A Population-Based Inference Framework for Feature-Based Attention in Natural Scenes

Fred H. Hamker

Allgemeine Psychologie, Psychologisches Institut II,
Westf. Wilhelms-Universität, 48149 Münster, Germany
fhamker@uni-muenster.de
http://wwwpsy.uni-muenster.de/inst2/lappe/Fred/FredHamker.html

Abstract. Vision is a crucial sensor. It provides a very rich collection of information about our environment. However, not everything in a visual scene is relevant for the task at hand. Feature-based attention has been suggested for guiding vision towards the objects of interest in a visual search situation. Computational models of visual attention have implemented different concepts of feature-based attention. We will discuss these approaches and present a solution which is based on population-based inference. We illustrate the proposed mechanism with simulations using real world-scenes.

1 Introduction

Visual Search and other experimental approaches have demonstrated that attention plays a crucial role in human perception. Understanding attention and human vision in general could be beneficial to computer vision, especially in vision tasks that are not limited to specific and constrained environments. Previous models of attention have suggested different underlying computational mechanisms of how feature cues (e.g., color) affect visual processing. In most models attention is solely defined by determining the locus of a unique spatial focus [24,13,28,1,19,10]. Feature-based attention is left to only guide the selection process by weighting the input into the saliency map [16,18]. For example, the search for the blue lighter is typically implemented by enhancing the input into the saliency map for cells encoding the target color (Fig. 1A). The selective tuning model implements feature-based attention by enhancing the value of the interpretive nodes which in turn biases the winner-take-all (WTA) competition for projection into the next layer [26]. A cascade of top-down directed WTA processes prune away all irrelevant connections within successively smaller receptive fields. As a result, features such as the color blue allow to segment a target object in the scene (Fig. 1B). Technically the top-down biasing nodes form an independent top-down path, but present implementations of the selective tuning model do not distinguish between feature and spatial attention in the sense that feature-based attention induces competition only through the spatially selective WTA.

Treue and Martínez Trujillo [25] have proposed a Feature-Similarity Theory of attention. Their single cell recordings in area MT revealed that directing attention to one stimulus enhances the response of a second stimulus presented elsewhere in the visual field, but only if the features of both stimuli match (e.g. upward motion). They proposed

M. De Gregorio et al. (Eds.): BVAI 2005, LNCS 3704, pp. 147–156, 2005.

that attending towards a feature could provide a global, spatially non-selective feedback signal. The same effect has been found in a similar experiment using fMRI [22]. In an earlier experiment that presumably revealed feature-based attention as well, the knowledge of a target feature increased the activity of V4 cells [17].

Inspired by these findings, computational approaches have been used to investigate the mechanisms of feature-based attention [14,12,27,4,21,3]. We have developed a model to investigate the putative feedforward and feedback interactions between area V4, TE and the frontal eye field [6,8]. In this model attention emerges by interactions in the vision process. To find an object in a crowded scene our model predicts a feature-specific component that highlights all cells encoding target features in parallel and a spatially directed, serial component that is linked to the planning of an eye movement. This prediction of our model has been recently confirmed in neural cell recordings [2]. However, only little has been done to demonstrate that the proposed mechanisms even hold for large networks, e.g. for natural scene processing.

Thus, we have further developed our aprochach and extended it to a large scale network for natural scene processing [7,9] (Fig. 1C). We now explain the population-based inference framework and its relation to feature-based attention. Then, the model is introduced and specifically its feature-based attention effects are illustrated.

2 Population-Based Inference

Population coding has been frequently used as a theoretical basis for describing computation in the brain. Much emphasis has been given to investigate how a population encodes a stimulus. Our population-based inference approach provides a framework to continuously update the conspicuity of an internal variable using prior knowledge in form of generated expectations. The population is represented by a set of cells. The selectivity of each cell is defined by its location $i \in \{1..20\}$ in the population and its activity r_i reflects the conspicuity of its preferred stimulus. Each cell is simulated by an ordinary differential equation, that governs its average firing rate over time. Thus, the model allows to describe the temporal change of activity induced by top-down inference. In abstract terms, the top-down signal represents the expectation $\hat{r}$ to which the input (observation) $r^{\uparrow}$ is compared. If the observation is similar to the expectation the conspicuity is increased. This increase is implemented as a gain control mechanism on the feedforward signal. The population-based inference approach has been proven to be a suitable computational framework for simulating spatial [5] and feature-based attention effects [6]. As far as feature-based attention is concerned a cell's response over time $r_{d,i,\mathbf{x}}(t)$ at location $\mathbf{x}$, selective dimension d and preferred feature i can be computed by a differential equation (with a time constant τ):

