Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 19 (2006) 1395–1407
Neural Networks
www.elsevier.com/locate/neunet

2006 Special Issue

# Modeling attention to salient proto-objects

Dirk Walther[a,*], Christof Koch[b]

[a] *Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 N. Mathews Ave., Urbana, IL 61801, USA*
[b] *Division of Biology, California Institute of Technology, MC 216-76, Pasadena, CA 91125, USA*

**Abstract**

Selective visual attention is believed to be responsible for serializing visual information for recognizing one object at a time in a complex scene. But how can we attend to objects *before* they are recognized? In coherence theory of visual cognition, so-called proto-objects form volatile units of visual information that can be accessed by selective attention and subsequently validated as actual objects. We propose a biologically plausible model of forming and attending to proto-objects in natural scenes. We demonstrate that the suggested model can enable a model of object recognition in cortex to expand from recognizing individual objects in isolation to sequentially recognizing all objects in a more complex scene.


*Keywords:* Visual attention; Proto-objects; Object recognition; Attention model

## 1. Introduction

Attention as a selective gating mechanism is often compared to a spotlight (Posner, 1980; Treisman & Gelade, 1980), enhancing visual processing in the attended ("illuminated") region of a few degrees of visual angle (Sagi & Julesz, 1986). In a modification to the spotlight metaphor, the size of the attended region can be adjusted depending on the task, making attention similar to a zoom lens (Eriksen & St. James, 1986; Shulman & Wilson, 1987). Neither of these theories considers the shape and extent of the attended object for determining the attended area. This may seem natural, since commonly attention is believed to act *before* objects are recognized. However, experimental evidence suggests that attention can be tied to objects, object parts, or groups of objects (Duncan, 1984; Egly, Driver, & Rafal, 1994; Roelfsema, Lamme, & Spekreijse, 1998). How can we attend to objects before we recognize them?

Several computational models of visual attention have been suggested. Tsotsos et al. (1995) use local winner-take-all networks and top-down mechanisms to selectively tune model neurons at the attended location. Deco and Schürmann (2000) modulate the spatial resolution of the image based on a top-down attentional control signal. Itti, Koch, and Niebur (1998) introduced a model for bottom-up selective attention based on serially scanning a saliency map, which is computed from local feature contrasts, for salient locations in the order of decreasing saliency. Making extensive use of feedback and long-range cortical connections, Hamker (2005a, 2005b) models the interactions of several brain areas involved in processing visual attention, which enables him to fit both physiological and behavioral data in the literature. Closely following and extending Duncan's Integrated Competition Hypothesis (Duncan, 1997), Sun and Fisher (2003) developed and implemented a common framework for object-based and location-based visual attention using "groupings". Presented with a manually preprocessed input image, their model replicates human viewing behavior for artificial and natural scenes. However, none of these models provides a satisfactory solution to the problem of attending to objects even before they are recognized.

Rensink (2000a, 2000b) introduced the notion of proto-objects in his interpretation of apparent blindness of observers to fairly dramatic changes in a scene when the original and the modified scenes were separated by a blank screen for a few milliseconds (Rensink, Oregan, & Clark, 1997; Simons & Levin, 1998). Rensink described proto-objects as volatile units of visual information that can be bound into a coherent and stable object when accessed by focused attention.

* Corresponding author. Tel.: +1 217 333 9961; fax: +1 217 333 2922.
*E-mail address:* walther@uiuc.edu (D. Walther).

0893-6080/$ - see front matter © 2006 Published by Elsevier Ltd
doi:10.1016/j.neunet.2006.10.001

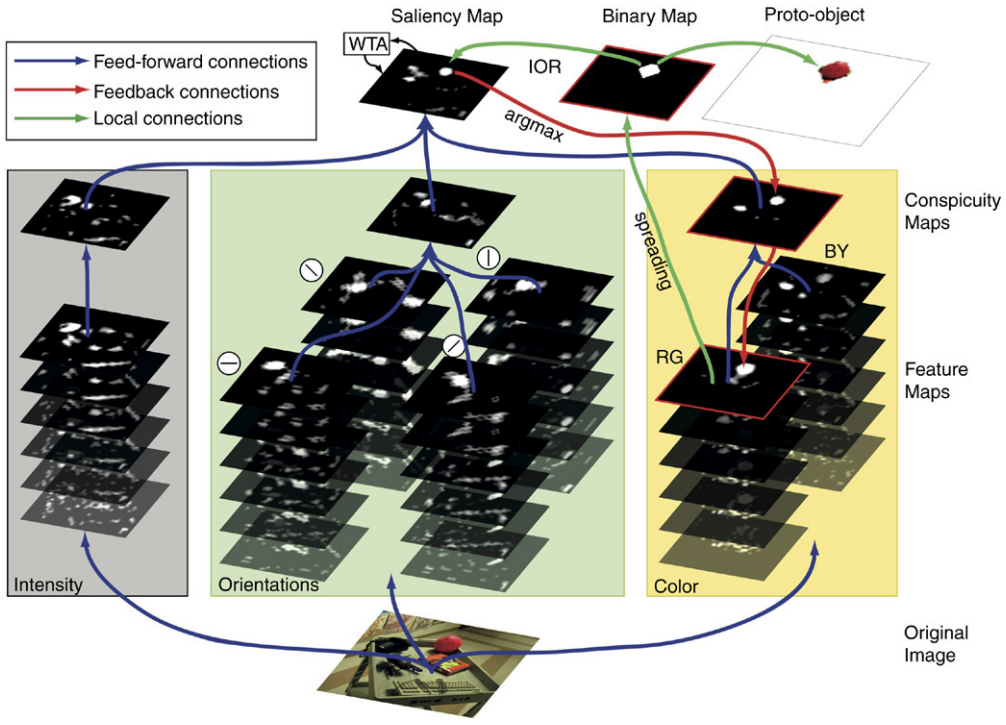


Fig. 1. Illustration of the processing steps for obtaining the attended region. The input image is processed for low-level features at multiple scales, and center-surround differences are computed (Eq. (6)). The resulting feature maps are combined into conspicuity maps (Eq. (9)) and, finally, into a saliency map (Eq. (10)). A winner-take-all (WTA) neural network determines the most salient location, which is then traced back through the various maps (marked in red) to identify the feature map that contributes most to the saliency of that location (Eqs. (11) and (12)). Spreading of attention in this winning feature map around the most salient location (Eq. (14)) yields a binary map that is used as a mask for obtaining the proto-object region as well as for object-based inhibition of return.

In a related concept, Kahneman and Treisman (1984) introduced "object files" as a term for object-specific collections of features in an analogy to case files at a police station. The main difference between proto-objects and object files is the role of location in space. Kahneman and Treisman treat the spatial location of an object as just another property of the object, as just another entry in the related object file. In coherence theory, on the other hand, spatial location has a prominent role as an index for binding together various low-level features into proto-objects across space and time (Rensink, 2000b). See Serences and Yantis (2006) for a recent review of coherence theory and its connections to selective attention.

In this paper we describe a biologically plausible model for generating and attending to proto-object regions. Furthermore, we demonstrate that the model of object recognition in cortex by Riesenhuber and Poggio (1999b) can indeed use these proto-objects successfully to serialize object recognition in multi-object scenes.

## 2. Model architecture

Our attention system is based on the Itti et al. (1998) implementation of the saliency map-based model of bottom-up attention by Koch and Ullman (1985), which models selective attention to salient *locations* in a given image. We extend this model by a process of inferring the extent of a proto-object at the attended location from the maps that are used to compute the saliency map (Fig. 1). In order to explain our extensions in a consistent notation, we first review the Itti et al. (1998) model briefly.

The input image $\mathcal{I}$ is sub-sampled into a dyadic Gaussian pyramid by convolution with a linearly separable Gaussian filter and decimation by a factor of two. Conventionally, convolution in the $x$ direction is followed by decimation in the $x$ direction, and then the procedure is repeated for the $y$ direction (Burt & Adelson, 1983; Itti et al., 1998). By computing convolution results only for pixels that survive subsequent decimation we were able to improve the efficiency of the procedure, reducing the number of multiplications required by a factor of two. For subsampling we use the $6 \times 6$ separable Gaussian kernel $[1\,5\,10\,10\,5\,1]/32$.

