To appear in: *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos (editors), Academic Press, 2004.

Probabilistic Models of Attention based on Iconic Representations and Predictive Coding

Rajesh P. N. Rao

Department of Computer Science and Engineering
University of Washington
Seattle WA 98195

Dana H. Ballard

Department of Computer Science
University of Rochester
Rochester, NY 14627

Abstract

We describe two models of attention that utilize probabilistic principles to compute task-relevant variables. In the first model, objects and visual scenes are represented iconically using spatial filters at multiple scales. A maximum likelihood-based approach is used to compute the location of a target in a given scene. The eye movements generated by such a strategy are shown to be similar to human eye movement patterns elicited during visual search in naturalistic scenes. The second model is based on the statistical concept of predictive coding. It assumes that top-down feedback from higher cortical areas conveys predictions of expected activity at lower levels while the errors in prediction are conveyed through feedforward connections. The model explains how multiple objects in a scene can be recognized sequentially without an explicit spotlight of attention. An extension of the model provides an interpretation of object-based versus spatial attention in terms of interactions between "what" and "where" networks in the visual pathway.

Key words: Visual Search, Spatial Filters, Eye Movements, Saliency Maps, Generative Models, Robust Kalman Filters, Object-Based Attention, Spatial Attention

Email addresses: rao@cs.washington.edu (Rajesh P. N. Rao), dana@cs.rochester.edu (Dana H. Ballard).

1 Introduction

Animals receive a vast amount of sensory information in their interactions with the natural world. The brain's limited processing resources permits only a fraction of this information to be processed at any given moment in time. Furthermore, this information is typically noisy and the animal's knowledge of its world is almost always incomplete. The fundamental challenge in such an environment is to be able to select and process only those portions of the sensory inputs that are relevant to the particular task at hand and to the animal's continued survival. Attention is nature's answer to this challenge.

Attention is often classified as being either overt or covert. Overt visual attention typically involves making eye movements to shift gaze to "interesting" or task-relevant parts of a scene. Covert attention, on the other hand, involves the ability to preferentially process an object or location in a visual scene without shifting gaze. A useful metaphor for understanding attention has been the notion of a *spotlight* or "search light" that can be focused on specific portions of a visual scene (see (Desimone and Duncan, 1995; Newsome, 1996) for reviews). Numerous models have been proposed for simulating an attentional spotlight, two prominent examples being saliency maps and hierarchical routing circuits (Hinton, 1981; Koch and Ullman, 1985; Olshausen et al., 1993; Tsotsos et al., 1995; Niebur and Koch, 1996; Itti and Koch, 2000).

In this review article, we describe models of attention formulated at two different levels of abstraction: the first model explains overt attention during visual search in terms of saliency maps and iconic representations. The second model, which is formulated closer to the neural implementation level, provides an interpretation of covert shifts of attention without the use of an explicit spotlight. The two models share a common foundation in that both acknowledge the noisy and uncertain nature of the environment by utilizing probabilistic principles for achieving their goals.

2 Probabilistic Control of Attention using Iconic Representations

Human vision relies extensively on the ability to make saccadic eye movements to orient the high-acuity foveal region of the eye over targets of interest in a visual scene. Many studies have shown that this overt form of attention is controlled by the ongoing cognitive demands of the task at hand (see (Rao et al., 2002) for references). A key problem in most visual tasks is saccadic targeting: how are points of interest selected as targets for eye movements?

The targeting problem can be better understood within the context of a task

such as visual search in which a subject executes eye movements to find a memorized target in the current visual scene. Three important computational problems need to be solved: (a) the target object and the visual scene need to be represented using an efficient visual code, (b) the contents of the visual scene need to compared with the memorized target object to find potential matches, and (c) an eye movement needs to be executed to the location deemed most likely to contain the target object. We discuss below a model proposed in (Rao et al., 1996, 2002) that addresses these three problems using iconic representations, saliency maps, and maximum likelihood estimation respectively.

2.1 Iconic Representation of Objects

The naive method of representing objects as grey-level images is clearly impractical, given the high dimensionality of such a representation and the lack of invariance to transformations and view changes. A more efficient alternative is to encode objects iconically using a set of basis functions, or spatial filters (Jones and Malik, 1992; Lades et al., 1993; Rao and Ballard, 1995; Itti and Koch, 2000). Such a representation approximates the transformations imposed by the receptive fields of neurons in the primary visual cortex. The model proposed in (Rao and Ballard, 1995; Rao et al., 1996, 2002) utilizes a set of oriented derivatives of Gaussians (Figure 1A, top panel) (Freeman and Adelson, 1991):

$$G_i^{\theta_j}, i = 1, 2, 3, \theta_j = 0, \ldots, m\pi/(i+1), m = 1, \ldots, i \tag{1}$$

where i denotes the order of the derivative and θ_j refers to the preferred orientation of the filter. The response of an image patch I centered at (x_0, y_0) to a particular basis filter $G_i^{\theta_j}$ can be obtained by convolving the image patch with the filter:

$$r_{i,j}(x_0, y_0) = \iint G_i^{\theta_j}(x_0 - x, y_0 - y) I(x, y) dx\, dy \tag{2}$$

