Biologically Motivated Computer Vision Workshop 2002,
22.-24. November 2002, Tübingen, Germany. Springer-Verlag,
to appear. ©Springer-Verlag

Object detection in natural scenes by feedback

Fred H. Hamker and James Worcester

California Institute of Technology, Division of Biology 139-74,
Pasadena, CA 91125, USA
fred@klab.caltech.edu
http://www.klab.caltech/∼fred.html

Abstract. Current models of object recognition generally assume a bottom-up process within a hierarchy of stages. As an alternative, we present a top-down modulation of the processed stimulus information to allow a goal-directed detection of objects within natural scenes. Our procedure has its origin in current findings of research in attention which suggest that feedback enhances cells in a feature-specific manner. We show that feedback allows discrimination of a target object by allocation of attentional resources.

1 Introduction

The majority of biologically motivated object recognition models process the visual image in a feedforward manner. Specific filters are designed or learned to allow recognition of a subset of objects. In order to facilitate recognition, an attentional module was proposed to pre-select parts of the image for further analysis. This is typically done by applying a spotlight or window of attention that suppresses input from outside the window. Such an approach results in two major disadvantages: i) A spotlight selects a region but not object features. Even when the whole image is reduced to a region of interest, object recognition algorithms still have to cope with clutter, different backgrounds and with overlapping from other objects, which modify the filter responses. ii) Object recognition follows attentional selection. If a task requires the detection of a specific item such an approach calls for serially scanning the scene and sending the content of each selected location to a recognition module until the target is found. The use of simple target cues, like color, can reduce the search space, but the serial scan is unavoidable.

We suggest a top-down approach for a goal-directed search. Instead of specialized learned or designed features, we use a general set of features that filter the image and construct a population of active cells for each scene. The information about a target is sent top-down and guides the bottom-up processing in a parallel fashion. This top-down modulation is implemented such that the features of the object of interest are emphasized through a dynamic competetive/cooperative process. Related ideas have been suggested in the past [1] [2] [3] [4] but not further implemented for a model of vision in natural scenes.

We have been working out this concept with a computational neuroscience approach. The starting point was to understand the role of goal-directed visual attention [5] [6] [7]. Experimental findings support the concept of prioritized processing by a biased competition [8]. For example, an elevated baseline activity was observed in IT cells after a cue was presented [9]. This effect could be a priming in order to prepare the visual system for detecting the target in a scene. Further evidence for a feature-selective feedback signal is found in V4 [10] and in the motion system [11].
Although some scenes allow the detection of categories during very brief presentations even in the near absence of spatial attention [12], ambiguities in IT cell populations encoding features within the same receptive field limits recognition in natural images [13]. We use feedback to clean up the population activity in higher stages from all unimportant stimuli so that a full recognition can take place. In the following we describe how feedback modulates the feedforward process, which allows for a goal-directed detection of an object in a natural scene.

2 Model

We combine stimulus-driven saliency, which is primarily a bottom-up process, with goal-directed attention, which is under top-down control (Fig. 1). The fact that features that are unique in their environment 'pop-out' is to a first degree achieved by computing center-surround differences. In this regard, our saliency module mostly follows the approach of Itti, Koch and Niebur [14]. However, their purely salience-driven approach continues in combining the center-surround maps into conspicuity maps and then into a final saliency map. We suggest combining the feature value with its corresponding saliency into a population code which feeds V4 cells (Fig. 2). This approach allows us to create a parallel encoding of different variables and achieve the dynamic enhancement of relevant variables by feedback connections. The hierarchy of the model is motivated through a computational neuroscience study of attention [7]. Features of the target template are sent downwards in parallel and enhance features in the scene that match the template. Feedback from the premotor map enhances all features at a specific location. Such an approach unifies recognition and attention as interdependent aspects of one network.