By repeating the subsampling and decimation process, the next levels $\sigma = [0, \ldots, 8]$ of the pyramid are obtained. The resolution of level $\sigma$ is $1/2^{\sigma}$ times the original image resolution, i.e., the eighth level has a resolution of $1/256$th of the input image's $\mathcal{I}$ and $(1/256)^2$ of the total number of pixels.

If $r$, $g$, and $b$ are the red, green, and blue values of the color image, then the intensity map is computed as

$$\mathcal{M}_I = \frac{r + g + b}{3}. \tag{1}$$

This operation is repeated for each level of the input pyramid to obtain an intensity pyramid with levels $\mathcal{M}_I(\sigma)$.

Each level of the image pyramid is furthermore decomposed into maps for red–green (RG) and blue–yellow (BY)

opponencies:

$$\mathcal{M}_{\mathrm{RG}} = \frac{r-g}{\max(r,g,b)} \tag{2a}$$

$$\mathcal{M}_{\mathrm{BY}} = \frac{b-\min(r,g)}{\max(r,g,b)}. \tag{2b}$$

To avoid large fluctuations of the color opponency values at low luminance, $\mathcal{M}_{\mathrm{RG}}$ and $\mathcal{M}_{\mathrm{BY}}$ are set to zero at locations with $\max(r,g,b) < 1/10$, assuming a dynamic range of [0, 1]. Note that the definitions in Eq. (2) deviate from the model by Itti et al. (1998) in the definition of yellow and in normalizing by $\max(r,g,b)$ rather than the average $(r+b+g)/3$. See Walther (2006) for the rationale for this definition of color opponency.

Local orientation maps $\mathcal{M}_\theta(\sigma)$ are computed by convolving the levels of the intensity pyramid with Gabor filters:

$$\mathcal{M}_\theta(\sigma) = \|\mathcal{M}_I(\sigma) * G_0(\theta)\| + \|\mathcal{M}_I(\sigma) * G_{\pi/2}(\theta)\|, \tag{3}$$

where

$$G_\psi(x,y,\theta) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\delta^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \tag{4}$$

is a Gabor filter with aspect ratio $\gamma$, standard deviation $\delta$, wavelength $\lambda$, phase $\psi$, and coordinates $(x', y')$ transformed with respect to orientation $\theta$:

$$x' = x\cos(\theta) + y\sin(\theta) \tag{5a}$$

$$y' = -x\sin(\theta) + y\cos(\theta). \tag{5b}$$

We use $\gamma = 1$, $\delta = 7/3$ pixels, $\lambda = 7$ pixels, and $\psi \in \{0, \pi/2\}$. Filters are truncated to $19 \times 19$ pixels.

Center-surround receptive fields are simulated by across-scale subtraction $\ominus$ between two maps at the center ($c$) and the surround ($s$) levels in these pyramids, yielding "feature maps":

$$\mathcal{F}_{l,c,s} = \mathcal{N}(|\mathcal{M}_l(c) \ominus \mathcal{M}_l(s)|) \quad \forall l \in L = L_I \cup L_C \cup L_O \tag{6}$$

with

$$L_I = \{I\}, \tag{7a}$$

$$L_C = \{\mathrm{RG}, \mathrm{BY}\}, \tag{7b}$$

$$L_O = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}. \tag{7c}$$

$\mathcal{N}(\cdot)$ is an iterative, nonlinear normalization operator, simulating local competition between neighboring salient locations (Itti & Koch, 2001). Each iteration step consists of self-excitation and neighbor-induced inhibition, implemented by convolution with a "difference of Gaussians" filter, followed by rectification. For the simulations in this paper, between one and five iterations are used.

The feature maps are summed over the center-surround combinations using across-scale addition $\oplus$, and the sums are normalized again:

$$\bar{\mathcal{F}}_l = \mathcal{N}\left(\bigoplus_{c=2}^{4}\bigoplus_{s=c+3}^{c+4}\mathcal{F}_{l,c,s}\right) \quad \forall l \in L. \tag{8}$$

For the general features color and orientation, the contributions of the sub-features are summed and normalized once more to yield "conspicuity maps". For intensity, the conspicuity map is the same as $\bar{\mathcal{F}}_I$ obtained in Eq. (8):

$$\mathcal{C}_I = \bar{\mathcal{F}}_I, \tag{9a}$$

$$\mathcal{C}_C = \mathcal{N}\left(\sum_{l\in L_C}\bar{\mathcal{F}}_l\right), \tag{9b}$$

$$\mathcal{C}_O = \mathcal{N}\left(\sum_{l\in L_O}\bar{\mathcal{F}}_l\right). \tag{9c}$$

All conspicuity maps are combined into one saliency map:

$$\mathcal{S} = \frac{1}{3}\sum_{k\in\{I,C,O\}}\mathcal{C}_k. \tag{10}$$

The locations in the saliency map compete for the highest saliency value by means of a winner-take-all (WTA) network of integrate-and-fire neurons. Details of the implementation of the WTA network are explained in Section 3.1.

The winning location $(x_w, y_w)$ of this process is attended to, and the saliency map is inhibited. Continuing WTA competition produces the second most salient location, which is attended to subsequently and then inhibited, thus allowing the model to simulate a scan path over the image in the order of decreasing saliency of the attended locations.

In the following paragraphs and in Section 3 we demonstrate a mechanism for extracting an image region around the focus of attention (FOA) that corresponds to the approximate extent of a proto-object at that location. Aside from its use to facilitate further visual processing of the attended proto-object, this mechanism enables object-based inhibition of return (IOR), thereby eliminating the need for a fixed-radius disc as an IOR template as used by Itti et al. (1998).

In order to estimate the proto-object region based on the maps and salient locations computed so far, we introduce feedback connections in the saliency computation hierarchy (Fig. 1). Looking back at the conspicuity maps, we find the one map that contributes the most to the activity at the most salient location:

$$k_w = \operatorname*{argmax}_{k\in\{I,C,O\}} \mathcal{C}_k(x_w, y_w). \tag{11}$$

In Section 3.2 we demonstrate how the *argmax* function, which is critical to this step, can be implemented in a neural network of linear threshold units (LTUs). For practical applications a functionally equivalent generic implementation of the argmax function may be used for its higher computational efficiency.

Examining the feature maps that gave rise to the conspicuity map $\mathcal{C}_{k_w}$, we find the one that contributes most to its activity at the winning location:

$$(l_w, c_w, s_w) = \operatorname*{argmax}_{l\in L_{k_w}, c\in\{2,3,4\}, s\in\{c+3,c+4\}} \mathcal{F}_{l,c,s}(x_w, y_w), \tag{12}$$

with $L_{k_w}$ as defined in Eq. (7).

In the "winning" feature map $\mathcal{F}_{l_w,c_w,s_w}$, activation spreads from the winning location $(x_w, y_w)$ over the shape of the proto-object at this location, which is defined by a contiguous 4-connected neighborhood of above-threshold activity. In

Section 3.3 we present an LTU network capable of this operation.

In image processing terms, the same operation can be realized by thresholding $\mathcal{F}_{l_w,c_w,s_w}$:

$$\mathcal{B}(x,y) = \begin{cases} 1 & \text{if } \mathcal{F}_{l_w,c_w,s_w}(x,y) \geq 0.1 \cdot \mathcal{F}_{l_w,c_w,s_w}(x_w,y_w) \\ 0 & \text{otherwise,} \end{cases} \quad (13)$$

and labeling the resulting binary map $\mathcal{B}$ around $(x_w, y_w)$:

$$\hat{\mathcal{F}}_w = \text{label}(\mathcal{B}, (x_w, y_w)). \quad (14)$$

For the *label* function we use the classical algorithm by Rosenfeld and Pfaltz (1966) as implemented in the Matlab `bwlabel` function. After a first pass over the binary map for assigning temporary labels, the algorithm resolves equivalence classes and replaces the temporary labels with equivalence class labels in a second pass.

The segmented feature map $\hat{\mathcal{F}}_w$ is used to deploy spatial attention to subsequent processing stages such as object detection. Furthermore, $\hat{\mathcal{F}}_w$ is used as a template to trigger object-based inhibition of return (IOR) in the WTA network. Spreading of inhibition over the surface of objects has been reported by Jordan and Tipper (1999) and Tipper, Jordan, and Weaver (1999).

