Vision Research xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

What and where: A Bayesian inference theory of attention

Sharat Chikkerur *, Thomas Serre, Cheston Tan, Tomaso Poggio

McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States

ARTICLE INFO

Article history:
Received 21 October 2009
Received in revised form 30 April 2010
Available online xxxx

Keywords:
Computational model
Attention
Bayesian inference
Object recognition

ABSTRACT

In the theoretical framework of this paper, attention is part of the inference process that solves the visual recognition problem of *what is where*. The theory proposes a computational role for attention and leads to a model that predicts some of its main properties at the level of psychophysics and physiology. In our approach, the main goal of the visual system is to infer the identity *and* the position of objects in visual scenes: spatial attention emerges as a strategy to reduce the uncertainty in shape information while feature-based attention reduces the uncertainty in spatial information. Featural and spatial attention represent two distinct modes of a computational process solving the problem of recognizing *and* localizing objects, especially in difficult recognition tasks such as in cluttered natural scenes.

We describe a specific computational model and relate it to the known functional anatomy of attention. We show that several well-known attentional phenomena – including bottom-up pop-out effects, multiplicative modulation of neuronal tuning curves and shift in contrast responses – all emerge naturally as predictions of the model. We also show that the Bayesian model predicts well human eye fixations (considered as a proxy for shifts of attention) in natural scenes.

1. Introduction

Much of the recent work in visual recognition both in computer vision and physiology focused on the 'what' problem: which object is in the image. Face detection and identification are typical examples. Recognition is, however, more than the mere detection of a specific object or object class: everyday vision routinely solves the problem of *what is where*. In fact, David Marr defined vision as "knowing what is where by seeing" (Marr, 1982).

In somewhat of an oversimplification, it has been customary to describe processing of visual information in the brain along two parallel and concurrent streams. The ventral ('what') stream processes visual shape appearance and is largely responsible for object recognition. The dorsal ('where') stream encodes spatial locations and processes motion information. In an extreme version of this view, the two streams underlie the perception of 'what' and 'where' concurrently and relatively independently of each other (Ungerleider & Haxby, 1994; Ungerleider & Mishkin, 1982). Lesions in a key area of the ventral ('what') stream (the inferior temporal cortex) cause severe deficits in visual discrimination tasks without affecting performance on visuospatial tasks such as visually guided reaching tasks or tasks that involve judgments of proximity between an object and a visual landmark. In contrast, parietal lesions in the dorsal ('where') stream cause severe deficits on visuospatial performance tasks while sparing visual discrimination ability. In everyday life, the identity and location of objects must somehow be integrated to enable us to direct appropriate actions to objects. Thus a hypothetical segregation of the two streams raises the question of how the visual system combines information about the identities of objects and their locations. The central thesis of this paper is that visual attention performs this computation (see also Deco & Rolls, 2004; Van Der Velde & De Kamps, 2001).

The past four decades of research in visual neuroscience have generated a large and disparate body of literature on attention (see Supplementary Online Information, Sections 3.1–3.3). Although several computational models have been developed to describe specific phenomena, a theoretical framework that explains the computational role of attention, while predicting and being consistent with known biological effects, is lacking. It was recently suggested that visual perception may be interpreted as a Bayesian inference process whereby top-down priors help disambiguate noisy bottom-up sensory input signals (Dayan, Hinton, & Neal, 1995; Dayan & Zemel, 1999; Dean, 2005; Epshtein, Lifshitz, & Ullman, 2008; Friston, 2003; George & Hawkins, 2005; Hinton, 2007; Kersten & Yuille, 2003a; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Lee & Mumford, 2003; Mumford, 1992; Murray & Kreutz-Delgado, 2007; Rao, 2004; Rao, Olshausen, & Lewicki, 2002; Weiss, Simoncelli, & Adelson, 2002).

In this paper, we build on earlier work (Rao, 2005; Yu & Dayan, 2005) to extend the Bayesian inference idea and propose that the computational role of attention is to answer the *what is where* question. Our model predicts several properties of attention at

* Corresponding author.
E-mail addresses: sharat@mit.edu (S. Chikkerur), serre@mit.edu (T. Serre), cheston@mit.edu (C. Tan), tp@ai.mit.edu (T. Poggio).

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.05.013

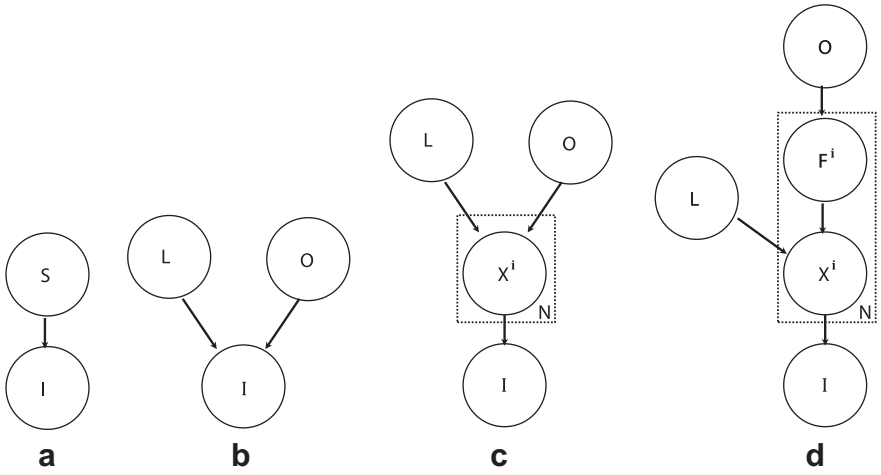


Fig. 1. The figure illustrates the progression of graphical models corresponding to the sequence of factorizations given in Eqs. (1)–(5) induced by the three main assumptions.

the level of visual psychophysics and physiology. It can also be regarded as extending existing models of the 'what' pathway, such as hierarchical feedforward models of the ventral stream in the visual cortex (*e.g.*, Amit & Mascaro, 2003; Fukushima, 1980; Mel, 1997; Riesenhuber & Poggio, 1999b; Serre, Kouh, et al., 2005; Thorpe, 2002; Ullman, Vidal-Naquet, & Sali, 2002; Wallis & Rolls, 1997; Wersing & Koerner, 2003) by taking into account some of the back-projections present throughout cortex.

In our framework, visual recognition corresponds to estimating posterior probabilities of visual features for specific object categories and their locations in an input image (*i.e.*, knowing what is where). Algorithmically, this computation can be performed using belief propagation (Pearl, 1988), variational methods (Wainwright & Jordan, 2008) or sampling-based approaches, such as MCMC (Gilks & Spiegelhalter, 1996; Neal, 1993). We use here the belief propagation algorithm for which implementations have been proposed that are biologically plausible (Beck & Pouget, 2007; Deneve, 2008; George, 2008; Lee & Mumford, 2003; Litvak & Ullman, 2009; Rao, 2004; Steimer, Maass, & Douglas, 2009; Zemel, Dayan, & Pouget, 1998) (see also Supplementary Online Information, Section 3.2).

Within our framework, spatial attention is related to spatial priors, while feature-based attention depends on task-based priors for objects and features. The posterior probability over location variables serves as a 'saliency map'. Attentional phenomena such as pop-out, multiplicative modulation and change in contrast response, which have been sometimes described in the recent literature as different and, in some cases, even conflicting findings, are all predicted by the same model. We also show that the Bayesian model predicts well human eye fixations (considered as a proxy for shifts of attention) in natural scenes in task-free as well as in task-dependent situations.

This paper proposes a formal interpretation of attention as a process that has the computational goal of inferring simultaneously the form and location of objects in the visual world. As we will discuss later (see Section 7), the generative model described below is based on a few assumptions, the first of which is related to the traditional 'computational bottleneck' (Tsotsos, 1997) theories of attention.

2. Computational model

2.1. Model preliminaries

A generative model $S \rightarrow I$ specifies how an image I (represented as either raw pixel intensities or as a collection of topographic feature maps) is determined by the scene description S (*e.g.,* vectorial description of properties such as global illumination, scene identity, and objects present). The product of the likelihood $P(I|S)$ and the prior probability of the scene description $P(S)$ determines the generative model (Kersten et al., 2004):

$$P(S,I) = P(I|S)P(S) \tag{1}$$

The generative model also specifies the probabilistic relationship between observed variables (object and image) and unobserved (latent) variables such as lighting, depth, and viewpoint. that influence the observed data. Following recent work (Kersten & Yuille, 2003b), we decompose the description of a scene in n components which in our case are just objects (including the background) $\{O_1, O_2, \ldots, O_n\}$ and their locations $\{L_1, L_2, \ldots, L_n\}$ in the scene.[1]

Thus, $S = \{O_1, O_2, \ldots, O_n, L_1, L_2, \ldots, L_n\}$. In the most general case, every random variable influences every other one. We show how a few key assumptions lead to a simple factorization of the generally complex joint probability – corresponding to simplifications of the original graphical model (see Fig. 1).

As we mentioned, one of the main tasks of vision is to recognize and localize objects in the scene. Here we assume that

(a) *To achieve this goal, the visual system selects and localizes objects, one object at a time.*

Since the requirements of the task split S into those variables that are important to estimate accurately for the task and those that are not, we write in this case $P(S,I) = P(O_1, L_1, O_2, L_2, \ldots, O_n, L_n, I)$. We can then integrate out the confounding variables (*i.e.*, all objects except one – labeled, without loss in generality, O_1):

$$P(O_1, L_1, I) = \sum_{O_2 \cdots O_n, L_2 \cdots L_n} P(O_1, L_1, O_2, \ldots, O_n, L_2, \ldots, L_n, I) \tag{2}$$

We further assume that

(b) *the object location and object identity are independent, leading to the following factorization*:

$$P(O, L, I) = P(O)P(L)P(I|L, O) \tag{3}$$

[1] The probabilistic model can be extended to generalize to scenes with an arbitrary number of objects.

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013

In Eq. (3) and in following equations, we replace, for simplicity the single object O_1 with O and its location L_1 with L.

