatial attention was allocated to unitary regions in space, which is typical for the model. We now show an example where a split of spatial attention occurs (Fig. 8). Due to the similarity of the target features obtained from the deo stick with the features of the shaving cream, high expectations at two noncontiguous regions are generated. They are almost equally strong 49 ms after onset (please bear in mind that a comparison with primate data requires to add the time a stimulus needs to reach V4). However, the different regions compete with each other and at the time of the eye movement at 79 ms the expectation around of the shaving cream is almost dissolved. Since there

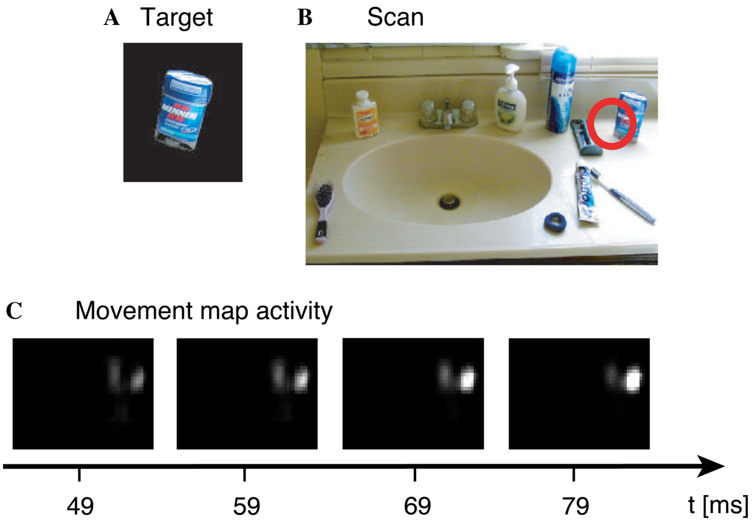


Fig. 8. Illustration of a split of spatial attention in the model. (A) Presented target. (B) Target detection. (C) Expectation in the movement map. In this specific case the model develops two centers of high expectation and thus two foci of spatial attention. Initially they are almost equal but over time one is favored against the other.

is still some expectation left, the location of the eye movement is slightly shifted to the left of the target. Mozer and Sitton [14] proposed an 'elastic' spotlight model which initially selects all regions with input and as time increases regions with less strong input drop out and the remaining area shrinks towards a compact region. They further assume that these selection units gate processing for recognition. This is different from our approach. First, expectation only biases processing [42]. Second, not all areas with input are selected, only areas with the strongest input build up a high expectation in space. Third, spatial attention is late, it needs time to develop.

Our own experimental data best fits with the idea that the observed split of attention correlates with such a transient competitive state in which spatial attention is first distributed to more than a single region and then settles onto a single region [89].

Can such a behavior be beneficial to vision or is this an artefact of competition over time? Attention has been often claimed to be necessary to reduce the computational burden of parallel processing beyond simple features. However, at present we have no final answer about the amount of possible parallelism [93]. A split of spatial attention can indeed be valuable, e.g., to track two important displays in parallel. Another example is a same/different comparison of two objects, like two faces or the shape of two red objects. This could either be done in parallel or sequentially. In the parallel approach one could directly compare the activation of features, whether they overlap or show separate clusters. In the sequential solution one has to memorize the first set of features and then make a comparison with the second set. A split of attention could also be beneficial to identify a ranking of potential targets. All interesting objects could then be analyzed in more detail to reveal whether they fulfill the needs of a task. If we use the initially high expectation as an entry into a spatial memory, we could implement a purely spatial search to localize the potential target before we start with the analysis. Thus, only a single feature-based search would be necessary. All other potential targets could be localized by a faster spatial "search." Without this ranking one would have to initiate for each potential target a feature-based search and ensure to avoid visiting regions twice. Thus, spatial cognition seems to be easier if one could highlight more than one region if appropriate.

4.2.4. Feature-based attention

The final example illuminates the role of feature-based attention in object recognition. Computational solutions for object recognition have been shown to be quite robust if we have cues available that help to segment the scene prior to recognition. However, for a general purpose vision system this is a chicken-egg problem. How can I segment an object before identification? Direct low level cues like color, boundary and motion are only reliable in specific tasks such as the detection of single part objects [29]. Spatial attention could help to localize a potential target but it does not offer robust tools to improve recognition in cluttered scenes. Quite the contrary, a purely spatial focus like a spotlight can be misleading if it does not sufficiently cover the object of interest.

We now demonstrate that feature-based attention can be beneficial for object recognition. The steerable filter responses of the lighter in isolation are vertical at the left and right corner (red color) and close to horizontal in the middle (green color) (Fig. 9A). The algorithm has obviously no problem in detecting the lighter in a cluttered scene (Fig. 9B), although it memorized only the slightly tilted orientation (Fig. 9C). However, the spatial focus (Fig. 9D) increases the conspicuity of all features within its area, so that the vertical edge of the cigarette box gets dominant as well (Figs. 9C and E). The level I conspicuity in the orientation channel initially exhibits a dominance for horizontal edges due to the top-down guided feature-based search (Fig. 9E). The emergence of a spatial focus (Fig. 9D), 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.

Feature-based attention helps, 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 attention serves as a cue to "segment" the object features in feature-space similar in effect to cue based region segmentation. However, a crucial difference is that the feature-based attention mechanism is not limited to low-level cues—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. In our model we have demonstrated the influence of feature-based attention using just simple features. By learning appropriate feedforward and feedback connections any complex feature detector can be used to enhance

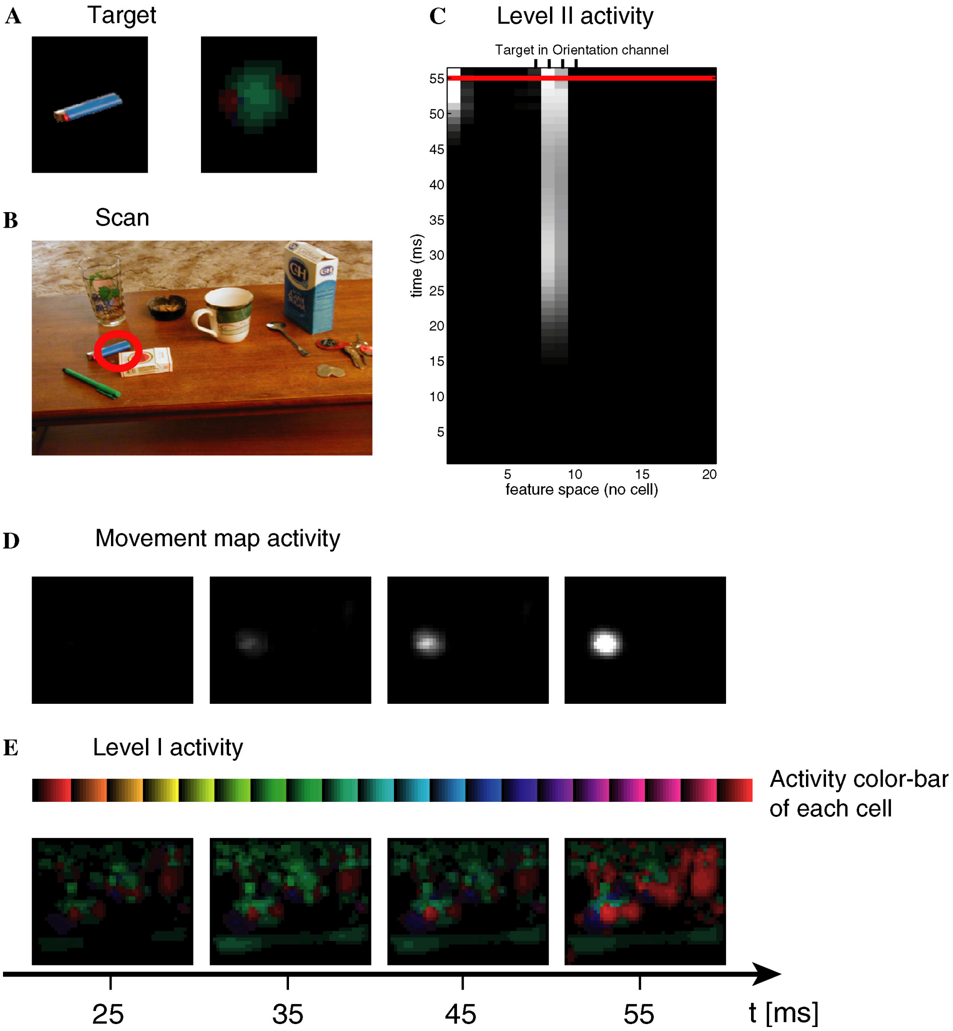


Fig. 9. Illustration of feature-based and spatial attention. (A) Target and the conspicuity of its features in the level I orientation channel during cue presentation (at 70 ms). At each point in space we show the most conspicious orientation out of 20 by means of a color code as explained in (E). (B) Target detection in the scene. (C) Conspicuity in level II. Initially the conspicuity of target features is represented in the orientation channel. Prior to the correct detection a non-target feature raises its conspicuity. (D) The spatial expectation is directed towards the location of the target. (E) Color code of the level I conspicuity within the map. Each of the 20 units is assigned a hue value. We only show the highest conspicuity at each location. The strength of the conspicuity is scaled by the intensity.

object related features. With such an extension we would get close to the idea of object-based attention (Section 2.7), where the object is supposed to group as a whole, even when it consists of a collection of different low-level features.