We have implemented our model of salient proto-object selection as part of the SaliencyToolbox for Matlab (http://www.saliencytoolbox.net) and as part of the iLab Neuromorphic Vision (iNVT) C++ toolkit (http://ilab.usc.edu/toolkit/). In the Matlab toolbox we provide both versions of the segmentation operation, the fast image processing implementation, and the LTU network version. They are functionally equivalent, but the LTU network simulation runs much slower than the fast image processing version.

## 3. Biological plausibility of model operations

In the previous section we have outlined the general model architecture. Many of the steps involved are obviously biologically realistic, such as convolution with Gabor filters corresponding to spatial filtering by neurons in the primary visual cortex, or extracting RG and BY color opponencies as done by retinal ganglion cells. Three operations that are crucial to the model, however, require closer scrutiny as far as their biological plausibility is concerned: the winner-take-all network used for identifying the most salient image location, the *argmax* function required for determining the feature map with the highest saliency contribution, and finally spreading of activation over the attended proto-object in that feature map.

### 3.1. Winner-take-all

Winner-take-all neural networks have been extensively discussed in the literature as a way of making decisions. The idea of using mutually inhibiting networks of model neurons for this purpose goes back, at least, to Grossberg (1976a, 1976b). Feldman and Ballard (1983) introduced the notion of a winner-take-all network built from such units. The first use of a WTA network for selective visual attention goes back to Koch and Ullman (1985). In his selective tuning model of visual attention, Tsotsos (1990) uses WTA networks at multiple levels of visual processing to overcome computational complexity barriers for visual search (see also Tsotsos et al. (1995)). Finally, WTA networks were embraced by the analog VLSI community as a means of stimulus selection (see, e.g., Indiveri (1997); Liu (2002)).

For our model, we use the leaky integrate-and-fire neuron implementation of the WTA network as outlined in Itti et al. (1998). The parameters of the model neurons are chosen such that they are physiologically realistic, and such that the ensuing time course of the competition for saliency results in shifts of attention in approximately 30–70 ms simulated time (Saarinen & Julesz, 1991). For further details see Itti and Koch (2000) and Itti (2000).

We now move on to explaining biologically plausible implementations of the *argmax* and *spreading* operations.

### 3.2. Argmax

In Eq. (11) the model requires the identification of the conspicuity map with the highest contribution to the saliency map activity at the winning location. The required operation amounts to an *argmax* function.

Here we demonstrate a network of linear threshold units (LTUs) for this operation for one particular image location. A conspicuity map location is represented by a small local network consisting of four units (Fig. 2): a feedforward unit $f$, a feedback unit $b$, a competition unit $c$, and an inhibitory interneuron $r$. Furthermore, corresponding locations across all conspicuity maps share a pool of inhibitory interneurons $A$.

For $k \in \{I, C, O\}$ let $f_k$ be the activations in the conspicuity maps for intensity ($I$), color ($C$), and orientations ($O$), and $f_{\text{SM}}$ the activation of the saliency map at the same location. With $l \in L_k$ as defined in Eq. (7), let $f_l$ be the activation of the feature maps contributing to conspicuity map $f_k$. Furthermore, let $c_k$ be the activation of a competition unit for conspicuity map $k$, $r_k$ that of an inhibitory interneuron, $b_k$ the activation of a feedback unit for conspicuity map $k$, and $b_{\text{SM}}$ for the saliency map. Finally, let $A$ be the activation of a pool of inhibitory interneurons shared among all conspicuity maps at the corresponding map locations. Then the dynamics of the argmax network in Fig. 2 is

$$f_k \leftarrow \phi\left(\sum_{l \in L_k} f_l\right) \quad (15a)$$

$$f_{\text{SM}} \leftarrow \phi\left(\sum_{k \in \{I,C,O\}} f_k\right) \quad (15b)$$

$$b_k \leftarrow \phi\,(b_{\text{SM}} - r_k) \quad (15c)$$

$$r_k \leftarrow \phi\,(1 - c_k) \quad (15d)$$

$$c_k \leftarrow \phi\,(c_k + f_k - A) \quad (15e)$$

$$A \leftarrow \phi\left(\frac{1}{2}A + \frac{1}{2}\sum_{k \in \{I,C,O\}} c_k\right). \quad (15f)$$

As the activation function $\phi(\cdot)$ we use the half-wave rectifying function:

$$\phi(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

Let us go through these connections one by one and compare with Fig. 2. In Eq. (15a), feedforward (blue) units $f_C$, $f_I$, and $f_O$ compute conspicuity maps for color, intensity, and orientation by pooling activity from the respective sets of feature maps as described in Eqs. (8) and (9), omitting the normalization operator $\mathcal{N}$ here for clarity. In Eq. (15b), the saliency map activation $f_{\mathrm{SM}}$ is computed in a similar fashion (see also Eq. (10)), and $f_{\mathrm{SM}}$ participates in the spatial WTA competition for the most salient location. The feedback (red) unit $b_{\mathrm{SM}}$ receives a signal from the WTA only when this location is attended to, and it relays the signal to the $b$ units in the conspicuity maps (Eq. (15c)). Competition units $c$ together with a pool of inhibitory interneurons $A$ (black) form an across-feature WTA network with input from the $f$ units of the respective conspicuity maps (Eqs. (15e) and (15f)). Only the most active $c$ unit will remain active due to WTA dynamics, allowing it to unblock the respective $b$ unit via an inhibitory interneuron $r$ (Eqs. (15c) and (15d)).

As a result, the activity pattern of the $b$ units represents the outcome of the *argmax* function in Eq. (11). This signal is relayed further to the constituent feature maps, where a similar network selects the feature map with the largest contribution to the saliency of this location (Eq. (12)).

### 3.3. Spreading of activation over proto-objects

Once the feature map with the highest contribution at the attended location is found, the approximate extent of the proto-object at that location must be determined. The proto-object is defined as a contiguous region of high activity in that feature map.

Spreading of attention over objects was first reported by Egly et al. (1994). They showed that invalidly cueing a position for subsequent target detection is still effective when the cue and the target location are located on the same object, but not when a different object is cued, although the distance between (invalid) cue and target locations was the same in both cases. In these experiments, the objects were defined by an outline based on luminance contrast. The same effect has been replicated for objects defined by color (Mitchell, Stoner, Fallah, & Reynolds, 2003; Reynolds, Alborzian, & Stoner, 2003) and illusory contours (Moore, Yantis, & Vaughan, 1998).

Fig. 3 shows a network for spreading the activity over a proto-object. At each image location $(i, j)$, the map activity is given by $f_{i,j}$ and the select signal from the argmax function by $b_{i,j}$. The pooling unit $P_{i,j}$ receives inhibitive input from all neighboring spreading units $S$. The spreading unit $S_{i,j}$ combines excitatory and inhibitory influences into the output activity, which is propagated to the neighboring $P$ units. $u_{i,j}$ and $v_{i,j}$ are local interneurons. The network dynamics is described by:

$$u_{i,j} \leftarrow \phi\left(0.1 - f_{i,j}\right) \quad (17a)$$

$$v_{i,j} \leftarrow \phi'\left(1 - P_{i,j}\right) \quad (17b)$$

$$P_{i,j} \leftarrow \phi\left(1 - S_{i-1,j} - S_{i+1,j} - S_{i,j-1} - S_{i,j+1} - b_{i,j}\right) \quad (17c)$$

$$S_{i,j} \leftarrow \phi\left(u_{i,j} - v_{i,j}\right). \quad (17d)$$

With the exception of Eq. (17b), the activation function $\phi$ is the same as in Eq. (16). The nonlinearity for $v$ (Eq. (17b)) is the Heaviside function $\phi'(x) = \{0, x \leq 0; 1, x > 0\}$.