The iconic representation for the local image patch centered at (x_0, y_0) is formed by combining into a high-dimensional vector the responses from all basis filters above at different scales:

$$\mathbf{r}(x_0, y_0) = [r_{i,j,s}(x_0, y_0)] \tag{3}$$

where i denotes the order of the filter, j denotes the orientation, and $s = s_{min}, \ldots, s_{max}$ denotes the scale of the filter. For computational efficiency, a Gaussian pyramid representation of the image was used to generate multi-scale

responses from a set of basis filters at a fixed scale. As an example, Figure 1A shows the filter-based responses at a given location in a cluttered scene for a set of five filters at five spatial scales. It can be shown that the filter response vector at an image location provides an almost unique representation of the local image region surrounding that location when compared with response vectors from other locations or images (Rao and Ballard, 1995).

2.2 Targeting Eye Movements in Visual Search

We now summarize the model proposed in (Rao et al., 1996, 2002) for characterizing human eye movements in visual search. Suppose that objects of interest are represented by a set of memorized filter response vectors $\mathbf{r}_s^m$ where m denotes a particular target object in memory and s denotes the scale of the filters. Given a new input image and a target object T, the model computes a "saliency map" $S(x, y)$ that stores, at each image location (x, y), the similarity between the response vector for that location and the memorized target response vector $\mathbf{r}_s^T$. Furthermore, the model assumes that the computation of the saliency map proceeds in a coarse-to-fine fashion: responses from larger spatial scale filters are compared before the smaller scale responses. Finally, the most likely location of the target object is chosen probabilistically according to the Boltzmann distribution computed from the similarity values in the saliency map (see below; for those familiar with the Boltzmann distribution, the "energy function" is assumed to be given by the saliency map outputs). The entire targeting process can be summarized as follows:

(1) Set the initial scale of analysis k to the largest scale i.e. $k = max$. Set $S(x, y) = 0$ for all (x, y).

(2) Compute the current saliency map across all locations (x, y) based on filter responses from the current scale k up to the maximum scale:

$$S(x, y) = \sum_{s=k}^{max} ||\mathbf{r}_s(x, y) - \mathbf{r}_s^m||^2 \tag{4}$$

$S(x, y)$ is the square of the Euclidean distance between the filter response vector $\mathbf{r}_s$ for image location (x, y) and the memorized target response vector $\mathbf{r}_s^m$, summed over the scales $s = k, \ldots, max$.

(3) The location for the next eye movement is given by a weighted average determined from the following maximum likelihood scheme (cf. (Nowlan, 1990)):

$$(\hat{x}, \hat{y}) = \sum_{(x,y)} (x, y) \cdot \frac{e^{-S(x,y)/\lambda(k)}}{\sum_{(x,y)} e^{-S(x,y)/\lambda(k)}} \tag{5}$$