Remarks

- Attention, as described later, emerges as the inference process implied by Eq. (3). In a sense, our framework with the key assumption (a), "predicts" attention and – with the approximations to Eq. (3) described in the rest of the section – several of its properties.
- Bayesian models of object recognition – but, emphatically, not of attention – assume different (w.r.t Eq. (3)) factorizations of $P(S,I)$, such as $P(S,I) = P(O_1,L_1,\ldots,O_n,L_n,I)$ (Sudderth, Torralba, Freeman, & Willsky, 2005) or $P(S,I) = P(O,L,I) = P(O,\ L|I)P(I)$ (Torralba, 2003a), in which location and identity of an object are modeled jointly. In Eq. (3), I corresponds to an entire array of measurements (every feature at every location). Eq. (3), dictated by the generative model and the requirements of the task, leads to a simpler approximation with $P(O,L,I) = P(O)P(L)P(I|O,L)$ – as a model of attention.
- The key assumption (a) characterizes the task of attention as selecting a single object – for recognition and localization – in the scene. This is a formalization of the standard *spotlight hypothesis* of attention, in which attention focuses processing to a region of the image. One can speculate about the reasons for this constraint. Previous proposals based on the bottleneck and salience hypotheses (Bruce & Tsotsos, 2006; Itti, Koch, & Niebur, 1998; Tsotsos, 1997) postulate that the role of attention is to prioritize the visual scene, where limited visual processing resources are directed towards 'relevant' regions. These hypotheses correspond to the assumption that the visual system needs attention in order to reduce the *computational complexity* of recognition. We prefer a related hypothesis to justify attention and our factorization. Our *hypothesis is that attention is needed to reduce the sample complexity* of learning the relevant probability distributions over objects, features and locations. We believe that it would take too much data, and therefore an unreasonably long time, unless one makes assumptions about the parametric form of the distributions – assumptions that are arguably as strong as ours.[2]
- Eq. (3) is not a *strong* generative model (Kersten et al., 2004) because it takes into account a generative model *and* the assumed constraints of the task of attention. It cannot produce images containing many objects, such as typical scenes used in our experiments (see for instance Fig. 6). It can synthesize images containing either no object or one object such as a single car. It corresponds to visual scenes 'illuminated by a spotlight of attention'. Note that from the inference point of view, if the task is to find a car in the image, there will always be either no car or one car which is more car-like than other ones (because of image "noise").
- Although, assumption (a) posits that the core model of attention should find a single object in the image, the process can be iterated, looking for other objects in other locations, one at a time. This assumption motivates most (extended) models of attention (Miau & Itti, 2001; Rutishauser, Walther, Koch, & Perona, 2004; Walther & Koch, 2007, Chapter: Attention in Hierarchical Models of Object Recognition) and also motivates mechanisms such as "inhibiton of return" (Itti & Koch, 2001). The full strategy of calling multiple times the core attention module to recognize and localize one object at a time is not Bayesian. It is an interesting question for future work how to model in a fully Bayesian framework the sequential process of recognizing and localizing objects (Lovejoy, 1991; Monahan, 1982; Smallwood & Sondik, 1973).

[2] Let us assume that ϵ is the error with which the distribution is learned, s is a measure of smoothness of the density being approximated, N is the number of objects, O is the number of object classes and L is the dimensionality of the location grid. As an example to give a feeling for the issue, we consider the joint probabilities: Learning joint probabilities of all the N objects and their locations would take in the order of $\epsilon^{-NOL/s}$ examples whereas learning a single object and its location would take in the order of $\epsilon^{-OL/s}$ examples. Whereas it would take in the order of $\epsilon^{-O/s} + \epsilon^{-L/s}$ examples for our factorization. There can be many orders of magnitude difference between the required examples (for instance take $\epsilon = 0.1$)!

2.2. Model

Consider the generative model specified in Eq. (3). We assume that the image of an object is generated through a set of relatively complex object features. In particular, *(c) we assume that each of N features is either present or absent and that they are conditionally independent, given the object and its location*. A similar approach can be found in other part-based object recognition frameworks (Crandall, Felzenszwalb, & Huttenlocher, 2005; Felzenszwalb & Huttenlocher, 2005; Fergus, Perona, & Zisserman, 2003). We use intermediate latent variables $\{X^1,X^2,\ldots,X^N\}$ to encode the position of the N object features; if feature i is not present, then $X^i = 0$. These intermediate variables can be considered as feature maps which depend on the object and its location. We model the joint probability of the object identity O, its location L, the feature maps $\{X^i, i = 1,2,\ldots,N\}$ and the image I. Eq. (3) takes the form

$$P(O,L,X^1,\ldots,X^N,I) = P(O)P(L)P(X^1,\ldots,X^N|L,O)P(I|X^1,\ldots,X^N) \tag{4}$$

We then take the variables to be discrete, because of computational considerations and because images (and arrays of neurons) can be represented on discrete grids. Because of the assumed conditional independence $P(X^1,\ldots,X^N|L,O)$ is given by the following factorization:

$$P(X^1,\ldots,X^N|L,O) = \prod_{i=1}^{i=N}\{P(X^i|L,O)\} \tag{5}$$

Applying Eq. (5), Eq. (4) leads to our final probabilistic model

$$P(O,L,X^1,\ldots,X^N,I) = P(O)P(L)\left\{\prod_{i=1}^{i=N}\{P(X^i|L,O)\}\right\}P(I|X^1,\ldots,X^N) \tag{6}$$

The model consists of a location encoding variable L, object encoding variable O, and feature-map variables $\{X^i, i = 1,\ldots,N\}$, that encode position-feature combinations. The object variable O is modeled as a multinomial random variable with $|O|$ values corresponding to objects known by the model. The location variable L is modeled as a multinomial random variable with $|L|$ distinct values that enumerate all possible location and scale combinations. The variable X^i is a multinomial variable with $|L| + 1$ values $(0,1,\ldots,L)$.

As we discuss later (Section 4), it is easier to map the model onto the functional cortical anatomy (see Fig. 2) of attention by introducing the (dummy) variables $(F^i)_{i=1\ldots N}$, which are not strictly needed but can be interpreted directly in a biological perspective. Each feature-encoding unit F^i is modeled as a binary random variable that represents the presence or absence of a feature irrespective of location and scale. The location (X^i) of feature i depends on the feature variable F^i and on the location variable L. This relation, and the definition of F^i, can be written as $P(X^i|L,O) = P(X^i|F^i,L)P(F^i|O)$. With the auxiliary variables $(F^i)_{i=1\ldots N}$ the factorization of Eq. (6) can be rewritten as

$$P(O,L,X^1,\ldots,X^N,F^1,\ldots,F^N,I) = P(O)P(L)\left\{\prod_{i=1}^{i=N}\{P(X^i|L,F^i)P(F^i|O)\}\right\}P(I|X^1,\ldots,X^N) \tag{7}$$

The conditional probability $P(X^i|F^i,L)$ is such that when feature F^i is present ($F^i = 1$), and $L = l^*$, the feature map is activated at either

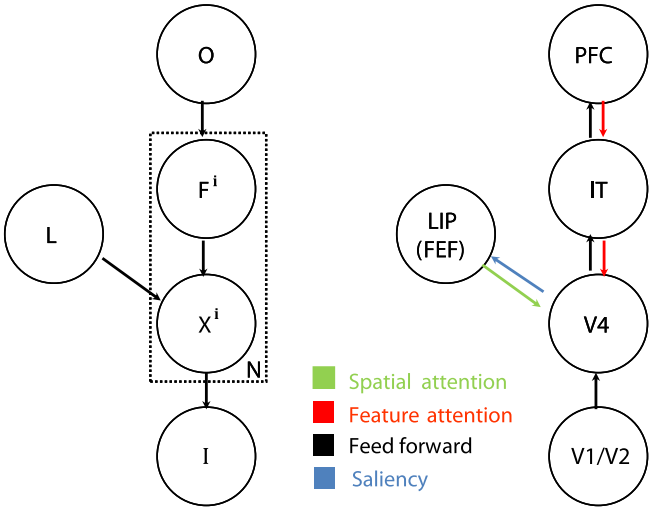


Fig. 2. Left: Proposed Bayesian model. Right: A model illustrating the interaction between the parietal and ventral streams mediated by feedforward and feedback connections. The main additions to the original feedforward model (Serre, Kouh, et al., 2005) (see also Supplementary Online Information) are (i) the cortical feedback within the ventral stream (providing feature-based attention); (ii) the cortical feedback from areas of the parietal cortex onto areas of the ventral stream (providing spatial attention); and (iii) feedforward connections to the parietal cortex that serves as a 'saliency map' encoding the visual relevance of image locations (Koch & Ullman, 1985).

$X^i = l^*$ or a nearby location with high probability (decreasing in a gaussian manner). However, when the feature F^i is absent ($F^i = 0$), only the 'null' state of X^i ($X^i = 0$) is active. Thus, when location $L = l^*$ is active, the object features are either near location l^* or absent from the image. In addition to this top-down generative constraint, bottom-up evidence $P(I|X^1 \cdots X^N)$ is computed from the input image. $P(I|X^1 \cdots X^N)$ obtained from the image is not a normalized probability. In practice, it is proportional to the output of a feature detector. However, this does not adversely affect the inference process. See Table 2 for further details.

Visual perception here corresponds to estimating posterior probabilities of visual features $(F^i)_{i=1\ldots N}$, object O and location L following the presentation of a new stimulus. In particular, $P(L|I)$ can be interpreted as a saliency map (Koch & Ullman, 1985), that gives the saliency of each location in a feature-independent manner. $P(F^i|I)$ and $P(O|I)$ can be thought of as location independent readout of object features and object identity respectively.

Remarks 1. The probabilistic model of Eq. (7) encodes several constraints resulting from our three assumptions:

- *Each feature F^i occurs at a single location/scale in the feature map.* This apparently strong constraint follows from assumption (a) and (c). Assumption (c) is suggested directly by the assumption that the features are relatively complex (such as V4-like features). Our model implements the constraint above through the automatically enforced mutual exclusion of different states of X^i. We emphasize that there is no mutual exclusion among the different features – multiple features can be active at the same location. This is in contrast to earlier probabilistic models (Rao, 2005) where features were assumed to be mutually exclusive as well.
- *Objects can be represented in terms of a single set of universal features* $(F^1, \ldots, F^N)$. Although some objects may have diagnostic features, a large variety of objects can be represented using a shared set of primitive shape features (Mutch & Lowe, 2006; Ranzato, Huang, Boureau, & LeCun, 2007; Serre, Wolf, Bileschi, Reisenhuber, & Poggio, 2007; Torralba, Murphy, & Freeman, 2004).