2.1 Low level stimulus-driven salience

We largely follow the implementation of Itti et al. [14] to obtain feature and contrast maps from a color image (Fig. 2). We currently use color, intensity and orientation as basic features. Our approach differs from Itti et al. [14] in how saliency influences processing. Itti et al. suggest to compute saliency in the 'where' system, select the most salient part of the image and then preferably process this part in the 'what' pathway. We compute feature conspicuity maps within the 'what' pathway that directly modulate the features according to their saliency in parallel without any spatial focus. Thus, salient features do

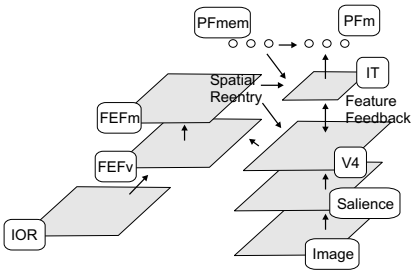

Fig. 1. Model for top-down guided detection of objects. First, information about the content and its low level stimulus-driven salience is extracted. This information is sent further upwards to V4 and to IT cells which are broadly tuned to location. The target template is encoded in PFmem. PFm cells indicate by comparison of PFmem with IT whether the target is actively encoded in IT. Feedback from PFmem to IT increases the strength of all features in IT matching the template. Feedback from IT to V4 sends the information about the target downwards to cells with a higher spatial tuning. FEFv combines the feature information across all dimensions and indicates salient or relevant locations in the scene. A winner-take-all process in FEFm (premotor) cells selects the strongest location. Even during this competition a reentry signal from this map to V4 and IT enhances all features at locations of activity in FEFm. The IOR map memorizes recently visited locations and inhibits the FEFv cells.

not have to be routed to higher areas by spatial attention. However, after 100ms spatial attention starts to implement a gain enhancement in order to prioritize processing at a certain location.

Feature maps: Starting from the color image, we extract orientation $O(\sigma, \theta)$ with varying resolution σ and orientation θ, intensity I, red-green $RG = R - G$ and blue-yellow $BY = B - Y$ information [14].

Contrast maps: Contrast maps determine the conspicuity of each feature and implement the known influence of lateral excitation and surround inhibition by center-surround operations '$\ominus$'. We construct orientation contrast $\mathcal{O}(c, s, \theta)$, intensity contrast $\mathcal{I}(c)$ as well as red-green $\mathcal{RG}(c)$ and blue-yellow $\mathcal{BY}(c)$ double opponency [14].

Feature conspicuity maps: For each variable or feature, we combine the feature information into an attribute $\mathbf{V}$ and its corresponding contrast value into a gain factor P of a population code. This dual coding principle is a very important characteristic. A feature is represented by the location of cell activity, and the conspicuity of this feature is represented by the strength of activity. At each location x_1, x_2 we construct a space, whose axes are defined by the represented features and by one additional conspicuity axis (Fig. 2). The population is then defined by a set of neurons $i \in N$ sampling the feature space, with each neuron tuned around its preferred value $\mathbf{u}_i$. For each neuron y_i we obtain an activity value:

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

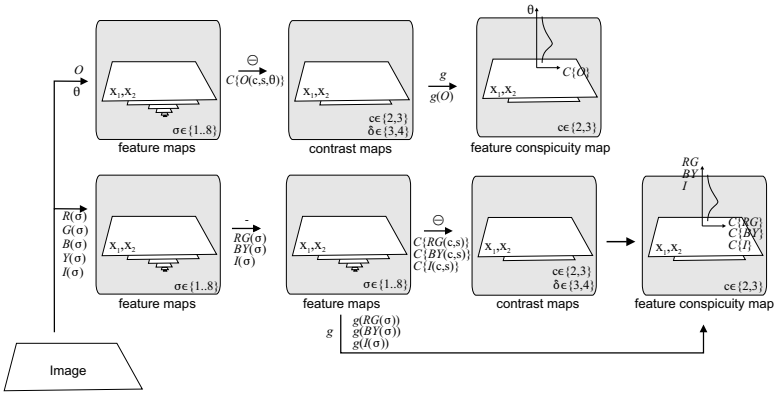

Fig. 2. Construction of a population indicating a feature and its stimulus-driven salience at each location in the image. Starting from a color image we construct broadly tuned color channels *R*ed, *G*reen, *B*lue and *Y*ellow, as well as an *I*ntensity channel. Each of these is represented by a Gaussian pyramid with the scale σ. The color channels are transferred into on opponency system RG and BY. By applying Gabor wavelets on the intensity image I with the scale σ and orientation θ we achieve for each orientation a pyramid that represents the match of the image with the filter. We use center-surround or contrast operations $\ominus$ for each of those feature maps to determine the location of conspicuous features. Both the feature maps and the contrast maps are then combined into feature conspicuty maps, which indicate the feature and its corresponding conspicuity value at each location x_1, x_2.

