# Attention in Hierarchical Models of Object Recognition

Dirk B. Walther
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
405 N. Mathews Ave., Urbana, IL 61801, USA
walther@uiuc.edu

Christof Koch
Division of Biology
California Institute of Technology, MC 216-76
Pasadena, CA 91125, USA

March 20, 2007

## Abstract

Object recognition and visual attention are tightly linked processes in human perception. Over the last three decades, many models have been suggested to explain these two processes and their interactions, and in some cases these models appear to contradict each other. We suggest a unifying framework for object recognition and attention and review the existing modeling literature in this context. Furthermore, we demonstrate a proof-of-concept implementation for sharing complex features between recognition and attention as a mode of top-down attention to particular objects or object categories.

*"At first he'd most easily make out the shadows; and after that the phantoms of the human beings and the other things in water; and, later, the things themselves."* — Socrates describing the visual experience of a man exposed to the richness of the visual world outside his cave for the first time (Plato, The Republic).

# 1 Introduction

Vision is the sense we humans rely on most for our everyday activities. But what does it mean to see? When light reflected by an object hits the photoreceptors of the retina, electrical impulses are created by retinal ganglion cells and sent out to other parts of the brain. How do these electrical impulses give rise to the percepts of the visual world surrounding us?

Two important parts of visual perception are the recognition of objects and the gating of visual information by attention. By object recognition we mean our visual system's ability to infer the presence of a particular object or member of an object category from the retinal image. By attention we mean the process of selecting and gating visual information based on saliency in the image itself (bottom-up), and on prior knowledge about scenes or a particular task (top-down) (Desimone and Duncan, 1995; Itti and Koch, 2001).

The algorithms required for object recognition are governed by the complementary forces of *specificity* and *invariance*. The activation level of a particular cone in the retina has a very *specific* spatial location, but it is likely to have the exact same activation level for a wide range of objects. Activation of object-selective cells in inferior temporal cortex, on the other hand, *specifically* indicates the presence of a particular object (or a member of a particular object category). To a large extent, this representation is *invariant* to where the object is located (at least in areas near the fovea), which way it is oriented, if it appears to be large or small, or whether it is brightly illuminated or in the shadow.

The majority of models of object recognition have at their heart a hierarchy of processing steps that more or less gradually increase both *specificity* to the structure of the stimulus and *invariance* to translation, rotation, size, and illumination. There is less agreement about the details of these steps, the tuning of units at intermediate levels, and the representation of their spatial relations.

Gating by attention may occur at any level of processing. Typically, attention is modeled as the preferred processing of some visual information selected by spatial location and/or the encoded feature(s). We analyze the various modes of attention in more detail in section 4.

Virtually all models of object recognition in cortex start with filtering the incoming image with orientation-sensitive filters. They approximate the receptive fields of the simple and complex cells found by Hubel and Wiesel (1962) in cat striate cortex with Gabor filters, steerable filters, or other orientation-tuned filters.

Details of the subsequent steps of processing are much more contentious. One of the controversies is about whether the three-dimensional structure of objects is represented by an explicit description of its components and their spatial relations, or whether it is inferred by interpolation between several two-dimensional views. We will briefly review both approaches in the next section and then show how they can be described by a unifying framework in section 3. This framework will also allow us to explain roles of attention in object recognition in section 4.

# 2 Hierarchical Models of Object Recognition

The two main ideas for implementing recognition of three-dimensional objects are recognition by components and view-based recognition. In this section we survey models for both approaches and highlight their differences and the common elements.

## 2.1 Recognition by Components

Marr postulated a primal sketch for inferring the presence of surfaces, which are then combined into a relief-like $2\frac{1}{2}$d sketch and eventually into 3d models of objects, which can be indexed and later recalled and referenced (Marr and Nishihara, 1978; Marr, 1982). Building on Marr's work, Biederman (1987) suggested that the next processing step should be fitting simple 3d shapes (generalized cones and cylinders termed "geons") to the surfaces, and that spatial relationships between geons are encoded explicitly. Crucial to the correct recognition of 3d objects are non-accidental properties such as T-junctions and line intersections.

Experimental evidence for this recognition-by-components (RBC) model comes from a study demonstrating successful priming for object identity by degraded line drawings, even when the priming and the primed stimulus have no lines in common (Biederman and Cooper, 1991). A recent study supports these results by finding stronger fMRI adaptation to line drawings with local features deleted than for line drawings with entire components (geons) removed (Hayworth and Biederman, 2006). Demonstration of a computational implementation of the model, however, was limited to carefully selected line drawings (Hummel and Biederman, 1992). In fact, the biggest criticism of RBC points at the difficulty of fitting geons with a potentially large variety of parameters to images of natural objects (e.g., Edelman, 1997).

## 2.2 View-based Recognition in HMAX

An alternative approach to recognition by components is the recognition of 3d objects by interpolating between 2d views of the objects at various rotations. Almost all models of view-based recognition trace their origins to the "Neocognitron", a hierarchical network developed by Fukushima (1980). The Neocognitron consists of alternating S and C layers, in an allusion to the simple and complex cells found by Hubel and Wiesel (1962) in cat visual cortex. The first layer of S units consists of Gabor-like edge detectors, and their output is pooled spatially by the corresponding C layer, leaving the shape tuning unaffected. The next S layer is responsible for recombining activations from the preceding layer into new, more complex patterns. The output is again pooled spatially by the next C layer and so forth. Typically, three such S and C layer sandwiches are stacked into a hierarchy, providing increasing complexity of the features and increasing invariance to stimulus translation as well as to slight deformations.

The network performs well on the recognition of hand-written digits, which are inherently two-dimensional.

How does this help in recognizing 3d objects? Ullman (1979) showed that it is possible to infer the 3d structure of objects from as few as two planar views, given the correspondence of feature points between the views. Poggio and Edelman (1990) applied this idea to the recognition of 3d objects from a series of 2d views in a network that uses Generalized Radial Basis Functions to match the spatial coordinates of sets of feature points between the views. These results inspired a study of the tuning properties of cells in inferior temporal (IT) cortex of macaque monkeys by Logothetis et al. (1994), who found that IT cells show a high degree of invariance to size changes and translations of previously learned objects, but that tuning to rotations of these objects is fairly narrow, thereby supporting a view-based theory of the perception of 3d objects. Further support for view-based object recognition comes from behavioral studies of the ability to learn and recognize novel objects from different angles (Tarr and Bülthoff, 1995; Gauthier and Tarr, 1997).

In their HMAX model of object recognition, Riesenhuber and Poggio (1999) combined Fukushima's idea of gradual build-up of invariance and complexity with the insight of view-based recognition of 3d objects. In its original form, HMAX consists of a sequence of two sets of S and C layers and so-called view-tuned units (VTUs), which are fed by the C2 layer. The S1 layer consists of edge detectors using Gabor filters. Layer C1 pools over spatial location (and scale) of S1 activity using a maximum (*max*) operation. The *max* operation was chosen instead of a linear sum in order to retain feature specificity of the signal across the pooling step. Recombination of C1 activations into S2 activity is achieved by hard-wired connections of all possible combinations of orientation-specific C1 units in a $2 \times 2$ neighborhood. Layer C2 pools over the remaining spatial locations, so that the spatial receptive field of C2 units encompasses the entire visual field of the model. Patterns of C2 activity, learned from pre-labeled training examples, are associated with specific object views, and VTUs belonging to different aspects of the same object are pooled into object sensitive cells. See figure 1 (feed-forward connections only) for a schematic of HMAX.

[Figure 1 about here.]

The model was shown to successfully learn and recognize computer-generated images of 3d wire frame ("paperclip") stimuli, faces, and rendered cats and dogs (Riesenhuber and Poggio, 1999; Freedman et al., 2003). Serre et al. (2005) endowed the model with a method for learning the connections from layer S1 to C2 from natural scene statistics and later added an additional set of S and C cells for better correspondence with functional areas of the primate brain. With these additions, the model (now frequently called the "standard model") is able to learn and recognize a large variety of real-world object categories in photographs (Serre et al., 2007a; Serre et al., 2007b).

## 2.3 Other View-based Models

The idea of view-based object recognition was also followed by others. Wallis and Rolls (1997), for instance, presented a hierarchical model of object recognition based on closely matching the receptive field properties of simple and complex cells. They report good performance of a computational implementation of their network for detecting "L", "T", and "+" stimuli as well as faces.

LeCun et al. (1998) refined the ideas of Fukushima in their back-propagation neural network for character recognition ("Le Net"). In his SEEMORE model of object recognition, Mel (1997) employed a rich set of low-level features, including oriented edge filters, color filters, and blobs. Edge filter responses were combined into contours and corners. A neural network trained on the output of all these filters was able to recognize simple objects in photographs with good performance and, as expected, decreasing performance for degraded images.

The model by Amit and Mascaro (2003) for combining object recognition and visual attention is also view-based, and it also employs features of increasing complexity. Translation invariant detection is achieved by pooling over the output of detectors at multiple locations using an *or* operation, the binary equivalent to Riesenhuber and Poggio's *max* operation. Basic units in their model consist of feature-location pairs, where location is measured with respect to the center of mass. Detection proceeds at many locations simultaneously, using hypercolumns with replica units that store copies of some image areas. Complex features are defined as combinations of orientations. There is a trade-off between accuracy and combinatorics: more complex features lead to a better detection algorithm, but more features are needed to represent all objects, i.e., the dimensionality of the feature space increases.