These assumptions limit the range and kind of "images" that can be generated by this model. The relevant question, however, is whether such a simplified model of the visual world, imposed by the objective constraint of sample complexity, actually describes what is used by the visual system.

2.3. Comparison to prior work

The model is closely related to the Bayesian model of spatial attention proposed by Rao (2005). The previous model was modified to include the following significant extensions: (i) The model includes both feature and object priors. This allows us to implement top-down feature-based attention in addition to spatial attention. (ii) The model allows conjunction of features that share common spatial modulation, while prior work modeled a single feature dimension (*e.g.*, orientation) with mutually exclusive features. (iii) Spatial attention is extended to include scale/size information in addition to just location information. Our new model can account not only for visual searches in artificial search arrays but also for searches in real-world natural images for which it predicts well human eye movements under bottom-up and top-down attention (see Section 6).

3. Model properties

In the following, we describe the properties of the model. For simplicity, we assume that the model consists of a single feature variable F^i and its corresponding feature-map variable X^i.

3.1. Translation invariance

The F^i units encode the presence or absence of individual features in a translation/scale invariant manner. The invariance is achieved by pooling responses from all locations. The posterior probability of the feature F^i is given by:

$$P(F^i|I) \propto P(F^i) \sum_{L,X^i} P(X^i|F^i, L)P(L)P(I|X^i) \tag{8}$$

Here, $P(F^i) = \sum_O P(F^i|O)P(O)$. Spatial invariance is achieved by marginalizing (summing over) the L variables (see Fig. 3a).

3.2. Spatial attention

In our model, spatial attention follows from setting a prior $P(L)$ concentrated around the location/scale of interest (see Fig. 3b). Consider the posterior estimate of the feature unit F^i. Ignoring the feature prior, the estimate is given by:

$$P(F^i|I) \propto \sum_{L,X^i} P(X^i|F^i, L)P(L)P(I|X^i) \tag{9}$$

The corresponding unit response can be considered as a weighted sum of the evidence $P(I|X^i)$. Under spatial attention, regions inside the spotlight of attention are weighed more, while those outside the spotlight are suppressed. As a consequence, the receptive fields of the non-retinotopic F^i units at the next stage are effectively *shrunk*.

3.3. Feature-based attention

As illustrated in Fig. 3c, when a single isolated object/feature is present, it is possible to read out its location from the posterior probability $P(L|I)$. However when multiple objects/features are

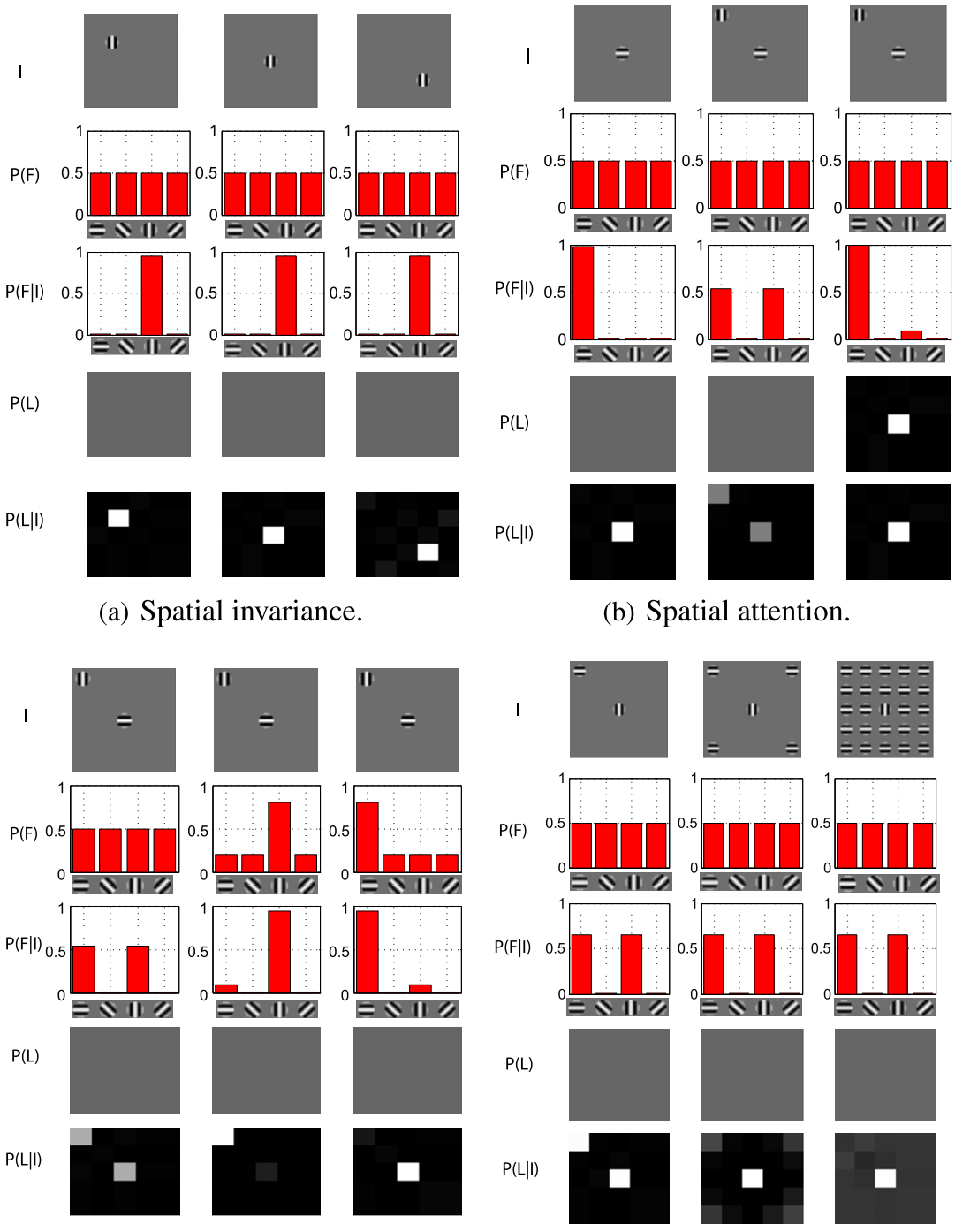


Fig. 3. An illustration of some of the key model properties. Here $P(L)$, $P(F)$ represent the prior that is set before the image is seen. $P(F|I)$, $P(L|I)$ represent the posterior probabilities after the image is observed. (a) Spatial invariance: The posterior probability $P(F|I)$ is independent of the stimulus position. (b) Illustration of how spatial attention contributes to solving the 'clutter' problem associated with the presentation of multiple stimuli. (c) Illustration of how feature-based attention contributes to solving the 'clutter' problem associated with the presentation of multiple stimuli. (d) The feature pop-out effect: The relative strength of the saliency map $P(L|I)$ increases as more and more identical distractors are being added increasing the conspicuity of the unique stimulus with its surround.

present (see Fig. 3c, first column), it is no longer possible to readout this information. To solve this problem, parallel feature-based attention results from concentrating the priors $P(F^i)$ ($P(F^i|O)$ for an object search) around the features of interest (*e.g.*, red and square features when searching for a red square). The value of the saliency map is given by:

$$P(L|I) \propto P(L) \prod_i \left\{ \sum_{F^i, X^i} P(X^i|F^i, L) P(F^i) P(I|X^i) \right\} \tag{10}$$

Increasing the concentration of the prior around the target feature F^i enhances the preferred feature at *all* locations while low priors on

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013

Table 1
Bayesian model units and tentative mapping to brain areas.

Model (brain)	Representation (biological evidence)
L (LIP/FEF)	This variable encodes the location and scale of a target object. It is modeled as a discrete multinomial variable with $\|L\|$ distinct values Prior studies (Colby & Goldberg, 1999) have shown that the parietal cortex maintains several spatial maps of the visual environment (eye-centered, head-centered, etc.). Studies also show that response of LIP neurons is correlated with the likelihood ratio of a target object (Bisley & Goldberg, 2003). In this paper, we hypothesize that the saliency map (corresponding to the variable L) is represented in the parietal cortex.
O (PFC)	This variable encodes the identity of the object. It is modeled as a discrete multinomial variable that can take $\|O\|$ distinct values The preferred stimulus of neurons tend to increase in complexity along the ventral stream: from simple oriented bars in area V1 (Hubel & Wiesel, 1959) to combinations of orientations and features of moderate complexity in intermediate visual areas V2 (Hegde & Van Essen, 2000; Ito & Komatsu, 2004) and V4 (Desimone & Schein, 1987; Gallant et al., 1996; Pasupathy & Connor, 2001), to parts and objects in area IT (Logothetis & Sheinberg, 1996; Tanaka, 1996). It has been shown that object category information is represented in higher areas such as the prefrontal cortex (PFC) (Freedman et al., 2001).
F^i (IT)	Each feature variable F^i encodes the presence of a specific shape feature. Each such unit is modeled as a discrete binary variable that can be either on or off. The presence/absence of a given feature is computed in a position/scale invariant manner (see Serre, Kouh, et al. (2005) for details). In practice, for the visual tasks described in this paper, we have used a dictionary of features of about $10 \sim 100$ such features. Neurons in the inferotemporal (IT) cortex are typically tuned to objects and parts (Tanaka, 1996) and exhibit some tolerance with respect to the exact position and scale of stimulus within their receptive fields (typically on the order of a few degrees for position and on the order of one octave for size (Logothetis et al., 1995).
X^i (V4)	This variable can be thought of as a feature map that encodes the joint occurrence of the feature (F^i) at location $L = l$. It is modeled as a discrete multinomial variable with $\|L\| + 1$ distinct values $(0, 1 \ldots, L)$. Values $(1 \cdots L)$ correspond to valid locations while $X^i = 0$ indicates that the feature is completely absent from the input. Feature-based attention is found to modulate the response of V4 neurons at all locations (Bichot et al., 2005). Under spatial attention, V4 neurons that have receptive fields overlapping with the locus of attention are enhanced (McAdams & Maunsell, 1999). Thus V4 neurons are involved in feature-based attention as well as spatial attention marking V4 as the likely area of interaction between ventral and parietal cortices.
I (V2)	This is the feed-forward evidence obtained from the lower areas of ventral stream model. Given the image I, for each orientation and location, $P(I\|X^i)$ is set proportional to the output of the filter. The neurons in area V2 are found to be sensitive to conjunction of orientations, curvature and grating-like stimuli (Hegde & Van Essen, 2000; Ito & Komatsu, 2004). We use the computational model of the ventral stream (Serre, Wolf, et al., 2007) to derive V2-like features from the image.

other features suppress activity from distracting objects. Thus, the evidence $P(I|X^i)$ is now modulated by the preference for the feature given by $P(F^i)$. The location of the preferred feature can be read out from the posterior probability $P(L|I)$, which can be interpreted as a saliency map.