4.2.5. Overview of all tested scenes

We have tested our approach on detecting different target objects on 16 trials in three different natural scenes (Fig. 10). In 13 of them it found the target among the first four covert or overt shifts. However, the model's task was not to make an overt shift of attention to a non-target. In this regard, any overt shift to a distractor would count as an error. Nevertheless, in only two of the 13 examples a wrong eye movement to a non-target occured. Among those 11 valid examples the model made covert shifts of attention in three cases before if detected the target. It immediately detected the target object in eight trials. The model never rejected a target object.

Among the three cases in which the model did not detect the target within the first four shifts (Figs. 10D, M, and N) are a ball and an ashtray. Both of them are not salient in the scene and the model does not have anything like a detector for round objects. The girl (Fig. 10N) was almost detected by the first eye movement but the conspicuity of the man in white was considerably higher so that we decided not to count the first eye movement as a valid trial. In a similar case (Fig. 10J) where the target was the street lamp, we found a high conspicuity in level I for the features of the lamp, so that we count this trail as valid.

The aim of this overview is to demonstrate that the proposed mechanisms robustly work in real world environments (Table 1). The emphasis is not on object recognition. Given the simple feature space and sparse target template the model works very well. Please note that the target is in most cases presented on a different background during cue presentation than during search.

5. Discussion

We have presented a new approach to modeling vision in a distributed architecture. Vision in this model is inherently driven by internal goals, it operates in parallel and is top-down guided.

The model's principles are strongly routed in neuroscience and psychology [41,59,42,64,91]. Regarding anatomy and temporal dynamics, level I can be compared to V4 and level II to IT, the perceptual map to frontal eye field (FEF) visuomovement cells and the movement map to FEF movement cells. The present approach, however, does not account for the complexity of the feature space in V4 and IT. With reference to the biological point of view, the activation in the level II units is very sparse. This seems to contradict our earlier simulations [41] of a visual search experiment performed by Chelazzi et al. [44] who placed objects on a plain background. In those simulations we used an artificial input for which we could control the target-distractor similarity. Due to the low level features used here target and distractors can be very similar. In order to ensure the rejection of attended non-targets we had to implement a strong competition in order to suppress the conspicuity of features at level II which are supported by the target template. However, other experiments that used natural scenes also find highly selective responses in IT [94,95], so that the strong selectivity in our model does not necessarily contradict with experimental data.

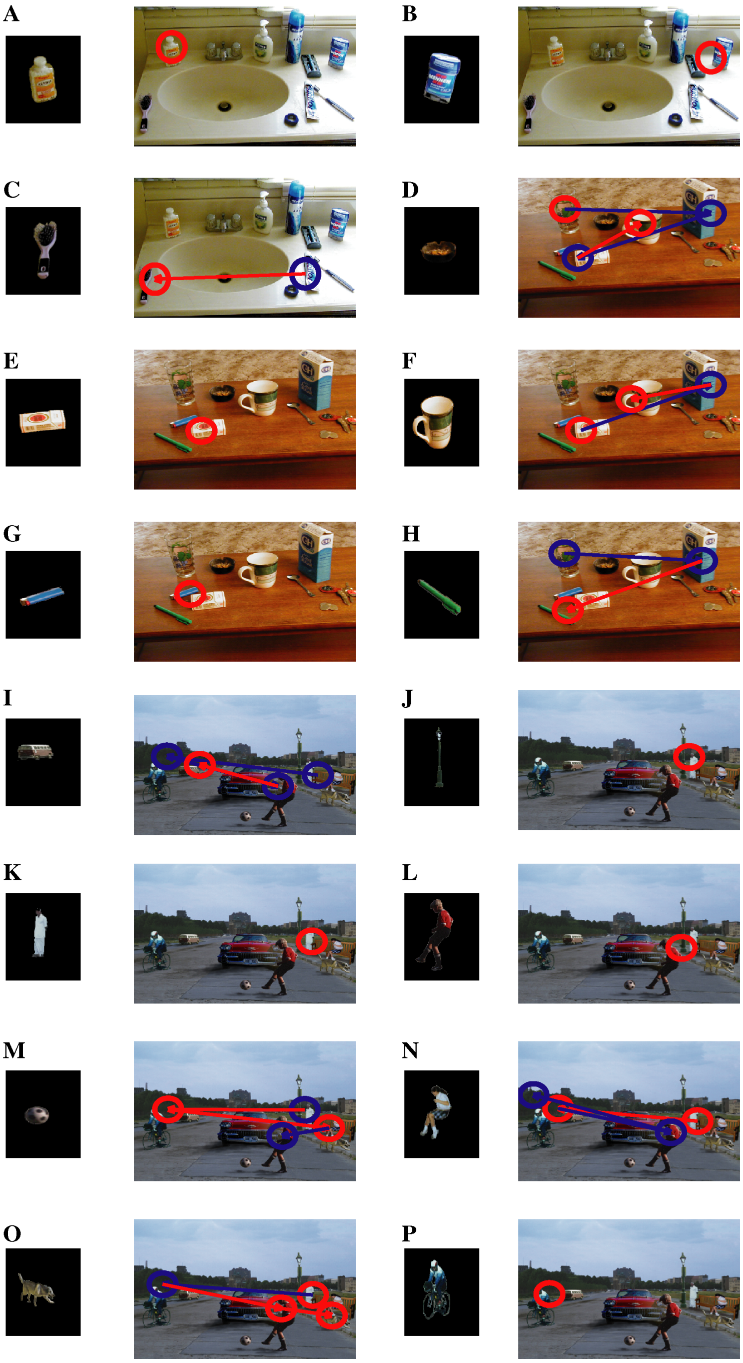


Fig. 10. Overview of the target detection and localization in all tested scenes. A covert shift of spatial attention is indicated in blue and an overt shift in red (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.).

Table 1
Overview of the performance on 16 test scenes (Fig. 10)

Criterion	Performance (%)
Target detection (within four shifts)	81
Immediate selection (no scan)	50
Distractor selection	25
Rejection of the target	0

A correct target detection is defined as a selection of the target object by an overt shift. An immediate selection refers to a parallel search. The criterion distractor selection considers all cases in which a distractor has been selected by an overt shift. The rejection of the target is defined as a covert shift towards a target. In this case the target would have been inspected but not recognized as such.

We have demonstrated that the earlier proposed neurobiological principles also hold for object detection in natural scenes. We are confident that this neurobiological approach provides a high potential for future computer vision tasks. Especially robots that are supposed to operate in the same environment as humans, must be equipped with a highly flexible vision unit that is under cognitive control. Our model provides an interface to cognition, in which cognition determines the templates that guide vision. In our demonstrations, we presented a target object to the model and it memorized some of the objects features. However, in future systems we have to develop cognitive architectures that provide these internal cues. If a task requires to detect a certain person, the knowledge about this person, global shape, face or clothing will be loaded into working memory. Once this information is uploaded it can automatically guide vision without describing and defining each step, like shifting attention, making eye movements, etc., by a central control unit. In our design, the complex problem of scene understanding is transformed into the generation of an appropriate target template. Once a template is generated, we have shown that a system can detect an object by an efficient parallel search as compared to purely saliency-driven models which rely on a sequential search strategy by rapidly selecting parts of the scene and analyzing these conspicuous regions in detail. Classically, a selection on a saliency map defines which stimulus is gated into later processing stages (Section 2.3). Most of the present computational models and machine vision approaches follow this idea. A potential problem can arise from an erroneous segmentation. A poor segmentation will affect all following processes, such as recognition. In our approach, an expectation about an object of interest will increase the conspicuity of the relevant features and filter out the irrelevant ones, which leads to several potential candidate locations. The spatial inference then narrows down the physical search space. Thus, in our approach selection and recognition are ultimately connected with each other.