$$\tau \frac{d}{dt} r^{\mathrm{V4}}_{d,i,\mathbf{x}} = I^{\uparrow}_{d,i,\mathbf{x}} + I^{N}_{d,i,\mathbf{x}} + I^{A}_{d,i,\mathbf{x}} - I^{\mathrm{inh}}_{d,\mathbf{x}} \qquad (1)$$

The activity of a V4 cell is primarily driven by its bottom-up input $I^{\uparrow}$. Inhibition $I^{\mathrm{inh}}_{d,\mathbf{x}}$ introduces competition among cells and normalizes the cell's response by a shunting term. $I^{N}_{d,i,\mathbf{x}}$ describes the lateral influence of other cells in the population. Feature-based attention is a result of the bottom-up signal $I^{\uparrow}_{d,i,\mathbf{x}}$ modulated by the feedback

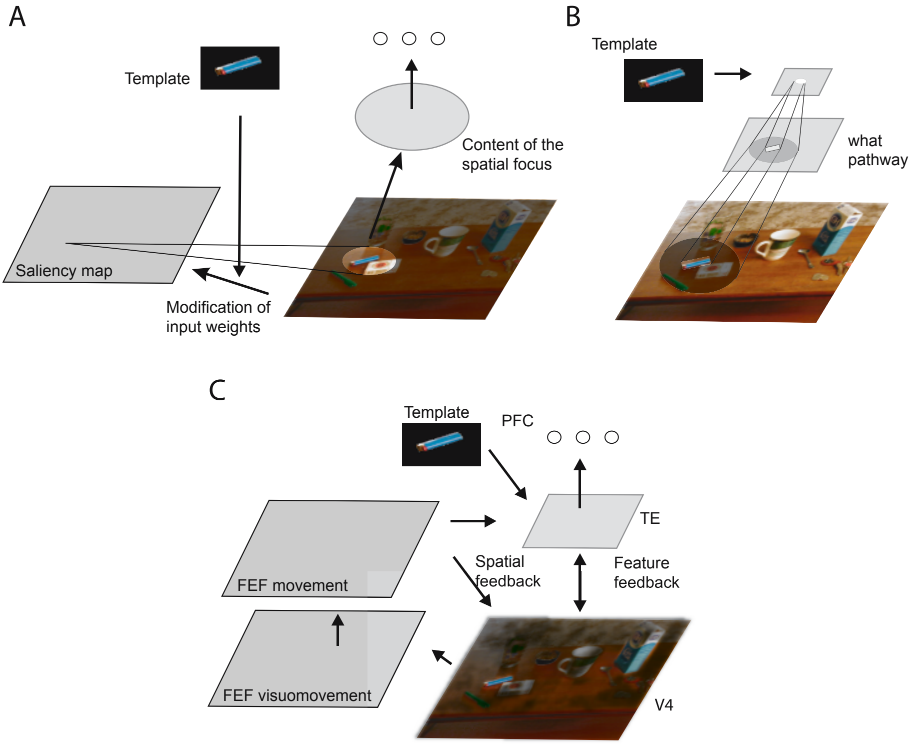

Fig. 1. Three models of attention for real world scenes and their implementation of feature-based attention. The goal directed search for the blue lighter requires some knowledge of the target object, called a template, to be represented. Most models assume that just simple, ”preattentive” features (e.g. color, orientation) are part of such a template. A) In the classical approach of visual attention, feature-based attention only modifies the input of the saliency map. For example, all weights into the saliency map of cells encoding blue are globally increased, such that the lighter has a higher chance being selected. A neural correlate of feature-based attention would therefore only be visible in a pronounced activation in the saliency map. A winner-takes-all process then determines the location of the highest activity, which in turn can be used to compute a focus of attention such that the area around the blue lighter is processed preferably . B) The selective tuning model uses top-down directed feature cues to guide competition in the what pathway. Present implementations of this model, however, do not distinguish feature-based and spatial attention, since a cascade of winner-take-all processes immediately generates an attentional beam that segments the lighter from its background and generates an inhibitory surround. C) A model of distributed processing with spatial and feature feedback. Here, attention emerges by the interactions in the network. A template, which can contain any object information, is send downwards, enhances the sensitivity of specific populations encoding the features of interest and lateral interactions normalize the activity. As a result, the model shows feature-based attention. For example, the search template of the lighter selectively enhances cells encoding blue in parallel prior to any spatial selection, as indicated by the brighter parts of the image. Other parts are relatively suppressed as illustrated by the darkened areas in the scene. This modulated activity in V4 guides areas responsible for eye movements, which in turn send a spatially selective signal back to enhance populations encoding stimuli at a specific location - spatial attention emerges.