3.4. Feature pop-out

Since the X^i units are mutually exclusive $(\forall i,\ \sum_{X^i} P(X^i|F^i, L) = 1)$, increasing the activity (probability) at one location in an image typically reduces the likelihood of the stimulus being present at other locations (see Fig. 3d). In a sense, this is similar to the extra-classical receptive field effects observed throughout the visual cortex (see Carandini, Heeger, & Movshon (1997) for instance). As a result, a unique feature that is active at only one location tends to induce a higher likelihood, concentrated at that location, than a common feature, present at multiple locations, for each of the corresponding locations. This predicts a 'pop-out' effect, whereby a salient items immediately draw attention (the model shows a strong bias of the saliency map towards the location of the salient or 'surprising' item (see Fig. 3d)).

It is important to realize that the pop-out effect in our model is directly due to our assumption that different locations within the same feature compete for representation at a higher level. This is a novel explanation for the pop-out phenonmenon. In contrast to our model, traditional approach to pop-out has been based on image saliency or breakdown of feature homogeneity. In (Itti et al., 1998), center-surround difference across color, intensity and orientation dimensions is used as measure of saliency. In (Gao & Vasconcelos, 2007), self information of the stimuli $(-\log(P(I)))$ is used as measure of visual saliency (Zhang, Tong, Marks, Shan, & Cottrell, 2008). In (Rosenholtz, 1985), the normalized deviation from mean response is used instead. Li and Snowden (2006) proposed a computational model based on spatially organized V1-like units showing that they are sufficient to reproduce attentional effects such as pop-out and search asymmetries. In addition, this model can reproduce effects such as contour completion and cross-orientation inhibition that is currently not possible with the proposed model.

4. Neural interpretation

Prior work has shown that perception under uncertainty can be modeled well using Bayesian inference (Kersten et al., 2004; Knill & Richards, 1996; Mumford, 1992; Rao et al., 2002). However, how the brain represents and combines probabilities at the level of neurons is unclear. Computational models have attempted to model probabilities using populations codes (Pouget, Dayan, & Zemel, 2000), spiking models of neurons (Deneve, 2008; Pouget et al., 2000), recurrent networks (Rao, 2004), etc. The properties of the model of attention described so far do not depend on how probabilities are mapped to neural activities. In the following neural interpretation of the model we assume, however, that probabilities are represented as firing rates of populations of neurons (the physiology experiments typically measure firing rates averaged across "identical" neurons over a series of recordings).

4.1. Tentative mapping to brain areas

The graphical model can be tentatively mapped – in a way which is likely to be an oversimplification – into the basic functional anatomy of attention, involving areas of the ventral stream such as V4 and areas of the dorsal stream such as LIP (and/or FEF), known to show attentional effects (see Table 1, *Supplementary Discussion*, Section 3.2, and Fig. 2). Thus, following the organization of the visual system (Ungerleider & Haxby, 1994), the proposed model consists of two separate visual processing streams: a 'where' stream, responsible for encoding spatial coordinates and a 'what' stream for encoding the identity of object categories. Our model describes a possible interaction between intermediate areas of the ventral ('what') stream such as V4/PIT (modeled as X^i variables) where neurons are tuned to shape-like features of moderate complexity (Kobatake & Tanaka, 1994; Logothetis & Sheinberg, 1996; Tanaka, 1996) and higher visual areas such as AIT where retinotopy is almost completely lost (Logothetis, Pauls, & Poggio, 1995; Oram

Table 2
Description of the model conditional probabilities.

Conditional probability	Modeling
$P(L)$	Each scene, with its associated viewpoint, places constraints on the location and sizes of objects in the image. Such constraints can be specified explicitly (*e.g.*, during spatial attention) or learned using a set of training examples (Torralba, 2003b)
$P(F^i\|O)$	The probability of each feature being present or absent given the object; it is learned from the training data
$P(X^i\|F^i,L)$	When the feature F^i is present and location $L = l^*$ is active, the X^i units that are nearby unit $L = l^*$ are most likely to be activated. When the feature F^i is absent, only the $X^i = 0$ location in the feature map is activated. This conditional probability is given by the following table

	$F^i = 1, L = l$	$F^i = 0, L = l$
$X^i = 0$	$P(X^i\|F^i, L) = \delta_1$	$P(X^i\|F^i, L) = 1 - \delta_2$
$X^i \neq 0$	$P(X^i\|F^i, L) \sim$ Gaussian centered around $L = l$	$P(X^i\|F^i, L) = \delta_2$

δ_1 and δ_2 are small values ($\sim$0.01), chosen to ensure that $\sum P(X^i|F^i, L) = 1$

$P(I\|X^i)$	For each location within the feature map, $P(I\|X^i)$ provides the likelihood that X^i is active. In the model, this likelihood is set to be proportional to the activations of the shape-based units (see Serre, Wolf, et al., 2007)

& Perrett, 1992) (modeled as F^i units). Prior (non-Bayesian) attempts to model this interaction can be found in Grossberg (1999) and Van Der Velde and De Kamps (2001).

In our interpretation, the L variable, which encodes position and scale independently of features, may correspond to the LIP area in the parietal cortex. In the model, the L variable is represented as a multinomial variable. Each X^i variable corresponds to a collection of V4 neurons, where each neuron can be interpreted as encoding one of the mutually exclusive states of X^i. The posterior probability $P(X^i = x|I)$ is then interpreted as the response of a V4 neuron encoding feature i and at location x. Thus, in the neural interpretation, $P(X^i = 1|I), P(X^i = 2|I) \ldots, P(X^i = |L||I)$ can be mapped to the firing rates of the neuron encoding feature F^i at location $1, 2 \ldots, |L|$ respectively.

The F^i units correspond to non-retinotopic, spatial and scale invariant cells found in higher layers of the ventral stream such as AIT and IT. In feedforward models (Riesenhuber & Poggio, 1999b; Serre, Kouh, et al., 2005), such invariance (over a limited range) is obtained via a max pooling operation. The original motivation for a max operation was that the max is a natural *selection operation*: when a feature is active at multiple locations within the receptive field of a unit, the max operation selects the strongest active location while ignoring other locations. Within the Bayesian framework, the individual locations within a feature map are mutually exclusive and thus a strong activation at one location suppresses the likelihood of activation at other locations. Interestingly, the Bayesian model of attention is also performing a selection akin to the max operation – by using the 'sum-product' algorithm for belief propagation.

4.2. Inference using belief propagation

Within the Bayesian network, inference can be done using any of several inference algorithms such as junction tree, variable elimination, Markov-chain Monte Carlo (MCMC) and belief propagation (Gilks & Spiegelhalter, 1996; Wainwright & Jordan, 2008). Sampling-based approaches such as MCMC and belief propagation lend themselves more easily to biological interpretations. In the simulations of this paper, the inference mechanism used is the 'belief propagation' algorithm (Pearl, 1988), which aims at propagating new evidence and/or priors from one node of the graphical model to all other nodes. We can regard some of the messages passed between the variables during belief propagation as interactions between the ventral and dorsal streams. Spatial attention and feature attention can then be interpreted within this message passing framework. A formal mathematical treatment of the messages passed between nodes is sketched below. For simplicity we consider the case of a model based on a single feature F and adopt the notation used in Rao (2005), where the top-down messages, $\pi()$ and bottom-up messages $\lambda()$ are replaced by a uniform $m()$ term.

$$m_{O\to F^i} = P(O) \tag{11}$$

$$m_{F^i\to X^i} = \sum_O P(F^i|O)P(O) \tag{12}$$

$$m_{L\to X^i} = P(L) \tag{13}$$

$$m_{I\to X^i} = P(I|X^i) \tag{14}$$

$$m_{X^i\to F^i} = \sum_L \sum_{X^i} P(X^i|F^i, L)(m_{L\to X^i})(m_{I\to X^i}) \tag{15}$$

$$m_{X^i\to L} = \sum_{F^i} \sum_{X^i} P(X^i|F^i, L)(m_{F^i\to X^i})(m_{I\to X^i}) \tag{16}$$

The first three messages correspond to the priors imposed by the task. The rest correspond to bottom-up evidence propagated upwards within the model. The posterior probability of location (saliency map) is given by

$$P(L|I) \propto (m_{L\to X^i})(m_{X^i\to L}) \tag{17}$$

The constant of proportionality can be resolved after computing marginals over all values of the random variable. Thus, the saliency map is influenced by task dependent prior on location $P(L)$, prior on features $P(F^i|O)$ as well as the evidence from the ventral stream $m_{X^i\to L}$. Note that the summations in the message passing equations are performed over all the discrete states of the variable. Thus, L is summed over its states, $\{1, 2 \ldots, |L|\}$, F^i is summed over $\{0, 1\}$ and X^i, over states $\{0, 1, \ldots, |L|\}$. Note that the belief propagation inference converges (to the posterior) after one bottom-up and one top-down cycle.