The spreading unit $S_{i,j}$ receives excitatory input from interneuron $v_{i,j}$ and inhibitory input from interneuron $u_{i,j}$ (Eq. (17d)). The excitatory interneuron $v_{i,j}$ gets inhibited by the pooling unit $P_{i,j}$ (Eq. (17b)), unless $P_{i,j}$ itself receives inhibitory input (Eq. (17c)). Inhibitory input for $P_{i,j}$ arrives either from the spreading units of the neighboring pixels, or from the select signal $b_{i,j}$ (red in Fig. 3), which travels the hierarchy of maps from the saliency map down to the winning feature map at the attended location (see Section 3.2). As a result, activity of $S$ units starts at the selected location and propagates to its neighbors and to their neighbors and so on, contingent on the respective $u$ units being blocked (Eq. (17a)) by sufficiently high map activation $f_{i,j}$ (blue).

The propagation of activity will stop at the border of the proto-object, where map activity falls below the threshold (0.1) required to block the inhibitory $u$ units. The pattern of activity of all $S$ units (green) represents the shape of the resulting proto-object, which is used for modulating object recognition and for object-based IOR.

## 4. Application to biologically plausible object recognition

### 4.1. Introduction

In the previous sections we have described our model for bottom-up attention to salient proto-objects. In the introduction we claimed that such a system would enable learning and recognition of one object at a time in a scene with multiple objects. In this section we set out to prove that this can indeed be achieved when we couple the system with a biologically plausible model of object recognition in cortex.

We adopt the hierarchical model of object recognition by Riesenhuber and Poggio (1999b). While this model works well for individual paper-clip objects, its performance deteriorates quickly when it is presented with scenes that contain several such objects because of erroneous binding of features (Riesenhuber & Poggio, 1999a). To solve this feature binding problem, we supplement the model with a mechanism of modulating the activity of the S2 layer, which has roughly the same receptive field properties as area V4, or the S1 layer, whose properties are similar to simple cells in areas V1 and V2, with an attentional modulation function obtained from our model for saliency-based region selection described in Section 2 (Fig. 4).

Note that only the shape selectivity of neurons in V1/V2 and V4, is captured by the model units. Other aspects such as motion sensitivity of area V1 or color sensitivity of V4 neurons are not considered here. Moreover, only a simple approximation of the shape selectivity of V4 cells is captured. The model has seen further improvements for learning the S2 receptive field from natural scene statistics (Serre, Wolf, & Poggio, 2005), and for extending to a third set of S and C layers,

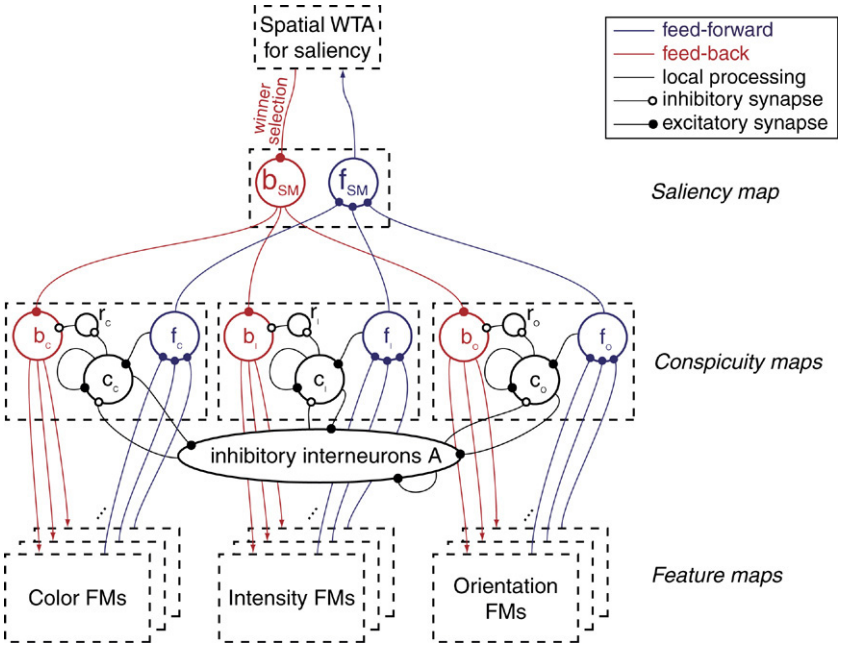


Fig. 2. A network of linear threshold units (LTUs) for computing the *argmax* function in Eq. (11) for one image location. See main text for a detailed description of the network.

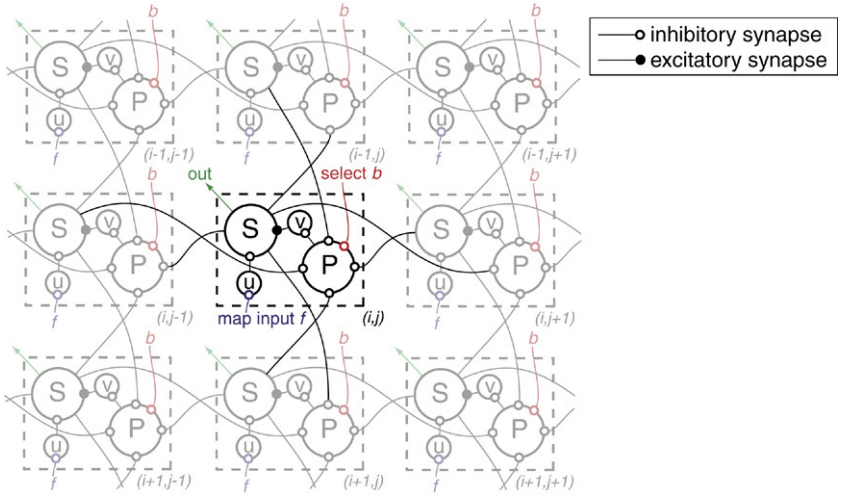


Fig. 3. An LTU network implementation of the segmentation operation in Eqs. (13) and (14).

enabling the model to detect real objects (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, in press). However, for the purpose of demonstrating the effect of attentional selection of proto-objects on recognition performance, we chose the simplest form of the model with hard-wired S2 features.

### 4.2. Recognition model

The hierarchical model of object recognition in cortex by Riesenhuber and Poggio (1999b) starts with S1 simple cells, which extract local orientation information from the input image by convolution with Gabor filters, for the four cardinal orientations at 12 different scales. S1 activity is pooled over local spatial patches and four scale bands using a maximum operation to arrive at C1 complex cells. While still being orientation selective, C1 cells are more invariant to space and scale than S1 cells.

In the next stage, activities from C1 cells with similar positions but different orientation selectivities are combined in a weighted sum to arrive at S2 composite feature cells that are tuned to a dictionary of more complex features. The dictionary we use in this section consists of all possible combinations of the four cardinal orientations in a $2 \times 2$ grid of neurons, i.e., $(2 \times 2)^4 = 256$ different S2 features. This choice of features limits weights to being binary, and, for a particular location in the C1 activity maps, the weight for one and only one of the orientations is set to 1. The S2 layer retains some spatial resolution, which makes it a suitable target for spatial attentional modulation detailed in the next subsection.

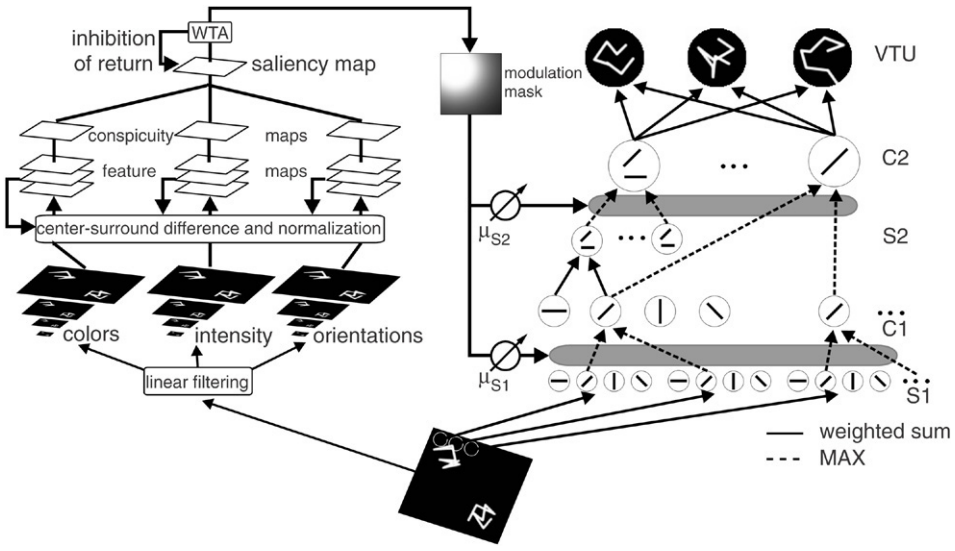


Fig. 4. Sketch of the combined model of bottom-up attention (left) and object recognition (right) with attentional modulation at the S2 or S1 layer as described in Eq. (19). (Adapted from Riesenhuber and Poggio (2003).)