Further support for the suitability of a hierarchy with increasing feature complexity comes from a study by Ullman et al. (2002), who showed that features (in their case rectangular image patches) of intermediate complexity carry more information about object categories than features of low or high complexity (see also Ullman, 2007).

## 2.4 Representation of Spatial Relations

An important difference between the view-based and the component-based models is the way in which spatial relations between parts are encoded (see table 1). Biederman (1987) describes a system, in which the relations between the detected components (geons) are encoded in an explicit way, either qualitatively (e.g., using spatial relations such as "above", "to the right" etc.) or quantitatively (e.g., "2 cm to the right").

In view-based models, spatial relations are represented implicitly by the design of the increasingly complex features. Their tuning with respect to the feature representations in the preceding layer inherently includes the spatial relations between these earlier, simpler features.

[Table 1 about here.]

There has been a long debate in the literature over those two contrary views of the representation of spatial relations, with experimental evidence presented for both (e.g., Biederman and Cooper, 1991; Bülthoff et al., 1995). At the current stage, component-based vision in both experiments and models appears to be limited to carefully prepared line drawings, while view-based models generalize well to photographs of real-world objects (Mutch and Lowe, 2006; Serre et al., 2007b; Ullman, 2007).

Regardless of the nature of the encoding of spatial relations, all models have in common the notion of increasing receptive field size and increasing complexity of tuning due to recombinations of simpler features. These traits can be summarized in a unifying formal framework. In the following section we describe such a framework, which we will use to explain roles of attention in object recognition in section 4.

# 3 A Unifying Framework for Attention and Object Recognition (UNI)

In a very broad sense, object recognition is a correspondence between the space of all possible images and the abstract space of all possible objects, linking objects with the images in which they appear. Note that this correspondence is not a function, since many images may be linked to the same object, and one image may contain several objects. For instance, an image of granny with a hat is associated with the representation of grandmother as well as that of grandmother's hat. Likewise, many different images of grandma are associated with the same grandmother representation.

The specifics of this correspondence may depend on the individual observer's prior experience, the task the individual is involved in, the state of alertness, and many other factors. Here we attempt to break down object recognition into intermediate steps that are in approximate agreement with neurophysiological and psychophysical evidence. Subdividing object recognition in this way also allows for the injection of attentional biases at various stages of processing.

## 3.1 Some Definitions

We model object recognition as a hierarchy of operations. Each level of the hierarchy consists of a bundle of retinotopic maps. By "map" we mean a retinotopically organized array with scalar values, encoding a particular feature. Bundles of maps encode several features, one map for each feature.

In the brain, these bundles of maps could be implemented in two general types of arrangements: as spatially separate maps, or as an array of hypercolumns, where each hypercolumn contains the activations of all maps at this location in retinotopic space. The particular type of arrangement does not matter for the computational principles outlined in the following section. In fact, a combination of the two types is often the most likely scenario.

As we move upward in the hierarchy, the complexity of the features encoded in the bundles of maps as well as the receptive field size will increase, until we arrive at object-selective units whose receptive field spans large parts of the visual field.

We start out by providing a formal way of encoding one such layer of the hierarchy. Then we will show how activity is transformed from layer to layer. Finally, we will construct a general formal framework for object recognition from stacking multiple layers in a hierarchy.

### 3.1.1 Bundles of Maps

Let us start with the definition of a bundle of maps. Each map corresponds to a feature $k \in \{1, ..., K\}$ and assigns an activation level $m(x, y, k)$ to map coordinates $(x, y)$. Thus, we can write a bundle of $K$ maps as a set of quadruples:

$$\mathbf{M} = \left\{(x, y, k, m) \,|\, (x, y) \in \mathbb{N}^2, k \in \{1, ..., K\}, m = m(x, y, k) \in \mathbb{R}\right\}. \quad (1)$$

We can construct feature maps recursively. The first layer (**M**) may encode the image, and $k$ may index the color channels for red, green, and blue. The second layer (**M'**) may encode center-surround color contrasts from the colors in the first layer, taking into account the structure of the features within a neighborhood of a given location in the first layer (e.g., Mel, 1997; Itti et al., 1998). This neighborhood is the spatial receptive field of the second layer.

### 3.1.2 Spatial Receptive Fields

The spatial receptive field can be described by an index function:

$$r(x, y, x_0, y_0) = \begin{cases} 1 & \text{if } (x, y) \text{ is part of the receptive field at } (x_0, y_0), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The receptive field around $(x_0, y_0)$ is the support of $r$ with $x_0$ and $y_0$ fixed, i.e., the set of all pairs $(x, y)$, for which $r(x, y, x_0, y_0)$ is larger than zero:

$$RF(x_0, y_0) = \sup\left(r(\cdot, \cdot, x_0, y_0)\right) = \{(x, y) \,|\, r(x, y, x_0, y_0) > 0\}. \quad (3)$$

Typically, the receptive field is defined as a contiguous disc-shaped region with center $(x_0, y_0)$. However, the definitions in eqs. 2 and 3 are general enough to also allow for non-contiguous regions, e.g., for units that are selective to the co-occurrence of two stimuli, such as geons, at separate locations.

### 3.1.3 Feature Recombination Functions

The features in layer **M'** can be modeled as a feature recombination function $\phi$, which maps combinations of activations for all $K$ features within the receptive field $RF$ at $(x_0, y_0)$ to the new features $K'$ at $(x_0, y_0)$:

$$\phi : \mathbb{R}^{K \cdot \|RF(x_0, y_0)\|} \rightarrow \mathbb{R}^{K'}. \quad (4)$$

In an exclusively linear model, $\phi$ would be a matrix. For example, the columns of $\phi$ could contain Gabor convolution kernels for mapping local image regions to the respective filter responses. More generally, $\phi$ can be any kind of function, including non-linearities such as sigmoidal functions as used in back-propagation networks (e.g., LeCun et al., 1998).

#### 3.1.4 Spatial and Feature-based Attentional Modulation

We would like to have the ability to modulate activity according to spatial and feature-based attentional biases. We achieve this by introducing a spatial modulation function $s$ and a feature modulation function $f$. The spatial modulation function assigns a non-negative real modulation factor to each coordinate pair $(x, y)$:

$$s : \mathbb{N}^2 \to [0, \infty). \tag{5}$$

Locations with $s = 1$ are not modulated; locations with $s = 0$ are entirely suppressed; locations with $0 < s < 1$ are slightly suppressed; and locations with $s > 1$ are enhanced due to spatial attention.

Similarly, the feature modulation function assigns a non-negative modulation factor to each feature index:

$$f : \{1, ..., K\} \to [0, \infty). \tag{6}$$

Both $s$ and $f$ can be used in a binary mode with only the values 0 for *not selected* and 1 for *selected*, or they can be used for more fine-grained attentional modulation.

It may seem like an omission to reduce the effects of attention to gain modulation only. In fact, other effects such as increased baseline activity, shifting of tuning curves, or biasing competitive interactions have been suggested in models and demonstrated experimentally (McAdams and Maunsell, 1999; Rees et al., 1997; Desimone and Duncan, 1995). In our unifying framework, these effects can be implemented by gain modulation of the afferent connections, i.e., by modulating activity in the preceding layer.

#### 3.1.5 Linking Layers

With these definitions, the feature activation $m'(x', y', k')$ at location $(x', y')$ for feature $k'$ in **M'** is given by:

$$m' = \phi_{k'} \left( \{ m(x, y, k) \cdot s(x, y) \cdot f(k) \,|\, (x, y) \in RF(x', y'), k \in \{1, ...K\} \} \right), \tag{7}$$

where $\phi_{k'}$ denotes the $k'$th component of the vector-valued function $\phi$.

Finally, we can write the bundle of maps **M'** as:

$$\mathbf{M'} = \left\{ (x', y', k', m') \,|\, (x', y') \in \mathbb{N}^2, k' \in \{1, ..., K'\}, m' \in \mathbb{R} \right\}, \tag{8}$$

with $m'(x', y', k')$ as defined in eq. 7.