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

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

To apply the same range of selectivity parameters $\sigma_g^2 \in \{0.05 \dots 0.2\}$ for all channels we normalize the feature values $\mathbf{V} \in \{I, RG, BY, \theta, \sigma\}$ of each channel between zero and one. The cell activity within the population should typically lie within the range of zero and one. Thus, we also normalize the contrast values to $\widetilde{\mathcal{I}}$, $\widetilde{\mathcal{RG}}$, $\widetilde{\mathcal{BY}}$, $\widetilde{\mathcal{O}}$. We finally receive the populations for each channel with scale c at each location $\mathbf{x}$:

$$\begin{aligned}
y_i^I(c,\mathbf{x}) &= \widetilde{\mathcal{I}}(c,\mathbf{x}) \cdot g(u_i - I(c,\mathbf{x})) \\
y_i^{RG}(c,\mathbf{x}) &= \widetilde{\mathcal{RG}}(c,\mathbf{x}) \cdot g(u_i - RG(c,\mathbf{x})) \\
y_i^{BY}(c,\mathbf{x}) &= \widetilde{\mathcal{BY}}(c,\mathbf{x}) \cdot g(u_i - BY(c,\mathbf{x})) \\
y_i^{\theta}(c,\mathbf{x}) &= \max_{\theta}\left(\widetilde{\mathcal{O}}(c,\theta,\mathbf{x}) \cdot g(u_i - \theta)\right) \\
y_i^{\sigma}(c,\mathbf{x}) &= \max_{\theta}\left(\widetilde{\mathcal{O}}(c,\theta,\mathbf{x}) \cdot g(u_i - \sigma)\right)
\end{aligned} \tag{3}$$

We now have $\#c$ maps, where $\#c$ is the number of center scales, with a population at each point $\mathbf{x}$ for a different center scale c. To combine these maps across space into one map with the lowest resolution (highest c) we use a maximum operation ($\max_{c,\mathbf{x}' \in RF(\mathbf{x})}$).

2.2 Goal-directed control

In order to compute the interdependence of object recognition and attention we need a continuous dynamic approach. Specifically, we use a population code simulated by differential equations. Each map in the model represents a functional area of the brain [7]. It contains at each location $\mathbf{x}$ a population of i cells encoding feature values (eq. 4), with the exception of the maps in the frontal eye field and IOR which only encode space ($i = 1$). In addition V4 and IT have separate maps for different dimensions d (RG, BY, etc.). The population of cells is driven by its input $y^{\uparrow}_{d,i,\mathbf{x}}$. Feedback implements an input gain control to enhance the representation of certain features and biases the competition [8] among active populations. Feature specific feedback (I^L) operates within the ventral pathway and enhances cell populations whose input matches the feedback signal. Spatial reentry (I^G) arrives from the frontal eye field and boosts features at a certain location, generally the target of the next saccade. $I^{f\,inh}$ induces competition among cells and I^{inh} causes a normalization and saturation. Both terms have a strong short range and weak long range inhibitory effect.

$$\tau \frac{d}{dt} y_{d,i,\mathbf{x}} = y^{\uparrow}_{d,i,\mathbf{x}} + I^L + I^G - y_{d,i,\mathbf{x}} \cdot I^{inh}_{d,\mathbf{x}} - I^{f\,inh}_{d,\mathbf{x}} \tag{4}$$

The following maps use implementations of the general equation quoted above (eq. 4).

V4: Each V4 layer receives input from a different dimension (d) in the feature conspicuity maps: $y^{\theta}_{i,\mathbf{x}}$ for orientation, $y^{I}_{i,\mathbf{x}}$ for intensity, $y^{RG}_{i,\mathbf{x}}$ for red-green opponency, $y^{BY}_{i,\mathbf{x}}$ for blue-yellow opponency and $y^{\sigma}_{i,\mathbf{x}}$ for spatial frequency. V4 cells receive feature specific feedback from IT cells ($I^L = I^L(y^{IT})$) and spatial reentry from the frontal eye field ($I^G = I^G(y^{FEFm})$).