signal from TE $r^{\mathrm{TE}}_{d,j,\mathbf{x}'}$ with $w^{\mathrm{IT,V4}}_{i,j,\mathbf{x},\mathbf{x}'}$ as the strength of the feedback connection:

$$I^{A}_{d,i,\mathbf{x}} = I^{\uparrow}_{d,i,\mathbf{x}}\sigma(\alpha - r^{\mathrm{V4}}_{d,i,\mathbf{x}}) \cdot \max_{j,\mathbf{x}'}(w^{\mathrm{TE,V4}}_{i,j} \cdot r^{\mathrm{TE}}_{d,j,\mathbf{x}'}) \tag{2}$$

$\sigma(\alpha - y^{\mathrm{V4}}_{d,k,\mathbf{x}})$ implements a saturation of the gain for salient stimuli [7]. Consistent with the Feature-Similarity Theory, the enhancement of the gain depends on the similarity between the input and the feedback signal.

3 Large Scale Approach for Modeling Attention

In this model, neural populations are defined in a space spanned by the feature selectivity i and spatial selectivity $\mathbf{x}$ of the cells. The variable d refers to different channels computed from the image such as orientation (O), intensity (I) or red-green (RG), blue-yellow (BY), or spatial resolution (σ). The conspicuity of each encoded feature is altered by the target template. A target encoded in prefrontal cortex defines the expected features $\hat{r}^{\mathrm{PFC}}_{d,i}$ (Fig. 2). We infer the conspicuity of each feature in TE denoted as $r^{\mathrm{TE}}_{d,i,\mathbf{x}}$ by comparing the expected features $\hat{r}^{\mathrm{PFC}}_{d,i}$ with the observation, i.e. the bottom-up input $r^{\mathrm{TE}\uparrow}_{d,i,\mathbf{x}}$. If the observation is similar to the expectation we increase the conspicuity. Such a mechanism enhances in parallel the conspicuity of all features in TE which are similar to the target template. The same procedure is performed in V4 to compute the conspicuity $r^{\mathrm{V4}}_{d,i,\mathbf{x}}$ where the expected features are the ones encoded in TE.

In order to detect an object in space the conspicuities $r^{\mathrm{V4}}_{d,i,\mathbf{x}}$ and $r^{\mathrm{TE}}_{d,i,\mathbf{x}}$ are combined across all channels d and encoded in the frontal eye field visuomovement cells. The projection from the visuomovement cells to the movement cells generates an expectation in space $\hat{r}^{\mathrm{FEFm}}_{\mathbf{x}}$. Thus, a location with high conspicuity in different channels d tends to have a high expectation in space $\hat{r}^{\mathrm{FEFm}}_{\mathbf{x}}$. Analogous to the inference in feature space the expected location $\hat{r}^{\mathrm{FEFm}}_{\mathbf{x}}$ is iteratively compared with the observation $r^{\uparrow}_{d,i,\mathbf{x}}$ in $\mathbf{x}$ and the conspicuity of a feature with a similarity between expectation and observation is enhanced. The conspicuity is normalized across each map by competitive interactions. Such interative mechanisms finally lead to a preferred encoding of the features and space of interest.