Observe that we used four functions to derive **M'** from **M**: the receptive field index function $r$, the feature recombination function $\phi$, the spatial modulation

function $s$, and the feature modulation function $f$. To indicate this fact, we will use the following notation:

$$\mathbf{M} \xrightarrow{(r,\phi,s,f)} \mathbf{M'} \tag{9}$$

## 3.2 The Recognition Hierarchy

The definitions of a bundle of maps and of the mapping from one bundle to the next provide us with the building blocks for assembling the framework for hierarchical object recognition. We will build our framework with four basic layers: the *image* **IM**, a *simple features layer* **SF**, a *complex features layer* **CF**, and an *object layer* **OB**. An *abstract object layer* $\mathbb{A}$ on top of the hierarchy as a fifth layer is a placeholder for a variety of cognitive functions, such as the realization of the percept, i.e., the awareness that particular objects are present, or the answer to the question whether a certain object is contained in the scene, or committing the perceived objects to memory. The entire feed-forward hierarchy looks like this:

$$\begin{array}{c} \mathbb{A} \\ \uparrow f_{\mathrm{ob}} \\ \mathbf{OB} \\ \uparrow (r_{\mathrm{ob}}, \phi_{\mathrm{ob}}, s_{\mathrm{cf}}, f_{\mathrm{cf}}) \\ \mathbf{CF} \\ \uparrow (r_{\mathrm{cf}}, \phi_{\mathrm{cf}}, s_{\mathrm{sf}}, f_{\mathrm{sf}}) \\ \mathbf{SF} \\ \uparrow (r_{\mathrm{sf}}, \phi_{\mathrm{sf}}, s_{\mathrm{im}}, f_{\mathrm{im}}) \\ \mathbf{IM} \end{array} \tag{10}$$

Let us now discuss the details of each of these layers.

The recognition process starts out with an image, which can be expressed as a bundle of maps:

$$\mathbf{IM} = \left\{ (x, y, k_{\mathrm{im}}, m_{\mathrm{im}}) \,|\, (x,y) \in \mathbb{N}^2, k_{\mathrm{im}} \in \{1, ..., K_{\mathrm{im}}\}, m_{\mathrm{im}} \in \mathbb{R} \right\}. \tag{11}$$

$k_{\mathrm{im}}$ enumerates the color channels of the image, and $m_{\mathrm{im}}(x, y, k_{\mathrm{im}})$ is the pixel value of color channel $k_{\mathrm{im}}$ at location $(x, y)$.

In the case of an RGB image, for example, we would have $K_{\mathrm{im}} = 3$, and $m_{\mathrm{im}}(3, 4, 1)$ would be the value of the red channel at the location with coordinates $x = 3$ and $y = 4$. Depending on the color space of the image, $K_{\mathrm{im}}$ can have different values: $K_{\mathrm{im}} = 1$ (e.g., a gray-level image), $K_{\mathrm{im}} = 3$ (e.g., RGB, LMS, HSV, YUV, or CIELAB color spaces), $K_{\mathrm{im}} = 4$ (e.g., CMYK), or other values for any kind of color space.

When used in a binary mode, the spatial modulation function $s_{\mathrm{im}}$ allows for spatial selection of a sub-region of the image for further processing. This mechanism is frequently described as an attentional window sliding over the image (e.g., Olshausen et al., 1993; Rutishauser et al., 2004; Walther et al., 2005a), and it is sometimes seen as the equivalent of eye movements selecting parts of a complex scene (e.g., Rybak et al., 1998).

Color channels can be selected with the feature modulation function $f_{\mathrm{im}}$. When using HSV color space, for instance, processing could be limited to the luminance (value) channel.

In the first processing step, simple features are derived from the image pixels. Typically, simple features are modeled as convolutions with Gabor filters of various orientations, spatial frequencies, and orientations (e.g., Fukushima, 1980; LeCun et al., 1998; Riesenhuber and Poggio, 1999). Other possibilities include color center-surround contrasts at various scales (e.g., Mel, 1997; Itti et al., 1998) or texture detectors (e.g., Ullman et al., 2002). All these operations are consolidated in the feature recombination function $\phi_{\mathrm{sf}}$, and their spatial receptive field properties are encoded in $r_{\mathrm{sf}}$.

These operations result in a bundle of maps **SF** with $K_{\mathrm{sf}}$ simple feature maps. $K_{\mathrm{sf}}$ can become quite large when considering Gabor filters with different spatial frequencies and phases at several orientations and spatial scales, for instance. The feature modulation function $f_{\mathrm{sf}}$ provides the mechanism for feature-based attention, allowing for modulation of activity or even restricting processing to only a few of these features. Spatial attention is expressed in the spatial modulation function $s_{\mathrm{sf}}$.

Complex features are encoded in the bundle of maps **CF**. They can encompass a variety of structures, such as corners and line intersections, parts of objects (patches) of an intermediate size, or 3d geometric shapes (geons). Their construction from simple features is described by $\phi_{\mathrm{cf}}$, and the spatial receptive field properties are given by $r_{\mathrm{cf}}$. The feature recombination function $\phi_{\mathrm{cf}}$ can be hard wired into the model architecture (e.g., Mel, 1997; Riesenhuber and Poggio, 1999), or it can be learned from image statistics (e.g., LeCun et al., 1998; Ullman et al., 2002; Serre et al., 2007b).

In some models of object recognition, several layers of increasingly complex features follow before objects are recognized or categorized (e.g., Fukushima, 1980; LeCun et al., 1998; Serre et al., 2007b). In a gross simplification we collapse all these intermediate complexity layers into the one complex feature layer **CF**. The activation in this layer can be modulated by spatial ($s_{\mathrm{cf}}$) and feature-based attention ($f_{\mathrm{cf}}$).

We assume that objects are recognized based on the information present in **CF**, activating object units **OB** according to the rules in $\phi_{\mathrm{ob}}$. The **OB** layer is functionally approximately equivalent to cells in the monkey inferior temporal cortex (IT). While **OB** units respond very specifically to an object category, a particular object, or a particular view of an object, they have very little spatial resolution. This means that their receptive fields ($r_{\mathrm{ob}}$) typically encompass large regions of the visual field, and that their feature activations $m_{\mathrm{ob}}(x, y, k)$ respond specifically to the presence of a particular object anywhere within that receptive

field. Many models replace the maps in **OB** with only one unit for each of the $K_{\mathrm{ob}}$ features.

Note that we start out with high spatial resolution (i.e., high spatial specificity) in **IM**, but with only one to four features (color channels). As information is processed in the hierarchy of eq. 10, spatial specificity decreases, but the number of features, and therefore their specificity, increases. The **OB** layer, finally, is fairly insensitive to the location of objects in the visual field, but it contains a potentially very large number of object-specific maps – up to tens of thousands in a fully trained model, according to estimates of the number of visual categories (Biederman, 1987).

It should be pointed out that the formalism described so far is agnostic to the way spatial relations between object parts are encoded. The definition of the receptive field index function (eq. 2) and the feature recombination function (eq. 4) are sufficiently general to encompass both explicit encoding of spatial relations, as in the RBC model, and implicit encoding by increasingly complex features, as in view-based recognition (see subsection 2.4). In fact, once encoded in the feature recombination function $\phi$, explicit and implicit encoding become almost indistinguishable.

The specifics of the mapping from the object-sensitive units of layer **OB** to the abstract object space $\mathbb{A}$ depend highly on task and context. A typical instantiation of $\mathbb{A}$ in computational models would be the look-up and report of an object label. In a behaving human or animal, this could be the behavioral response required by the task or situation. Other potential instantiations of $\mathbb{A}$ include committing the percept to memory or assigning an emotional value to the perceived objects. The feature modulation function $f_{\mathrm{ob}}$ allows for preferential perception of particular objects or object categories.

# 4 Mechanisms of Attention

Now that we have mapped out this general formal framework, we use it to review a number of ways of integrating attention with object recognition. Modes of attention can be characterized by the origin of the attentional signal and by the way attention is deployed.

Bottom-up attention is derived only from low-level image properties, typically determining the salience of target regions by some sort of feature contrast. This mode of attention is fast, automatic and task-independent. Top-down attention, on the other hand, is driven by a task or an intention. Its deployment is comparatively slow and volition-controlled. Most of the models surveyed in this section have provisions for both bottom-up and top-down attention.

The most common way to deploy attention is spatial. In this mode, an attentional "spotlight" selectively enhances processing at a particular location in the visual field (Posner, 1980; Treisman and Gelade, 1980). Occasionally, attention is compared to a zoom lens (Eriksen and St. James, 1986), adapting the size of the spotlight to the attended object.

In feature-based attention, processing of particular features is biased in a

way that is optimal for detecting a known target. Feature-based attention is deployed either directly to the object recognition hierarchy, or it is deployed spatially by biasing the computation of a spatial saliency map.

Object-based attention captures a variety of effects that range from spatially limiting attention to the attended object to setting optimal feature biases for a particular search target.

The unifying framework (UNI) described in the previous section provides the means to model deployment of both spatial and feature-based attention via the spatial and feature modulation functions at each processing level. In this section we review various models of visual attention and show that most attention effects can be modeled within the UNI framework.

## 4.1 The Saliency Map

The saliency map is a retinotopic map whose activation strength is an indicator for how much a particular image region should attract attention based solely on bottom-up, image-based information. This notion of saliency was introduced by Koch and Ullman (1985).

A computational implementation of the model was given by Itti et al. (1998). In this implementation, several simple features (**SF**) are extracted from the input image at multiple scales in feature pyramids: red-green and blue-yellow color opponencies, luminance, and the four canonical orientations. Center-surround contrasts in these features are computed as differences between scales in the respective feature pyramids and, after normalization, stored in "feature maps". Feature maps can be written as a bundle of maps **FM**, and the center-surround and normalization operations for computing **FM** from the simple features as $\phi_{\text{fm}}$.

