CHAPTER

98

How the Detection of Objects in Natural Scenes Constrains Attention in Time

Fred H. Hamker

ABSTRACT

If we want to explain attention, we ultimately have to explore how we perceive natural scenes, the environment primates typically encounter. The task of detecting an object in a natural scene constrains the involvement of attention differently than in artificial scenes. I suggest that attention emerges through time on the systems level based on three general principles, and I demonstrate their feasibility in a computational model. In this model, attention itself is not a prerequisite for object recognition but feedback constrains feedforward processing and improves target discrimination. As a result, a state evolves that allows the linking of areas involved in planning to early areas responsible for scene analysis.

I. INTRODUCTION

Attention has been investigated in numerous tasks using displays with isolated items. In those experiments and related models, the spotlight model of attention has been shown to explain a number of findings. In natural scenes, however, a subject faces the additional problem of object detection and segmentation. A pure spatially based form of attention (e.g., a spotlight of attention) hardly improves the discrimination of the object of interest against the background. I suggest the following three principles that allow the detection of objects in natural scenes: (1) High-level processing and object detection are possible without spatial attention; that is, spatial attention does not gate processing; (2) a form of feature-based attention enhances features of interest in parallel and thus enhances an object of interest against the background; and (3) focal processing needs time to develop and emerges through reentrant processing from occulomotor areas.

These three principles are routed in one general rule of perception: Convergent zones in one brain area provide an expectation about feature or space. When an expectation is sent to other areas, it enhances the gain, given that a match with the input of this area occurs. This concept of a population-based inference is related to Bayesian inference, but avoids the computation of probabilities. Competition cleans up the population activity in higher stages from all unimportant stimuli so that a full recognition can take place. As a result, attention emerges on the network level. There is no area in the brain that is solely devoted to computing attention.

II. THE MODEL

In order to explain attention as a distributed, competitive resource I have suggested that attention emerges through interactions (Hamker, 1999). The presented computational model consisted of an area with large receptive fields inferior temporal cortex (IT) that is responsible for a largely location-invariant scene description, an area with small receptive fields (V1–V4) that encodes the features of objects within a small to intermediate spatial scale and an area of spatial processing (PP, FEF, SC) that encodes information about the location of objects. Similar as in the Selective Tuning Model (Chapter 92) top-down connections within the ventral pathway play an important role (Chapters 25, 83, and 107), but here a selection in space is localized outside the ventral stream. Such a tri-modular architecture has been recently shown to account for various attention effects based on competition (Hamker, 2000; Chapter 97).

As an extension to this general model, I here describe a new population-based computational approach that aims at modeling specific areas of the brain, including their temporal dynamics (Fig. 98.1). This model has been developed to be consistent with a range of electrophysiological findings. The model V4 area (Hamker, 2004a) has been demonstrated to quantitatively account for receptive field competition in V4 (Reynolds et al., 1999) and for multiplicative effects on the tuning curve (Chapter 49). A slightly simplified version of the proposed systems model (Hamker, 2003) has been shown to match the time course of IT and V4 activity in visual search (Chelazzi et al., 1993, 1998). The model is also consistent with findings in the FEF (Schall, 2002) and psychophysical data (Hamker, 2004b), and it predicts that the FEF provides a spatially organized reentry signal to extrastriate visual areas (Hamker, 2001).