where $\lambda(k)$ is a "temperature" parameter that is decreased with k. De-

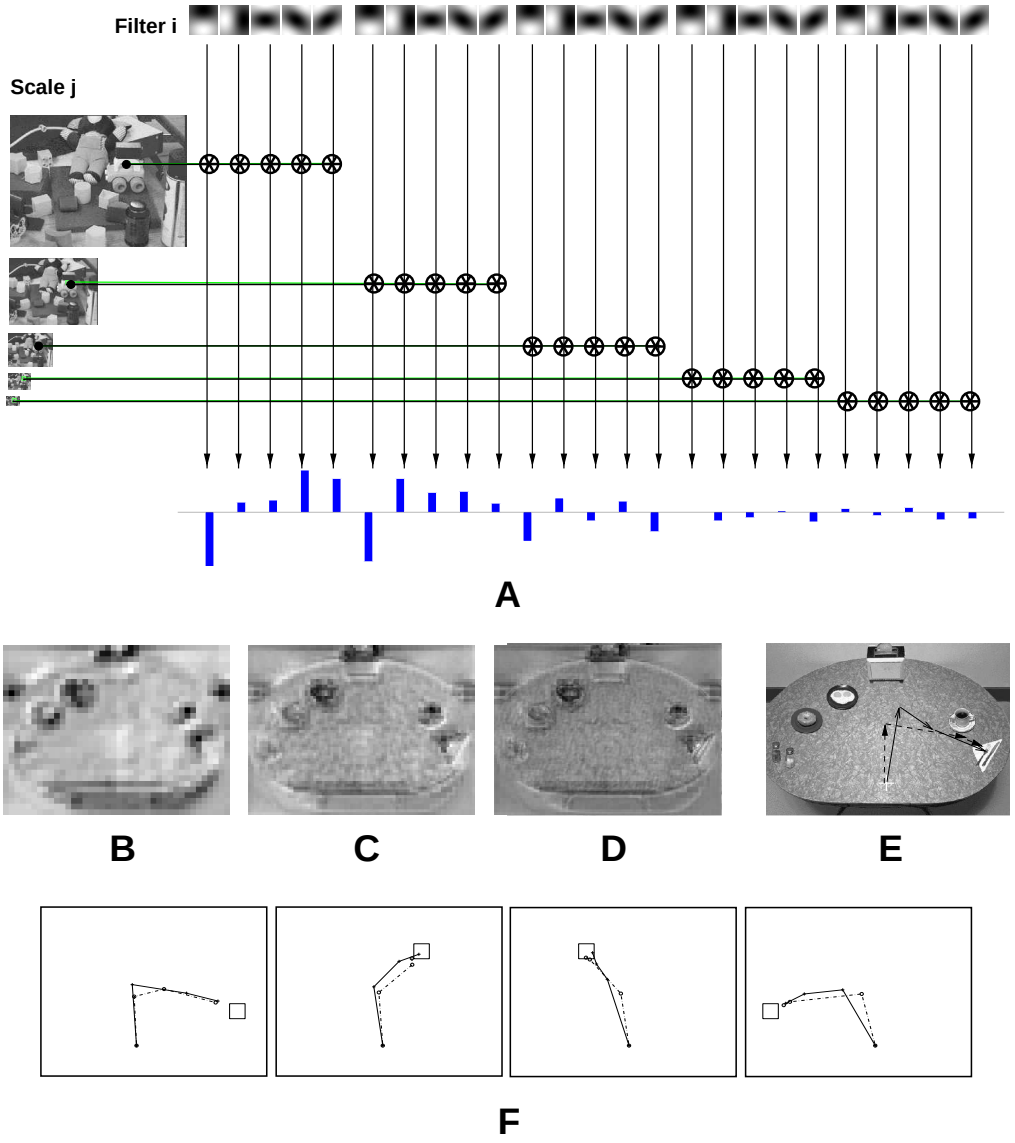

Fig. 1. **Iconic Representations and Saliency Maps for Overt Attention**. (A) Iconic representation for characterizing local image patches. A set of five oriented filters (repeated five times) is shown at the top. A cluttered image is shown on the left at five different spatial scales (Gaussian "pyramid" representation). The iconic representation for a given image location is the vector of 25 spatial filter responses obtained at that location. This representation is depicted at the bottom as a histogram. Positive responses are represented as upward bars and negative responses as downward bars. (B), (C), and (D) show the saliency map $S(x, y)$ after the inclusion of the largest, intermediate, and smallest scale filter responses respectively (only 3 scales were used in these simulations). The brightest points are the closest matches to the target object (in this case, a fork and knife on a napkin). (E) shows three successive eye movements as determined from maximum-likelihood weighted averaging of the saliency maps in (B), (C), and (D) respectively. For comparison, saccades from a human subject are depicted as dashed lines with arrows. (F) Comparison of model and human eye movements to four different target locations averaged over subjects and target objects (see text). The square box denotes a one degree region centered around each target. The dashed line segments correspond to human data while the solid line segments correspond to model data.

creasing $\lambda(k)$ allows the search to evolve from an initial state where all target locations compete equally for an eye movement to a final state where only a few most likely target locations remain.

(4) Repeat steps (2) and (3) for $k = max\text{-}1, max\text{-}2, \ldots$ until either the target object has been foveated or the number of scales has been exhausted. In the former case, a recognition process signals the termination of the search. In the latter case, successive eye movements are made using saliency maps computed from an increasing number of finer scales.

2.3 Comparison with Human Eye Movement Patterns

The model described above was tested in a series of eye tracking experiments involving human subjects performing visual search in naturalistic scenes (Rao et al., 1996; Zelinsky et al., 1997; Rao et al., 2002). Figure 1B, 1C, and 1D show the saliency maps for one such scene (a dining table scene) after including one, two, and three different spatial scales in the iconic representation. Figure 1E shows the sequence of fixations generated by the model for this image, together with those recorded from a human subject. The target (composed of the fork and the knife) was the same in both cases. As can be seen, the locations predicted by the model for successive eye movements are similar to those seen in the fixation pattern that the human subject generated for this image.

A detailed comparison of the model to human data can be found in (Rao et al., 2002). We briefly summarize the results here. The comparison was based on 480 search trials pooled over four subjects. An *average path* to each of six possible target locations was computed by averaging the fixations over subjects and search scenes. The model data was averaged over the different targets for each location. A comparison of the average paths generated by the model and by human subjects is shown in Figure 1F. The box in each subfigure represents a one degree region centered on each target location. As is evident, there is good agreement between the model and human data for each location. The number of errors made by the model was found to be close to the number of errors made by human subjects (Rao et al., 2002). The average standard deviation for the subjects, averaged over all fixations, was 1.5 degrees whereas the deviation between model and the average subject fixations was 0.7 degrees, indicating that the model's behavior is within the profile expected of an individual subject.

These results suggest that human eye movement patterns during visual search can be understood in terms of a maximum likelihood procedure for computing the most likely location of a target in a coarse-to-fine manner. This model can be viewed as a systems-level model of overt attention in that it is formulated in terms of higher-level abstractions such as saliency maps. In the next section,

we describe a probabilistic model of attention that attempts to bridge the gap between systems-level modeling and neural modeling.