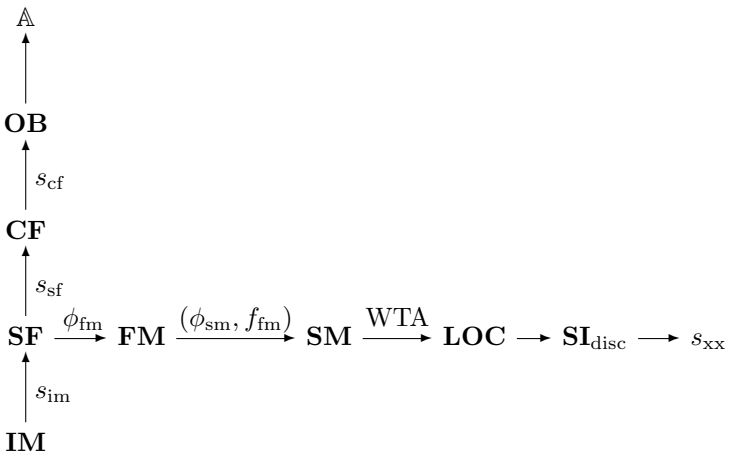
In the original implementation by Itti et al. (1998) the number of feature maps is $K_{\text{fm}} = 42$ (4 orientations, 2 color contrasts, 1 luminance contrast, all at 6 spatial scale combinations). In later versions of the model, features such as flicker, motion, and extended contours were added (Itti, 2005; Peters et al., 2005; Carmi and Itti, 2006).

In a series of normalization and across-scale pooling steps, feature maps are eventually combined into a saliency map, itself a bundle of one map **SM** with $K_{\text{sm}} = 1$. We model the contribution of each feature to the saliency map with a feature modulation function $f_{\text{fm}}$.

A winner-take-all (WTA) network of integrate-and-fire neurons determines the most active location in the saliency map, which is then attended to. The attended location is inhibited (inhibition of return, IOR), and competition continues for the next most salient location. The succession of attended locations can be written as a bundle of maps **LOC**, where each map is 0 everywhere except for the attended location, where it is 1. **LOC** contains as many maps as there are successive fixations on the image.

In order to arrive at a spatial modulation function with spatially extended regions, Itti et al. (1998) convolve the maps in **LOC** with a disc-shaped kernel of fixed size, arriving at a bundle of binary spatial modulation maps $\mathbf{SI}_{\text{disc}}$. The

entire flow of processing can be summarized as:




(12)

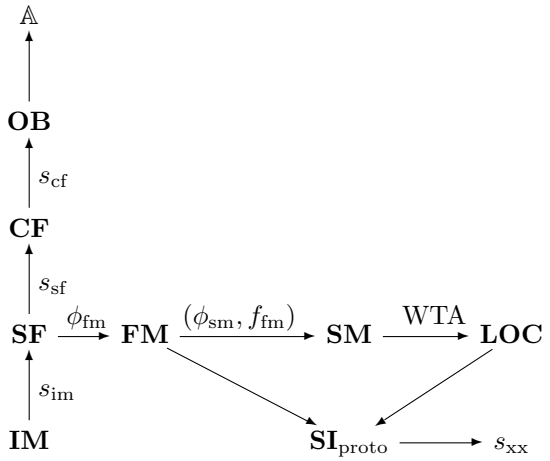
The individual spatial modulation maps $m_{\mathrm{disc}}(x, y, k)$ in $\mathbf{SI}_{\mathrm{disc}}$ can be applied at any of the processing stages in the object recognition hierarchy.

Miau et al. (2001) demonstrated the use of these maps as binary spatial modulation functions $s_{\mathrm{im}}$ for the first processing step from **IM** to **SF**. They implemented this deployment of spatial attention by cutting out rectangular sections of the image around the attended locations and limiting object recognition to these regions.

Building on the work of Itti et al. (1998), we have developed a mechanism for attending to salient proto-objects (Walther et al., 2002; Walther and Koch, 2006). This work was inspired by Rensink's coherence theory, in which low-level "proto-objects" are formed rapidly and in parallel across the visual field prior to attention. When focused attention accesses a proto-object, it becomes available to higher-level perception as a coherent object, but it loses its coherence once attention is released (Rensink, 2000a; Rensink, 2000b).

In our version of the saliency model, feedback mechanisms within the saliency computations identify the feature map with the strongest contribution to the saliency at the attended location. Spreading of attention from the attended location over a contiguous region of high activation within that feature map yields the shape of the attended proto-object. This provides us with a first estimate of the size and extent of an attended object, object part, or group of

objects[1]. In a modification of eq. 12, our approach can be summarized as:




(13)

We (Walther and Koch, 2006) have modeled the deployment of spatial attention to proto-objects at the level of $s_{\mathrm{cf}}$ in the HMAX model. In a task that required the successive recognition of two objects in an image, we varied the strength of the spatial modulation between 0 % (no attentional modulation, corresponding to $s_{\mathrm{cf}} = 1$ at all loactions) and 100 % (binary modulation with total suppression of regions outside of the attended proto-object, corresponding to $s_{\mathrm{cf}} = 0$ there). We found that a modulation strength of 20 % suffices for successful deployment of attention for the recognition of simple wire frame object, and 40 % for the recognition of faces. These values are in good agreement with attentional modulation found in the response of neurons in area V4 of macaques (Spitzer et al., 1988; Connor et al., 1997; Luck et al., 1997; Reynolds et al., 2000; Chelazzi et al., 2001; McAdams and Maunsell, 2000). Deploying spatial attention at the level of $s_{\mathrm{sf}}$ yielded very similar results.

Working with a non-biological object recognition algorithm (Lowe, 2004), we applied binary versions of the maps in $\mathbf{SI}_{\mathrm{proto}}$ to $s_{\mathrm{im}}$ and enabled learning and recognition of multiple objects in cluttered scenes (Rutishauser et al., 2004; Walther et al., 2005a). Using binary versions of the spatial modulation functions and applying them directly to the image instead of later steps is sensible in computer vision, because computationally expensive processing can be restricted to the support of $s_{\mathrm{im}}$, i.e., those image regions $(x, y)$, for which $s_{\mathrm{im}}(x, y) > 0$.

The idea of a saliency map was used by others as well. Milanese et al. (1994), for instance, describe a method for combining bottom-up and top-down information through relaxation in an associative memory. Frintrop et al. (2005) included depth information from a laser range finder in their version of a saliency map.

[1]Matlab code for this model is available online at `http://www.saliencytoolbox.net`.

## 4.2 Other Models of Spatial Attention

With their MORSEL model of object recognition and selective attention, Mozer (1991) and Mozer and Sitton (1998) implemented a connectionist network and tested it successfully with stylized letters in four standard positions in the display. Spatial attention is deployed as gain modulation of the activity of their first layer ("retina"), corresponding to the **IM** bundle of maps in our UNI framework.

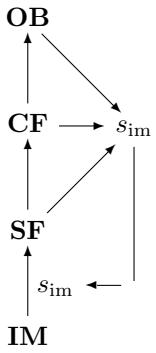
In the "shifter circuit" model by Olshausen et al. (1993), spatial attention is deployed as gain modulation at various levels of the visual processing hierarchy. Modulation is controlled such that visual information is selectively "re-routed" from the initial layer to the final object-processing area, implemented as an associative shape memory (Hopfield, 1984). In combination with the Hopfield network (the $\mathbb{A}$ layer in the UNI framework), the model by Olshausen and colleagues is capable of detecting objects invariant to translation and scale.

The idea of dynamic re-routing of visual information was also used by Heinke and Humphreys (1997) in their SAIM model (selective attention for identification model). Instead of an associative Hopfield network, SAIM uses a "content layer" for matching the visual information in the focus of attention (FOA) with stored object templates. The model is able to predict a range of experimental results such as reaction times in detection tasks and behavior of the system when lesioned in a systematic way (Heinke and Humphreys, 2003).

The "Selective Tuning" (ST) model of visual attention by Tsotsos et al. (1995) tightly integrates visual attention and object detection (Rothenstein and Tsotsos, 2006). In the first feed-forward pass through this hierarchical system, features of increasing complexity compete locally in WTA networks. After top-down selection of a particular feature or feature combination, competition is biased in a feedback pass such that the stimulus with the selected property is enhanced, and the activity around it is suppressed (inhibitive surround). Once spatially isolated in this manner, the stimulus is processed in another feed-forward pass for ultimate detection or identification. In the UNI framework, this is equivalent to selectively tuning the spatial modulation functions $s_{\mathrm{xx}}$ for particular maps. The ST model has been demonstrated successfully for motion-defined shapes (Tsotsos et al., 2005).

Rybak et al. (1998) proposed a model for learning and recognizing objects as combinations of their parts, with the relations between the parts encoded in saccade vectors. Their attention window coincides with $s_{\mathrm{im}}$ in our UNI framework, their primary feature detection and invariant transformation modules to layers **SF** and **CF**, and their "what" structure to **OB**. All processing steps in the model of Rybak and colleagues contribute to the formation of the next

saccade, i.e. the formation of the next spatial modulation function:

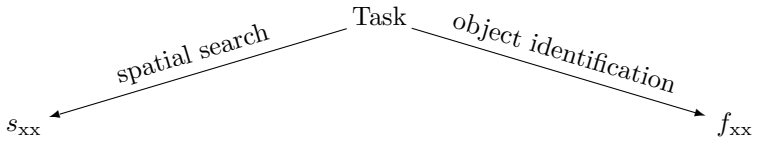



(14)

The model of Rybak and colleagues can memorize and recognize objects and faces in gray-level photographs.