Tsotsos [96] has pointed out, that attention is rarely mentioned in the computer vision literature. Thus, we would like to stress the function of attention from our point of view. In many computer vision problems the search space has been already seriously reduced before the visual processing takes place [96]. For example, the object detection task simulated by our model could have been also solved by defining an appropriate filter and repeatedly convolving the image with the filter and then determining the maximal response. Nevertheless, this can still result in an expensive search. Attention is often introduced in models to speed up such serial processing. In this respect, a cheap and fast filter is applied first and expensive computations are only performed at some candidate regions. However, is there any more fundamental need for attention that goes beyond spatial guidance? Shifter-circuit models [15,17] suggest a location-based attention mechanism in which a subimage is shifted into a central module to facilitate invariant recognition. A drawback is the enormous amount of image copies required. In parallel architectures translation invariant object recognition could be solved in a hierarchical fashion [97–99]. These models, however, fail in natural scenes. The problem lies in the interference inherent to a convergent projection. Purely hierarchical, bottom-up approaches require that sufficiently complex features are employed, possibly at the level of complexity of the objects themselves, which in turn would lead to a combinatorial explosion of the number of units required [96,17]. Imposing a top-down expectation with a good model seems to be a necessary consequence to reduce the influence of clutter [100]. We propose a solution in which attention filters out irrelevant information within a hierarchy of processing stages. If hierarchical models fail because of the occurrence of interference within a convergent projection, attention could provide a solution if it successfully reduces the interference. For example, if we apply a feedforward approach and compare the image with the target template in parallel, we would almost always receive a match, since the to be searched features are somewhere, distributed in the scene. Our population-based inference operates as a dynamic filter. It allows a successful confirmation or rejection of an object being the target, even with a very simple template. In addition to the spatial component, feature-based attention has been shown to enhance the robustness against clutter within the vicinity of the target object (Fig. 9). In our model attention is not necessarily a prerequisite for object recognition, but it is equivalent to resolving ambiguities over time.

Thus, in our model attention goes much beyond a mere selection of a region or location in the image. Attention allows to transfer the goal of the task flexibly into the process of vision in order to emphasize the relevant aspects within a scene. Attention reduces interference in a hierarchical architecture to facilitate object recognition. And attention enables a distributed, concurrent processing.

Our approach is strongly biologically motivated, but we have to ask for possible limitations that could hinder the transfer of knowledge into computer vision. Current simulations have shown that even very simple information about an object can be used in a parallel multi-cue approach to detect and focus an object. Although our approach seems to be largely invariant against background changes (please note that we initially present the object to the model on a black background and then let the model search for the object within a natural scene), it would be misleading to compare the performance of the model on this specific object detection and localization task with optimized solutions in computer vision (e.g. [101]). We do not claim that the present version is a better approach for object detection in natural scenes than existing computer vision solutions. Our emphasis is on giving insight into the function of attention to pave the way for a transfer of approaches routed in computational neuroscience into computer vision [102].

First of all, vision in this model is based on very simple features. This limits performance in two respects: (i) the representation on which object detection is made does hardly allow for real object recognition tasks and (ii) the guidance of vision can only be based on simple color and orientation cues. However, it is possible to extend our approach by learning feedback and feedforward transformations into feature spaces of different complexity considering image statistics. This would generate a detailed representation of an object and facilitate the guidance of vision by using more complex templates.

Second, the model generates a very simple template from the object which comprises just a single population in each channel. This can lead to very similar target templates of objects even when they appear quite different to us. For example, the man (Fig. 10K) and the girl (Fig. 10N) show identical target templates in three channels and strongly overlapping expectations in the other two channels. Given this very similar template it is surprising that the tiny difference in the template resulted not in the selection of identical target regions. However, this factor clearly limits the models detection performance. The memorization of more than one target feature per channel could potentially improve performance. However, a stronger improvement is to be expected if the target template would not encode simple features but a local arrangement of features, e.g., the spatial relationship of orientations activated by the lighter (Fig. 9).

Third, the computation of the model is expensive on serial computers since vision is defined as an iterative process. Nevertheless, integrated into a large-scale framework and connected to a camera, the model has been demonstrated to operate in real environments [103]. For real-time computer vision tasks, however, the model would require dedicated hardware that makes use of the inherent parallel architecture.

Fourth, the model does not account for invariant recognition. If the model would have to search for a vertical object that is lying on the table, it would reject it since the orientation does not match the template. Invariant object recognition is still an open issue and we do not aim to provide a solution here. At present, we would have to compute more invariant representations or apply a view-based approach. Both approaches are consistent with our approach. By modifying target templates dynamically in the vision process invariant recognition can potentially be facilitated by searching for appropriate templates in time.

Fifth, we integrate the conspicuity across channels to determine the saliency at each location. Although we already consider the reliability of each observation, we give all channels equal weight. However, for a given task one channel can carry more information than others. Thus, it would be more adequate to define the integration across channels as a general problem of cue integration or sensor fusion for which several solutions have been proposed [29,104,105].

Although we can identify limitations of the present implementation, none of those seems to be a limitation inherent to our approach. We are confident that our novel, integrative approach of attention in visual scene perception offers much room for improvements so it can be further developed to provide solutions for present and future computer vision problems.

Acknowledgments

The main part in the development of the present approach has been done at Caltech. I thank James Worcester for his enthusiasm in helping to extend the model for computing natural scenes, Christof Koch for his strong support, Laurent Itti, and Christof Koch for providing the source code of their saliency model from which parts have been used for image preprocessing, Dirk Walther and Rufin VanRullen for helpful discussions, and Rufin VanRullen for providing the test scenes.

Appendix A. Model equations

A.1. Feature maps

To construct the color channels R, G, B, Y, the color values r, g, and b from the image are normalized by $I = (r + g + b)/3$ in order to decouple hue from intensity. As opposed to Itti et al. [21], we do not apply an additional constraint to the normalization, which sets all values in the color opponent channels RG and BY to zero at locations with I smaller than 1/10 of its maximum over the entire image. There are two reasons not to do this. First, in this stage of processing, we are interested in the color values and not how easily they are perceived. Second, we do not use the normalization to enhance low contrast values if no high contrast value is in the map, and therefore we do not have to erase low contrast values to prevent their increase.

For each pixel in the pyramid we generate the color channels $R = r–(g + b)/2$ for red, $G = g–(r + b)/2$ for green, $B = b–(r + g)/2$ for blue, and $Y = (r + g)/2 - |r–g|/2 - b$ for yellow (negative values are set to zero). The brain represents colors within an opponency system $RG = R - G$ and $BY = B - Y$. The values can be negative and positive, but for an easy visualization we shift and rescale the values to 0–255.

A.2. Contrast maps

Contrast maps represent the conspicuity of each feature. For each pixel of the resolution σ we create intensity contrast maps $\mathcal{I}(c,s)$ by subtracting the map with the coarse scale s from the one with center scale c.

$$\mathcal{I}(c,s) = |I(c) \ominus I(s)|_{\delta \in \{3,4\}.}^{c \in \{2,3\},} \tag{10}$$

Similarly, we create the color double opponent values by

$$\begin{aligned} \mathcal{RG}(c,s) &= |RG(c) \ominus RG(s)|, \\ \mathcal{BY}(c,s) &= |BY(c) \ominus BY(s)|. \end{aligned} \tag{11}$$

Since the contrast in the color channels is small for natural images we stretch the scale by a non-linear function s:

$$\begin{aligned}\widehat{\mathscr{RG}}(c,s) &= s(\mathscr{RG}(c,s)),\\ \widehat{\mathscr{BY}}(c,s) &= s(\mathscr{BY}(c,s)).\end{aligned} \tag{12}$$

with

$$s(x) = \begin{cases} k_C \cdot x & \text{if } k_C \cdot x \leqslant 255 \\ 255 & \text{if } k_C \cdot x > 255 \end{cases} \quad \text{and} \quad k_C = 3. \tag{13}$$

This specific scaling function s can equal very high contrast values up to 255, but since the hue is not very high we typically do not run into this situation.

We average the maps obtained by a different course scale $s = \sigma + \delta$ to receive one contrast value per channel and center scale:

$$\begin{aligned}\mathscr{I}(c) &= \frac{1}{\#s} \bigoplus_s \mathscr{I}(c,s),\\ \mathscr{RG}(c) &= \frac{1}{\#s} \bigoplus_s \widehat{\mathscr{RG}}(c,s),\\ \mathscr{BY}(c) &= \frac{1}{\#s} \bigoplus_s \widehat{\mathscr{BY}}(c,s).\end{aligned} \tag{14}$$

Itti et al. [21] suggested a normalization to enhance salient locations before combining maps from different scales. However, we did not find a negative effect on the signal-to-noise ratio by averaging across s, since maps with the same center scale but different surround scales just represent the influence of a different surround size.

Orientation contrast maps are computed for each orientation θ using the center surround operation with a fine scale c and a course scale $s = \sigma + \delta$. However, if the center orientation is different from the surround, the use of absolute values, suggested by Itti et al. [21], would result in a high contrast at the center location in the map with the same orientation as the surround. This interferes with the construction of the population code, since the orientation at the center location is not in the image. To get rid of this effect we only use the values with a higher center than surround input.

$$\mathscr{O}(c,s,\theta) = \begin{cases} O(c,\theta) \ominus O(s,\theta) & \text{if } O(c,\theta) > O(s,\theta), \\ 0 & \text{else.} \end{cases} \tag{15}$$

We observed that this contrast measurement has a favorable effect on the signal-to-noise ratio, since it is less sensitive to highly textured image parts such as bushes and trees, which are indeed perceived as not very salient.