3 Predictive Coding Model of Attention

The model in the previous section focused on overt attentional control using spatial filter-based representations. The choice of the filters themselves was arbitrary. One may be inclined to ask whether there exist methods to *learn* appropriate representations of objects and natural scenes directly from their images. This can be accomplished through the probabilistic notion of generative models of images. We show in this section that various forms of attention may be regarded as emergent properties of predictive coding networks that utilize generative models for representing images. Such networks also suggest functional roles for feedback and feedforward connections in the dorsal and ventral visual pathways in the mammalian brain.

3.1 Generative Models and Predictive Coding

Assume that an image, denoted by a vector $\mathbf{I}$ of n pixels, can be represented as a linear combination of a set of k basis vectors $U_1, U_2, \ldots, U_k$:

$$\mathbf{I} = \sum_{j=1}^{k} U_j r_j + \mathbf{n} \tag{6}$$

$$= U\mathbf{r} + \mathbf{n} \tag{7}$$

where $\mathbf{n}$ is a zero-mean Gaussian white noise process, U is the $n \times k$ matrix whose columns consist of the basis vectors U_j, and $\mathbf{r}$ is the $k \times 1$ vector consisting of coefficients r_j. In a neurobiological setting, the values in the ith row of U can be regarded as synaptic strength of the ith model neuron while the coefficients r_j denote the pre-synaptic activities received by these neurons.

Our goal is to estimate the coefficients $\mathbf{r}$ for any given image and on a longer time scale, learn appropriate basis vectors in U directly from the input image stream. Consider the following squared-error optimization function for minimization:

$$E = (\mathbf{I} - U\mathbf{r})^T S(\mathbf{I} - U\mathbf{r}) \tag{8}$$

where the superscript T denotes vector (or matrix) transpose and S is a diagonal weighting matrix. Given that $\mathbf{n}$ is Gaussian, it can be shown that

minimizing E is equivalent to *maximizing the log likelihood* of the observed data $\mathbf{I}$ with respect to model parameters U and $\mathbf{r}$ (see, for example, (Rao and Ballard, 1997; Rao, 1999)). For the purposes of modeling attention, it is useful to choose the diagonal entries in the matrix S as:

$$S^{i,i} = \min\left\{1, c/(\mathbf{I}^i - U^i\mathbf{r})^2\right\}$$

Here, $\mathbf{I}^i$ is the ith pixel of $\mathbf{I}$, U^i is the ith row of U, and c is a threshold parameter. Note that S effectively clips the ith summand in E to a constant saturation value c whenever the squared error $(\mathbf{I}^i - U^i\mathbf{r})^2$ exceeds c. Thus, statistical outliers (i.e. image regions containing distracting, irrelevant, or unknown objects) are prevented from influencing the optimization process due to the large errors that they produce (see below for an example).

One can minimize E with respect to $\mathbf{r}$ and U using gradient descent to obtain the following differential equations:

$$\dot{\mathbf{r}} = -\frac{k_1}{2}\frac{\partial E}{\partial \mathbf{r}} = k_1 U^T G(t)(\mathbf{I} - U\mathbf{r}) \tag{9}$$

$$\dot{U} = -\frac{c_1}{2}\frac{\partial E}{\partial U} = c_1 G(t)(\mathbf{I} - U\mathbf{r})\mathbf{r}^T \tag{10}$$

where $\dot{\mathbf{r}}$ and $\dot{U}$ denote the temporal derivatives of $\mathbf{r}$ and U respectively, and k_1 and c_1 are positive time constants that determine the rate of descent towards a minimum of E. For a given static image, U is typically kept fixed until $\mathbf{r}$ converges to a stable value; this value of $\mathbf{r}$ is then used to update U as specified in Equation 10. The matrix $G(t)$ in the equations above is an $n \times n$ diagonal matrix whose diagonal entries at time t are given by:

$$G^{i,i}(t) = \begin{cases} 0 \text{ if } (\mathbf{I}^i(t) - U^i\mathbf{r}(t))^2 > c(t) \\ 1 \text{ otherwise} \end{cases} \tag{11}$$

G can be regarded as the sensory residual gain or "gating" matrix. It determines the gain on the various components of the incoming sensory residual error $(\mathbf{I} - U\mathbf{r})$. By effectively excluding any high residual errors, G allows the model to ignore the corresponding outliers (occluding objects or clutter) in the input $\mathbf{I}$, thereby enabling it to robustly estimate $\mathbf{r}$. In fact, Equation 9 can be interpreted as implementing an approximate form of the *robust Kalman filter* (Rao, 1998).

Figure 2A depicts a recurrent network that implements Equation 9. The network can be regarded as a "predictive coding" circuit wherein feedback connections carry predictions $(U\mathbf{r})$ of lower level inputs $(\mathbf{I})$ while feedforward

connections carry filtered error signals ($(\mathbf{I}(\mathbf{x}) - U\mathbf{r})$). Predictive coding has previously been used to model visual cortical response properties such as contextual and non-classical receptive field effects (see (Rao and Ballard, 1999)).

