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

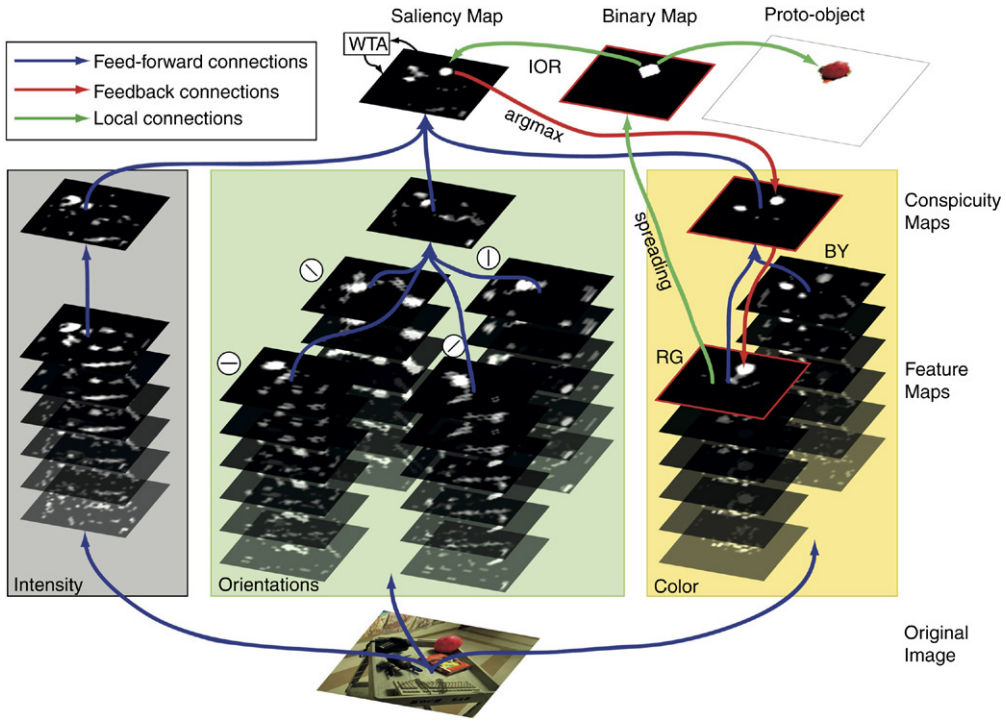


Fig. 1. Illustration of the processing steps for obtaining the attended region. The input image is processed for low-level features at multiple scales, and center-surround differences are computed (Eq. (6)). The resulting feature maps are combined into conspicuity maps (Eq. (9)) and, finally, into a 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)