A.3. Feature conspicuity maps

The feature conspicuity maps combine the feature information $\mathbf{V}$, like orientation θ or intensity I, with its gain P, obtained from the conspicuity such as $\mathscr{O}$ or $\mathscr{I}$, into a population code. The feature information is encoded by the location of the cell i in feature space and the conspicuity value ($P \in \{\mathscr{O}, \mathscr{I}, \mathscr{RG}, \mathscr{BY}\}$) determines the firing rate r_i:

$$r_i = P \cdot g(\mathbf{u}_i - \mathbf{V}), \tag{16}$$

Specifically we use a Gaussian tuning curve with the selectivity parameter σ_g:

$$g(\mathbf{u}_i - \mathbf{V}) = \exp\left(-\frac{\|\mathbf{u}_i - \mathbf{V}\|^2}{\sigma_g^2}\right). \tag{17}$$

To apply the same range of selectivity parameters $\sigma_g^2 \in \{0.05\ldots0.2\}$ for all channels we normalize the feature values $\mathbf{V}$ of each channel between 0 and 1 and get $\tilde{\mathrm{I}}$, $\widetilde{RG}$, $\widetilde{BY}$, $\tilde{\theta}$, and $\tilde{\sigma}$. The initial conspicuity value should typically lie within the range of 0 and 1. Thus, we also normalize contrast values to $\tilde{\mathscr{I}}$, $\widetilde{\mathscr{RG}}$, $\widetilde{\mathscr{BY}}$, and $\tilde{\mathscr{O}}$. We finally obtain the populations for each channel with scale c at each location $\mathbf{x}$:

$$\begin{aligned} r_{I,i}(c,\mathbf{x}) &= \tilde{\mathscr{I}}(c,\mathbf{x}) \cdot g(u_i - I(c,\mathbf{x})), \\ r_{RG,i}(c,\mathbf{x}) &= \widetilde{\mathscr{RG}}(c,\mathbf{x}) \cdot g(u_i - RG(c,\mathbf{x})), \\ r_{BY,i}(c,\mathbf{x}) &= \widetilde{\mathscr{BY}}(c,\mathbf{x}) \cdot g(u_i - BY(c,\mathbf{x})). \end{aligned} \tag{18}$$

The orientation information is transferred into two channels, one for scale or spatial frequency σ and one for orientation θ. The orientation channel reads:

$$r_{\theta,i}(c,\theta,\mathbf{x}) = \tilde{\mathscr{O}}(c,\theta,\mathbf{x}) \cdot g(u_i - \theta). \tag{19}$$

Since it is not feasible to represent orientations in different maps within a population code, we combine the maps across orientations:

$$r_{\theta,i}(c,\mathbf{x}) = \max_{\theta}\left(r_{\theta,i}(c,\theta,\mathbf{x})\right). \tag{20}$$

In order to further reduce the information we ignore the different center scales using a convergent mapping (Eq. 2):

$$\begin{aligned} r_{\theta,i}(\mathbf{x}) &= \max_{c,\mathbf{x}' \in RF(\mathbf{x})} r_{\theta,i}(c,\mathbf{x}), \\ r_{I,i}(\mathbf{x}) &= \max_{c,\mathbf{x}' \in RF(\mathbf{x})} r_{I,i}(c,\mathbf{x}), \\ r_{RG,i}(\mathbf{x}) &= \max_{c,\mathbf{x}' \in RF(\mathbf{x})} r_{RG,i}(c,\mathbf{x}), \\ r_{BY,i}(\mathbf{x}) &= \max_{c,\mathbf{x}' \in RF(\mathbf{x})} r_{BY,i}(c,\mathbf{x}). \end{aligned} \tag{21}$$

The 5th conspicuity map is gained from the spatial resolution of the steerable filters. Thus, the orientation information is also transferred into features encoding spatial frequency σ:

$$r_{\sigma,i}(c,\theta,\mathbf{x}) = \mathscr{O}(c,\theta,\mathbf{x}) \cdot g(u_i - \sigma). \tag{22}$$

As for orientation, we combine the maps across spatial frequencies:

$$r_{\sigma,i}(\theta,\mathbf{x}) = \max_{c}\left(r_{\sigma,i}(c,\theta,\mathbf{x})\right) \tag{23}$$

and repeat the same process across the orientation:

$$r_{\sigma,i}(\mathbf{x}) = \max_{\theta}\left(r_{\sigma,i}(\theta,\mathbf{x})\right). \tag{24}$$

A.4. Level I

For each channel $d \in \{\theta, I, RG, BY, \sigma\}$ we use a one-dimensional space $\mathbb{R}$ to encode the features with $i \in N$ units at each location $\mathbf{x}$. Level I units receive input from five channels (d): $r_{\theta,i,\mathbf{x}}$ for orientation, $r_{I,i,x}$ for intensity, $r_{RG,i,\mathbf{x}}$ for red–green opponency, $r_{BY,i,\mathbf{x}}$ for blue–yellow opponency and $r_{\sigma,i,x}$ for spatial frequency. A feature-specific expectation of level I units originates in level II and a location-specific expectation in the movement map.

$$\Delta r^{\mathrm{I}}_{d,i,\mathbf{x}} = G - H, \quad \tau^{\mathrm{I}} = 0.012\ \mathrm{s}, \tag{25}$$

$$\begin{aligned} G = {} & w \cdot r_{d,i,\mathbf{x}} + w \cdot r_{d,i,\mathbf{x}} \cdot \left(\sum_j w^d_{ij} r^{\mathrm{I}}_{d,j,\mathbf{x}} + \sum_{\mathbf{x}'} w^d_{\mathbf{x},\mathbf{x}'} r^{\mathrm{I}}_{d,i,\mathbf{x}'} \right) \\ & + \Gamma\left(A - \max_i (r^{\mathrm{I}}_{d,i,\mathbf{x}}) \right) \cdot \left(w^L r_{d,i,\mathbf{x}} \cdot r^m_{\mathbf{x}} + w^F \max_{\mathbf{x}^{\mathrm{II}}} (r_{d,i,\mathbf{x}} \cdot r^{\mathrm{II}}_{d,i,\mathbf{x}^{\mathrm{II}}}) \right), \end{aligned} \tag{26}$$

$$\begin{aligned} & H = r^{\mathrm{I}}_{d,i,\mathbf{x}} \cdot I^{inh}_{d,i,\mathbf{x}} + I^{f\,inh}_{d,\mathbf{x}} \text{ for orientation,} \\ & H = r^{\mathrm{I}}_{d,i,\mathbf{x}} \cdot I^{inh}_{d,\mathbf{x}} + I^{f\,inh}_{d,\mathbf{x}} \text{ for other channels,} \\ & I^{inh}_{d,i,\mathbf{x}} = w_{inh} \sum_j r^{\mathrm{I}}_{d,j,\mathbf{x}} + w^{RF}_{inh} \max_{\mathbf{x} \in \mathbf{x}^{\mathrm{II}}} z^{RF}_{d,\mathbf{x}^{\mathrm{II}}} + w^{\mathrm{map}}_{inh} z^{\mathrm{map}}_{d,i}, \\ & I^{inh}_{d,\mathbf{x}} = w_{inh} \sum_j r^{I}_{d,j,\mathbf{x}} + w^{RF}_{inh} \max_{\mathbf{x} \in \mathbf{x}^{\mathrm{II}}} z^{RF}_{d,\mathbf{x}^{\mathrm{II}}} + w^{\mathrm{map}}_{inh} z^{\mathrm{map}}_{d}, \\ & I^{f\,inh}_{d,\mathbf{x}} = w_{f\,inh} \sum_j r^{\mathrm{I}}_{d,j,\mathbf{x}}, \end{aligned} \tag{27}$$

where z is an inhibitory unit which receives its input from all cells in the map. The lateral weights support similar features with high conspicuity at the same location. They are computed by a Gaussian

$$w_{ij} = 0.2 \cdot \exp\left(-\frac{(i-j)^2}{0.1} \right). \tag{28}$$

For orientation, the lateral weights are circular to account for the similarity properties of orientations.

$$w_{ij} = 0.2 \cdot \exp\left(\frac{\min((i-j)^2, (I - |i-j|)^2)}{0.04} \right). \tag{29}$$

Identical features at neighboring locations support each other as well.

$$w_{\mathbf{x},\mathbf{x}'} = 0.1 \cdot \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^2}{0.005} \right). \tag{30}$$

The lateral interactions of orientations in space follows not a simple similarity rule, since a dissimilar arrangement can lead to a perception of a pop-out [106].