In a final non-linear pooling step over all positions and scale bands, activities of S2 cells are combined into C2 units using the same maximum operation as used from the S1 to the C1 layer. While C2 cells retain their selectivity for the complex features, this final step makes them entirely invariant to location and scale of the preferred stimulus. The activity patterns of the 256 C2 cells feed into view-tuned units (VTUs) with connection weights learned from exposure to training examples. VTUs are tightly tuned to object identity, rotation in depth, illumination, and other object-dependent transformations, but show invariance to translation and scaling of their preferred object view.

In their selectivity to shape, S1 and C1 layers are approximately equivalent to simple and complex cells in areas V1 and V2, S2 to area V4, and C2 and the VTUs to areas in posterior infero-temporal cortex (PIT) with a spectrum of tuning properties ranging from complex features to full object views.

It should be noted that this is a model of fast feedforward processing in object detection. The time course of object detection is not modeled here, which means in particular that such effects as masking or priming are not explained by the model. In this section we introduce feedback connections for deploying spatial attention, thereby introducing some temporal dynamics due to the succession of fixations.

### 4.3. Attentional modulation

Attentional modulation of area V4 has been reported in monkey electrophysiology (Chelazzi, Miller, Duncan, & Desimone, 2001; Connor, Preddie, Gallant, & van Essen, 1997; Luck, Chelazzi, Hillyard, & Desimone, 1997; McAdams & Maunsell, 2000; Moran & Desimone, 1985; Motter, 1994; Reynolds, Pasternak, & Desimone, 2000) as well as human psychophysics (Braun, 1994; Intriligator & Cavanagh, 2001). Other reports find attentional modulation in area V1 using fMRI in humans (Gandhi, Heeger, & Boynton, 1999; Kastner, De Weerd, Desimone, & Ungerleider, 1998) and electrophysiology in macaques (McAdams & Reid, 2005). There are even reports of the modulation of fMRI activity in LGN due to selective attention (O'Connor, Fukui, Pinsk, & Kastner, 2002). See Fig. 7 for an overview of attentional modulation of V4 units in electrophysiology work in macaques.

Here we explore attentional modulation of layers S2 and S1, which correspond approximately to areas V4 and V1, by gain modulation with variable modulation strength (Walther, Itti, Riesenhuber, Poggio, & Koch, 2002). We use the bottom-up salient region selection model introduced in Section 2 to attend to proto-object regions one at a time in order of decreasing saliency. We obtain a modulation mask $\mathcal{F}_M$ by rescaling the winning segmented feature map $\hat{\mathcal{F}}_w$ from Section 3.3 (or Eq. (14)) to the resolution of the S2 or S1 layer, respectively, smoothing it, and normalizing it such that:

$$\mathcal{F}_M(x,y) = \begin{cases} 1 & (x,y) \text{ is inside the object region;} \\ 0 & (x,y) \text{ is far away from the object region;} \\ \text{between 0 and 1} & \text{around the border of the object region.} \end{cases} \tag{18}$$

If $S(x,y)$ is the neural activity at position $(x,y)$, then the modulated activity $S'(x,y)$ is computed according to

$$S'(x,y) = [1 - \mu(1 - \mathcal{F}_M(x,y))] \cdot S(x,y), \tag{19}$$

with $\mu$ being a parameter that determines the modulation strength ($0 \le \mu \le 1$).

This mechanism leads to inhibition of units away from the attended region by an amount that depends on $\mu$. For $\mu = 1$, S2 activity far away from the attended region will be suppressed entirely; for $\mu = 0$, Eq. (19) reduces to $S' = S$, canceling any attention effects.

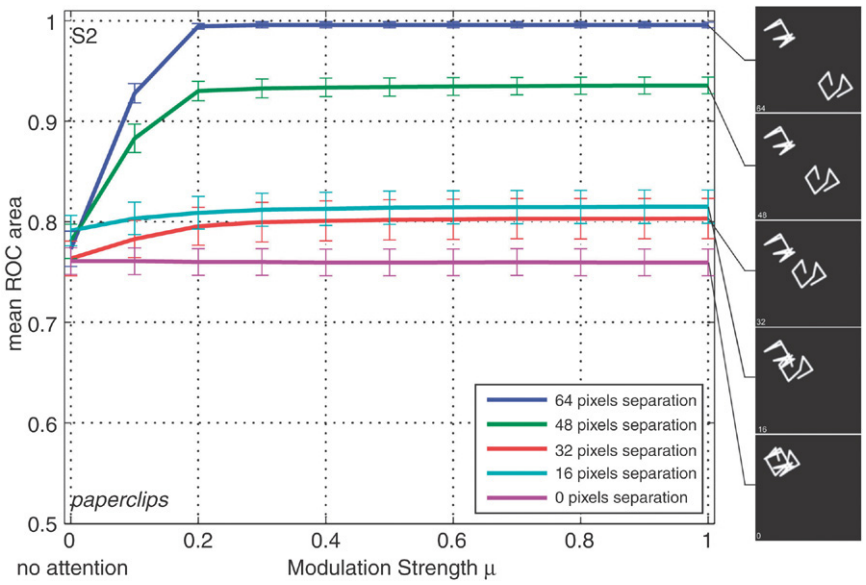


Fig. 5. Mean ROC area for the detection of two paper-clip stimuli. Without attentional modulation ($\mu = 0$), detection performance is around 0.77 for all stimulus separation values. With increasing modulation of S2 activity, individual paper clips can be better distinguished if they are spatially well separated. Performance saturates around $\mu = 0.2$, and a further increase of attentional modulation does not yield any performance gain. Error bars are standard error of the mean. On the right, example displays are shown for each of the separation distances.

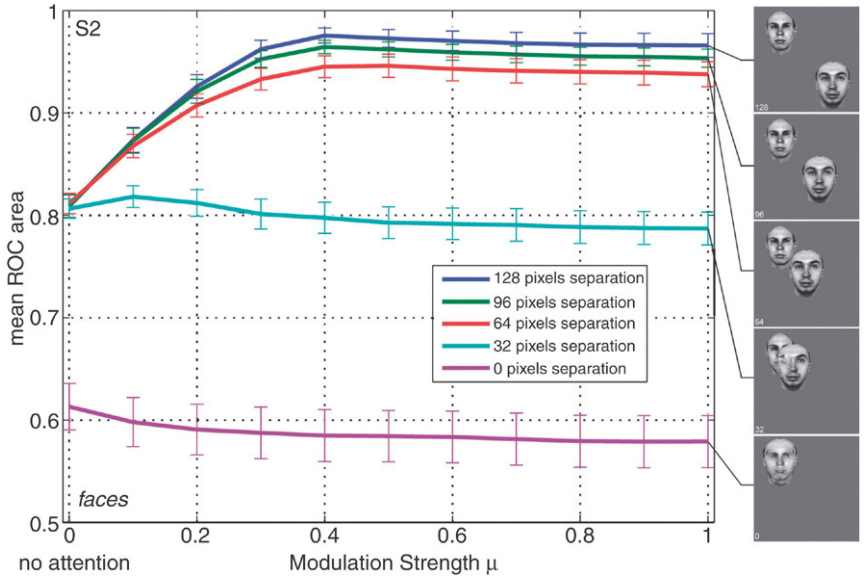


Fig. 6. Performance for detection of two faces in the display as a function of attentional modulation of S2 activity. As in Fig. 5, performance increases with increasing modulation strength if the faces are clearly separated spatially. In this case, mean ROC area saturates at about $\mu = 0.4$. Error bars are the standard error of the mean. Example displays are shown on the right.