Multiple features. When considering multiple features, the Bayesian inference proceeds as in a general polytree (Pearl, 1988). Most messages remain identical. However, the message $m_{L\to X^i}$ is influenced by the presence of other features and is now given by:

$$m_{L\to X^i} = P(L) \prod_{j\neq i} m_{X^j\to L} \tag{18}$$

Remarks:

- The mapping between the multinomial nodes/units in the model and neurons in the cortex is neither obvious nor unique. Consider a multinomial variable Y that takes states $y_1, y_2, \ldots, y_S$. A possible mapping is to S individual binary indicator variables $I_1, I_2 \ldots, I_S$, with the constraint that $(I_1 + I_2 \cdots I_S) = 1$. Then we

would map each variable I_i to an individual neuron whose firing rate is proportional to its posterior probability of being on. The constraint that only a single neuron is active may be implemented through lateral inhibition in terms of a form of divisive normalization. In this interpretation, a multinomial random variable Y corresponds to a collection of S laterally inhibited neurons such that the firing rate of neuron i represents a value proportional to its posterior probability. For binary random variables, the mapping is more direct. Each binary variable can be interpreted as a single neuron with its firing rate proportional to the posterior probability of the variable being on.

5. "Predicting the effects of attention

Here we show that the proposed model is consistent with neurophysiology experiments about the effects of feature-based and spatial attention (Bichot, Rossi, & Desimone, 2005; McAdams & Maunsell, 1999; Reynolds & Heeger, 2009). We also find that, surprisingly, several key attentional phenomena such as pop-out, multiplicative modulation and change in contrast response emerge directly, without any further assumptions or parameter tuning, as properties of the Bayesian model.

5.1. Attentional effects in V4

Within our model, V4 neurons are represented with variables $\{X^1, X^2 \ldots, X^N\}$. For analysis, we assume a single feature for simplicity. Now, consider the response of the model unit X^i given a stimulus I, which is given by

$$P(X^i|I) = \frac{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}} \quad (19)$$

Here, the term $P(I|X^i)$ represents the excitatory component – the bottom-up evidence from the input I. For example, assume that when features F^i correspond to different orientations, given the image I, for each orientation and location, $P(I|X^i)$ is set proportional to the output of an oriented Gabor filter. $P(L)$ and $P(F^i)$ serve as the attentional modulation. We make the assumption that features and location priors can be set independently based on the search task. The conditional probabilities $P(X^i|F^i,L)$ may then be interpreted as synaptic strengths, indicating how strongly locations on the feature map are affected by attentional modulation. The sum over all X^i (used to generate normalized probabilities) in the denominator can be regarded as a divisive normalization factor.

Thus, Eq. (19) may be rewritten in terms of three components: (i) an excitatory component $E(X^i) = P(I|X^i)$ (image I is observed and fixed); (ii) an attentional modulation component $A(L,F^i) = P(L)P(F^i)$; and (iii) a divisive normalization factor $S(L,F^i)$. With this notation, Eq. (19) can be rewritten as:

$$P(X^i|I) = \frac{A(F^i,L)E(X^i)}{S(F^i,L)} \quad (20)$$

Eq. (20) turns out to be closely related to a phenomenological model of attention recently proposed by Reynolds and Heeger (2009). They integrated the normalization model of neural response (Carandini et al., 1997; Simoncelli & Schwartz, 1999; Heeger, 1991; Heeger, 1992) with an early divisive inhibition model of attention (Reynolds, Chelazzi, & Desimone, 1999) to account for a variety of different effects of attention. The response of a neuron at location x and tuned to orientation θ is given by:

$$R(x,\theta) = \frac{A(x,\theta)E(x,\theta)}{S(x,\theta)+\sigma} \quad (21)$$

Here, $E(x,\theta)$ represents the excitatory component of the neuron response. $S(x,\theta)$ represents the suppressive component of the neuron response derived by pooling activity over a larger area and across all features. $A(x,\theta)$ represents the attentional modulation that enhances specific orientations and locations, based on the search task – a factor absent in the earlier normalization models. The free parameter σ determines the slope of the contrast response. Reynolds and Heeger showed that the model of Eq. (21) can reproduce key physiological effects of attention such as contrast gain behavior under different stimulus conditions. A comparison of Eq. (20) with Eq. (21) suggests that the normalization model of Reynolds and Heeger model is a special case of our model, e.g. Eq. (20). Normalization in our model emerges directly from the Bayesian formulation, instead of being an ad hoc assumption. The most notable difference between the models is based on how the pooling term $S()$ is computed. In our model, the pooling is done across all locations but within the same feature. However, in the normalization model, pooling is done across features as well.

5.2. Multiplicative modulation

5.2.1. Spatial attention

In (McAdams & Maunsell, 1999), it was observed that the tuning curve of a V4 neuron is enhanced (multiplicatively) when attention is directed to its receptive field. We observe that this effect occurs in the model. Recall that the response of a simulated neuron encoding feature i and at location x, is given by

$$P(X^i = x|I) \propto \sum_{F^i,L} P(X^i = x|F^i,L)P(I|X^i)P(F^i)P(L) \quad (22)$$

Under normal conditions, $P(L)$ and $P(F^i)$ can be assumed to have a uniform distribution and thus the response of the neuron is largely determined by the underlying stimulus ($P(I|X^i)$). Under spatial attention, the location priors are concentrated around $L = x$. This leads to a multiplicative change (from $P(L = x) = 1/|L|$ to $P(L = x) \approx 1$) that enhances the response, even under the same stimulus condition (see Fig. 4).

Reinterpreting in terms of the message passing algorithm, spatial attention corresponds to concentrating the prior $P(L)$ around the location/scale of interest (see Fig. 6b). Such a change in the prior is propagated from L to X^i (through messages in the Bayesian network). This results in a selective enhancement of all feature maps X^i for $i = 1 \cdots n$ at locations $l_1 \cdots l_m$ that overlap with the attentional spotlight $P(L)$ and in suppression everywhere else (see Fig. 6c). The message passing is initiated at the level of the L units assumed to be in parietal cortex) and should manifest itself

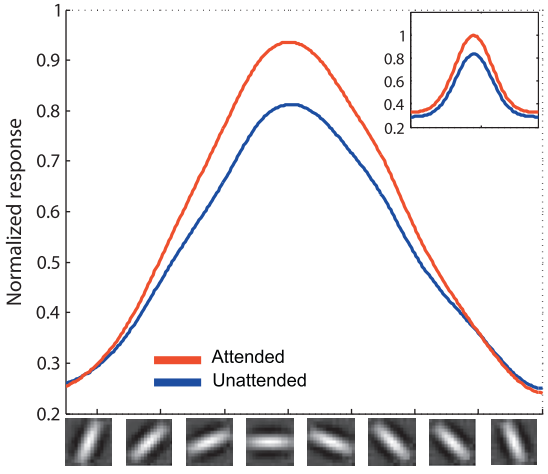


Fig. 4. Effect of spatial attention on tuning response. The tuning curve shows a multiplicative modulation under attention. The inset shows the replotted data from McAdams and Maunsell (1999).

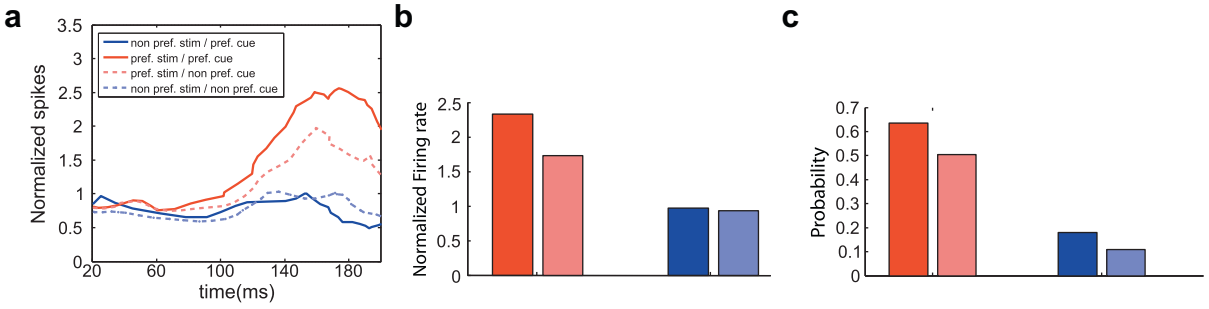


Fig. 5. (a) Effect of feature attention on neuron response (replotted from Bichot et al., 2005). (b) The time course of the neuron response is sampled at 150 ms. (c) The model predicts multiplicative modulation of the response of X^i units under attention.

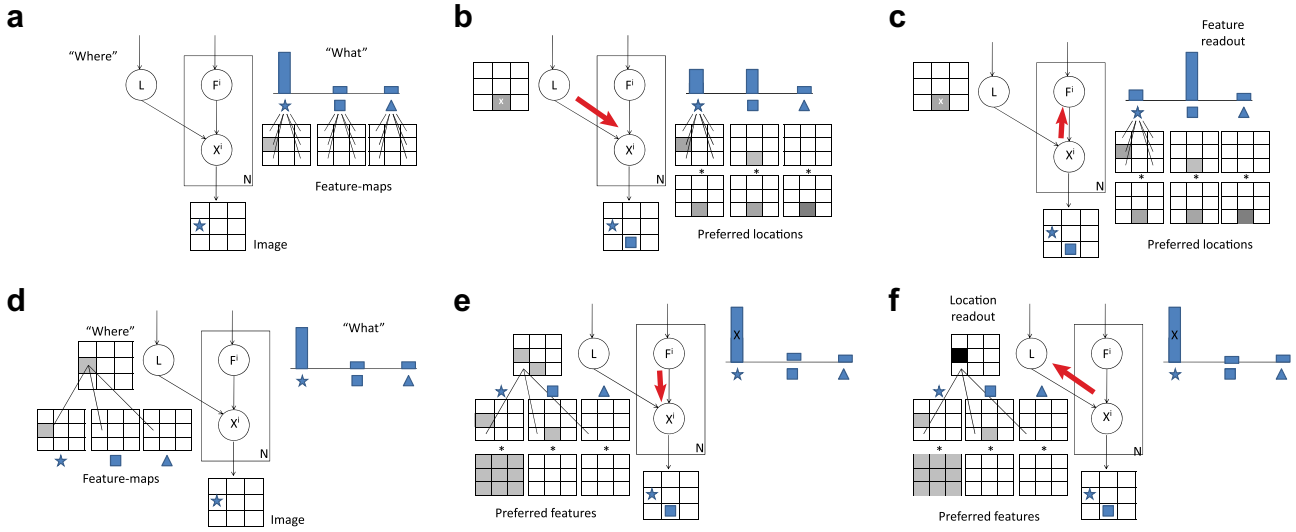


Fig. 6. Spatial and feature attention re-interpreted using message passing within the model. Spatial attention: (a) Each feature unit F^i pools across all locations from the corresponding X^i unit. (b) Spatial attention here solves the 'clutter' problem by concentrating the prior $P(L)$ around a region of interest (the *attentional spotlight*, marked '**X**') via a message passed between the L nodes in the 'where' stream and the X^i nodes in the 'what' stream. (c) Following this message passing, the feature within the spotlight can be read out from the posterior probability $P(F^i|I)$. Feature-based attention (d) Each location represented in the L unit output from all features at the same location. (e) Feature attention can be deployed by altering the priors $P(F^i)$ such that $P(F^i)$ is high for the preferred feature and low for the rest. The message passing effectively enhances the preferred features at *all* locations while suppressing other features from distracting objects. (f) The location of the preferred feature can be read out from the posterior probability $P(L|I)$.