IT: The populations from different locations in V4 project to IT, but only within the same dimension. We simulate a map containing 9 populations with overlapping receptive fields. We do not increase the complexity of features from V4 to IT. Thus, our model IT populations represent the same feature space as our model V4 populations. The receptive field size, however, increases in our model, so that several populations in V4 converge onto one population in IT: $y^{\uparrow}_{i,d,\mathbf{x}} = w^{\uparrow} \max_{\mathbf{x}' \in RF(\mathbf{x})} y^{\mathrm{V4}}_{i,d,\mathbf{x}'}$. IT receives feature specific feedback from the prefrontal memory ($I^L = I^L(y^{PFmem})$) and location specific feedback from the frontal eye field ($I^G = I^G(y^{FFEm})$).

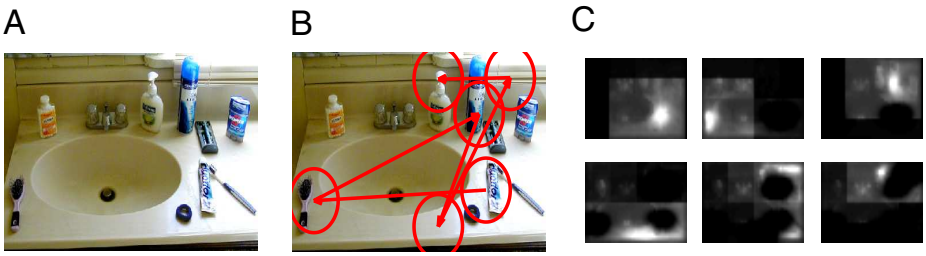

Fig. 3. Results of a free viewing task. (A) Natural scene. (B) Scanpath. It starts on the toothpaste, visits the hairbrush, the shaving cream, two salient edges and then the soap. (C) Activity of FEFv cells prior to the next scan. By definition they represent locations which are actively processed in the V4 and IT map and thus represent possible target locations. An IOR map inhibits FEFv cells at locations that were recently visited (causing the black circles).

FEFv: The perceptual map (FEFv) neurons receive convergent afferents from V4 and IT $y^{\uparrow a}_{\mathbf{x}} = w^{\mathrm{V4}} \sum_d \max_i y^{\mathrm{V4}}_{d,i,\mathbf{x}} + w^{\mathrm{IT}} \sum_d \max_{i,\mathbf{x}' \in RF(x)} y^{\mathrm{IT}}_{d,i,\mathbf{x}'}$. The information from the target template additionally enhances the locations that result in a match between target and encoded feature $y^{\uparrow b}_{\mathbf{x}} = w^{\mathrm{PFmem}} \prod_d \max_i y^{\mathrm{PFmem}}_{d,i} \cdot y^{\mathrm{V4}}_{d,i,\mathbf{x}}$ at all locations simultaneously. This allows the biasing of specific locations by the joint probability that the searched features are encoded at a certain location. The firing rate of FEFv cells represent the saliency of locations, whereas the saliency of each feature is encoded in the ventral pathway.

FEFm: The effect of the perceptual map on the premotor cells (FEFm) is a slight surround inhibition: $y^{\uparrow}_{\mathbf{x}} = w^{\mathrm{FEFv}} y^{\mathrm{FEFv}}_{\mathbf{x}} - w^{\mathrm{FEFv}}_{inh} \sum_{\mathbf{x}} y^{\mathrm{FEFv}}_{\mathbf{x}}$. Thus, by increasing their activity slowly over time premotor cells compete for the selection of the strongest location.