$$w_{\mathbf{x},\mathbf{x}'} = \begin{cases} 0.126 \cdot \exp^{-\left(\frac{\beta}{d}\right)^2 - 2\left(\frac{\beta}{d}\right)^7 - \frac{d^2}{\pi/2}} & \text{if} \quad (0 < d \leqslant 10) \text{ and } \left(\left(\beta < \frac{\pi}{2.69}\right) \text{ or} \right. \\ & \left(\left(\beta < \frac{\pi}{1.1}\right) \text{ and } \left(|\theta_1| < \frac{\pi}{5.9}\right) \text{ and} \right. \\ & \left.\left.\left(|\theta_2| < \frac{\pi}{5.9}\right)\right)\right) \\ -0.8 \cdot \left(1 - \exp^{-0.4\left(\frac{\beta}{d}\right)^{1.5}}\right) \exp^{\left(\frac{|\Delta\theta|}{\pi/4}\right)^{1.5} - \frac{d^2}{\pi/4}} & \text{if} \quad (0 < d \leqslant 10) \text{ and} \left(\beta \geqslant \frac{\pi}{1.1}\right) \text{ and} \\ & \left(|\Delta\theta| < \frac{\pi}{3}\right) \text{ and} \left(|\theta_1| \geqslant \frac{\pi}{11.999}\right). \\ 0 & \text{else.} \end{cases} \tag{31}$$

The parameters d, θ_1, θ_2, $\Delta\theta$, and β are determined as follows. The distance between two interacting populations is $d = |\mathbf{x} - \mathbf{x}'|$. The angles between the encoded orientations and the line from $\mathbf{x}$ to $\mathbf{x}'$ are denoted as θ_1 and θ_2, β as $\beta = 2|\theta_1| + 2\sin(|\theta_1 + \theta_2|)$ and $\Delta\theta$ as $\Delta\theta = \theta - \theta'$.

Parameters used for the above equations are: $w = 0.7$; $w^L = 50$; $w^F = 15$; $w_{inh} = \frac{1}{\#i}$; $w_{f\,inh} = \frac{1}{\#i}$; $w_{inh}^{RF} = \frac{72}{\#\mathbf{x}}$; $w_{inh}^{\text{map}} = \frac{3.2}{\#\mathbf{x}}$.

A.5. Level II

The conspicuity at level I is transferred by a convergent transformation into a representation with larger receptive fields in level II. The size at level II does not adapt to the image size. We use a 3×3 map with overlapping receptive fields.

$$\Delta r_{d,i,\mathbf{x}}^{\text{II}} = G - H, \quad \tau^{\text{II}} = 0.012 \text{ s}, \tag{32}$$

$$G = \max_{i,\mathbf{x}' \in RF(x)} (F(w \cdot r_{d,i,\mathbf{x}'}^{\text{I}})) \tag{33}$$

$$\begin{aligned} &+ \max_{i,\mathbf{x}' \in RF(x)} (F(w \cdot r_{d,i,\mathbf{x}'}^{\text{I}})) \cdot \left(\sum_j w_{ij}^d r_{d,j,\mathbf{x}}^{\text{II}} + \sum_{\mathbf{x}''} w_{\mathbf{x},\mathbf{x}''}^d r_{d,i,\mathbf{x}''}^{\text{II}}\right) \\ &+ \Gamma\left(A - \max_i (r_{d,i,\mathbf{x}}^{\text{II}})\right) \\ &\cdot \left(w^L \max_{i,\mathbf{x}' \in RF(x)} (F(r_{d,i,\mathbf{x}'}^{\text{I}}) \cdot r_{\mathbf{x}'}^{\text{m}}) + w^F \max_{i,\mathbf{x}' \in RF(x)} (F(r_{d,i,\mathbf{x}'}^{\text{I}}) \cdot r_{d,i}^{\text{T}})\right), \end{aligned} \tag{34}$$

$$\begin{aligned} H &= r_{d,i,\mathbf{x}}^{\text{I}} \cdot I_{d,\mathbf{x}}^{inh} + I_{d,\mathbf{x}}^{f\,inh}, \\ I_{d,\mathbf{x}}^{inh} &= w_{inh} \sum_j y_{d,j,\mathbf{x}}(t) + w_{inh}^{\text{map}} z_d^{\text{map}}(t), \\ I_d^{f\,inh}(t) &= w_{f\,inh}^{\text{map}} z_d^{\text{map}}. \end{aligned} \tag{35}$$

$w_{ij} = 0.35 \cdot \exp(\frac{\min((i-j)^2,(I-|i-j|)^2)}{0.05})$ for orientation $w_{ij} = 0.35 \cdot \exp(\frac{(i-j)^2}{0.05})$ for other populations $w_{inh} = \frac{6}{\#i}$; $w_{inh}^{\text{map}} = \frac{3}{\#\mathbf{x}}$; $w_{f\,inh}^{\text{map}} = 5 \cdot w_{inh}^{\text{map}}$; $w = 1.6$;$w^F = 8$; $w^L = 40$.

A.6. Perceptual map

The perceptual map receives afferents from levels I and II at the same retinotopic location irrespective of the feature information and thus, encodes the conspicuity of locations or often referred to as saliency. Inhibition of return suppresses the saliency at locations that have been visited recently. We define an additional influence from the working memory to further bias those locations that match the target template in all channels. The feedback from the movement map reinforces the selection process.

$$\Delta r^{\mathrm{v}}_{\mathbf{x}} = G - H, \quad \tau^{\mathrm{v}} = 0.012\ \mathrm{s}, \tag{36}$$

$$\begin{aligned} G = {} & \sum_d \left(w^{\mathrm{I}} \max_i r^{\mathrm{I}}_{d,i,\mathbf{x}} + w^{\mathrm{II}} \max_{i,\mathbf{x}' \in RF(x)} r^{\mathrm{II}}_{d,i,\mathbf{x}'} \right) + w^{\mathrm{T}} \prod_d \max_{i,\mathbf{x}' \in RF(\mathbf{x})} r^{\mathrm{T}}_{d,i} \cdot r^{\mathrm{I}}_{d,i,\mathbf{x}'} \\ & + \sum_{\mathbf{x}'} w_{\mathbf{x},\mathbf{x}'} r_{\mathbf{x}'} + w^{\mathrm{m}} r^{\mathrm{m}}_{\mathbf{x}} - w^{\mathrm{IOR}} r^{\mathrm{IOR}}_{\mathbf{x}}, \end{aligned} \tag{37}$$

$$H = r_{\mathbf{x}} (w_{inh} \max_{\mathbf{x}} r_{\mathbf{x}} + w^{\mathrm{map}}_{inh} z^{\mathrm{map}} + B), \tag{38}$$

$w_{\mathbf{x},\mathbf{x}'} = 0.1 \cdot \exp(\frac{(\mathbf{x}-\mathbf{x}')^2}{0.004})$; $B = 0.3$; $w^{\mathrm{I}} = 0.2$; $w^{\mathrm{II}} = 0.04$; $w^{\mathrm{IOR}} = 1$; $w_{inh} = 2$; $w^{\mathrm{map}}_{inh} = \frac{3}{\#\mathbf{x}}$; $w^{\mathrm{m}} = 0.4$; $w^{\mathrm{T}} = 50$.

A.7. Eye movement map

The movement map reads out the saliency and transfers it into an expectation for level I and level II maps. It also indicates an eye movement if the expectation is sufficiently high. The expectation is inhibited by fixation cells. In our example, the task allows an eye movement only towards a target not towards another object. We consider this by tracking the match units.

$$\begin{aligned} & \Delta r^{\mathrm{m}}_{\mathbf{x}} = G - H, \quad \tau^{\mathrm{m}} = 0.015\ \mathrm{s}, \\ & G = w^{\mathrm{v}} r^{\mathrm{v}}_{\mathbf{x}} - w^{\mathrm{v}}_{inh} \sum_{\mathbf{x}} r^{\mathrm{v}}_{\mathbf{x}} + \sum_{\mathbf{x}'} w_{\mathbf{x},\mathbf{x}'} r^{\mathrm{m}}_{\mathbf{x}'} - w^{\mathrm{f}} r^{\mathrm{f}}, \\ & H = r^{\mathrm{m}}_{\mathbf{x}} w^{\mathrm{map}}_{inh} \sum_{\mathbf{x}} r_{\mathbf{x}}, \end{aligned} \tag{39}$$

$w_{\mathbf{x},\mathbf{x}\prime} = 0.1 \cdot \exp(\frac{(\mathbf{x}-\mathbf{x}\prime)^2}{0.004})$; $w^{\mathrm{v}} = 0.3$; $w^{\mathrm{f}} = 0.7$; $w^{\mathrm{map}}_{inh} = 0.075$; $w^{\mathrm{v}}_{inh} = \frac{0.4}{\#\mathbf{x}}$.