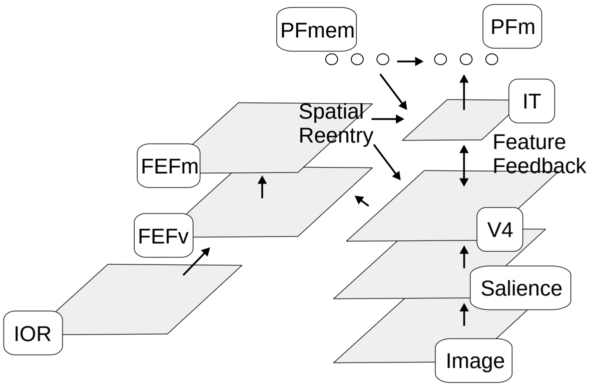

FIGURE 98.1 Model for top-down guided detection of objects. First, information about the content and its low-level stimulus-driven salience is extracted. This information is sent further upward to V4 and to IT cells that are broadly tuned to location. The target template is encoded in prefrontal memory cells (PFmem). Prefrontal match cells (PFm) 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 expected features. Feedback from IT to V4 sends the information about the target downward to cells with a higher spatial tuning. Frontal eye-field visuomovement cells (FEFv) combine the feature information across all dimensions and indicate salient or relevant locations in the scene. A competition among frontal eye-field movement cells (FEFm) determines the expected location of a target. Even during this competition, the movement cells provide a reentry signal to V4 and IT, which enhances the gain for all features at locations where the receptive field overlaps with the movement field. The inhibition of return (IOR) map memorizes recently visited locations and inhibits the FEFv cells.

Evidence for the latter prediction has been given by stimulating the FEF, which influenced stimulus-related activity in V4 (Moore and Armstrong, 2003). By fitting the model data with the experimental data gained by Chelazzi et al. (1998), I identified movement cells in the FEF as a possible convergent zone, which provides the spatial reentry signal (Hamker, 2003). Their timing and selectivity corresponds with the observed target discrimination in the ventral stream. In such a movement plan model, activity of the movement cells is required to produce a reentry signal. A potential problem could arise in explaining covert attention. During fixation, movement neurons might be inhibited by fixation cells and thus are presumably inactive, whereas visual neurons are not inhibited and therefore can provide both a reentry signal that modulates visual processes in extrastriate cortex and the target selection signal to the movement neurons. However, no experiment has clearly ruled out that the movement cells are inactive during covert attention. It is possible that fixation cell activity is reduced, which in turn allows movement cells to be active but below the level that elicits an eye movement. Others have proposed a visual selection model (Chapter 22). A potential problem of the visual selection model is its low signal-to-noise ratio. Although the visual cells show a target selection, distractor activity is initially almost equally strong. If these activities are directly fed back, spatial attention would be initially distributed to all stimuli. In addition, the target selection in visual cells appears very early as compared to the late occurrence of spatial attention in some psychophysical experiments. At present, it is not possible to rule out either model describing how the FEF might be involved in attention.

The model (Fig. 98.1) consists of the following components. Consistent with the idea of stimulus-driven salience (Chapter 39), a saliency module extracts features from the natural scene and weights their initial conspicuity by computing center-surround differences in parallel. Please note, such center-surround differences have been used by Itti and Koch (2000) to compute a saliency map. However, a saliency map that selects a location on a very fast time scale, which would be necessary to explain a popout perception on basis of a spatial selection, has not been found in the brain. Thus, according to principle 1, I suggest an alternative to an external saliency map: Stimulus-driven saliency emphasizes, but does not select, unique features in parallel within the ventral stream. For simplicity, the model computes the center-surround differences prior to V4 (see also Chapters 45 and 93). The model is consistent with the idea that stimulus-driven conspicuity can be determined at higher levels as well (Hochstein and Ahissar, 2002). I combine the feature value with its corresponding conspicuity into a population code (Hamker and Worcester, 2002), which is then continuously modified. At each location x, I construct a space, whose axes are defined by the represented features and by one additional conspicuity axis. The encoded feature is then defined by the subset of active cells within a set of neurons i sampling the feature space. The present version computes five parallel channels: intensity, orientation, red-green opponency, blue-yellow opponency, and spatial frequency.

Each V4 layer receives the obtained features, weighted by the initial conspicuity value, as input. Feature-specific feedback from IT cells and spatial reentry from the frontal eye-field movement cells both control the gain of the bottom-up input. V4 cells compete in representing their encoded stimuli.