### 4.4. Experimental setup

Closely following the methods in Riesenhuber and Poggio (1999b), we trained VTUs for the same 21 paper-clip views that they used. The bent paper-clip objects were first used in an electrophysiology study by Logothetis, Pauls, Bülthoff, and Poggio (1994). Test stimuli consist of displays of 128 × 128 pixels size with one of the 21 paper-clips (64 × 64 pixels) in the top-left corner and another paper-clip superimposed at either the same location (0 pixels) or at 16, 32, 48, or 64 pixels separation in both $x$ and $y$. All combinations of the 21 paper-clips were used, resulting in 441 test displays for each level of object separation. See Fig. 5 for example stimuli.

Rosen (2003) showed that, to some extent, the simple recognition model described above is able to detect and identify faces. To test attentional modulation of object recognition beyond paper clips, we also tested stimuli consisting of synthetic faces rendered from 3D models, which were obtained by scanning the faces of human subjects (Vetter & Blanz, 1999). Again, we trained VTUs on 21 unique face stimuli and created 441 test stimuli of size 256×256 pixels with one face (128×128 pixels) in the top-left corner and a second one at $x$ and $y$

distances of 0, 32, 64, 96, and 128 pixels separation. Example stimuli are shown in Fig. 6.

Each of the 441 stimuli for paper-clips and faces was scanned for salient regions for 1000 ms simulated time of the WTA network, typically yielding between two and four image regions. The stimulus was presented to the VTUs modulated by each of the corresponding modulation masks $\mathcal{F}_M^{(i)}$, and the maximum response of each VTU over all attended locations was recorded. VTUs corresponding to paper-clips or faces that are part of the test stimulus were designated "positive" VTUs, and the others "negative". Based on the responses of positive and negative VTUs an ROC curve was derived, and the area under the curve was recorded as a performance measure. This process was repeated for all 441 paper-clip and all 441 face stimuli for each of the separation values and for $\mu \in \{0, 0.1, 0.2, \ldots, 1\}$.

### 4.5. Results

In Fig. 5 we show the mean ROC area for the detection of paper-clips in our displays composed of two paper-clip objects at a separation distance between 0 pixels (overlapping) and 64 pixels (well separated) for varying attentional modulation strength $\mu$ when modulating S2 activations. In the absence of attention ($\mu = 0$), the recognition system frequently confuses features of the two stimuli, leading to mean ROC areas between 0.76 and 0.79 (mean 0.77). Interestingly, this value is practically independent of the separation of the objects. Already at $\mu = 0.1$, a clear performance increase is discernible for displays with clearly separated objects (64 and 48 pixels separation), which increases further at $\mu = 0.2$ to 0.99 for 64 pixels separation and to 0.93 for 48 pixels separation. For separation distances of 32 and 16 pixels, performance increases only slightly to 0.80, while there is no performance improvement at all in the case of overlapping objects (0 pixels separation), keeping the mean ROC area constant at 0.76. Most importantly, there is no further performance gain beyond $\mu = 0.2$ for any of the stimulus layouts. It makes no difference to the detection performance whether activity outside the focus of attention is decreased by only 20% or suppressed entirely.

Detection performance for faces shows similar behavior when plotted over $\mu$ (Fig. 6), with the exception of the case of overlapping faces (0 pixels separation). Unlike with the mostly transparent paper-clip stimuli, bringing faces to an overlap largely destroys the identifying features of both faces, as can be seen in the bottom example display on the right hand side of Fig. 6. At $\mu = 0$, mean ROC area for these kinds of displays is at 0.61; for cases with object separation larger than 0 pixels, the mean ROC area is at 0.81, independent of separation distance. For the well separated cases (64 or more pixels separation), performance increases continuously with increasing modulation strength until saturating at $\mu = 0.4$ with mean ROC areas of 0.95 (64 pixels), 0.96 (96 pixels), and 0.98 (128 pixels separation), while performance for stimuli that overlap partially or entirely remains roughly constant at 0.80 (32 pixels) and 0.58 (0 pixels), respectively. Increasing $\mu$ beyond 0.4 does not change detection performance any further.

The general shape of the curves in Fig. 6 is similar to those in Fig. 5, with a few exceptions. First and foremost, saturation is reached at a higher modulation strength $\mu$ for the more complex face stimuli than for the fairly simple bent paper-clips. Secondly, detection performance for completely overlapping faces is low for all separation distances, while detection performance for completely overlapping paper-clips for all values of $\mu$ is on the same level as for well separated paper-clips at $\mu = 0$. As can be seen in Fig. 5, paper-clip objects hardly occlude each other when they overlap. Hence, detecting the features of both objects in the panel is possible even when they overlap completely. If the opaque face stimuli overlap entirely, on the other hand, important features of both faces are destroyed (see Fig. 6) and detection performances drops from about 0.8 for clearly separated faces at $\mu = 0$ to about 0.6. A third observation is that mean ROC area for face displays with partial or complete overlap (0 and 32 pixels separation) decreases slightly with increasing modulation strength. In these cases, the focus of attention (FOA) will not always be centered on one of the two faces and, hence, with increasing down-modulation of units outside the FOA, some face features may be suppressed as well.

The results for modulating activity of units at the V1-equivalent S1 layer are almost identical with the results for modulating at the S2 layer for both paper-clips (Fig. 5) and faces (Fig. 6).

### 4.6. Discussion

In our computer simulations, modulating neural activity by as little as 20%–40% is sufficient to effectively deploy selective attention for detecting one object at a time in a multiple object display, and even 10% modulation is effective to some extent. This main result is compatible with a number of reports of attentional modulation of neurons in area V4: Spitzer, Desimone, and Moran (1988), 18%; Connor et al. (1997), 39%; Luck et al. (1997), 30%–42%; Reynolds et al. (2000), 51%; Chelazzi et al. (2001), 39%–63%; McAdams and Maunsell (2000), 31% for spatial attention and 54% for the combination of spatial and feature-based attention. See Fig. 7 for a graphical overview.

While most of these studies used oriented bars (Connor et al., 1997; Luck et al., 1997; Spitzer et al., 1988) or Gabor patches (McAdams & Maunsell, 2000; Reynolds et al., 2000) as stimuli, Chelazzi et al. (2001) use cartoon drawings of real-world objects for their experiments. With these more complex stimuli, Chelazzi et al. (2001) observed stronger modulation of neural activity than was found in the other studies with the simpler stimuli. We observe a similar trend in our simulations, where performance for detecting fairly simple bent paper-clips saturates at a modulation strength of 20%, while detection of the more complex face stimuli only reaches its saturation value at 40% modulation strength. Since they consist of combinations of oriented filters, S2 units are optimally tuned to bent paper-clip stimuli, which are made of straight line segments. Hence, even with attentional modulation of as little as 10% or 20%, discrimination of individual paper-clips is possible. These

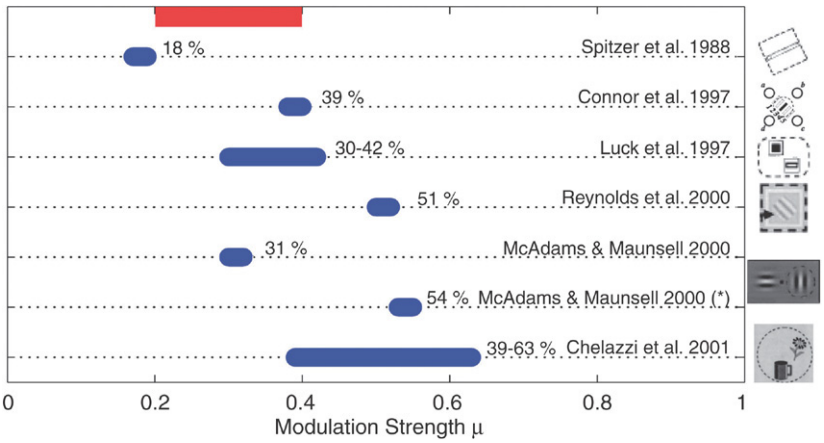


Fig. 7. Modulation of neurons in macaque area V4 due to selective attention in a number of electrophysiology studies (blue). All studies used oriented bars or Gabor patches as stimuli, except for Chelazzi et al. (2001), who used cartoon images of objects. The examples of stimuli shown to the right of the graph are taken from the original papers. The modulation strength necessary to reach saturation of the detection performance in two-object displays in our model is marked in red on the top of the graph. The first number reported by McAdams and Maunsell (2000) is for spatial attention only, the second one (*) is for combined spatial and feature-based attention. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

features are not optimal for the face stimuli, however. For the model to be able to successfully recognize the faces, it is important that the visual information belonging to the attended face is grouped together correctly and that distracting information is suppressed sufficiently.