If the expectation exceeds the threshold Γ^{m}_0 at the time t_0, we calculate the center of gravity to indicate the location of an eye movement. Thus, in cases with a split of attention the overt shift differs from the covert shift.

$$\mathbf{x}_c = \frac{\sum_{\mathbf{x}} r^{\mathrm{m}}_{\mathbf{x}}(t_0) \cdot \mathbf{x}}{\sum_{\mathbf{x}} r^{\mathrm{m}}_{\mathbf{x}}(t_0)}. \tag{40}$$

A.8. Fixation unit

Some tasks demand an eye movement only when a target is in the scene, but not when the scene contains only distractors. Thus, we define a fixation unit, that is under control of a very simple cognitive process (r^{c}). The fixation unit also resets the movement units: after the expectation in the movement map exceeds the threshold Γ_0^{m} and thus an eye movement is initiated at the time t_0, the fixation unit gets activated for a brief period T^{SAC}.

$$\begin{aligned}
&\Delta r^{\mathrm{f}} = G - H, \quad \tau^{\mathrm{f}} = 0.012\ \mathrm{s},\\
&G = w^{\mathrm{m}} I^{\mathrm{m}} + w^{\mathrm{c}} r^{\mathrm{c}},\\
&H = r^{\mathrm{f}},\\
&I^{\mathrm{m}} = \begin{cases} 1 & \text{if } r^{\mathrm{m}}(t_0) > \Gamma_0^{\mathrm{m}}\ \&\ t < t_0 + T^{\mathrm{SAC}},\\ 0 & \text{else.}\end{cases}
\end{aligned} \tag{41}$$

$T^{\mathrm{SAC}} = 50\,ms$; $\Gamma_0^{\mathrm{m}} = 0.8$; $w^{\mathrm{m}} = 4$; $w^{\mathrm{c}} = 0.6$.

A.9. Inhibtion of return (IOR)

The IOR map serves as a buffer to memorize recently visited locations. Recently visited locations are overt and covert shifts of spatial attention. We regard each location $\mathbf{x}$ as inspected, dependent on the selection of an eye movement at time t_0 and location $\mathbf{x}_c$ or when the attended item at location $\mathbf{x}_m$ does not sufficiently match the target template. The latter case is calculated in the control units and expressed by the variable I^{c}. In this case the IOR cells are charged at the location of the highest expectation in the movement map for a period of time T^{IOR}. The IOR buffer slowly decays with a low weight w_{inh}.

$$\begin{aligned}
&\Delta r_{\mathbf{x}}^{\mathrm{IOR}} = G - H, \quad \tau^{\mathrm{IOR}} = 0.01\ \mathrm{s},\\
&G = (1 - r_{\mathbf{x}}^{\mathrm{IOR}})(I_{\mathbf{x}}^{\mathrm{SC}} + w^{\mathrm{m}} I_{\mathbf{x}}^{\mathrm{m}} \cdot I^{\mathrm{c}}),\\
&H = w_{inh} r_{\mathbf{x}}^{\mathrm{IOR}}\\
&I_{\mathbf{x}}^{\mathrm{SC}} = \begin{cases} \exp(-\frac{(\mathbf{x}-\mathbf{x_c})^2}{0.01}) & \text{if } t < t_0 + T^{\mathrm{IOR}},\\ 0 & \text{else,}\end{cases}\\
&I_{\mathbf{x}}^{\mathrm{m}} = \exp\left(-\frac{(\mathbf{x}-\mathbf{x_m})^2}{0.01}\right) \quad r_{\mathbf{x_m}}^{\mathrm{m}} = \max_{\mathbf{x}}(r_{\mathbf{x}}^{\mathrm{m}}),
\end{aligned} \tag{42}$$

$w^{\mathrm{m}} = 1$; $w_{inh} = 0.02$; $\Gamma_0^{\mathrm{m}} = 0.8$; $T^{\mathrm{IOR}} = 50$ ms; $\Gamma_c^{\mathrm{m}} = 0.4$.

A.10. Target template

We model a simple recurrent local circuit for working memory (T) to encode the expected features of level II units, i.e., the target template. The memorization of a pattern is achieved through recurrent excitation. Whether a pattern should be memorized depends on the task. The variable $I^{\mathrm{store}}(t) \in \{0,1\}$ determines when a pattern should be memorized. It is set externally according to the task instruction. If a pattern is memorized ($r_{d,j}^{\mathrm{T}}$ is high), the term $\Gamma(r_{\Gamma\mathrm{mem}} - w^{\mathrm{cue}} \max_j r_{d,j}^{\mathrm{T}})$ ensures that no other stimulus in level II can penetrate the memory. In the conjunction search experiments we defined the target template externally with I_d^{Target}, instead of showing a cue. r^{mem} indicates by an activity of one that a pattern is in memory.

$$\begin{aligned} \Delta r_{d,i}^{\mathrm{T}} &= G - H, \quad \tau^{\mathrm{T}} = 0.012\ \mathrm{s}, \\ G &= \Gamma(\Gamma_{\mathrm{mem}} - \max_j r_{d,j}^{\mathrm{T}}) \max_{\mathbf{x}} \Gamma(r_{d,i,\mathbf{x}}^{\mathrm{II}} - C) + I_{d,i}^{\mathrm{Target}} + \sum_j w_{ij} r_{d,j}^{\mathrm{T}}, \\ H &= r_{d,i}^{\mathrm{T}} \cdot w_{inh} \sum_j r_{d,j}^{\mathrm{T}} + (0.7 - 0.6 \cdot S_d \cdot I^{\mathrm{store}}) z_d. \end{aligned} \tag{43}$$

For controlling the memorization and deletion we define the following variables:

$$\begin{aligned} r_d^{\mathrm{mem}}(t) &:= \begin{cases} 1 & \text{if } \max_i (r_{d,i}^{\mathrm{T}}) > \Gamma^{\mathrm{mem}}, \\ 0 & \text{else,} \end{cases} \\ S_d &= 1 \quad \text{if } \max(r_{d,i}^{\mathrm{T}}) > 0.1, \\ S_d &= 0 \quad \text{if } I^{\mathrm{mem}} = 0 \ \& \ r_d^{\mathrm{mem}} = 0. \end{aligned} \tag{44}$$

The lateral ij are computed from a Gaussian with $w_{ij} = 0.45 \cdot \exp(\frac{\min((i-j)^2,(I-|i-j|)^2)}{0.005})$ for orientation $w_{ij} = 0.45 \cdot \exp(\frac{(i-j)^2}{0.005})$ for other dimensions $w_{inh} = 0.25; C = 0.05;$ $\Gamma_{mem} = 0.35$.

A.11. Match detection

To determine whether a pattern in the visual scene is similar to the target template we define match units which compare the encoded conspicuity in level II with the expectation in memory. The values in the match population indicate the degree of match. This is implemented by multiplying the template $r_{d,i}^{\mathrm{T}}$ with the conspicuity at level II $r_{d,i,\mathbf{x}}^{\mathrm{II}}$.

$$\begin{aligned} \Delta r_{d,i}^{\mathrm{md}} &= G - H, \quad \tau^{\mathrm{md}} = 0.012\ \mathrm{s}, \\ G &= w \cdot r_{d,i}^{\mathrm{T}} \max_{\mathbf{x}} (r_{d,i,\mathbf{x}}^{\mathrm{II}}) + \sum_j w_{ij} r_{d,j}^{\mathrm{md}}, \\ H &= r_{d,i}^{\mathrm{md}} \cdot w_{inh} \sum_j r_{d,j}^{\mathrm{md}} + w_{f\,inh} \sum_j r_{d,j}^{\mathrm{md}}, \end{aligned} \tag{45}$$

$w_{ij} = 0.45 \cdot \exp(\frac{\min((i-j)^2,(I-|i-j|)^2)}{0.005})$ or orientation,$w_{ij} = 0.45 \cdot \exp(\frac{(i-j)^2}{0.005})$ for other dimensions, $w = 3.0$; $w_{inh} = 0.6$; $w_{finh} = 0.6$.