3.2 Visual Attention without a Spotlight

We now illustrate how the predictive coding model discussed above can be used to model attentional shifts. We assume a training phase in which objects are shown to the recognition system without occlusions or background clutter (e.g., Figure 2B). These objects are learned by alternating between Equations 9 and 10 during repeated exposures to the objects, until the basis matrix U stabilizes.

Now consider the case where a familiar object from the training database occurs with another occluding object or background clutter in an input image (Figure 2C, leftmost image). When the object vector $\mathbf{r}$ is calculated using Equation 9, the model predicts only the familiar object, causing relatively large errors in the areas of the image that do not match the predictions. These regions of the image are treated as outliers and the gating matrix G prevents these regions from influencing the estimation of $\mathbf{r}$. The system is thus able to "focus attention" on a familiar object despite occlusions and background clutter as shown in Figure 2C.

More interestingly, the outliers (shown in white) produce a crude *segmentation* of the occluder and background clutter, which can subsequently be used to focus "attention" on previously ignored objects and recover their identity. In particular, an *outlier mask* $\mathbf{m}$ can be defined by taking the complement of the diagonal of G (i.e., $\mathbf{m}^i = 1 - G^{i,i}$). By replacing the diagonal of G with $\mathbf{m}$ in Equation 9 and repeating the estimation process, the network can "attend to" the image region(s) that were previously ignored as outliers. As shown in Figure 2D, the network first recognizes the "dominant" object, typically the object occupying a larger area of the input image or possessing regions with higher contrast. The outlier mask $\mathbf{m}$ is subsequently used for "switching attention" and extracting the identity of the second object (Figure 2D, lower arrow and rightmost image).

3.3 Object-Based versus Spatial Attention

The predictive coding model discussed above can be extended to account for transformations of objects in an image using a generative model based on the Taylor series expansion of a new image $\mathbf{I}(\mathbf{x})$ in terms of a canonical image $\mathbf{I}$:

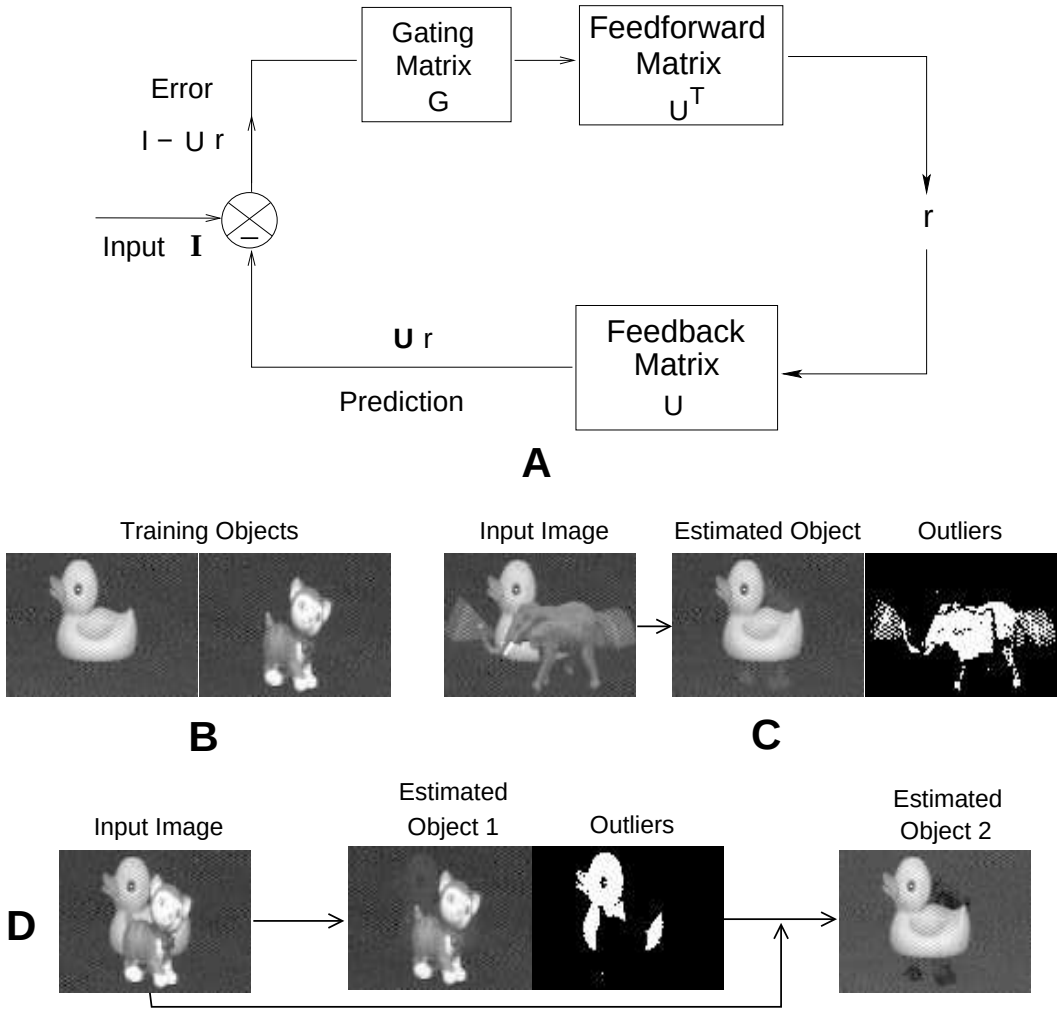

Fig. 2. **Attention in a Predictive Coding Model of Visual Processing**. (A) shows an implementation of the predictive coding model (Equation 9) in the form of a recurrent neural network. The matrices U and U^T are represented by the synaptic weights of linear feedback and feedforward neurons respectively. The gating matrix G is implemented by a set of threshold non-linear neurons with binary outputs. (B) Example images used to train a predictive coding network. (C) Given a cluttered image, the network treats occlusions and background objects as outliers (white regions in the third image, depicting the diagonal of the gating matrix G). This allows the network to "attend to" and recognize a training object ("duck") despite clutter, as indicated by the relatively accurate final reconstructed image ($U\mathbf{r}$) shown in the middle. (D) In the more interesting case of the training objects occluding each other, the network converges to one of the objects (the "dominant" one in the image - in this case, the object in the foreground). Having recognized one object, the second object is attended to and recognized by taking the complement of the outliers (diagonal of G) and repeating the estimation process (third and fourth images).

$$\mathbf{I}(\mathbf{x}) = \mathbf{I} + \frac{\partial \mathbf{I}}{\partial \mathbf{x}}\mathbf{x} + \mathbf{n} \tag{12}$$

$$= U\mathbf{r} + DI_d\mathbf{x} + \mathbf{n} \tag{13}$$

where D is a matrix of differential operators to be learned from data and I_d is a diagonal matrix containing the appropriate number of copies of the image $\mathbf{I} = U\mathbf{r}$ along the diagonal (Rao and Ballard, 1998).

The generative model in Equation 13 can be used to derive equations similar to Equations 9 and 10 for estimating both $\mathbf{r}$ and $\mathbf{x}$, and for learning U and D (see (Rao and Ballard, 1998) for details). Figure 3A shows an implementation of this model using two parallel but cooperating networks, one estimating object identity $\mathbf{r}$ ("what") and the other estimating object transformation $\mathbf{x}$ ("where"). This functional dichotomy between object recognition and transformation estimation is reminiscent of the well-known division of labor between the ventral and dorsal streams in the primate visual cortex (Felleman and Van Essen, 1991).

Given an input image, the pair of networks in Figure 3A simultaneously estimate an object and its transformation by jointly optimizing the generative model in Equation 13. Therefore, fixing a particular spatial location $\mathbf{x}$ in the transformation network should cause the object network to converge to the identity $\mathbf{r}$ of the object in that spatial location (a form of *spatial attention*). On the other hand, fixing an object's identity in the object network should cause the transformation network to converge to its most likely spatial location in the image (a form of *object-based attention*).

Figures 3B and 3C illustrate an example of spatial attention using an input image containing two training objects simultaneously, one in Location 1 and the other in Location 2. The networks were trained on images containing only one object in the center of an image. As shown in Figure 3C, when the transformation vector $\mathbf{x}$ is set to Location 1 (left panels, "Attending Location 1"), the object vector $\mathbf{r}$ converges to the canonical representation of the object in Location 1: the image predicted by the object network is that of object 1 in its central (canonical) position. Setting $\mathbf{x}$ to Location 2 causes the object network to converge to object 2 (right panels, "Attending Location 2"). In both cases, pixels containing the second object are treated as outliers (shown here in grayscale rather than in binary form). Thus, spatial attention emerges as a consequence of a top-down signal (for example, from short-term or working memory) that constrains the activity of the transformation network to be a memorized value. Likewise, object-based attention emerges in the network as a consequence of constraining the activity of the object network (not shown). These results suggest an interpretation of spatial and object-based attention in terms of specific constraints being placed on activities in the dorsal or ventral visual pathway by memory-related neurons in prefrontal cortex and