IOR: There is currently no clear indication where cells that ensure an inhibition of return are located. We regard each location $\mathbf{x}$ as inspected, dependent on the selection of an eye movement at $y^{\mathrm{FEFm}}_{\mathbf{x}}(t_e) > I^{\mathrm{FEF}}_o$ or when a match in the PFm cells is lost. In this case the IOR cells are charged at the location of the strongest FEFm cell for a period of time T^{IOR}. This causes a suppression of the recently attended location in the FEFv map. IOR cells get slowly discharged by decay with a low weight w_{inh}.

$$\tau \frac{d}{dt} y^{\mathrm{IOR}}_{\mathbf{x}} = (1 - y^{\mathrm{IOR}}_{\mathbf{x}})(w^{\mathrm{FEFm}} I^{\mathrm{FEFm}}_{\mathbf{x}} - w_{inh} y^{\mathrm{IOR}}_{\mathbf{x}})$$

$$I^{\mathrm{FEFm}}_{\mathbf{x}} = \begin{cases} \exp(-\frac{(\mathbf{x}-\mathbf{x_m})^2}{0.01}) & if \quad t < t_e + T^{\mathrm{IOR}} \\ 0 & else \end{cases} \; ; \; y^{\mathrm{FEFm}}_{\mathbf{x_m}} = \max_{\mathbf{x}}(y^{\mathrm{FEFm}}_{\mathbf{x}}) \quad (5)$$

3 Results

We first show how the model operates in a free viewing task, which is only driven by the stimulus saliency (Fig. 3). The overall scanning behaviour is similar to feedforward approaches (e.g. [14]). The major difference it that the saliency is actively constructed within the network as compared to a static saliency map (Fig. 3C). We could now generate prototypes of various objects and place them into the space spanned by the IT cells. By comparing the prototypes with IT activity during the scans we could then determine the selected object. However, this is not a very interesting strategy. Recognition fully relies on the stimulus-driven selection. According to our interpretation of findings in brain research, primates are able to perform a goal-directed search. The idea is that the brain might acquire knowledge about objects by learning templates. To mimic this strategy we present the model objects from which it generates very simple templates (Fig. 4).

If such an object is relevant for a certain task, the templates are loaded into the PFmem cells and IT cells get modulated by feature-specific feedback. When presenting the search scene, initially IT cells reflect salient features, but over time those features that match the target template get further enhanced (Fig. 5). Thus, the features of the object of interest are enhanced prior to any spatial focus of attention. The frontal eye field visual cells encode salient locations. Around 85-90ms all areas that contain objects are processed in parallel. Spatial attention then enhances all features at the selected location, in searching for the asprin bottle at around 110ms and for the hairbrush 130ms after scene onset. As a result the initial top-down guided information is extended towards all the features of the target object. For example, the very red color of the asprin bottle or the dark areas of the hairbrush are detected by spatial attention because those features were not part of the target template. This aspect is known as prioritized processing. In the beginning only the most salient and relevant features receive a high processing whereas later all features of a certain object are processed.

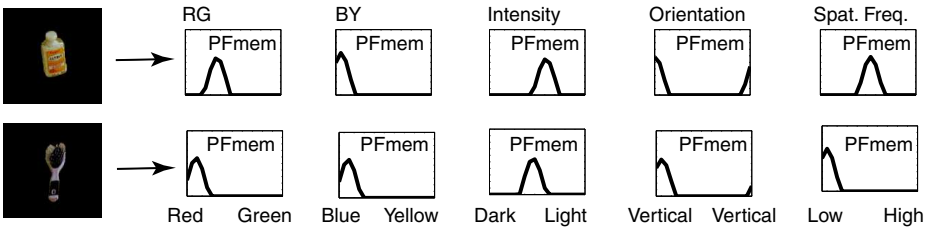

Fig. 4. We presented the asprin bottle and the hairbrush to the model and in each dimension the most salient feature was memorized in order to generate a target template. For the asprin bottle the stopper is most salient so that in the *RG*-dimension the memorized color is only slightly shifted to red. Altogether, we only use crude information about the object and do not generate a copy. Note, that the objects are placed on a black background whereas they appear in the image on a white background.