The populations from different locations in V4 project to IT, but only within the same channel. I simulate a map containing nine populations (sets of i neurons) with overlapping receptive fields. For simplicity, the complexity of features is not increased from V4 to IT. Thus, the model IT populations represent the same feature space as model V4 populations. The receptive field size, however, increases in the model so that several populations in V4 converge onto one population in IT. IT receives feature-specific feedback from the prefrontal memory and location-specific feedback from the frontal eye-field movement cells, which again control the gain. Principle 2 is implemented in the model by feedback from prefrontal memory to IT and further back to V4.

The FEFv neurons receive convergent afferents from V4 and IT and add up the activity across all channels. The information from the target template, in addition, enhances the locations that result in a match between target and encoded feature 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 and task relevance of location (Chapter 21), pooled over different channels, whereas the conspicuity of each feature is encoded in V4 and IT.

The effect of the FEFv cells on the FEFm cells is a feedforward excitation and surround inhibition. Thus, by increasing their activity slowly over time FEFm cells determine the expected location of the target. According to principle 3, the FEFm activity provides a delayed reentry signal to extrastriate areas.

There is currently no clear indication where cells that ensure an inhibition of return are located (chapter 16). We regard each location x as inspected, dependent on the selection of an eye movement or when a match in the PFm cells is lost. In this case, the inhibition of return (IOR) cells are charged at the location of the strongest FEFm cell for a period of time. This causes a suppression of the recently attended location in the FEFv map. IOR cells slowly decay.

I now show how the FEF and IT might contribute to the detection of an object in natural scenes.

III. RESULTS

I first demonstrate how the model operates in a free-viewing task, which is only driven by the stimulus saliency (Fig. 98.2). The overt scanning behavior is

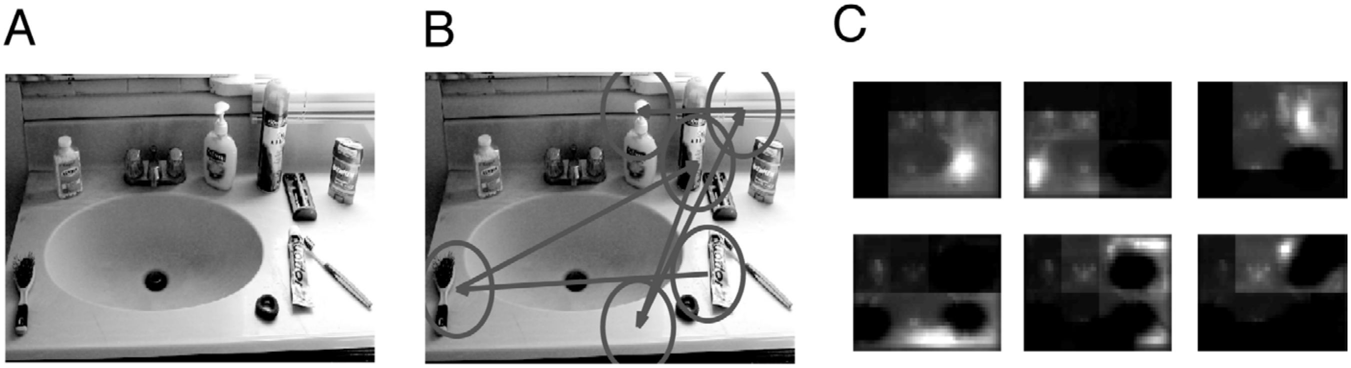

FIGURE 98.2 Results of a free-viewing task. (A) Natural scene. (B) Scanpath. The scan starts on the toothpaste and visits the hairbrush, the shaving cream, two salient edges, and then the soap. (C) Activity of FEFv cells prior to the next scan. By definition, they represent locations, which are actively 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 holes in the activity landscape).