We now briefly explain the simulated areas in the model. A detailed description can be found in [9].

Early visual processing: Feature maps for Red-Green opponency (RG), Blue-Yellow opponency (BY), Intensity (I), Orientation (O), and Spatial Resolution (σ) are computed. The initital conspicuity is determined by center-surround operations [10]. Center-surround operations calculate the difference of feature values in maps with a fine scale and a coarse scale and thus, the obtained conspicuity value is a measure of stimulus-driven saliency. The feature information and the conspicuity are used to determine a population code, so that at each location the features and their related conspicuities are encoded.

V4: V4 has d channels which receive input from the feature conspicuity maps: $r_{\theta,i,\mathbf{x}}$ for orientation, $r_{I,i,\mathbf{x}}$ for intensity, $r_{RG,i,\mathbf{x}}$ for red-green opponency, $r_{BY,i,\mathbf{x}}$ for blue-yellow opponency and $r_{\sigma,i,\mathbf{x}}$ for spatial frequency (Fig. 2). The expectation of features

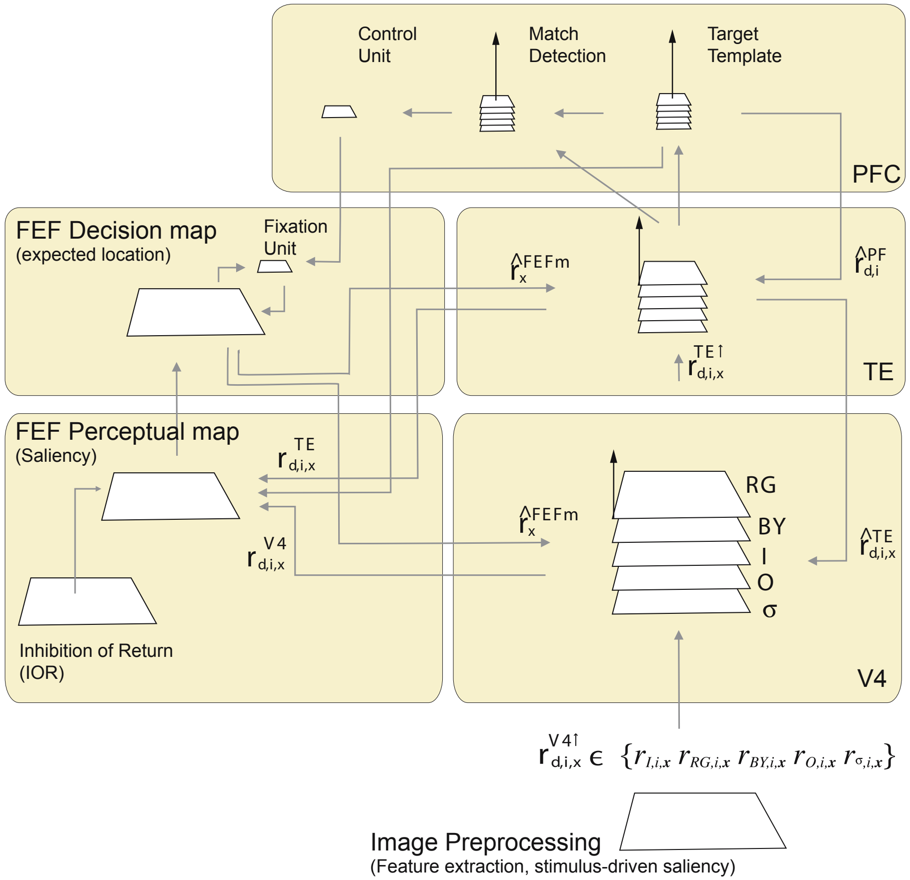

Fig. 2. Model for object detection in natural scenes. From the image, the features of 5 channels (RG, BY, I, O, σ) are obtained. For each feature we also compute its conspicuity as determined by the spatial arrangement of the stimuli in the scene and represent both aspects within a population code, so that at each location a feature and its related conspicuity is encoded. This initial, stimulus-driven conspicuity is now dynamically updated within a hierarchy of levels. From V4 to TE a pooling across space is performed to obtain a representation of features with a coarse coding of location. The target template encodes features of the target object by a population of sustained activated cells. It represents the expected features $\hat{r}_{d,i}^{\text{PFC}}$ which are used to compute the (posterior) conspicuity in TE. Similarly, TE represents the expectation for V4. As a result, the conspicuity of all features of interest is enhanced regardless of their location in the scene. In order to identify candidate objects by their saliency the activity across all 5 channels is integrated in the FEF perceptual map. The saliency is then used to compute the target location of an eye movement in the FEF decision map. The activity in this map $\hat{r}_{\mathbf{x}}^{\text{FEFm}}$ is fed back, which in turn enhances the conspicuity of all features in V4 and TE at the activated areas in the FEF decision map. Thus, objects at expected locations are preferably represented. By comparing the conspicious features in TE with the target template in the match detection units it is possible to continuously track if the object of interest is encoded in TE. Visited locations are being tagged by an inhibition of return. This allows the model to make repeated fixations while searching for an object.