In a different approach to deploying attention, Deco and Schürmann (2000) suggested a model for using spatial attention to selectively enhance resolution for object processing in order to iteratively test hypotheses about the object identity. In the UNI framework, this would amount to actually modifying the feature recombination functions $\phi_{\mathrm{xx}}$ based on the task.

In their physiologically detailed model of visual learning and recognition, Deco and Rolls (2004) implemented attention as biased competition between spatial locations (Duncan and Humphreys, 1989). Depending on the task (spatial search versus object identification), spatial or feature modulation functions are adjusted throughout the hierarchy of visual processing layers in order to bias competition toward the target:




(15)

## 4.3 Feature-based Attention

Many of these models of visual attention contain provisions for biasing particular features from the top down based on a task or intention. In the UNI framework, two basic approaches for feature-based attention are possible. First, features can be biased in the recognition hierarchy by using the mechanism of the feature modulation functions $f_{\mathrm{xx}}$ in eq. 10. In this manner, red targets, for instance, could be processed preferentially throughout the visual field. Experimental evidence for this kind of biasing was found for color and luminance in macaque area V4 (Motter, 1994), for motion in macaque MT (Treue and Martinez Trujillo, 1999), and for color and motion in human V4 and MT (Saenz et al., 2002).

This approach is followed by the selective tuning model by Tsotsos and colleagues when running the model in top-down search mode. A particular

property chosen at the top of the hierarchy is selectively enhanced throughout the hierarchy to find an object with this property (Tsotsos et al., 1995; Tsotsos et al., 2005; Rothenstein and Tsotsos, 2006).

The other possibility for feature-based attention is to bias the computation of the saliency map for particular features or combinations of features. This can be achieved with the feature modulation function $f_{\text{fm}}$ in eq. 12. Feature-based attention of this kind is deployed by way of the spatial modulation functions that are derived from the saliency map (eqs. 12 and 13). This mode of attention is the essence of Guided Search, a model proposed by Wolfe and colleagues to explain human behavior in visual search for targets that are defined by individual features or feature conjunctions (Wolfe et al., 1989; Wolfe, 1994; Cave, 1999).

Most models of attention and object recognition follow the second approach. Navalpakkam and Itti (2005), for instance, modeled the effects of task on visual attention by adjusting the weights for the combination of feature maps into the saliency map. The combinations of weights for particular targets were learned from training images. In the UNI framework, this corresponds to learning the feature modulation function $f_{\text{fm}}$ in eq. 13 from the **SF** map bundle:

$$\mathbf{SF} \longrightarrow f_{\text{fm}} \tag{16}$$

Schill et al. (2001) proposed a method for learning features that maximize the gain of information in each saccade in a belief propagation network using orientations only. Their system is tested on 24000 artificially created scenes which can be classified with a 80 % hit rate.

In his model of dynamic interactions between prefrontal areas, IT and V4, Hamker (2004) proposed a method for setting feature biases in order to guide spatial attention to target locations. This biologically detailed model was fitted to reproduce the neurophysiological data by Chelazzi et al. (1998) for a cued target selection task in rhesus monkeys (Hamker, 2003). Additionally, the model is capable of detecting objects in natural scenes (Hamker, 2005).

No matter how feature-based attention is deployed, there is always the question of how to choose the optimal modulation function for a particular task. Navalpakkam and Itti (2007) showed an elegant way of choosing the weights for a linear modulation function by maximizing the signal-to-noise ratio between search targets and distracters. Counterintuitively, it is sometimes optimal to enhance the activity of neurons tuned to an exaggerated property of the target instead of the perfectly tuned neuron.

## 4.4 Object-based Attention

Experimental evidence suggests that attention can be tied to objects, object parts, or groups of objects (Duncan, 1984; Egly et al., 1994; Roelfsema et al., 1998). In particular, Egly et al. (1994) reported that attention spreads over objects defined by luminance contrast. In their study, an invalid spatial cue for the location of a briefly flashed target is still effective when cue and target are

located on the same object, but not when they are on different objects. The effect has been replicated for objects defined by color (Reynolds et al., 2003; Mitchell et al., 2003) and illusory contours (Moore et al., 1998).

We have modeled this effect as spreading of attention over a contiguous region of high activity in the feature map that contributes most to the saliency of the attended location, thus obtaining an estimate for the size and shape of the attended objects (see eq. 13).

Rolls and Deco (2006) model object-based attention by shrinking the size of the receptive field of model IT neurons to match the size of the attended object. This idea of dynamically adjusting the size of the attended region to the attended object like a zoom lens was pioneered by Eriksen and St. James (1986). In a similar spirit, the shape of both the enhanced center-region and the suppressive surround in the selective tuning model adapt to the shape of the attended object (Tsotsos et al., 1995).

Objects can also be attended to when they are not clearly separated. The model by Lee and Lee (2000) is able to learn optimal biases for top-down attention using back-propagation in a multilayer perceptron network. Their system can segment superimposed handwritten digits on the pixel level. Treating attention as a by-product of a recognition model based on Kalman filtering, the system by Rao (1998) can attend to spatially overlapping (occluded) objects on a pixel basis as well.

Object-based attention does not need to be spatial, however. O'Craven et al. (1999) showed that human subjects could attend selectively to faces and houses when both stimuli were semi-transparently superimposed. In these fMRI experiments the attended object could also be defined by its coherent motion. In the UNI framework, these results can be interpreted as feature-based attention by adjusting $f_{\mathrm{sf}}$ for a particular direction of motion or $f_{\mathrm{ob}}$ for a particular object category in eq. 10.

If it is possible to effectively deploy top-down attention for particular objects as feature biases for the relatively unspecific simple features, then using complex, more object-specific features should make selection even easier. Object recognition as described in subsection 3.2 provides us with a set of complex features and with a mapping from these features to the object layer and finally to the abstract object space:

$$\mathbf{CF} \xrightarrow{\phi_{\mathrm{ob}}} \mathbf{OB} \longrightarrow \mathbb{A} \tag{17}$$

We propose a mode of object-specific attention that uses *feedback connections* from abstract object representations $\mathbb{A}$ via the object layer **OB** back to the complex features **CF** in order to infer suitable complex feature maps for the localization of a given object or object category. Both the complex features **CF** and the feature recombination function $\phi_{\mathrm{ob}}$ are acquired when the system is trained for object detection. We suggest that these same representations can be used for top-down attention as well. As a proof of concept for this idea, we show in our simulations in the next section that it is possible to share complex features between object detection and attention.

# 5 Sharing Features between Object Detection and Top-down Attention

In the hierarchy of processing steps for object recognition in eq. 10, the structure of objects is encoded in the feature recombination functions $\phi_{\mathrm{xx}}$. While the structure of these functions is fairly generic at early stages (e.g., Gabor filters for $\phi_{\mathrm{sf}}$), at later stages these functions are more complex and more specific for particular object categories. Some models of object recognition use hard-wired features at these higher levels (e.g., Riesenhuber and Poggio, 1999; Amit and Mascaro, 2003), others learn these features from their visual input (e.g., LeCun et al., 1998; Serre et al., 2005).

As mentioned in subsection 2.2, Serre et al. (2005) extended the HMAX model of Riesenhuber and Poggio (1999) by learning the structure of complex features (i.e., the details of $\phi_{\mathrm{cf}}$ in eq. 10) from large numbers of natural images. Once learned, these complex features are fixed and used as prototypes for feature recombination functions that resemble radial basis functions. With these functions, the model can be trained to categorize objects in photographs with high accuracy (Serre et al., 2007b).

Here we suggest a method of sharing these same complex feature representations between object detection and top-down attention. We propose that by cortical feedback connections, top-down processes can re-use these same features to bias attention to locations with a higher probability of containing the target object. We compare the performance of a computational implementation of such a model with pure bottom-up attention and, as a benchmark, with biasing for skin hue, which is known to work well as a top-down bias for faces.

## 5.1 Methods

The basic architecture of our model is shown in figure 1. Proto-types for the S2 features are randomly initiated from a set of training images. For the work presented in this chapter, we trained the model on detecting frontal views of human faces in photographs and investigated the suitability of the corresponding S2 features for top-down attention to faces.

For feature learning and training, we used 200 color images, each containing one face among clutter, and 200 distracter images without faces (see figure 2 for examples). For testing the recognition performance of the system, we used a separate test set of 201 face images and 2119 non-face distracter images. To evaluate top-down attention, we used a third set of 179 color images containing between 2 and 20 frontal views of faces (figure 2, third and fourth row). All images were obtained from the world wide web, and face images were labeled by hand, with the eyes, nose and mouth of each face marked.[2]

[Figure 2 about here.]

[2]We would like to thank Xinpeng Huang and Thomas Serre for collecting and labeling the images. The image database is available online at `http://web.mit.edu/serre/www/Resources.htm`.