A.12. Control unit

To meet different task demands we implemented a set of rules that control the memory and internal actions, like fixation. They define when and under what conditions a saccade has to be executed. In the example given below, the unit gets activated if there is a high expectation in the movement map ($\max(r_{\mathbf{x}}^{\mathrm{m}}) > \Gamma_c^{\mathrm{m}}$) and no match occurs ($\max_i(r_{i,d}^{\mathrm{md}}) < \Gamma_d^{\mathrm{OFF}} \forall d$). This rule is applied after $t > T_0$ to allow the presentation of previous images without affecting the rule. $I^{\mathrm{Fix}}(t)$ is set to 1 by the task definition if by no means an eye movement should occur. The planning of an eye movement is always allowed if $I^{\mathrm{mem}}(t) = 1$, to ensure that a loss of the match is not generating an inhibition of return.

$$\begin{aligned}\Delta r^{\mathrm{c}} &= G - H, \quad \tau^{\mathrm{c}} = 0.012\ \mathrm{s},\\ G &= I^{\mathrm{c}} + w_{\mathrm{Fix}} I^{\mathrm{Fix}},\\ H &= r^{\mathrm{c}},\end{aligned}$$

$$I^{\Delta\mathrm{move}}(t) = \begin{cases} 1 & \text{if } I^{\mathrm{no-item}}(t_0 + \Delta t) - I^{\mathrm{no-item}}(t_0) = 1 \ \&\ t < t_0 + T^{\mathrm{IOR}},\\ 0 & \text{else}\end{cases} \tag{46}$$

$$I^{\mathrm{no-item}}(t) = \begin{cases} 0 & \text{if } I^{\mathrm{mem}},\\ 1 & \text{if } \max(r^{\mathrm{m}}_{\mathbf{x}}) > \Gamma^{\mathrm{m}}_{c} \ \&\ \max\limits_{i}(r^{\mathrm{md}}_{i,d}) < \Gamma^{\mathrm{OFF}}_{d} \quad \forall d \text{ with } r^{\mathrm{mem}}_{d} = 1,\\ 0 & \text{else}\end{cases} \tag{47}$$

$$I^{\mathrm{c}} = \max(I^{\Delta\mathrm{move}}, I^{\mathrm{no-item}}), \tag{48}$$

$\Gamma^{\mathrm{m}}_{c} = 0.4$; $w_{\mathrm{Fix}} = 10.2$; $T^{\mathrm{IOR}} = 50$ ms.

References

[1] J. Tsotsos, Analyzing vision at the complexity level, Behav. Brain Sci. 13 (1990) 423–445.

[2] D. Ballard, Animate vision, Artif. Intell. 48 (1991) 57–86.

[3] P. Parodi, R. Lanciwicki, A. Vijh, J.K. Tsotsos, Empiricaly-derived estimates of the complexity of labeling line drawings of polyhedral scenes, Artif. Intell. 105 (1998) 47–75.

[4] M.I. Posner, C.R.R. Snyder, B.J. Davidson, Attention and the detection of signals, J. Exp. Psychol.: General 109 (1980) 160–174.

[5] A. Treisman, G. Gelade, A feature integration theory of attention, Cognit. Psychol. 12 (1980) 97–136.

[6] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Psychol. 4 (1985) 219–227.

[7] A. Treisman, S. Sato, Conjunction search revisited, J. Exp. Psychol.: Hum. Percept. Perform. 16 (1990) 459–478.

[8] J.M. Wolfe, Guided search 2.0: a revised model of visual search, Psychon. Bull. Rev. 1 (1994) 202–238.

[9] J. Duncan, G.W. Humphreys, R. Ward, Competitive brain activity in visual attention, Curr. Opin. Neurobiol. 7 (1997) 255–261.

[10] S. Ahmad, VISIT: an efficient computational model of human visual attention, Ph.D. Thesis, University of California, Berkley, 1991.

[11] J.K. Tsotsos, S.M. Culhane, W. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artif. Intell. 78 (1995) 507–545.

[12] F. Cutzu, J.K. Tsotsos, The selective tuning model of attention: psychophysical evidence for a suppressive annulus around an attended item, Vision Res. 43 (2003) 205–219.

[13] E.O. Postma, H.J. van den Herik, P.T. Hudson, SCAN: a scalable model of attentional selection, Neural Netw. 10 (1997) 993–1015.

[14] M.C. Mozer, M. Sitton, Computational modeling of spatial attention, in: H. Pashler (Ed.), Attention, Psychology Press, East Sussex, UK, 1998, pp. 341–393.

[15] 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 (1998) 4700–4719.

[16] D. Heinke, G.W. Humphreys, Attention, spatial representation and visual neglect: simulating emergent attention and spatial memory in the Selective Attention for Identification Model (SAIM), Psychol. Rev. 110 (2003) 29–87.

[17] Y. Amit, M. Mascaro, An integrated network for invariant visual detection and recognition, Vision Res. 43 (2003) 2073–2088.

[18] G. Deco, J. Zihl, A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system, Psychol. Rev. 10 (2001) 231–253.

[19] R. Milanese, Detecting salient regions in an image: from biological evidence to computer implementation, Ph.D. Thesis, University of Geneva, 1993.

[20] G. Backer, B. Mertsching, M. Bollmann, Data- and model-driven gaze control for an active-vision system, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 23 (2001) 1415–1429.

[21] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 20 (1998) 1254–1259.

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

[23] R. Milanese, H. Wechsler, S. Gil, J.-M. Bost, T. Pun, Integration of bottom-up and top-down cues for visual attention using non-linear relaxation, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, 1994, pp. 781–785.

[24] R. Milanese, S. Gil, T. Pun, Attentive mechanisms for dynamic and static scene analysis, Opt. Eng. 34 (1995) 2428–2434.

[25] H. Tagare, K. Toyama, J.G. Wang, A maximum-likelihood strategy for directing attention during visual search, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 23 (2001) 490–500.

[26] P. van de Laar, T. Heskes, S. Gielen, Task-dependent learning of attention, Neural Netw. 10 (1997) 981–992.

[27] J.M. Wolfe, K. Cave, S. Franzel, Guided search: an alternative to the feature integration model for visual search, J. Exp. Psychol.: Hum. Percept. Perform. 15 (1989) 419–433.

[28] F.H. Hamker, H.-M. Gross, Task relevant relaxation network for visuo-motory systems, in: Proc. Internat. Conf. on Pattern Recognition (ICPR'96), Vienna, 1996, pp. 406–410.

[29] S. Dickinson, H. Christensen, J. Tsotsos, G. Olofsson, Active object recognition integrating attention and viewpoint control, Comput. Vision Image Understand. 67 (1997) 239–260.

[30] F.H. Hamker, H.-M. Gross, Object selection with dynamic neural maps, in: Proc. Internat. Conf. on Artificial Neural Networks (ICANN'97), Springer-Verlag, Lausanne, pp. 919–924.

[31] A. Maki, P. Nordlund, J.-O. Eklundh, Attentional scene segmentation: integrating depth and motion, Comput. Vision Image Understand. 78 (2000) 351–373.

[32] 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, Springer Verlag, Berlin, Heidelberg, New York, pp. 472–479.

[33] B. Takacs, H. Wechsler, A dynamic and multiresolution model of visual attention and its application to facial landmark detection, Comput. Vision Image Understand. 70 (1998) 63–73.

[34] S. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields, 27 (1977) 77–87.

[35] K. Kopecz, Neural field dynamics provide robust control of attentional resources, in: B. Mertsching (Ed.), Aktives Sehen in technischen und natürlichen Systemen, AKA Akademische Verlagsgesellschaft, Berlin, 1996.

[36] A. Corradini, U.-D. Braumann, H.-J. Boehme, H.-M. Gross, Contour-based person localization by 3D neural fields and steerable filters, in: Proc. IAPR Workshop on Machine Vision Applications (MVA'98), Chiba (Japan), 1998, pp. 93–96.

[37] 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, Springer Verlag, London, 1999, pp. 252–261.

[38] F.H. Hamker, Distributed competition in directed attention, in: G. Baratoff, H. Neumann (Eds.), Proceedings of the Artificial Intelligence, vol. 9, Dynamische Perzeption, AKA Akademische Verlagsgesellschaft, Berlin, 2000, pp. 39–44.

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

[40] G. Deco, O. Pollatos, J. Zihl, The time course of selective visual attention: theory and experiments, Vision Res. 42 (2002) 2925–2945.

[41] F.H. Hamker, The reentry hypothesis: linking eye movements to visual perception, J. Vision 11 (2003) 808–816.

[42] F.H. Hamker, A dynamic model of how feature cues guide spatial attention, Vision Res. 44 (2004) 501–521.

[43] B.C. Motter, Neural correlates of attentive selection for color or luminance in extrastriate area V4, J. Neurosci. 14 (1994) 2178–2189.

[44] L. Chelazzi, J. Duncan, E.K. Miller, R. Desimone, Responses of neurons in inferior temporal cortex during memory-guided visual search, J. Neurophysiol. 80 (1993) 2918–2940.

[45] S. Treue, J.C. Martínez Trujillo, Feature-based attention influences motion processing gain in macaque visual cortex, Nature 399 (1999) 575–579.

[46] S. Ullman, Sequence seeking and counter streams: a computational model for bidirectional flow in the visual cortex, Cereb. Cortex 5 (1995) 1–11.

[47] E. Koechlin, Y. Burnod, Dual population coding in the neocortex: a model of interaction between representation and attention in the visual cortex, J. Cogn. Neurosci. 8 (1996) 353–370.

[48] K.L. Kirkland, G.L. Gerstein, A feedback model of attention and context dependence in visual cortical networks, J. Comput. Neurosci. 7 (1999) 255–267.

[49] F. van der Velde, M. de Kamps, From knowing what to knowing where: modeling object-based attention with feedback disinhibition of activation, J. Cogn. Neurosci. 13 (2001) 479–491.

[50] P.R. Roelfsema, V.A. Lammé, H. Spekreijse, H. Bosch, Figure-ground segregation in a recurrent network architecture, J. Cogn. Neurosci. 14 (2002) 525–537.