in V4 originates in TE $\hat{r}^{\mathrm{V4}_F}_{d,i,\mathbf{x}'} = r^{\mathrm{TE}}_{d,i,\mathbf{x}}$ and the expected location in the FEF decision map $\hat{r}^{V4_L}_{\mathbf{x}'} = r^{\mathrm{FEFm}}_{\mathbf{x}'}$. Please note that even TE has a coarse dependency on location.

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

FEF perceptual map: The FEF perceptual map indicates salient locations by integrating the conspicuity of V4 and TE across all channels. Its cells show a response which fits into the category of FEF visuomovement cells (FEFv). In addition to the conspicuity in V4 and TE the match of the target template with the features encoded in V4 is considered by computing the product $\prod\limits_{d} \max\limits_{i} r^{\mathrm{PFC}}_{d,i} \cdot r^{\mathrm{V4}}_{d,i,\mathbf{x}}$. This implements a bias to locations with a high joint probability of encoding all searched features in a certain area.

FEF decision map: The projection of the perceptual map to the decision map transforms the salient locations into a few candidate locations, which dynamically compete for determining the target location of an eye movement. This is achieved by subtracting the average saliency from the saliency at each location $w^{\mathrm{FEFv}} r^{\mathrm{FEFv}}_{\mathbf{x}} - w^{\mathrm{FEFv}}_{inh} \sum\limits_{\mathbf{x}} r^{\mathrm{FEFv}}_{\mathbf{x}}$. Thus, the cells in the decision map show none or only little response to the onset of a stimulus, such that their response fits into the category of the FEF movement cells (FEFm). Their activity provides the expected location for V4 and TE units.

4 Results

An object is presented to the model for 100 ms and the model memorizes some of its features as a target template. We do not give the model any hints which feature to memorize. The model's task is to make an eye movement towards the target (Fig. 3A,B). When presenting the search scene, TE cells that match the target template quickly increase their activity to guide perception on the level of V4 cells. Thus, the features of the object of interest are enhanced prior to any spatial focus of attention. This feature-based attention effect allows for a goal-directed planning of a saccade in the FEF. The planning of an eye movement provides a spatially organized reentry signal, which enhances the gain of all cells around the target location of the intended eye movement. As a result of these inference operations, the high-level goal description in PFC is bound to an object in the visual world. Further simulation results are discussed in [9].