similar to pure feedforward approaches (Chapter 94). The major difference is that the saliency is actively constructed within the network prior to each shift (Fig. 98.2C). I now demonstrate how the behavior of the model changes when it searches for a specific object in the scene. To mimic the activation of a search template, I present the model objects from which it generates very simple templates (Fig. 98.3A). This template, which is hold in PFmem cells, guides perception by changing the sensitivity of IT cells due to a feature-specific feedback. When presenting the search scene, initially IT cells reflect conspicuous features, but over time those features that match the target template get further enhanced (Fig. 98.3B). 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. At approximately 85–90 ms, all areas that contain objects are processed in parallel. Spatial reentry then enhances all features at the selected location at approximately 110 ms after scene onset. As a result, the initial top-down guided information is altered to process all the features of the target object. For example, the very red color of the asprin bottle is only encoded in IT after the emergence of the

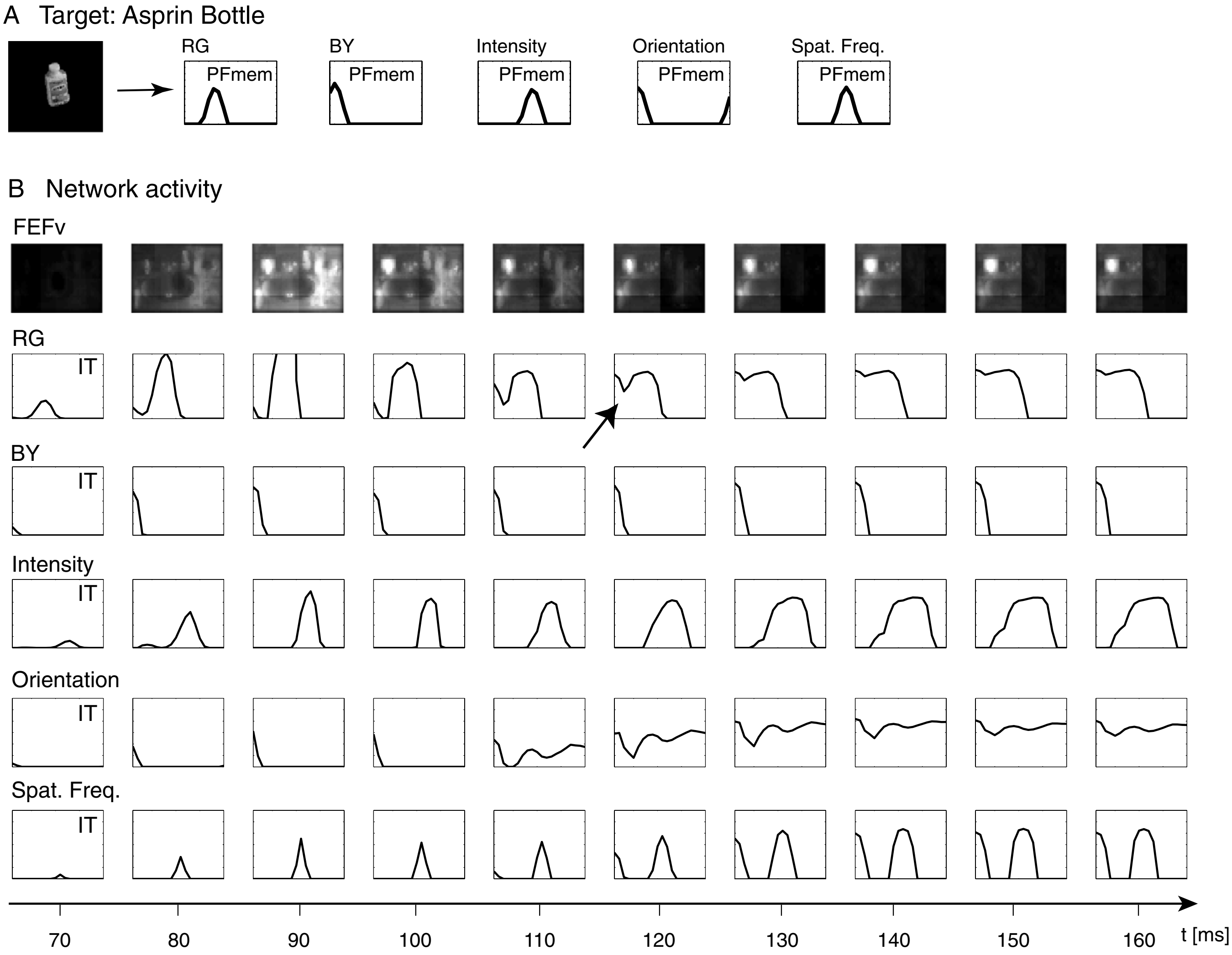

FIGURE 98.3 (A) After the presentation of a target object, PFmem cells memorize in each channel the most conspicous feature. (B) The temporal process of a goal-directed object detection task in a natural scene. The frontal eye-field visual cells indicate preferred processing, which is not identical with 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 is dominated by the emerging spatial focus and reflects other features of the object that were not searched for. The arrow indicates the enhancement of the cells encoding the red color due to spatial reentry.

spatial reentry signal because those features were not among of the expected features (arrow in Fig. 98.3B).

The model does not always search in parallel. If the target does not sufficiently discriminate from the background, reentrant processing can be misguided and the model automatically switches into a serial search mode.

IV. DISCUSSION

I have presented an approach to model perception in natural scenes based on three general principles of computation in the brain. (1) I postulate that high-level vision does not require spatial selection. This is consistent with the finding that some scenes allow the parallel detection of categories, such as animals, in the near absence of spatial attention (Li et al., 2002). (2) Because the spatial resolution of high-level vision is poor, the function of massive feedback projections is to provide cells in early location-specific areas information about the feature of interest. (3) Such enhanced activity in the "what" pathway is picked up by maps in the "where" pathway, which locates the object for action preparation. Reentrant activity, for example, from the FEF (Moore and Armstrong, 2003), then enhances all features of the object in order to allow a more detailed analysis. The model predicts that object identification begins before the eyes actually fixate on the object.