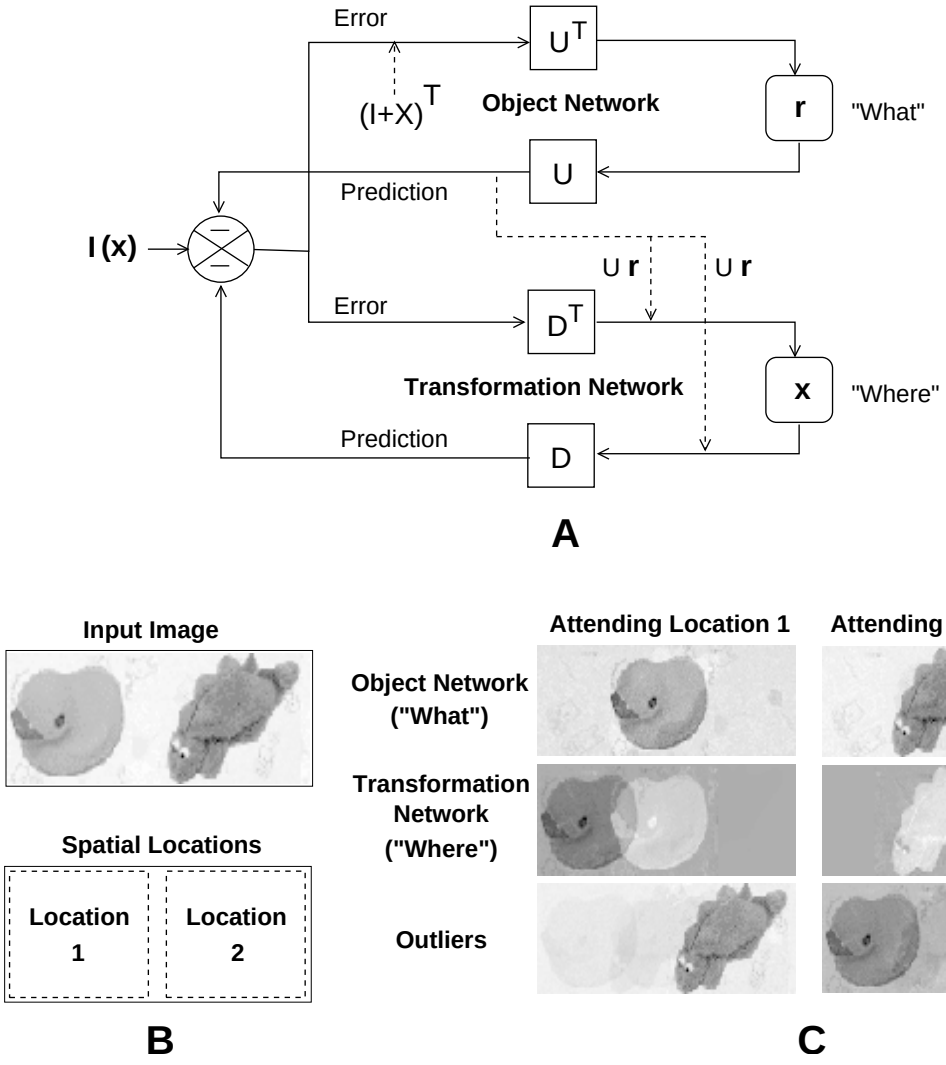

Fig. 3. **"What-Where" Networks and Spatial Attention**. (A) shows a pair of predictive coding networks: the one at the top computes the identity of objects through object features $\mathbf{r}$ ("What") while the one at the bottom computes the transformation $\mathbf{x}$ ("Where") (see (Rao and Ballard, 1998) for more details). The gating matrix G is not shown. Fixing $\mathbf{x}$ in the "Where" network causes the "What" network to converge to the identity of the object in that spatial location (a form of spatial attention), while fixing $\mathbf{r}$ in the "What" network causes the "Where" network to converge to the object's most likely spatial location (a form of object-based attention). (B) A test image containing two training objects in Locations 1 and 2 respectively. The network was trained on images containing only one object in the center of an image. (C) Fixing the spatial position vector $\mathbf{x}$ to Location 1 causes the object network to converge to the "duck" object (left panels, "Attending Location 1"). The pixels containing the "dinosaur" object are treated as outliers. On the other hand, fixing $\mathbf{x}$ to Location 2 "focuses attention" on the object in Location 2, i.e., the dinosaur object (right panels, "Attending Location 2").

other areas implicated in working memory.

4 Summary and Conclusions

In this article, we reviewed two models of attention, both based on probabilistic principles but formulated at two different levels of abstraction. The first model relies on iconic representations and the concept of saliency maps to predict eye movements during visual search in naturalistic scenes. Saliency maps have played an important role in models of attention, especially those focusing on extracting "interesting" locations in a scene based on bottom-up sensory information (Koch and Ullman, 1985; Niebur and Koch, 1996; Itti and Koch, 2000). The model discussed in this article combines both bottom-up scene representations and a top-down target representation to generate a saliency map. The model further assumes that the saliency map is computed in a coarse-to-fine manner such that larger scale filter responses are compared first. Motivation for coarse-to-fine computation of saliency maps comes from several studies that show that lower spatial frequencies influence visual perception earlier than higher spatial frequencies (e.g., (Navon, 1977; Schyns and Oliva, 1994)). For a given saliency map, the model computes the most likely target location as the weighted average of all locations, the weight being determined by the location's saliency. This procedure is motivated by previous work in probabilistic reasoning and learning based on the Boltzmann and related distributions (Hinton and Sejnowski, 1986; Nowlan, 1990). The saliency map and the weighted averaging scheme in the model may have correlates in the posterior parietal cortex and the superior colliculus respectively (Desimone and Duncan, 1995; McIlwain, 1991). The model explains experimental results showing that humans make successive eye-movements to the "center-of-gravity" of clusters of objects before landing on a most-likely object location (Zelinsky et al., 1997). The model assumes that the oculomotor system is ready to move before all the scales can be matched, and thus the eyes move to the current best target position, thereby increasing the chances of an early match. These results suggest that the human visual system utilizes a probabilistic method based on maximum likelihood (or more generally, maximum a posteriori) estimation to shift gaze to points of interest in a natural scene.