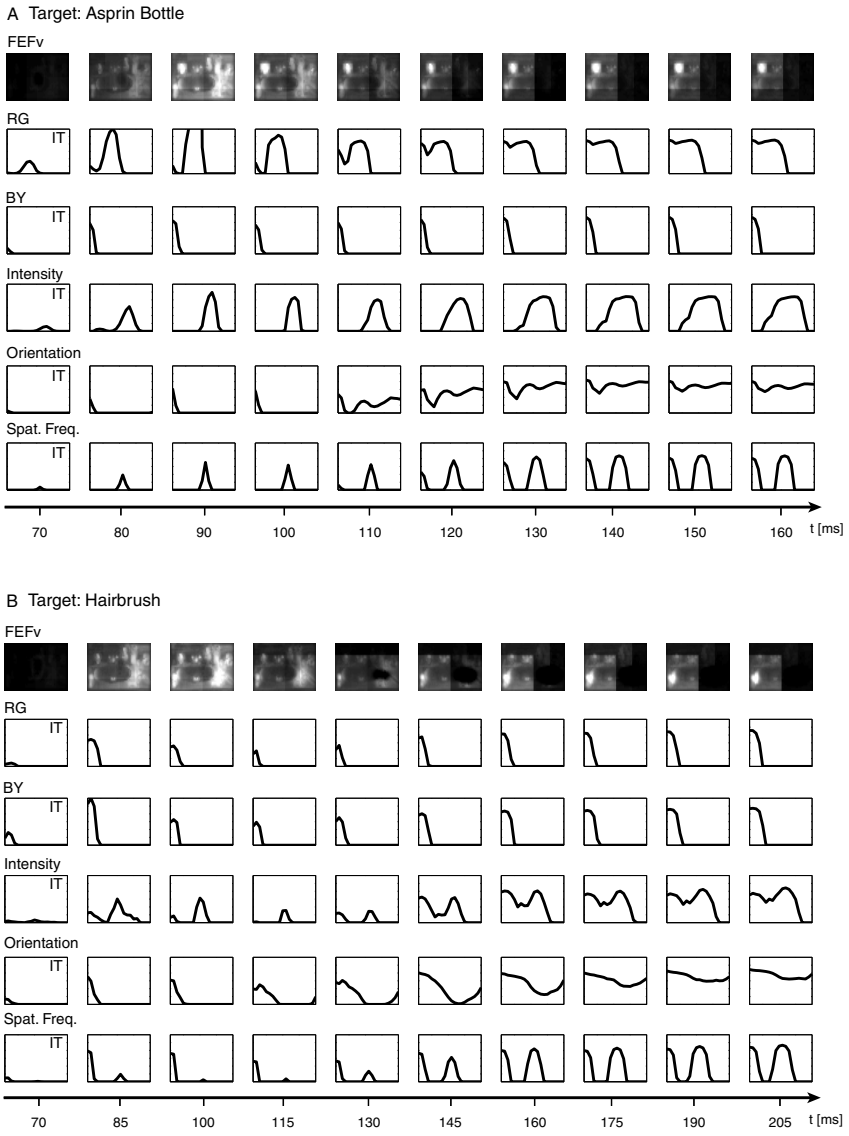

Fig. 5. The temporal process of a goal-directed object detection task in a natural scene. (A) Asprin bottle as target. (B) Hairbrush as target. The frontal eye field visual cells indicate preferred processing, which is not identical with a spatial focus of attention. At first they reflect salient locations whereas later they discriminate target from distractor locations. The activity of IT cell populations with a receptive field covering the target initially show activity that is inferred by the search template. Later activity also reflects other features of the object that were not searched for.

4 Discussion

We have presented a new approach to goal-directed vision based on feedback within the 'what'-pathway and spatial reentry from the 'where'-pathway. The complex problem of scene understanding is here transformed into the generation of an appropriate target template. Once a template is generated, we show that a system can detect an object by an efficient parallel search as compared to pure saliency-driven models which rely on a sequential search strategy by rapidly selecting parts of the scene and analyzing these conspicuous locations in detail. Our model only uses a sequential strategy if the parallel is not efficient to guide the frontal eye field cells toward the correct location. Stimulus-driven saliency is suggested to prioritize the processing in a parallel fashion as compared to an early selection. Regarding the finding that categories can be detected even in the near absence of spatial attention [12], it is important to notice that in our model spatial attention is not a prerequisite of object detection. If the target sufficiently discriminates from the background, the match with the template in PFm can be used for report before any spatial reentry occurs. The simulation results also provide an explanation for Sheinbergs and Logothetis' [13] finding of early activation of IT cells if the target is foveated by the next fixation. Classical models of scene analysis would predict that the process of identifying objects begins after each fixation. In our model the match with the target template increases the firing rate of cells in the 'what'-pathway indicating the detection of the object. Such enhanced activity is picked-up by maps in the 'where'-pathway which locate the object for action preparation. Reentrant activity then enhances all features of the object in order to allow a more detailed analysis. Thus, object identification begins before the eyes actually fixate on the object.
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. Future work should of course extend the model with more shape selective filters to perform object recognition tasks. We think that such an approach provides a serious alternative to present feedforward models of object recognition.