My model suggests that the brain uses reentrant processing to constrain processing in some areas by decisions in convergent areas. A match of the bottom-up input with the expectation increases the gain, which can alter the interpretation of a visual scene by tuning the population response. As a result, suppressive and facilitatory effects occur, commonly referred to as attention. My model is consistent with the idea of Biased Competition (Desimone and Duncan, 1995; Chapter 50), but population-based inference extends the idea of a mere competition toward the high-level guidance of low-level processing. The described mechanisms implement a dynamic filter that allows the connection of planning processes with the physical world and presumably the elaboration of the content of awareness (Chapter 29). Such an approach unifies recognition and attention as interdependent aspects of one network.

References

Chelazzi, L., Duncan, J., Miller, E. K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiol.* **80**, 2918–2940.

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature* **363**, 345–347.

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

Hamker, F. H. (1999). The role of feedback connections in task-driven visual search. *In* "Connectionist Models in Cognitive Neuroscience" (D. Heinke et al., Eds.) pp. 252–261. Springer Verlag, London.

Hamker, F. H. (2000). Distributed competition in directed attention. *In* "Proceedings in Artificial Intelligence" (G. Baratoff and H. Neumann, Eds.), Vol. 9, pp. 39–44. Akademische Verlagsgesellschaft, Berlin.

Hamker, F. H. (2001). Attention as a result of distributed competition. *Soc. Neurosci. Abstr.* **27**, 348.10.

Hamker, F. H. (2003). The reentry hypothesis: linking eye movements to visual perception. *J. Vis.* **11**, 808–816.

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

Hamker, F. H. (2004b). A dynamic model of how feature cues guide spatial attention. *Vis. Res.* **44**, 501–521.

Hamker, F. H., and Worcester, J. (2002). Object detection in natural scenes by feedback. *In* "Biologically Motivated Computer Vision" (H. H. Bülthoff et al., Eds.), pp. 398–407. Springer Verlag, New York.

Hochstein S, and Ahissar M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* **36**, 791–780.

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

Li, F.-F., VanRullen, R., Koch, C., and Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9596–9601.

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

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

Schall, J. D. (2002). The neural selection and control of saccades by the frontal eye field. *Phil. Trans. R. Soc. Lond. B* **357**, 1073–1082.