During feature learning, 100 patches of size $6 \times 6$ were extracted from the C1 maps for each presentation of a training image. Over five iterations of presenting the 200 training images in random order, 100 stable features were learned. Two separate sets of features were learned in this manner: set A was derived from patches that were extracted from any location in the training images (figure 2, top row); patch selection for set B was limited to regions around faces (figure 2, second row).

To evaluate feature sets A and B for top-down attention, the S2 maps were computed for the third set of 179 images containing multiple faces. These top-down feature maps were compared to the bottom-up saliency map of Itti et al. (1998) and to a skin hue detector for each of the images.

Skin hue is known to be an excellent indicator for the presence of faces in color images (Darrel et al., 2000). Here we use it as a benchmark. Since we want our model of skin hue to be independent of light intensity, we model it in intensity-normalized $(r', g')$ color space. If $(r, g, b)$ are the RGB values of a given color pixel, then we compute our $(r', g')$ color coordinates as

$$r' = \frac{r}{r+g+b} \text{ and } g' = \frac{g}{r+g+b}. \tag{18}$$

Note that it is not necessary to have a separate value for blue, because the blue content of the pixel can be inferred from $r'$ and $g'$ at any given light intensity $(r + g + b)$.

For the description of skin hue in this color space, we use a simple Gaussian model with mean $(\mu_r, \mu_g)$, standard deviations $(\sigma_r, \sigma_g)$, and correlation coefficient $\rho$. For a given color pixel with coordinates $(r', g')$, the model's hue response is given by

$$h(r', g') = \exp\left[-\frac{1}{2}\left(\frac{(r'-\mu_r)^2}{\sigma_r^2} + \frac{(g'-\mu_g)^2}{\sigma_g^2} - \frac{\rho(r'-\mu_r)(g'-\mu_g)}{\sigma_r \sigma_g}\right)\right]. \tag{19}$$

To estimate the parameters of the skin hue distribution, we used 1153 color photographs containing a total of 3947 faces from the world wide web[3] and fitted the hue distribution of the faces. The resulting parameters are shown in table 2. The images used for estimating the skin hue model are separate from the sets of images used elsewhere in this section.

[Table 2 about here.]

The images depict humans of many different ages and ethnicities, both female and male. There is a slight bias toward caucasian males, reflecting a general bias of images of humans in the world wide web. We observed that the skin *hue* does not vary much between different ethnicities, while brightness of the skin shows much more variations. Figure 3 shows the hue of the training faces and the fitted distribution.

[Figure 3 about here.]

[3]We would like to thank Pietro Perona for providing the images.

## 5.2 Results

After feature learning, the recognition model was trained to detect faces using the training images. Recognition performance on the test images was 98.9 % with feature set A and 99.4 % with set B (measured as area under the ROC curve).

For the purpose of testing the suitability of the features for top-down attention we use an analysis of fixations on faces based on the respective activation maps. The S2 feature maps for both feature sets, the bottom-up saliency map, and the skin hue bias map were computed for the 179 multiple-face images. Each map was treated like a saliency map, and the locations in the map were visited in order of decreasing saliency, neglecting spatial relations between the locations. While this procedure falls short of the full simulation of a winner-take-all network with inhibition of return as described in Koch and Ullman (1985), it nevertheless provides a simple and consistent means of scanning the maps.

For each map we determined the number of "fixations" required to find a face and, once the focus of attention leaves the most salient face, how many fixations are required to attend to each subsequent face. The fraction of all faces that required one, two, three, or more than three fixations was determined for each feature and used as a measure for the quality of the respective S2 feature map.

[Figure 4 about here.]

The results are shown in figure 4 for the best features from sets A and B, for bottom-up attention, and for skin hue detection. Feature set B shows slightly higher performance than our benchmark skin hue detection, followed by feature set A and bottom-up attention. Results are significantly better when feature selection is restricted to faces (set B) compared to unrestricted feature selection (set A).

Top-down attention based on S2 features by far outperform bottom-up attention in our experiments. While bottom-up attention is well suited to identify salient regions in the absence of a specific task, it cannot be expected to localize a specific object category as well as feature detectors that are specialized for this category.

## 5.3 Discussion

We show that features learned for recognizing a particular object category may also serve for top-down attention to that object category. Object detection can be understood as a mapping from a set of features to an abstract object representation. When a task implies the importance of an object, the respective abstract object representation may be invoked, and feedback connections may reverse the mapping, allowing inference as to which features are useful to guide top-down attention to image locations that have a high probability of containing the target object.

The important question of how to combine several S2 feature maps optimally for the search for a specific target remains unanswered in this section. It

is possible that feature combination strategies like the one that Navalpakkam and Itti (2007) applied to simple features are also viable for our more complex features. For now, however, this remains an open issue.

Note that this mode of top-down attention does not necessarily imply that the search for any object category can be done in parallel using an explicit map representation. Search for faces, for instance, has been found to be efficient (Hershler and Hochstein, 2005), although this result is disputed (VanRullen, 2006). We have merely shown a method for identifying features that can be used to search for an object category. The efficiency of the search will depend on the complexity of those features and, in particular, on the frequency of the same features for other object categories, which constitute the set of distracters in visual search.

To analyze this aspect further, it would be of interest to explore the overlap in the sets of features that are useful for multiple object categories. Torralba et al. (2004) have addressed this problem for multiple object categories as well as multiple views of objects in a machine vision context.

# 6 Conclusions

We have reviewed a number of models of object recognition and visual attention, showing that many seemingly disparate ideas can in fact be captured by a common formal framework. In this unifying framework for attention and object recognition we have explained recognition by components as well as view-based object recognition; we have covered many aspects of spatial, feature-based, and object-based attention, as well as the interactions between attention and object recognition. Furthermore, we have shown in a particular instantiation how complex features learned for the purpose of object detection can be shared with top-down attention.

In our review we have focused on the ideas for attention and object recognition that fit within our hierarchical framework. However, there are other ideas that are not captured by the UNI framework, either because they describe the neurobiological processes at a more detailed level than is provided by the UNI framework, or because they make use of information from sources that fall outside the framework, such as visual context.

Grossberg and Raizada (2000) and Raizada and Grossberg (2001), for instance, proposed a model of attention and visual grouping based on biologically realistic models of neurons and neural networks. Their model relies on grouping edges within a laminar cortical structure by synchronous firing, allowing it to extract real as well as illusory contours.

Recognizing individual objects is only part of visual perception. Objects are typically embedded in context with other objects or with the general layout of a scene (the "gist"). Interactions between object recognition and scene perception go both ways: gist provides a powerful top-down cue, restricting the possibilities for object identity; objects contribute to the general context and interpretation of a scene. Attention serves as a means of deploying this top-down information.

Some aspects of this interaction were modeled by Oliva et al. (2003) in a probabilistic framework, incorporate context information into the spatial probability function for seeing certain objects (e.g., people) at particular locations. Comparison with human eye tracking results show improvement over purely bottom-up saliency-based attention.

Work on scene perception has been progressing at a rapid pace over the last few years (see, for instance, Bar, 2004; Oliva and Torralba, 2006; Fei-Fei et al., 2007), and integrating scene and object information into a general framework is an interesting challenge for years to come. Exploring and modeling the interactions between scene perception, object recognition, and visual attention will bring us closer to understanding the rich and varied experience afforded to us by visual perception.

## Acknowledgments

Thomas Serre and Tomaso Poggio collaborated on parts of the work described in section 5. We would like to thank Karen F. Bernhardt-Walther, Kerstin Preuschoff, and Daniel Simons for helpful discussions and feedback on versions of the manuscript. This work was funded by DARPA, the NSF, the NIH, the NIMH, the ONR, the Keck Foundation, and a Beckman Postdoctoral Fellowship to D.B.W.

# References

Amit Y, Mascaro M (2003) An integrated network for invariant visual detection and recognition. *Vision Research* 43:2073–2088.

Bar M (2004) Visual objects in context. *Nature Review Neuroscience* 5:617–629.

Biederman I (1987) Recognition-by-components: a theory of human image understanding. *Psychological Review* 94:115–147.

Biederman I, Cooper EE (1991) Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cognitive Psychology* 23:393–419.

Bülthoff HH, Edelman SY, Tarr MJ (1995) How are three-dimensional objects represented in the brain? *Cerebral Cortex* 5:247–260.

Carmi R, Itti L (2006) Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research* 46:4333–45.

Cave KR (1999) The FeatureGate model of visual selection. *Psychological Research* 62:182–194.

Chelazzi L, Duncan J, Miller EK, Desimone R (1998) Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology* 80:2918–2940.

Chelazzi L, Miller EK, Duncan J, Desimone R (2001) Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex* 11:761–72.

Connor CE, Preddie DC, Gallant JL, van Essen DC (1997) Spatial attention effects in macaque area V4. *Journal of Neuroscience* 17:3201–3214.

Darrel T, Gordon G, Harville M, Woodfill J (2000) Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision* 37:175–185.

Deco G, Rolls ET (2004) A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research* 44:621–642.

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

Desimone R, Duncan J (1995) Neural mechanisms of selective visual-attention. *Annual Review of Neuroscience* 18:193–222.