[51] B.J. Scholl, Objects and attention: the state of the art, Cognition 80 (2991) 1–46.

[52] G.W. Humphreys, H. Müller, Search via recursive rejection (SERR): a connectionist model of visual search, Cognit. Psychol. 25 (1993) 43–110.

[53] S. Grossberg, E. Mingolla, W.D. Ross, A neural theory of attentive visual search: interactions of boundary, surface, spatial and object representations, Psychol. Rev. 101 (1994) 470–489.

[54] M. Behrmann, R.S. Zemel, M.C. Mozer, Object-based attention and occlusion: evidence from normal participants and a computational model, J. Exp. Psychol.: Hum. Percept. Perform. 24 (1998) 1011–1036.

[55] S. Grossberg, R. Raizada, Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex, Vision Res. 40 (2000) 1413–1432.

[56] Y. Sun, R. Fisher, Object-based visual attention for computer vision, Artif. Intell. 146 (2003) 77–123.

[57] C.J. McAdams, J.H. Maunsell, Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4, J. Neurosci. 19 (1999) 431–441.

[58] H. Nakahara, S. Wu, S. Amari, Attention modulation of neural tuning through peak and base rate, Neural Comput. 13 (2001) 2031–2048.

[59] F.H. Hamker, Predictions of a model of spatial attention using sum- and max-pooling functions, Neurocomputing C 56 (2004) 329–343.

[60] R. Desimone, J. Duncan, Neural mechanisms of selective attention, Annu. Rev. Neurosci. 18 (1995) 193–222.

[61] E. Niebur, C. Koch, A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons, J. Comput. Neurosci. 1 (1994) 141–158.

[62] M. Usher, E. Niebur, Modeling the temporal dynamics of IT neurons in visual search: a mechanism for top-down selective attention, J. Cogn. Neurosci. 8 (1996) 311–327.

[63] J.H. Reynolds, L. Chelazzi, R. Desimone, Competetive mechanism subserve attention in macaque areas V2 and V4, J. Neurosci. 19 (1999) 1736–1753.

[64] 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 15 (2005) 431–447.

[65] F.H. Hamker, J. Worcester, Object detection in natural scenes by feedback, in: Bülthoff H.H. (Ed.), Biologically Motivated Computer Vision, Lecture Notes in Computer Science, Springer Verlag, Berlin, Heidelberg, New York, 2002, pp. 398–407.

[66] A.P. Georgopoulos, R.E. Kettner, A.B. Schwartz, Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population, J. Neurosci. 8 (1988) 2928–2937.

[67] T.D. Sanger, Probability density estimation for the interpretation of neural population codes, J. Neurophysiol. 76 (1996) 2790–2793.

[68] A. Pouget, P. Dayan, R. Zemel, Information processing with population codes, Nat. Rev. Neurosci. 1 (2000) 125–132.

[69] E. Kobatake, K. Tanaka, Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex, J. Neurophysiol. 71 (1994) 856–867.

[70] J. Duncan, G.W. Humphreys, Beyond the search surface: visual search and attentional engagement, J. Exp. Psychol.: Hum. Percept. Perform. 18 (1992) 578–588.

[71] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vision Res. 42 (2002) 107–123.

[72] P.J. Burt, E.H. Adelson, The Laplacian pyramid as a compact image code, IEEE Trans. Commun. 3 (1983) 532–540.

[73] H. Greenspan, S. Belongie, P. Perona, R. Goodman, S. Rakshit, C. Anderson, Overcomplete steerable pyramid filters and rotation invariance, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1994, pp. 222–228.

[74] D. Kersten, A. Yuille, Bayesian models of object perception, Curr. Opin. Neurobiol. 13 (2003) 150–158.

[75] E. Koechlin, J.L. Anton, Y. Burnod, Bayesian inference in populations of cortical neurons: a model of motion integration and segmentation in area MT, Biol. Cybern. 80 (1999) 25–44.

[76] T. Moore, K.M. Armstrong, Selective gating of visual signals by microstimulation of frontal cortex, Nature 421 (2003) 370–373.

[77] J.H. Reynolds, T. Pasternak, R. Desimone, Attention increases sensitivity of V4 neurons, Neuron 26 (2000) 703–714.

[78] G. Rizzolatti, L. Riggio, I. Dascola, C. Umiltá, Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention, Neuropsychologica 25 (1987) 31–40.

[79] A.A. Kustov, D.L. Robinson, Shared neural control of attentional shifts and eye movements, Nature 384 (1996) 74–77.

[80] E. Kowler, E. Anderson, B. Dosher, E. Blaser, The role of attention in the programming of saccades, Vision Res. 35 (1995) 1897–1916.

[81] H. Deubel, W.X. Schneider, Saccade target selection and object recognition: evidence for a common attentional mechanism, Vision Res. 36 (1996) 1827–1837.

[82] J.M. Wolfe, What can 1 Million trials tell us about visual search? Psychol. Sci. 9 (1998) 33–39.

[83] E. Weichselgartner, G. Sperling, Dynamics of automatic and controlled visual attention, Science 238 (1987) 778–780.

[84] E. Bricolo, T. Gianesini, A. Fanini, C. Bundesen, L. Chelazzi, Serial attention mechanisms in visual search: a direct behavioral demonstration, J. Cogn. Neurosci. 14 (2002) 980–993.

[85] G.F. Woodman, S.J. Luck, Electrophysiological measurement of rapid shifts of attention during visual search, Nature 400 (1999) 867–869.

[86] B.C. Motter, E.J. Belky, The guidance of eye movements during active visual search, Vision Res. 38 (1998) 1805–1815.

[87] N.P. Bichot, K.R. Cave, H. Pashler, Visual selection mediated by location: feature-based selection of noncontiguous locations, Percept. Psychophys. 61 (1999) 403–423.

[88] E. Awh, H. Pashler, Evidence for split attentional foci, J. Exp. Psychol.: Hum. Percept. Perform. 26 (2000) 834–846.

[89] F.H. Hamker, R. VanRullen, The time course of attentional selection among competing locations [Abstract], J. Vision 2 (2002) 7a.

[90] M.M. Müller, P. MalinowskI, T. Gruber, S.A. Hillyard, Sustained division of the attentional spotlight, Nature 424 (2003) 309–312.

[91] F.H. Hamker, A computational model of visual stability and change detection during eye movements in real world scenes, Vis. Cognit., in press.

[92] C.W. Eriksen, Y.Y. Yeh, Allocation of attention in the visual field, J. Exp. Psychol.: Hum. Percept. Perform. 11 (1985) 583–597.

[93] G.A. Rousselet, S.J. Thorpe, M. Fabre-Thorpe, How parallel is visual processing in the ventral pathway? Trends Cogn. Sci. 8 (2004) 363–370.

[94] 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 (2001) 1340–1350.

[95] E.T. Rolls, N.C. Aggelopoulos, F. Zheng, The receptive fields of inferior temporal cortex neurons in natural scenes, J. Neurosci. 23 (2003) 339–348.

[96] J.K. Tsotsos, Motion understanding: task-directed attention and representations that link perception with action, Int. J. Comput. Vision 45 (2001) 265–280.

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

[98] G. Wallis, E.T. Rolls, Invariant face and object recognition in the visual system, Prog. Neurobiol. 51 (1997) 167–194.

[99] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nat. Neurosci. 2 (1999) 1019–1025.

[100] S.C. Zhu, R. Zhang, Z.W. Tu, Integrating top-down/bottom-up for object recognition by data driven Markov chain Monte Carlo, in: Proc. Internat Conf. on Computer Vision and Pattern Recognition (CVPR 2000), 2000, pp.738–745.

[101] A.J. Baerveldt, A vision system for object verification and localization based on local features, J. Robot. Auton. Syst. 34 (2001) 83–92.

[102] F.H. Hamker, Modeling attention: from computational neuroscience to computer vision, in: L. Paletta et al. (Eds.), Attention and Performance in Computational Vision, Second International Workshop on Attention and Performance in Computer Vision (WAPCV 2004), LNCS 3368, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 118–132.

[103] M. Arnold, E. Ros, O. Coenen, F. Hamker, C. Assad, M. Jabri, T. Sejnowski, From Visual Attention through to Motor Control, Nips 2003 Demonstration, http://www.nips.snl.salk.edu/Conferences/2003/Demonstrations.php

[104] J. Trieschĕ, C. vonder Malsburg, Democratic integration: self-organized integration of adaptive cues, Neural Comput. 13 (2001) 2049–2074.

[105] H. Wu, M. Siegel, R. Stiefelhagen, J. Yang, Sensor fusion using dempster-shafer theory, in: Proc. IEEE Instrumentation and Measurement Technology Conf., Anchorage, AK, USA, 2002.

[106] Z. Li, A saliency map in primary visual cortex, Trends Cogn. Sci. 6 (002) 9–16.