after a short delay in the F^i units (in the ventral stream), in agreement with physiological data (Buschman & Miller, 2007).

5.2.2. Feature-based attention

Recent findings in physiology (Bichot et al., 2005) show multiplicative modulation of neuronal response under attention (see Fig. 5a and b). Units in the PFC and higher areas seem to modulate arrays of "feature detector" cells in intermediate areas of the ventral stream (PIT and V4) according to how diagnostic they are for the specific categorization task at hand. The data suggest that this modulation is effective at all locations within the receptive field. An equivalent effect is also observed in the model (see Fig. 5c). Under normal conditions, $P(L)$ and $P(F^i)$ have a uniform distribution and thus the response of the neuron is largely determined by the underlying stimulus ($P(I|X^i)$). Under feature-based attention, the feature priors are modified to $P(F^i = 1) \approx 1$. This leads to a multiplicative change (from $P(F^i = 1) = 1/2$ to $P(F^i = 1) \approx 1$) enhancing the response at all locations. The response is more pronounced when the stimulus is preferred (*i.e.*, $P(I|X^i)$ is high (see Fig. 5a–c)).

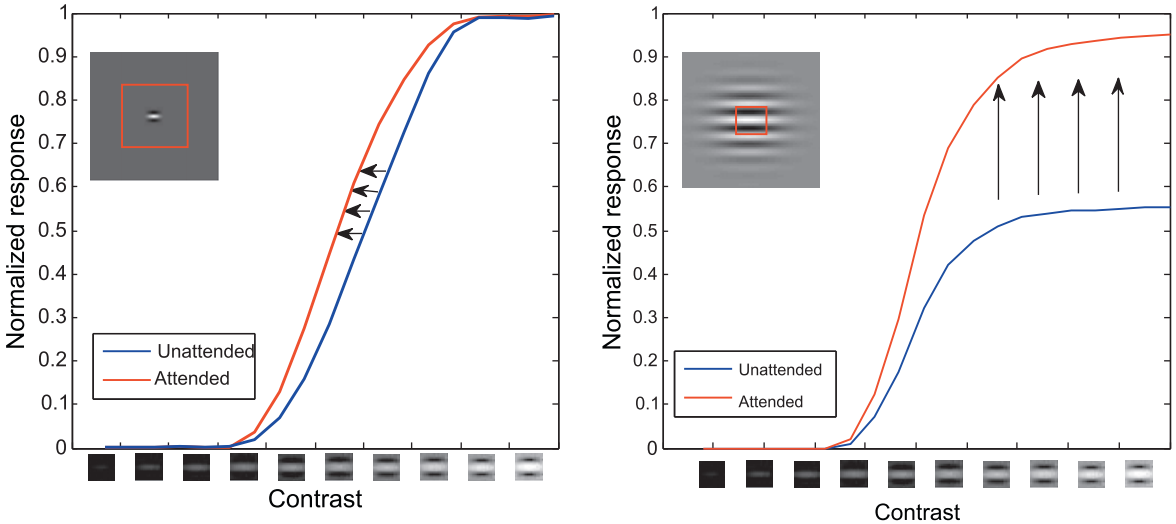
In terms of message passing, objects priors are first concentrated around the object(s) of interest (*e.g.*, (see Fig. 6d). 'pedestrian' when asked to search for pedestrians in street scenes). The change in object prior is propagated to the feature units, through the message $O \rightarrow F^i$. This results in a selective enhancement of the features that are typically associated with the target object (*e.g.*, vertical features when searching for pedestrians) and suppression of others (see Fig. 6e). This preference propagates to all feature-map locations through the message $m_{F^i \rightarrow X^i} = \sum_O P(F^i|O)P(O)$.

The L unit pools across all features X^j for $j = 1 \cdots n$ at a specific location l. However, because of the feature-based modulation, only the locations that contain features associated with the object are selectively enhanced (see Fig. 6f). Thus, priors on objects in the ventral stream activates units in the parietal cortex at locations that are most likely to contain the object of interest. The message passing is thus initiated in the ventral stream first and is manifested in the parietal cortex (L units) later, in agreement with the recent data by Buschman and Miller (2007).

5.3. Contrast response

The influence of spatial attention on the contrast response of V4 neurons has been studied extensively. Prior work showed two major, apparently contradictory, effects: in Martınez-Trujillo and

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013



(a) Shift in contrast response under attention.

(b) Response gain under attention.

Fig. 7. The model (a) exhibits shift in contrast response when the attentional spotlight is larger than the stimulus and (b) exhibits response gain modulations when the spotlight is smaller than the stimulus.

Treue (2002) and Reynolds et al. (2000) attention was shown to shift the contrast response of neurons, while in McAdams and Maunsell (1999) and Treue and Trujillo, 1999 attention was shown to induce a multiplicative gain in the contrast response of neurons. Reynolds and Heeger (2009) reconciled these differences by observing that these two experiments were performed under different stimulus conditions. In the experiments in which attention modulated contrast gain, the stimuli were smaller than those used in the experiments that showed a mixture of response and contrast gain effects. Further, the task demands in the latter experiments, which required the monkey to plan a saccade precisely to the target, may have required more focused attentional feedback.

In Fig. 7a and b we show that our Bayesian model, as expected given its close relation with Reynolds and Heeger's model, is also consistent with the observed dependency of contrast on attention. In this simulation, the response without attention is assumed to depend on contrast (the bottom-up evidence $P(I|X^1 \cdots X^n)$ is directly derived from the outputs of oriented Gabor filters operating on images of varying contrast). The Bayesian model "predicts" how the contrast response changes with attention.

6. Predicting eye movements during free viewing and visual search tasks

Human and animal studies (see Wolfe (2007) for a recent review) have isolated at least three main components used to guide the deployment of eye movements. First, studies have shown that image-based *bottom-up* cues can capture attention, particularly during free viewing conditions.[3] Second, task dependence also plays a significant role in visual search (Wolfe, 2007; Yarbus, 1967).[4] Third, structural associations between objects and their locations within a scene (*contextual cues*) have been shown to play a significant role in visual search and object recognition (Torralba, 2003b).

How the visual system combines these cues and what the underlying neural circuits are, remain largely unknown. Here we show that our model, which combines bottom-up as well as top-down cues within a probabilistic Bayesian framework, can predict well human eye movements in complex visual search tasks as well as in free viewing conditions.

6.1. Free viewing

Here we evaluate the performance of the model in a task-free scenario where attention is purely bottom-up and driven by image salience. We used images and eye-movement data provided by (Bruce & Tsotsos, 2006). The dataset consists of 120 images containing indoor and outdoor scenes with at least one salient object in each image. The images were presented to 20 human subjects in random order and all the eye movements made within the first four seconds of presentation were recorded using an infrared eye tracker. In their work, Bruce and Tsostos used low level filters derived by performing ICA (Bell & Sejnowski, 1995) on color image patches to generate feature maps. The visual salience of each position is derived from self information. In contrast to low level filters, our approach uses higher level shape-tuned features and color information (see Supplementary Online Methods Section 2.5).

There are at least two measures that have been used to compare models of attention to human fixations: normalized scan path saliency (*NSS*) from Peters and Itti (2007) and fixations in the most salient region (*FMSR*) from Bruce and Tsotsos (2006) and Torralba et al., 2006. For brevity, we only report results using the *FMSR* measure, but qualitatively similar results were obtained for *NSS*. For each stimulus and task, we calculated an *FMSR* value by first thresholding the computed saliency map, retaining only the most salient pixels (see Fig. 8). The *FMSR* index corresponds to the percentage of human fixations that fall within this most salient region. A higher value indicates better agreement with human fixations. We generated an ROC curve by continuously varying the threshold. The area under the ROC curve provides a summary measure of the agreement with human observers. We compare our Bayesian approach with two baseline algorithms (see Table 3).[5]

[3] A measure that has been shown to be particularly relevant is the local image salience (*i.e.*, the local feature contrast), which corresponds to the degree of conspicuity between that location and its surround (Itti & Koch, 2001).

[4] Evidence for *top-down feature-based* attention comes from both imaging studies in humans (Kanwisher & Wojciulik, 2000) as well as monkey electrophysiology studies (Maunsell & Treue, 2006).

[5] Since the fixation data were pooled from all subjects, it is not possible to compare inter-subject consistency or provide error intervals for this data.

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013

The results show that the Bayesian attention model using shape-based features can predict human eye movements better than approaches based on low level features.