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

Duncan J, Humphreys GW (1989) Visual search and stimulus similarity. *Psychological Review* 96:433–458.

Edelman S (1997) Computational theories of object recognition. *Trends in Cognitive Sciences* 1:296–304.

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

Eriksen CW, St. James JD (1986) Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics* 40:225–240.

Fei-Fei L, Iyer A, Koch C, Perona P (2007) What do we perceive in a glance of a real-world scene? *Journal of Vision* 7(1):1–29.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience* 23:5235–5246.

Frintrop S, Rome E, Nüchter A, Surmann H (2005) A bimodal laser-based attention system. *Computer Vision and Image Understanding* 100:124–151.

Fukushima K (1980) Neocognitron: A self-organizing neural network model for a machanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36:193–202.

Gauthier I, Tarr MJ (1997) Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vision Research* 37:1673–1682.

Grossberg S, Raizada RD (2000) Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research* 40:1413–1432.

Hamker FH (2003) The reentry hypothesis: linking eye movements to visual perception. *Journal of Vision* 3:808–816.

Hamker FH (2004) A dynamic model of how feature cues guide spatial attention. *Vision Research* 44:501–521.

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

Hayworth KJ, Biederman I (2006) Neural evidence for intermediate representations in object recognition. *Vision Research* 46:4024–4031.

Heinke D, Humphreys GW (1997) SAIM: a model of visual attention and neglect. In *7th international conference on artificial neural networks - ICANN 97*, pp. 913–918, Lausanne, Switzerland. Springer Verlag.

Heinke D, Humphreys GW (2003) Attention, spatial representation, and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (SAIM). *Psychological Review* 110:29–87.

Hershler O, Hochstein S (2005) At first sight: a high-level pop out effect for faces. *Vision Research* 45:1707–1724.

Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science of the USA* 81:3088–3092.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160:106–154.

Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. *Psychological Review* 99:480–517.

Itti L (2005) Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition* 12:1093–1123.

Itti L, Koch C (2001) Computational modelling of visual attention. *Nature Reviews Neuroscience* 2:194–203.

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* 20:1254–1259.

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

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* pp. 2278–2324.

Lee SI, Lee SY (2000) Top-down attention control at feature space for robust pattern recognition In *Biologically-Motivated Computer Vision*, Seoul, Korea.

Logothetis NK, Pauls J, Bülthoff HH, Poggio T (1994) View-dependent object recognition by monkeys. *Current Biology* 4:401–414.

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60:91–110.

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

Marr D (1982) *Vision* W. H. Freeman and Company, New York.

Marr D, Nishihara HK (1978) Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences* 200:269–294.

McAdams CJ, Maunsell JHR (1999) Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience* 19:431–441.

McAdams CJ, Maunsell JHR (2000) Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology* 83:1751–1755.

Mel BW (1997) SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation* 9:777–804.

Miau F, Papageorgiou C, Itti L (2001) Neuromorphic algorithms for computer vision and attention In *SPIE 46 Annual International Symposium on Optical Science and Technology*, Vol. 4479, pp. 12–23.

Milanese R, Wechsler H, Gill S, Bost JM, Pun T (1994) Integration of bottom-up and top-down cues for visual attention using non-linear relaxation In *International Conference on Computer Vision and Pattern Recognition*, pp. 781–785.

Mitchell JF, Stoner GR, Fallah M, Reynolds JH (2003) Attentional selection of superimposed surfaces cannot be explained by modulation of the gain of color channels. *Vision Research* 43:1323–1328.

Moore CM, Yantis S, Vaughan B (1998) Object-based visual selection: evidence from perceptual completion. *Psychological Science* 9:104–110.

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

Mozer MC (1991) *The perception of multiple objects: a connectionist approach* MIT Press, Cambridge, MA.

Mozer MC, Sitton M (1998) Computational modeling of spatial attention In Pashler H, editor, *Attention*, pp. 341–393. Psychology Press, London.

Mutch J, Lowe DG (2006) Multiclass object recognition with sparse, localized features In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 11–18, New York, NY.

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

Navalpakkam V, Itti L (2007) Search goal tunes visual features optimally. *Neuron* 53:605–17.

O'Craven KM, Downing PE, Kanwisher N (1999) fMRI evidence for objects as the units of attentional selection. *Nature* 401:584–587.

Oliva A, Torralba A (2006) Building the gist of a scene: The role of global image features in recognition In *Progress in Brain Research: Visual perception*, Vol. 155, pp. 23–36.

Oliva A, Torralba A, Castelhano M, Henderson J (2003) Top-down control of visual attention in object detection In *International Conference on Image Processing*.

Olshausen BA, Anderson CH, Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience* 13:4700–19.

Peters RJ, Iyer A, Itti L, Koch C (2005) Components of bottom-up gaze allocation in natural images. *Vision Research* 45:2397–2416.

Plato Book vii In *The Republic*, p. 195. Basic Books, New York, NY, 1968.

Poggio T, Edelman S (1990) A network that learns to recognize three-dimensional objects. *Nature* 343:263–266.

Posner MI (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology* 32:3–25.

Raizada RD, Grossberg S (2001) Context-sensitive bindings by the laminar circuits of V1 and V2: A unified model of perceptual grouping, attention, and orientation contrast. *Visual Cognition* 8:431–466.

Rao RPN (1998) Visual attention during recognition In *Advances in Neural Information Processing*.

Rees G, Frackowiak R, Frith C (1997) Two modulatory effects of attention that mediate object categorization in human cortex. *Science* 275:835–838.

Rensink RA (2000a) The dynamic representation of scenes. *Visual Cognition* 7:17–42.

Rensink RA (2000b) Seeing, sensing, and scrutinizing. *Vision Research* 40:1469–1487.

Reynolds JH, Alborzian S, Stoner GR (2003) Exogenously cued attention triggers competitive selection of surfaces. *Vision Research* 43:59–66.

Reynolds JH, Pasternak T, Desimone R (2000) Attention increases sensitivity of V4 neurons. *Neuron* 26:703–714.

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

Roelfsema PR, Lamme VAF, Spekreijse H (1998) Object-based attention in the primary visual cortex of the macaque monkey. *Nature* 395:376–381.

Rolls ET, Deco G (2006) Attention in natural scenes: Neurophysiological and computational bases. *Neural Networks* 19:1383–1394.

Rothenstein AL, Tsotsos JK (2006) Attention links sensing to recognition. *Image and Vision Computing* (in press).

Rutishauser U, Walther D, Koch C, Perona P (2004) Is attention useful for object recognition? In *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 37–44, Washington, DC.

Rybak IA, Gusakova VI, Golovan AV, Podladchikova LN, Shevtsova NA (1998) A model of attention-guided visual perception and recognition. *Vision Research* 38:2387–2400.

Saenz M, Buracas GT, Boynton GM (2002) Global effects of feature-based attention in human visual cortex. *Nature Neuroscience* 5:631–632.

Schill K, Umkehrer E, Beinlich S, Krieger G, Zetzsche C (2001) Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. *Journal of Electronic Imaging* 10:152–160.

Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T (2005) A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex, CBCL Paper #259/AI Memo #2005-036 Technical report, Massachusetts Institute of Technology.

Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T (2007a) A quantitative theory of immediate visual recognition In *(FIXME: this volume)*, p. xxx.

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007b) Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29:411–426.

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, San Diego, CA.

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

Tarr MJ, Bülthoff HH (1995) Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance* 21:1494–1505.

Torralba A, Murphy K, Freeman W (2004) Sharing features: efficient boosting procedures for multiclass object detection In *IEEE International Conference on Computer Vision and Pattern Recognition*, Washington, DC.

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

Treue S, Martinez Trujillo JC (1999) Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399:575–9.

Tsotsos JK, Culhane SM, Wai WYK, Lai YH, Davis N, Nuflo F (1995) Modeling visual-attention via selective tuning. *Artificial Intelligence* 78:507–545.

Tsotsos JK, Liu Y, Martinez-Trujillo JC, Pomplum M, Simine E, Zhou K (2005) Attending to motion. *Computer Vision and Image Understanding* 100:3–40.

Ullman S (1979) *The Interpretation of Visual Motion* MIT Press, Cambridge, Massachusetts.

Ullman S (2007) Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences* 11:58–64.

Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. *Nature Neuroscience* 5:682–687.

VanRullen R (2006) On second glance: Still no high-level pop-out effect for faces. *Vision Research* 46:3017–3027.

Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual system. *Progress in Neurobiology* 51:167–94.

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. Springer, Berlin, Germany.

Walther D, Koch C (2006) Modeling attention to salient proto-objects. *Neural Networks* 19:1395–1407.

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

Walther D, Serre T, Poggio T, Koch C (2005b) Modeling feature sharing between object detection and top-down attention [abstract]. *Journal of Vision* 5:1041a.

Wolfe JM (1994) Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review* 1:202–238.

Wolfe JM, Cave KR, Franzel SL (1989) Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance* 15:419–433.