We now take a close view on the feature-based attention effects of the model. In this respect we compare two conditions: attend towards the visual properties of the lighter (Fig. 3A) and attend towards the cigarettes (Fig. 3B). Fig. 3C shows the difference activity of both conditions in V4 prior to any spatial selection as determined by a low FEFm activity ($\max r^{\mathrm{FEFm}}_{\mathbf{x}}(t) < 0.05$). Our analysis clearly shows that feature-based attention selectively modulates the activity according to the task at hand. Thus, the model

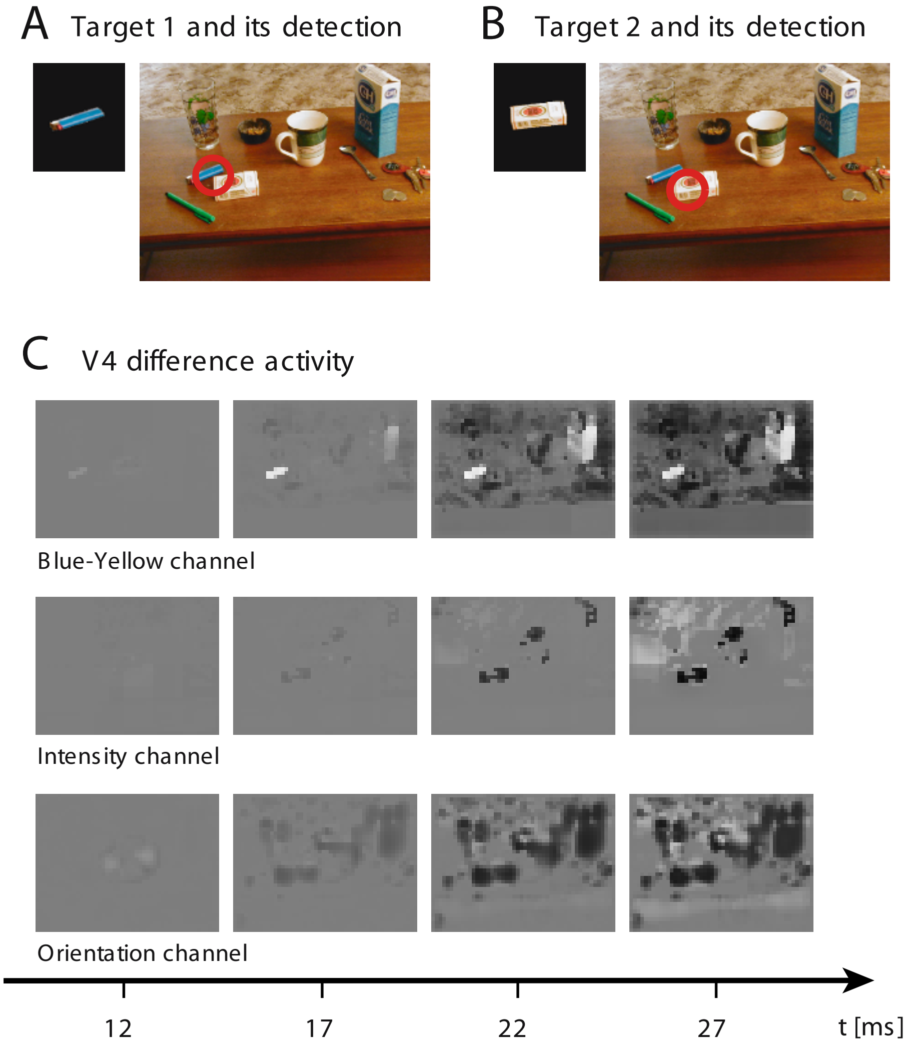

Fig. 3. Illustration of feature-based attention. A) Target object 1 and its detection in the visual scene. B) Target object 2 and its detection in the visual scene. C) Difference activity in V4 in three channels over time. For a comparison with cell recordings a latency of about 60 ms has to be added to the time axis. Only the difference of the maximal activity at each location is shown irrespective of the feature selectivity. Gray areas indicate equal (maximal) activity, light areas more activity in the first condition and dark areas more activity in the second condition. We can observe that parts of the scene are relatively enhanced or reduced according to the target template.

predicts feature-based attention effects independent of focused attention. Although the effect is global in space it can guide gaze towards the object of interest since it depends on the content encoded at each location.