The second model is based on the probabilistic notion of generative models. By hypothesizing a mathematical model for how images are synthesized using a set of basis functions, a network can be derived for learning these basis functions and estimating their coefficients. This network implements a form of predictive coding in which top-down feedback is used to predict a lower-level signal (e.g., an image) while the feedforward signals convey the prediction errors. The network computes an "optimal" set of coefficients that serve to represent the

contents of an image in a compact and efficient manner. Such predictive coding networks have proved useful in modeling visual cortical response properties (Rao and Ballard, 1999). We discussed how attention emerges in such networks as a consequence of selective filtering of predictive error signals. This allows the network to "focus attention" on a single object or "switch attention" to another object without using an explicit "spotlight of attention."

The predictive coding model can be extended to account for transformations of objects in images, resulting in "what-where" networks that can simultaneously recognize an object and estimate its pose. We discussed how object-based attention and spatial attention are emergent properties of such networks, caused by placing constraints on the "what" or "where" network respectively. The model thus provides a unifying explanation for well-known spatial attention results as well as more recent results on object-based attention showing that subjects can reliably track an object superimposed with a distractor object occupying the same spatial location (Kanwisher and Wojciulik, 2000).

The functional dichotomy in the predictive coding model between the "what" and the "where" networks resembles the well-known division of labor between the cortical networks in the ventral and dorsal visual pathway. This observation leads to several potentially testable predictions. Substantial connections exist between the dorsal and ventral visual pathways (Felleman and Van Essen, 1991), but their function is unknown. These connections form an integral part of the joint optimization process in the model (see Figure 3A and (Rao and Ballard, 1998)). The model predicts that damage to dorsal areas should produce noticeable effects in object-based attentional tasks, while damage to ventral areas should produce significant deficits in spatial attention tasks. This is interesting, given that dorsal and ventral areas are traditionally associated only with spatial and object-related perception respectively. Similarly, damage to frontal cortex areas, the presumed source of top-down constraints on the "what" and "where" networks, should adversely affect both spatial as well as object-based attention tasks.

The predictive coding model thus emphasizes the global nature of attention: the different forms of visual attention are interpreted as emerging from the constrained optimization of a generative model thought to be encoded jointly within the dorsal and ventral visual cortical pathways.

References

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

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47.

Freeman, W. T. and Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(9):891–906.

Hinton, G. (1981). A parallel computation that assigns canonical object-based frames of reference. In *7th International Joint Conference on Artificial Intelligence*, pages 683–685.

Hinton, G. and Sejnowski, T. (1986). Learning and relearning in Boltzmann machines. In Rumelhart, D. and McClelland, J., editors, *Parallel Distributed Processing*, volume 1, chapter 7, pages 282–317. MIT Press, Cambridge.

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

Jones, D. G. and Malik, J. (1992). A computational framework for determining stereo correspondence from a set of linear spatial filters. In *Proceedings of the Second European Conference on Computer Vision*.

Kanwisher, N. and Wojciulik, E. (2000). Visual attention: Insights from brain imaging. *Nature Reviews Neuroscience*, 1:91–100.

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

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42:300–311.

McIlwain, J. T. (1991). Distributed spatial coding in the superior colliculus: A review. *Visual Neuroscience*, 6:3–13.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9:353–383.

Newsome, W. T. (1996). Spotlights, highlights and visual awareness. *Current Biology*, 6(4):357–360.

Niebur, E. and Koch, C. (1996). Control of selective visual attention: Modeling the "where" pathway. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems 8*, pages 802–808. Cambridge, MA: MIT Press.

Nowlan, S. (1990). Maximum likelihood competitive learning. In Touretzky, D., editor, *Advances in Neural Information Processing Systems 2*, pages 574–582. San Mateo, CA: Morgan Kaufmann.

Olshausen, B. A., Essen, D. C. V., and Anderson, C. H. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13:4700–4719.

Rao, R. P. N. (1998). Correlates of attention in a model of dynamic visual recognition. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 80–86. Cambridge, MA: MIT Press.

Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, 39(11):1963–1989.

Rao, R. P. N. and Ballard, D. H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505.

Rao, R. P. N. and Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4):721–763.

Rao, R. P. N. and Ballard, D. H. (1998). Development of localized oriented receptive fields by learning a translation-invariant code for natural images. *Network: Computation in Neural Systems*, 9(2):219–234.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, 2(1):79–87.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., and Ballard, D. H. (1996). Modeling saccadic targeting in visual search. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems 8*, pages 830–836. Cambridge, MA: MIT Press.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., and Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463.

Schyns, P. G. and Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4):195–200.

Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545.

Zelinsky, G. J., Rao, R. P. N., Hayhoe, M. M., and Ballard, D. H. (1997). Eye movements reveal the spatio-temporal dynamics of visual search. *Psychological Science*, 8(6):448–453.