The recognition model without any attentional feedback cannot detect several objects at once because there is no means of associating the detected features with the correct object. Deploying spatial attention solves this binding problem by spatially grouping features into proto-objects based on their most salient feature.

With their "shifter circuit" model, Olshausen, Anderson, and Van Essen (1993) successfully demonstrated deployment of spatial attention using gain modulation at various levels of the visual processing hierarchy. In combination with an associative memory, their model is capable of object detection invariant to translation and scale. This model, however, has only a rudimentary concept of saliency, relying solely on luminance contrast, and the extent of the attended "blobs" is fixed rather than derived from image properties as done in our model.

Most reports of modulation of area V1 or LGN are fMRI studies (e.g., Gandhi et al. (1999); Kastner et al. (1998); O'Connor et al. (2002)) and do not allow a direct estimation of the level of modulation of neural activity. In a recent electrophysiology study, however, McAdams and Reid (2005) found neurons in macaque V1 whose spiking activity was modulated by up to 27% when the cell's receptive field was attended to.

While our simulation results for modulating the S1 layer agree with this number, we are cautious to draw any strong conclusions. The response of S2 model units is a linear sum of C1 activities, which in turn are max-pooled S1 activities. Therefore, the fact that the results for attentional modulation of activity of S1 units are very similar to the results for modulating S2 activity is not surprising.

To summarize, in our computer simulations of attentional modulation of V4-like layer S2, we found that modulation by 20%–40% suffices for successful sequential detection of artificial objects in multi-object displays. This range for modulation strength agrees well with the values found in several electrophysiological studies of area V4 in macaque monkeys.

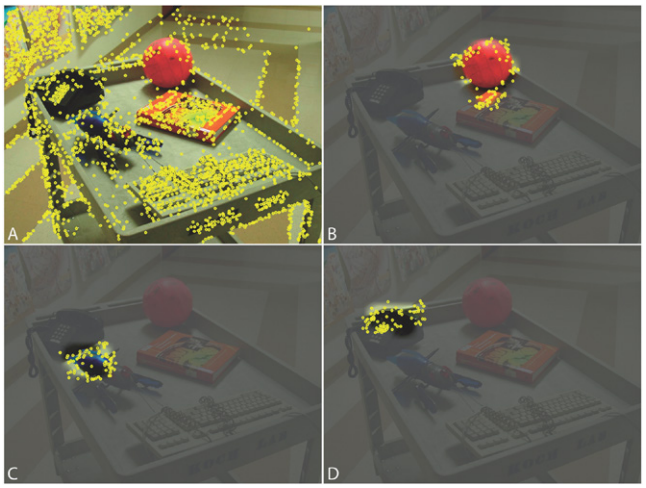


Fig. 8. Illustration of the reduction in complexity of object learning due to salient region selection. (A) Example training image with all keypoints marked in yellow; (B–D) the three most salient proto-object regions marked by contrast modulation, and the keypoints for only the attended regions marked in yellow. There is a clear reduction in complexity when matching the set of keypoints in a test image with sets for the individual proto-objects in (B), (C), and (D), compared to attempting to match with the entire set of keypoints in (A). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## 5. Application to computer vision

In the previous section we have shown how attending to salient proto-objects aids recognition of multiple objects in a biologically plausible way. We have shown there that modulation of neural activity due to attention by as little as 20% can be sufficient to successfully bias recognition toward the attended object.

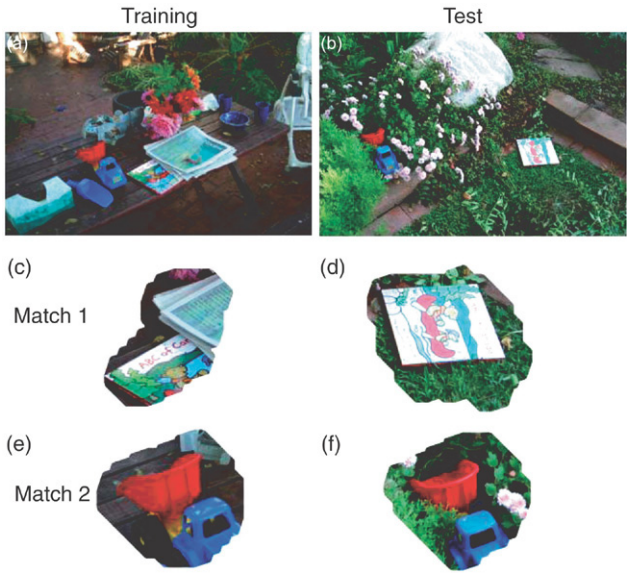


Fig. 9. Example for learning two objects (c) and (e) from the training image (a) and establishing matches (d) and (f) for the objects in the test image (b), in a different visual context, with different object orientations and occlusions. (Adapted from Rutishauser et al. (2004), supplementary material.)

In contrast, in machine vision it is desirable to omit processing of image parts outside the attended region in order to save computational resources. We have demonstrated the feasibility of this process in previously published work (Rutishauser, Walther, Koch, & Perona, 2004; Walther, Rutishauser, Koch, & Perona, 2005) by restricting the selection of scale invariant features (SIFT) in the object recognition model by Lowe (2004) to salient proto-object regions. Fig. 8 shows an example of an image with all SIFT features (A), and with the features restricted to the three most salient proto-objects (B–D). There is a clear reduction in complexity for any algorithm that is trying to match constellations of such keypoints.

In addition to improvements in complexity, selecting salient regions for preferred processing improves robustness to clutter. Most importantly, it becomes possible to learn several objects from an individual image. When asking an object recognition algorithm without any attentional region selection to learn all the objects that are present in the image in Fig. 9(a), for instance, the algorithm would have no notion of which feature points belong to which object. By selecting salient proto-objects, keypoints can be grouped into separate object models, and individual objects can be recognized in a test image (Fig. 9(b)), as shown with the successfully matched picture book ((c) and (d)) and toy truck ((e) and (f)).

To summarize, using selective visual attention for object recognition in the domain of computer vision has at least three major advantages: (i) improved efficiency by scanning the image starting with the regions most likely to contain objects rather than scanning the image from top-left to bottom-right; (ii) strongly improved robustness to visual clutter in large scenes with many distractors; (iii) enabling the learning of multiple object models from just one training image. For more details see Rutishauser et al. (2004) and Walther et al. (2005).

## 6. Conclusion

In this paper we have introduced a model for bottom-up attention to salient proto-objects. We have given a detailed description of biologically plausible implementations of the key processing steps in networks of linear threshold units. Furthermore, we have demonstrated how this model for attending to proto-objects can be used for serializing visual processing by the biologically plausible model of object recognition by Riesenhuber and Poggio (1999b).

Attended regions may not necessarily have a one-to-one correspondence to objects. Groups of similar objects, e.g., a bowl of fruits, may be segmented as one region, as may object parts that are dissimilar from the rest of the object, e.g., a skin-colored hand appearing to terminate at a dark shirt sleeve. These regions are termed “proto-objects” because they can lead to the recognition of the actual objects in further iterative interactions between the attention and recognition systems. See the work by Rybak, Gusakova, Golovan, Podladchikova, and Shevtsova (1998), for instance, for a model that uses the vector of saccades to code for the spatial relations between object parts.

The additional computational cost for region selection is minimal because the feature and conspicuity maps have already been computed during the processing for saliency. Note that although ultimately only the winning feature map is used to segment the attended image region, the interaction of WTA and IOR operating on the saliency map provides the mechanism for sequentially attending several salient locations.

There is no guarantee that the region selection algorithm will find objects. It is purely bottom-up, stimulus driven and has no prior notion of what constitutes an object. Also note that we are not attempting an exhaustive segmentation of the image, such as done by Shi and Malik (2000) or Martin, Fowlkes, and Malik (2004). Our algorithm provides us with a first rough guess of the extent of a salient region.