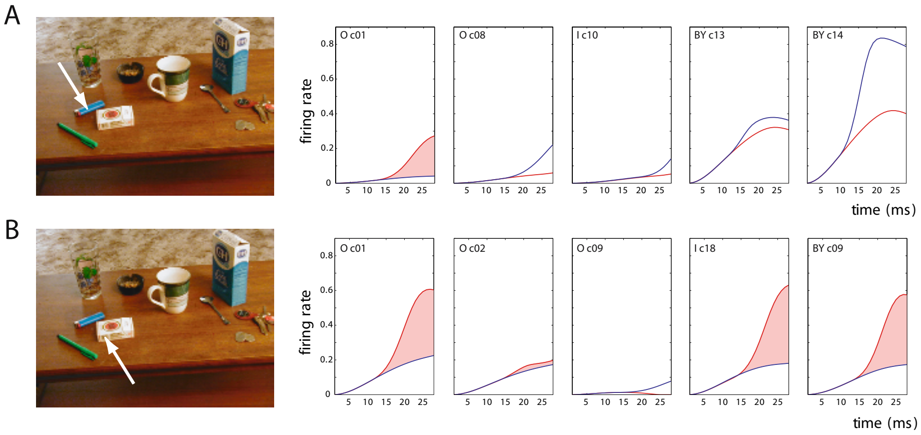

Fig. 4. Illustration of feature-based attention effects on the single cell level. The activity is shown in two conditions with time relative to search array onset (0 ms): attend towards the lighter (blue) and attend towards the cigarettes (red). The red shaded area between the curves appears when the activity in the second condition is higher. A) Selected cells in the orientation (O), intensity (I) and blue-yellow (BY) channel with the receptive field center located on the lighter. A) Selected cells in the orientation (O), intensity (I) and blue-yellow (BY) channel with the receptive field center located on the cigarette box.

To illustrate the effects of feature-based attention on the cell level we show their time course of activity. Fig. 4A shows the activity of cells with their receptive field centered on the lighter. A difference in activity between the attend lighter and attend cigarettes condition reflects the relative effect of feature-based attention. In the orientation channel (O) cell 01 shows an enhancement in the attend cigarettes condition whereas cell 08 an enhancement in the attend lighter condition. Thus, even cells with their receptive field on the lighter can be enhanced in the attend cigarettes condition. The target template for orientation in the attend lighter condition was close to horizontal and thus increased the activity of cell 08, whereas target template for orientation in the attend cigarettes condition was vertical and thus enhanced the sensitivity of cell 01 and adjacent cells. The blue color of the lighter primarily increased the activity of cells around cell 14 of the BY channel in the attend lighter condition. The white color of the cigarette box increased cell 18 of the intensity channel in the attend cigarettes condition. We observe also differences in the timing of the feature-based attention effect, which are based on recurrent interactions between V4 and TE as well as TE and PFC.

5 Discussion

We have introduced different models of attention and their implementation of feature-based attention. The classical approach, which defines attention solely by a selection of a location in the saliency map, predicts that target templates only guide the competition for spatial attention. Such guidance of spatial attention does also occur in the Selective

Tuning model as well as in our approach. These models use feature cues to enhance the activity of feature-sensitive cells. However, our approach seems to be closer to a neural correlate of feature-based attention, since we consider the temporal dynamics prior to any spatial selection. We predict that goal directed visual search first selectively modulates feature-sensitive cells prior to any spatial selection.

This prediction is consistent with cell recordings in visual search [2] and recent findings in which the learning of degraded natural scenes resulted in a selective enhancement of V4 cells [20]. According to this study V4 plays a crucial role in resolving an indeterminate level of visual processing by a coordinated interaction between bottom-up and top-down streams.

Our model further predicts that saliency is encoded as part of the variable itself through the dual coding property of a population code. Saliency is not encoded in a single map. Thus, attentional effects can be found throughout the visual system. The observation of an attentional modulation does therefore not allow to conclude that a stimulus has been selected by a spatially directed focus. For example, V4 also provides a spatially organized map encoding saliency (Fig. 3C), which is consistent with recent findings [15]. However, V4 cells are selective for location and specific features. Consistent with recordings in the FEF [23], the FEF visuomovement cells in our model are more related to the classical idea of a saliency map [11], since they solely encode location by integrating the activity across all channels and features. We assume that this information needs an additional, decisional stage of processing before it is feed back such that the saliency information is transformed into a dynamic, competitive representation of a few candidate regions.