6.2. Search for cars and pedestrians

We manually selected 100 images (containing cars and pedestrians) from the CBCL Street-scene database (Bileschi, 2006), while an additional 20 images that did not contain cars or pedestrians were selected from *LabelMe* (Russell, Torralba, Murphy, & Freeman, 2008). These 120 images were excluded from the training set of the model. On average, images contained 4.6 cars and 2.1 pedestrians. The images (640×480 pixels) were presented at a distance of about 70 cm, roughly corresponding to $16^\circ \times 12^\circ$ of visual angle.

We recruited eight human subjects (age 18–35) with normal or corrected-to-normal vision. Subjects were paid and gave informed consent. Using a block design (120 trials per block), participants were asked to either count the number of cars or the number of pedestrians. Task and presentation order were randomized for each subject. Every image was presented twice: once for pedestrians and once for cars. No instructions regarding eye movements were given, except to maintain fixation on a central cross in order to start each trial. Each image was then presented for a maximum of 5 s, and within this time observers had to count the number of targets (cars or pedestrians) and press a key to indicate completion.

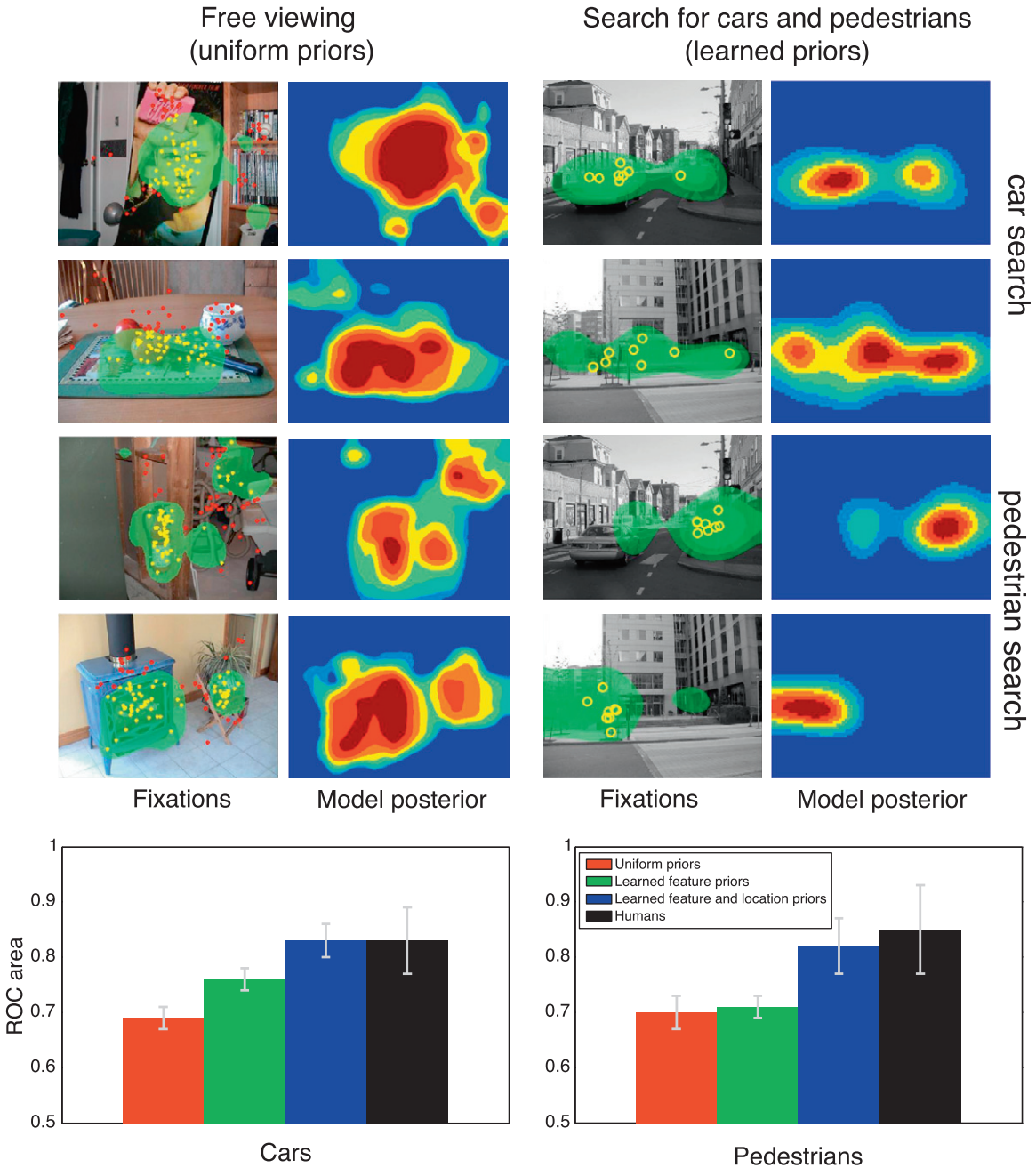


Fig. 8. Predicting human eye movements: (a) Agreement between the model and human eye fixations during free viewing (left) and a complex visual search for either cars or pedestrians. Sample images overlaid with most salient (top 20%) regions predicted by the model (green) along with human eye movements (yellow: agree with prediction, red: not predicted by model) and corresponding model posteriors (*i.e.*, predicted image saliency). (b) Model performance at predicting human eye fixations during visual searches. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013

Table 3
Comparison of the proposed Bayesian model with shape-based features with prior work that relies on low level features.

Models	Agreement with humans (ROC area)
Bruce and Tsotsos (2006)	0.728
Itti et al. (1998)	0.727
Proposed model	**0.779**

Table 4
Comparison between the performance of the various models to *localize* objects. The values indicate the area under the ROC.

	Car	Pedestrian
Bottom-up (Itti & Koch, 2001)	0.437	0.390
Context (Torralba et al., 2006)	0.800	0.763
Model/uniform priors	0.667	0.689
Model/learned spatial priors	0.813	0.793
Model/learned feature priors	0.688	0.753
Model/full	**0.818**	**0.807**

Subjects then verbally reported the number of targets present, and this was recorded by the experimenter. We verified that reported counts agreed well with the actual number of targets. We used an ETL 400 ISCAN table-mounted, video-based eye tracking system to record eye position during the course of the experiment. Eye position was sampled at a rate of 240 Hz with an accuracy of about 0.5° of visual angle.

Training the model to attend to specific objects or object classes corresponds to estimating the probability distribution $P(F^i|O)$). In practice, this is done by computing feature maps for a set of training images. The corresponding feature maps are discretized to maximize classification accuracy following Fleuret (2004). The feature F^i is said to be present if its detected at any location in the feature map. $P(F^i|O)$ is determined by simply counting the frequency of occurrence of each feature. Since scenes of streets obey strong constraints on where the objects of interest may be found, it is important to use not only feature priors but also priors over object location. We follow a procedure outlined in Torralba (2003a) for this purpose. Given the image, we compute the 'gist' (or global summary) of the scene in a deterministic manner. We use a mixture-of-regressors as in Murphy, Torralba, and Freeman, 2003 to learn the mapping between the context features and location/scale priors for each object. Details about how the model was trained for the task are provided in the Supplementary Online Methods, Section 2.5.

As assumed in several previous psychophysical studies (Itti and Koch, 2001; Rao et al., 2002; Torralba, Oliva, Castelhano, and Henderson, 2006), we treat eye movements as a proxy for shifts of attention. To calculate inter-subject consistency, we generated a saliency map by pooling fixations from all but one subject in a manner similar to Torralba et al. (2006), and then tested the left-out subject on this map. Thus, inter-subject consistency measures performance by a model constructed from human fixations, which is regarded here as an "ideal model".

Fig. 8b shows the agreement between the model (and how the location and feature priors influence performance) and human observers for the first fixation. Tables S2 and S3 provide comparisons for additional number of fixations and against other models of eye movements (Itti and Koch, 2001; Torralba et al., 2006). Our results suggest that our Bayesian model of attention accounts relatively well for the very first fixations (especially for cars, see Fig. 8b). Beyond the first saccade, the agreement between model and human fixations decreases while the inter subject agreement increases (see Tables S2 and S3). The higher relative contribution of the context (*i.e.,* learned location priors) to the overall prediction is not surprising, since street scenes have strong spatial constraints regarding the locations of cars and pedestrians. We found that using image based saliency cues, corresponding to setting all the priors to be uniform (see also the bottom-up saliency model (Itti & Koch, 2001) in Tables S2 and S3), does worse than chance. Learning either spatial priors or feature priors improves the agreement between models and humans significantly. In addition, learning priors for both cues does better than either in isolation. The model agrees at the 92% level with human eye fixations on both pedestrian and car search tasks (measured in terms of the overlap between ROC areas for the first three fixations). Recently, Ehinger, Hidalgo-Sotelo, Torralba, and Oliva (2009) used a combination of feature bias, *gist* and bottom-up saliency to achieve similar predictive performance. The inconsistency between human subjects and the model may be due to higher-level abstract information available to humans but not to the model. Humans routinely utilize higher level visual cues (*e.g.*, location of ground-plane) as well non-visual information (*e.g.*, pedestrians are found on pavements and cross walks) while examining a visual scene.

Previous work has shown that attention is useful in priming object detection (Navalpakkam and Itti, 2006; Torralba, 2003a), pruning interest points (Rutishauser et al., 2004), quantifying visual clutter (Rosenholtz and Mansfield, 2005) and predicting human eye movements (Oliva, Torralba, Castelhano, and Henderson, 2003). Here we provide a quantitative evaluation of the proposed model of attention for detecting objects in images as opposed to predicting human eye movements. Table 4 shows the percentage of object locations that are correctly predicted using different cues and models. An object was considered to be correctly detected if its center lay in the thresholded saliency map. An ROC curve can be obtained by varying the threshold on the saliency measure. The area under the ROC curve provides an effective measure of the predictive ability of the individual models. The context (gist) representation derived from shape-based units (Serre, Kouh, et al., 2005) perform better than the representation based on simple oriented features (Torralba et al., 2006). As expected, bottom-up cues derived using shape-based features perform better than bottom-up saliency obtained using simple oriented features (Itti and Koch, 2001).