Acknowledgements: We thank Laurent Itti for sharing his source code, Rufin VanRullen for providing the test scene and Christof Koch for his support. This research was supported by DFG HA2630/2-1 and by the NSF (ERC-9402726).

References

1. Mumford D.: On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern. **66** (1992) 241–251.
2. Tononi, G., Sporns, O., Edelman, G.: Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. Cereb. Cortex **2** (1992) 310–335.
3. Grossberg S. How does a brain build a cognitive code? Psychol Rev. **87** (1980) 1–51.

4. Ullman, S.: Sequence seeking and counter streams: A computational model for bidirectional flow in the visual cortex. Cerebral Cortex **5** (1995) 1–11.
5. Hamker, F.H.: The role of feedback connections in task-driven visual search. In: Connectionist Models in Cognitive Neuroscience. Springer Verlag, London (1999), 252–261.
6. Corchs, S., Deco, G.: Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data. Cereb. Cortex **12** (2002) 339–348.
7. Hamker, F.H.: How does the ventral pathway contribute to spatial attention and the planning of eye movements? Proceedings of the 4th Workshop Dynamic Perception, 14-15 November 2002, Bochum, Germany, to appear.
8. Desimone, R., Duncan, J., Neural mechanisms of selective attention. Annu Rev Neurosci **18** (1995) 193–222.
9. Chelazzi, L., Duncan, J., Miller, E.K., Desimone, R.: Responses of neurons in inferior temporal cortex during memory-guided visual search. J. Neurophysiol. **80** (1998) 2918–2940.
10. Motter, B.C.: Neural correlates of attentive selection for color or luminance in extrastriate area V4. J. Neurosci. **14** (1994) 2178–2189.
11. Treue, S., Martínez Trujillo, J.C.: Feature-based attention influences motion processing gain in macaque visual cortex. Nature **399** (1999) 575–579.
12. Li, F.-F., VanRullen, R., Koch, C., Perona, P.: Rapid natural scene categorization in the near absence of attention. Proc Natl Acad Sci USA **99** (2002) 9596-9601.
13. Sheinberg, D.L., Logothetis, N.K.: Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J Neurosci. **21** (2001) 1340–1350.
14. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **20** (1998), 1254–1259.

Appendix: Model equations

Feature specific topological feedback from the origin ν:

$$I^L_{d,i,\mathbf{x}}(y^\nu) = \max_{\mathbf{x}' \in RF(\mathbf{x})} w^{\downarrow L} y_{i,d,\mathbf{x}'} \cdot y^\nu_{i,d} \tag{6}$$

Location specific topographic feedback from the origin ν:

$$I^G_{d,i,\mathbf{x}}(y^\nu) = y^{\uparrow}_{d,i,\mathbf{x}} \cdot \max_{\mathbf{x}' \in RF(\mathbf{x})} w^{\downarrow G} y_{i,d,\mathbf{x}'} \cdot y^\nu_{\mathbf{x}'} \tag{7}$$

Inhibition for normalization and saturation:

$$I^{inh}_{d,\mathbf{x}} = w_{inh} \sum_j y_{d,j,\mathbf{x}}(t) + w^{map}_{inh} z^{map}_d \tag{8}$$

Inhibition for competition among cells:

$$I^{f\,inh}_{d,\mathbf{x}} = w^{map}_{f\,inh} z^{map}_d(t) \tag{9}$$

using

$$\tau^{map}_{inh} \frac{d}{dt} z^{map}_d = \sum_{\mathbf{x}} \max_j \left(y_{j,d,\mathbf{x}}\right) - z^{map} \tag{10}$$