Acknowledgements

Parts of this research had been supported by DFG HA2630/2-1 and the ERC Program of the NSF (EEC-9402726).

References

1. Ahmad, S. (1992) VISIT: a neural model of covert visual attention. In: J.E. Moody, et.al. (eds.) Advances in Neural Information Processing Systems, vol 4, 420-427, San Mateo, CA: Morgan Kaufmann.
2. Bichot, N.P., Rossi, A.F., Desimone, R. (2005) Parallel and serial neural mechanisms for visual search in macaque area V4. Science, 308:529-534.
3. Corchs, S., Deco, G. (2002) Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data. Cereb. Cortex, 12:339-348.
4. Grossberg, S., Raizada, R. (2000) Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. Vis. Research, 40:1413-1432.
5. Hamker, F.H. (2004) Predictions of a model of spatial attention using sum- and max-pooling functions. Neurocomputing, 56C, 329-343.
6. Hamker, F.H. (2004) A dynamic model of how feature cues guide spatial attention. Vision Research, 44, 501-521.

7. Hamker, F. H. (2005a) 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. Berlin, Heidelberg: Springer-Verlag, 118-132.
8. Hamker, F. H. (2005b) 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, 15:431-447.
9. Hamker, F. H. (in press) The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. Journal for Computer Vision and Image Understanding.
10. Itti, L., Koch, C. (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res., 40:1489-1506.
11. Itti L, Koch C. (2001) Computational modelling of visual attention. Nat Rev Neurosci. 2:194-203
12. Kirkland, K. L., Gerstein, G. L. (1999) A feedback model of attention and context dependence in visual cortical networks. J. Comput. Neurosci., 7:255-267.
13. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Human Psychology 4:219-227.
14. Koechlin E., Burnod Y. (1996) Dual population coding in the neocortex: A model of interaction between representation and attention in the visual cortex. J. Cog. Neurosci., 8:353-370.
15. Mazer, J.A., Gallant ,J.L. (2003) Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. Neuron, 40:1241-1250.
16. Milanese, R., Gil, S., Pun, T. (1995) Attentive mechanisms for dynamic and static scene analysis, Optical Engineer., 34:2428-2434.
17. Motter, B.C. (1994) Neural correlates of feature selective memory and pop-out in extrastriate area V4. J. Neurosci., 14, 2190-2199.
18. Navalpakkam V, Itti L. (2005) Modeling the influence of task on attention. Vision Res., 45:205-31.
19. Olshausen, B., Anderson, C., van Essen, D. (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci., 13:4700-4719.
20. Rainer G, Lee H, Logothetis NK. (2004) The effect of learning on the function of monkey extrastriate visual cortex. PLoS Biol., 2:275-283.
21. Roelfsema, P. R., Lammé, V. A., Spekreijse, H., Bosch, H. (2002) Figure-ground segregation in a recurrent network architecture. J. Cogn. Neurosci., 14:525-537.
22. Saenz M., Buracas G.T., Boynton G.M. (2002). Global effects of feature-based attention in human visual cortex. Nature Neuroscience. 5:631-632.
23. Schall, JD (2002) The neural selection and control of saccades by the frontal eye field. Phil Trans R Soc Lond B 357:1073-1082.
24. Treisman, A. (1988) Features and objects: The Fourteenth Bartlett Memorial Lecture. Quarterly Journal of Experimental Psychology, 40A:201-237.
25. Treue, S., Martínez Trujillo, J.C. (1999) Feature-based attention influences motion processing gain in macaque visual cortex. Nature, 399:575-579.
26. Tsotsos JK, Culhane SM, Wai W, Lai Y, Davis N, Nuflo F (1995) Modeling visual attention via selective tuning. Artificial Intelligence, 78:507-545.
27. van der Velde, F., de Kamps, M. (2001) From knowing what to knowing where: modeling object-based attention with feedback disinhibition of activation. J. Cogn. Neurosci., 13:479-491.
28. Wolfe, J. M. (1994) Guided search 2.0: A revised model of visual search. Psychonomic Bulletin & Review, 1, 202-238.