Being able to attend to salient proto-objects should only be the first step, the tie-breaker in an iterative back and forth between object recognition and selective visual attention. Once a proto-object region is selected, the object recognition system will be able to form hypotheses about the identity of the attended object. This will then in turn instruct the attention system to focus on features or regions that would provide information for the verification or falsification of those hypotheses.

There is much further room for modeling the close interactions between visual attention and object recognition in cortex. In the simplest case this would mean to share resources such as orientation filtering. In our combined attention and recognition model as illustrated in Fig. 4, for instance, the output of V1-like orientation filters should be shared by the two sub-systems. But interactions could reach much further and incorporate the learning of optimal search strategies for particular objects or object categories based on particular features (Navalpakkam & Itti, 2005) or spatial priors (Torralba, 2003).

We believe that our suggested model for attention to proto-objects bridges several concepts in visual cognition such as coherence theory, object-based attention, and spreading of

attention and inhibition over object surfaces. The model is meant to provide a first step to solving the chicken-and-egg problem of directing selective attention to object regions before objects are recognized.

## Acknowledgements

Parts of the work presented in this paper originated from collaborations with Laurent Itti, Maximilian Riesenhuber, Tomaso Poggio, Ueli Rutishauser, and Pietro Perona. Funding was provided by the NSF Engineering Research Center for Neuromorphic Systems Engineering at Caltech, by the NIH, the NIMH, the Keck Foundation, a Sloan-Swartz pre-doctoral Fellowship, and a Beckman postdoctoral Fellowship to D.W.

## References

Braun, J. (1994). Visual-search among items of different salience — removal of visual-attention mimics a lesion in extrastriate area V4. *Journal of Neuroscience*, *14*(2), 554–567.

Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, *COM-31*(4), 532–540.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex*, *11*(8), 761–772.

Connor, C. E., Preddie, D. C., Gallant, J. L., & van Essen, D. C. (1997). Spatial attention effects in macaque area V4. *Journal of Neuroscience*, *17*(9), 3201–3214.

Deco, G., & Schürmann, B. (2000). A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Research*, *40*(20), 2845–2859.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501–517.

Duncan, J. (1997). Integrated mechanisms of selective attention. *Current Opinion in Biology*, *7*, 255–261.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology General*, *123*(2), 161–177.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, *40*(4), 225–240.

Feldman, J. A., & Ballard, D. H. (1983). Connectionist models and their properties. *Cognitive Science*, *6*(3), 205–254.

Gandhi, S. P., Heeger, D. J., & Boynton, G. M. (1999). Spatial attention affects brain activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the USA*, *96*(6), 3314–3319.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*(3), 121–134.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*(4), 187–202.

Hamker, F. H. (2005a). The reentry hypothesis: The putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cerebral Cortex*, *15*(4), 431–447.

Hamker, F. H. (2005b). The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Computer Vision and Image Understanding*, *100*(1–2), 64–106.

Indiveri, G. (1997). Winner-take-all networks with lateral excitation. *Analog Integrated Circuits and Signal Processing*, *13*(1/2), 185–193.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology*, *43*(3), 171–216.

Itti, L. (2000). Models of bottom-up and top-down visual attention. Ph.D. thesis. California Institute of Technology.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506.

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*(1), 161–169.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jordan, H., & Tipper, S. P. (1999). Spread of inhibition across an object's surface. *British Journal of Psychology*, *90*(4), 495–507.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman, & D. A. Davies (Eds.), *Varieties of attention* (pp. 29–61). New York: Academic Press.

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, *282*(5386), 108–111.

Koch, C., & Ullman, S. (1985). Shifts in selective visual-attention — towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Liu, S. -C. (2002). A normalizing aVLSI network with controllable winner-take-all properties. *Analog Integrated Circuits and Signal Processing*, *31*(1), 47–53.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*(5), 401–414.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, *77*(1), 24–42.

Martin, D., Fowlkes, C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(5), 530–549.

McAdams, C. J., & Maunsell, J. H. R. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*(3), 1751–1755.

McAdams, C. J., & Reid, R. C. (2005). Attention modulates the responses of simple cells in monkey primary visual cortex. *Journal of Neuroscience*, *25*(47), 11023–11033.

Mitchell, J. F., Stoner, G. R., Fallah, M., & Reynolds, J. H. (2003). Attentional selection of superimposed surfaces cannot be explained by modulation of the gain of color channels. *Vision Research*, *43*(12), 1323–1328.

Moore, C. M., Yantis, S., & Vaughan, B. (1998). Object-based visual selection: Evidence from perceptual completion. *Psychological Science*, *9*(2), 104–110.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784.

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, *14*(4), 2178–2189.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205–231.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *5*(11), 1203–1209.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*(11), 4700–4719.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Rensink, R. A. (2000a). Seeing, sensing, and scrutinizing. *Vision Research*, *40*(10–12), 1469–1487.

Rensink, R. A. (2000b). The dynamic representation of scenes. *Visual Cognition*, *7*(1/2/3), 17–42.

Rensink, R. A., Oregan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368–373.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703–714.

Reynolds, J. H., Alborzian, S., & Stoner, G. R. (2003). Exogenously cued attention triggers competitive selection of surfaces. *Vision Research*, *43*(1), 59–66.

Riesenhuber, M., & Poggio, T. (1999a). Are cortical models really bound by the "binding problem"? *Neuron*, *24*(1), 87–93. 111–125.

Riesenhuber, M., & Poggio, T. (1999b). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Riesenhuber, M., & Poggio, T. (2003). How visual cortex recognizes objects: the tale of the standard model. In L. M. Chapula, & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1640–1653). Cambridge, MA: MIT Press.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–381.

Rosen, E. (2003). Face representation in cortex: Studies using a simple and not so special model, CBCL Paper #228/AI Memo #2003-010. Technical report. Massachusetts Institute of Technology. June.

Rosenfeld, A., & Pfaltz, J. L. (1966). Sequential operations in digital picture processing. *Journal of the Association for Computing Machinery*, *13*, 471–494.

Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is attention useful for object recognition? In *International conference on computer vision and pattern recognition*: *Vol. 2* (pp. 37–44).

Rybak, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N., & Shevtsova, N. A. (1998). A model of attention-guided visual perception and recognition. *Vision Research*, *38*(15–16), 2387–2400.

Saarinen, J., & Julesz, B. (1991). The speed of attentional shifts in the visual field. *Proceedings of the National Academy of Sciences of the USA*, *88*(5), 1812–1814.

Sagi, D., & Julesz, B. (1986). Enhanced detection in the aperture of focal attention. *Nature*, *321*, 693–695.

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences*, *10*(1), 38–45.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *IEEE international conference on computer vision and pattern recognition*: *Vol. 2* (pp. 994–1000). CA: San Diego.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2006). Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (in press).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Shulman, G. L., & Wilson, J. (1987). Spatial frequency and selective attention to spatial location. *Perception*, *16*(1), 103–111.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review*, *5*(4), 644–649.

Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science*, *240*(4850), 338–340.

Sun, Y., & Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial Intelligence*, *20*(11), 77–123.

Tipper, S. P., Jordan, H., & Weaver, B. (1999). Scene-based and object-centered inhibition of return: Evidence for dual orienting mechanisms. *Perception and Psychophysics*, *61*(1), 50–60.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A*, *20*(7), 1407–1418.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, *13*(3), 423–445.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., & Nuflo, F. (1995). Modeling visual-attention via selective tuning. *Artificial Intelligence*, *78*, 507–545.

Vetter, T., & Blanz, V. (1999). A morphable model for the synthesis of 3D faces. In *SIGGRAPH* (pp. 187–194).

Walther, D. (2006). Appendix A.2. Color opponencies for bottom-up attention. In Interactions of visual attention and object recognition: Computational modeling, algorithms, and psychophysics. Ph.D. thesis (pp. 98–99). California Institute of Technology. http://resolver.caltech.edu/CaltechETD:etd-03072006-135433.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., & Koch, C. (2002). Attentional selection for object recognition — a gentle way. In *Lecture notes in computer science*: *Vol. 2525* (pp. 472–479). Berlin, Germany: Springer.

Walther, D., Rutishauser, U., Koch, C., & Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, *100*(1–2), 41–63.