# List of Figures

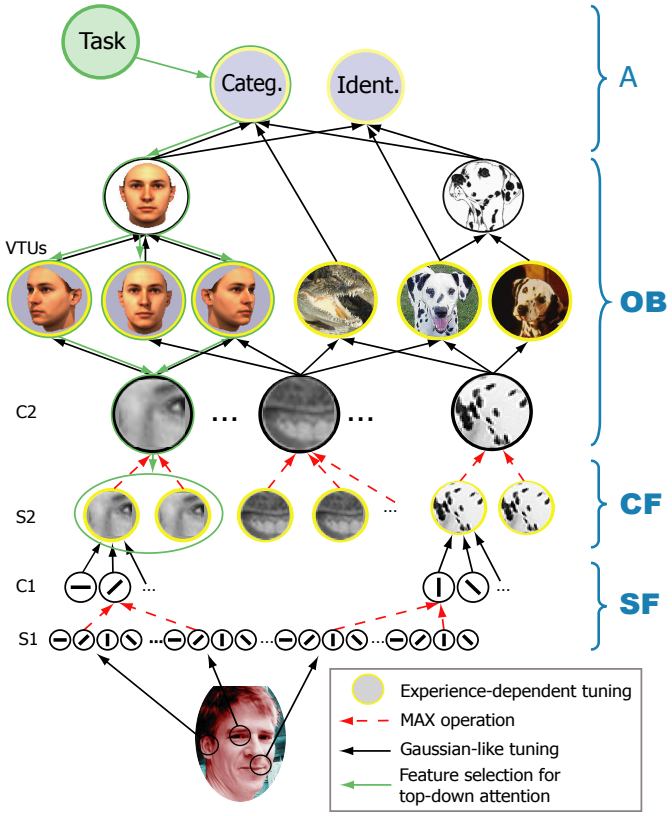



Figure 1: The basic architecture of the model of object recognition and top-down attention in section 5 (adapted from Walther et al., 2005b and Serre et al., 2005). In the feed-forward pass, orientation-selective S1 units filter the input image, followed by max-like pooling over space and scale in the C1 units. The S1 and C1 layers correspond to the "simple features map bundle" **SF** in eq. 10. S2 units (complex features **CF** in eq. 10) respond according to the Gaussian distance of each image patch from a set of prototypes. After another step of max-pooling over space and scale, C2 units (**OB** layer in eq. 10) respond to the presence of particular features anywhere in the visual field. View-tuned units (VTUs) and, finally, units selective to individual objects or object categories are acquired from labeled training data (abstract object space $\mathbb{A}$ in eq. 10). By association with a particular object or object category, activity due to a given task could traverse down the hierarchy (green) to identify a small subset of features at the S2 level that are indicative for the particular object category.


Figure 2: Examples for training stimuli for feature set A (top row), feature set B (second row), test stimuli with two or more faces (third and fourth row), and for non-face distracters (bottom row). See subsection 5.1 for details.

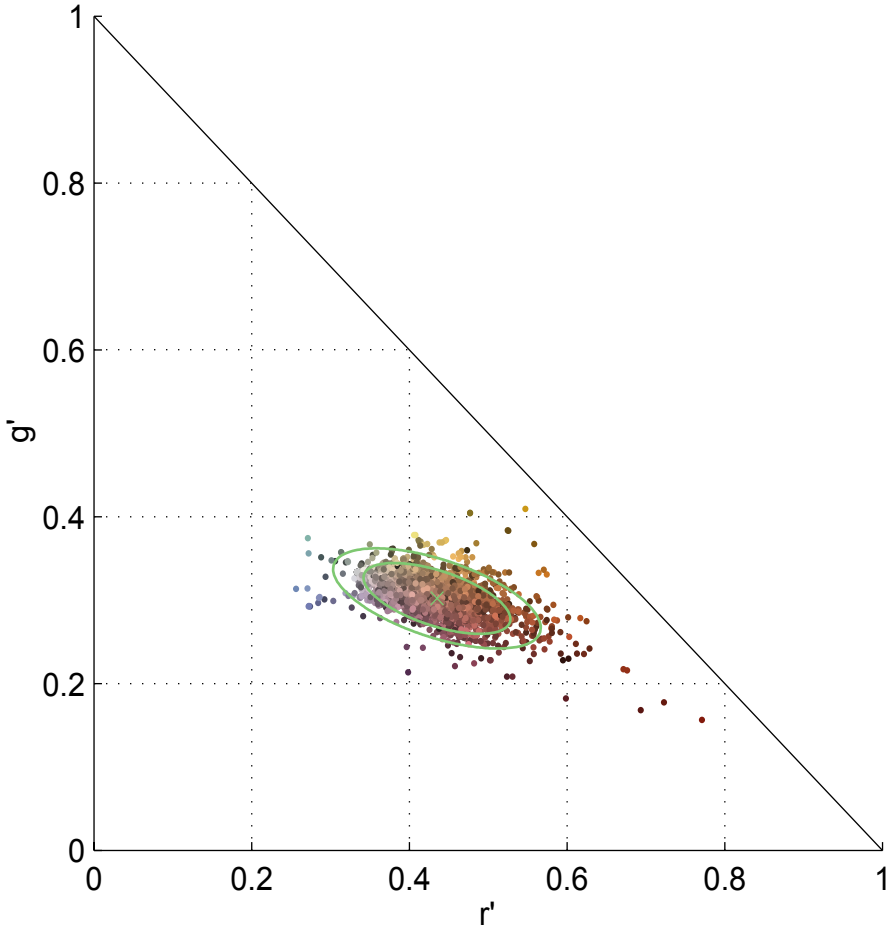



Figure 3: The Gaussian model for skin hue. The individual training points are derived from 3974 faces in 1153 color photographs. Each dot represents the average hue for one face and is plotted in the color of the face. The green cross represents the mean $(\mu_r, \mu_g)$, and the green ellipses the $1\sigma$ and $2\sigma$ intervals of the hue distribution.

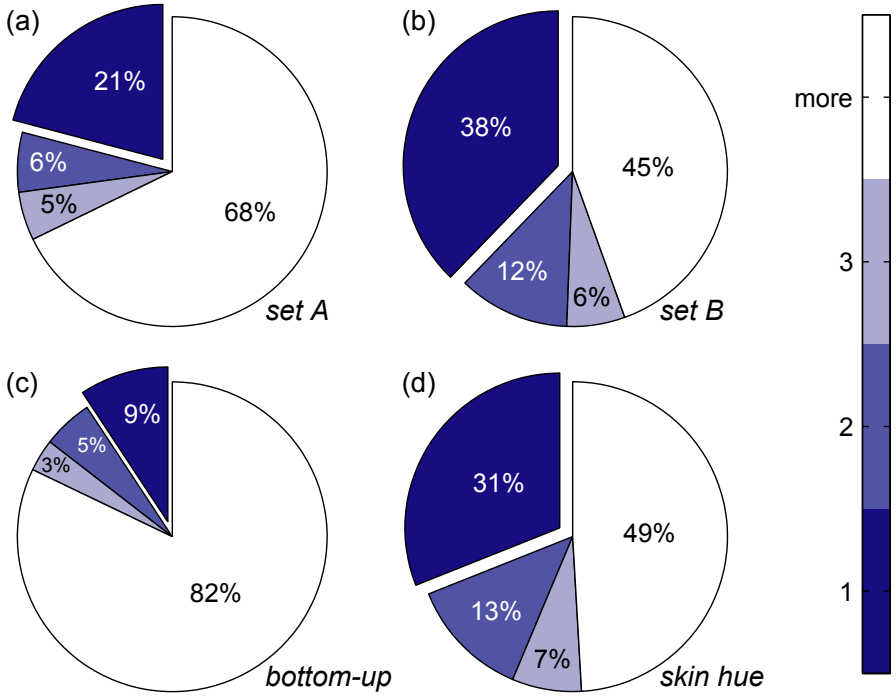



Figure 4: Fractions of faces in test images requiring one, two, three, or more than three fixations to be found when using top-down feature sets A or B, bottom-up attention, or biasing for skin hue.

# List of Tables

Table 1: Comparison of recognition by components and view-based object recognition.

| | **Recognition by Components** | **View-based recognition** |
|---|---|---|
| *early features* | orientations | orientations |
| *intermediate features* | 3d geometric shapes (geons) | 2d patterns of orientations |
| *object representation* | 3d assemblies of geons | multiple 2d views |
| *spatial relations* | explicit representation | implicit representation, given by the design of increasingly complex features |
| *translation and scale invariance* | given by explicit representation of spatial relations | pooling over space and scale bands |
| *rotation invariance* | mental rotation of 3d objects | established by interpolation between views |
| *key papers* | Marr and Nishihara (1978)<br>Biederman (1987)<br>Biederman and Cooper (1991)<br>Hummel and Biederman (1992)<br>Hayworth and Biederman (2006) | Ullman (1979)<br>Fukushima (1980)<br>Bülthoff et al. (1995)<br>Logothetis et al. (1994)<br>Riesenhuber and Poggio (1999)<br>Serre et al. (2007b) |

Table 2: Parameters of the distribution of skin hue in intensity-normalized (r',g') color space.

| *Parameter* | *Value* |
|---|---|
| $\mu_r$ | 0.434904 |
| $\mu_g$ | 0.301983 |
| $\sigma_r$ | 0.053375 |
| $\sigma_g$ | 0.024349 |
| $\rho$ | 0.5852 |