7. Discussion

7.1. Relation to prior work

A few theories and several specific models (see Tables S4 and S5 for an overview and comparison with our approach) have been proposed to explain the main functional roles of visual attention and some of its properties. An influential proposal by Tsotsos (1997) maintains that attention reflects evolution's attempt to fix the processing bottleneck in the visual system (Broadbent, 1958) by directing the finite computational capacity of the visual cortex preferentially to relevant stimuli within the visual field while ignoring everything else. Treisman and Gelade (1980) suggested that attention is used to *bind* different features (*e.g.*, color and form) of an object during visual perception. Desimone (1998) suggested that the goal of attention is to bias the choice between competing stimuli within the visual field. These general proposals, though correct and groundbreaking, do not yield detailed insights on how attention should be implemented in the visual cortex and do not yield direct predictions about the various behavioral and physiological effects of attention. Other, more specific models exist, each capable of modeling a different effect of attention. Behavioral effects include pop-out of salient objects (Itti et al., 1998; Rosenholtz and Mansfield, 2005; Zhang et al., 2008), top-down bias of target features (Navalpakkam and Itti, 2006; Wolfe, 2007), influence from scene context (Torralba, 2003b), serial vs.

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013

parallel-search effect (Wolfe, 2007), etc. Physiological effects include multiplicative modulation of neuron response under spatial attention (Rao, 2005) and feature-based attention (Bichot et al., 2005). This paper describes a possible unifying framework that defines a computational goal for attention, derives possible algorithmic implementations and predicts its disparate effects listed above.

7.2. Our theory

The theoretical framework of this paper assumes that one goal of vision is to solve the problem of *what is where*. Attention follows from the assumption that this is done sequentially, one object at a time. It is a reasonable conjecture that the sequential strategy is dictated by the intrinsic sample complexity of the problem. Solving the 'what' and 'where' problem is especially critical for recognizing and finding objects in clutter. In a probabilistic framework, the Bayesian graphical model that emerges from the theory maps into the basic functional anatomy of attention involving the ventral stream (V4 and PIT) and the dorsal stream (LIP and FEF). In this view, attention is not a visual routine, but is the inference process implemented by the interaction between ventral and dorsal areas within this Bayesian framework. This description integrates bottom-up, feature-based and context-based attentional mechanisms.

7.3. Limitations

In its current form, the model has several limitations: (i) The model we have implemented only accounts for effects of attention in areas V4, IT and LIP. The lower regions are assumed to be purely feedforward. However, studies have shown that some attentional effects can be found even in areas V1 (Hegde and Felleman, 2003). These effects may be accounted for by extending the Bayesian framework to lower areas at the expense of computational complexity. (ii) The model currently uses a static inference scheme and thus cannot model dynamic effects of attention. In particular, it is likely that the saliency map is updated after each shift of attention – currently not represented in the model. (iii) The model currently does not account for effects that can be explained by spatial organization (Li and Snowden, 2006). (iv) Understanding how the brain represents uncertainty is one of the open questions in neuroscience. Here our assumption (which is relevant for some of the comparisons with neural data) is that neurons represent uncertainty using probabilities – firing rates of neurons directly represent probability estimates. A similar but more indirect mappings (using population responses) have been assumed before (Rao, 2004; Zemel et al., 1998). It is surprising that a model with such limitations can predict such a variety of attentional phenomena.

7.4. Validation

We checked that the theory and the associated model predicts well human psychophysics of eye movements (which we consider a proxy for attention) in a task-free as well as in a search task scenario. In a task-free scenario the model, tested on real world images, outperforms existing 'saliency' models based on low-level visual features. In a search task, we found that our model predicts human eye movements better than other, simpler models. Finally the same model predicts – surprisingly – a number of psychophysical and physiological properties of attention that were so far explained using different, and somewhat *ad hoc* mechanisms.

Acknowledgments

The authors wish to thank Aude Oliva and Barbara Hidalgo-Sotelo for the use of, as well as, help with the eye tracker. We also thank the reviewers for their helpful suggestions in improving this paper. This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by Grants from DARPA (FA8650-06-C-7632 and FA8650-09-1-7946). Additional support was provided by: Honda Research Institute USA, NEC, Sony and especially by the Eugene McDermott Foundation.

The views, opinions, and/or findings contained in this article/presentation are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.visres.2010.05.013.

References

Amit, Y., & Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vision Research, 43*(19), 2073–2088.
Beck, J., & Pouget, A. (2007). Exact inferences in a neural implementation of a hidden Markov model. *Neural Computation, 19*(5), 1344–1361.
Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159.
Bichot, N., Rossi, A., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science, 308*(5721), 529–534.
Bileschi, S. M. (2006). StreetScenes: Towards scene understanding in still images. Ph.D. Thesis, MIT.
Bisley, J., & Goldberg, M. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science, 299*(5603), 81–86.
Broadbent, D. E. (1958). *Perception and communication*.
Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems, 18*, 155.
Buschman, T., & Miller, E. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science, 315*(5820), 1860.
Carandini, M., Heeger, D., & Movshon, J. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience, 17*(21), 8621–8644.
Colby, C., & Goldberg, M. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience, 22*(1), 319–349.
Crandall, D., Felzenszwalb, P., & Huttenlocher, D. (2005). Spatial priors for part-based recognition using statistical models. In *Proc. IEEE computer vision and pattern recognition* (Vol. 1).
Dayan, P., & Zemel, R. (1999). Statistical models and sensory attention. In *Proceedings of the international conference on artificial neural networks* (p. 2).
Dayan, P., Hinton, G. E., & Neal, R. M. (1995). The Helmholtz machine. *Neural Computation, 7*, 889–904.
Dean, T. (2005). A computational model of the cerebral cortex. In *National conference on artificial intelligence* (Vol. 20, p. 938).
Deco, G., & Rolls, E. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research, 44*(6), 621–642.
Deneve, S. (2008). Bayesian spiking neurons I: Inference. *Neural Computation, 20*(1), 91–117.
Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society*.
Desimone, R., & Schein, S. J. (1987). Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *Journal of Neurophysiology, 57*(3), 835–868.
Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*.
Epshtein, B., Lifshitz, I., & Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. In *Proc. of the national academcy of sciences*.
Felzenszwalb, P., & Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision, 61*(1), 55–79.
Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE computer vision and pattern recognition*.
Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *The Journal of Machine Learning Research, 5*, 1531–1555.
Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*, 312–316.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16*(9), 1325–1352.
Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.
Gallant, J. L., Connor, C. E., Rakshit, S., Lewis, J. W., & Van Essen, D. C. (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of Neurophysiology, 76*(4), 2718–2739.
Gao, D., & Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In *Proc. of the international conference on computer vision*.
George, D. (2008). How the brain might work: A hierarchical and temporal model for learning and recognition. Ph.D. Thesis, Stanford University.
George, D., & Hawkins, J. (2005). A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In *International joint conference on neural networks* (Vol. 3).
Gilks, W., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.
Grossberg, S. (1999). How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision, 12*(2), 163–185.
Heeger, D. J. (1991). Nonlinear model of neural responses in cat visual cortex. *Computational models of visual processing, 2*, 119–133.
Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*(2), 181–197.
Hegde, J., & Felleman, D. J. (2003). How selective are V1 cells for pop-out stimuli? *Journal of Neuroscience, 23*(31).
Hegde, J., & Van Essen, D. (2000). Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience, 20*(5), 61.
Hinton, G. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*(10), 428–434.
Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology, 148*, 574–591.
Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *Journal of Neuroscience, 24*(13), 3313–3324.
Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews on Neuroscience, 2*(3), 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*(11).
Kanwisher, N., & Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nature Reviews on Neuroscience, 1*(2), 91–100.
Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference.
Kersten, D., & Yuille, A. (2003a). Bayesian models of object perception. *Current Opinion in Neurobiology, 13*(2), 150–158.
Kersten, D., & Yuille, A. (2003b). Bayesian models of object perception. *Current Opinion in Neurobiology, 13*(2), 150–158.
Knill, D., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge Univ. Pr.
Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71*, 856–867.
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.
Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*.
Li, Z., & Snowden, R. (2006). A theory of a saliency map in primary visual cortex (V1) tested by psychophysics of colour–orientation interference in texture segmentation. *Visual Cognition, 14*(4), 911–933.
Litvak, S., & Ullman, S. (2009). Cortical circuitry implementing graphical models. *Neural Computation*.
Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*, 552–563.
Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19*, 577–621.
Lovejoy, W. (1991). A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research, 28*(1), 47–65.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt and Co., Inc..
Martınez-Trujillo, J., & Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron, 35*(2), 365–370.
Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neuroscience, 29*(6), 317–322.
McAdams, C., & Maunsell, J. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience, 19*(1), 431–441.
Mel, B. W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation, 9*, 777–804.
Miau, F., & Itti, L. (2001). A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what. In *Proc. IEEE engineering in medicine and biology society* (pp. 789–792).
Monahan, G. (1982). A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 1–16.
Mumford, D. (1992). On the computational architecture of the neocortex – II: The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.
Murphy, K., Torralba, A., & Freeman, W. (2003). Using the forest to see the trees: A graphical model relating features, objects and scenes. *Advances in Neural Information Processing Systems, 16*.
Murray, J. F., & Kreutz-Delgado, K. (2007). Visual recognition and inference using dynamic overcomplete sparse learning. *Neural Computation, 19*(9), 2301–2352.
Mutch, J., & Lowe, D. (2006). Multiclass object recognition with sparse, localized features. In *Proc. IEEE computer vision and pattern recognition*.
Navalpakkam, V., & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Proc. IEEE computer vision and pattern recognition*.
Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical report, University of Toronto.
Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *International conference on image processing*.
Oram, M., & Perrett, D. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology, 68*, 70–84.
Pasupathy, A., & Connor, C. (2001). Shape representation in area V4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology, 86*(5), 2505–2519.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers.
Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proc. IEEE computer vision and pattern recognition*, Minneapolis, MN.
Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews on Neuroscience, 1*(2), 125–132.
Ranzato, M., Huang, F., Boureau, Y., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. IEEE computer vision and pattern recognition*.
Rao, R. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation, 16*(1), 1–38.
Rao, R. (2005). Bayesian inference and attentional modulation in the visual cortex. *NeuroReport, 16*(16), 1843–1848.
Rao, R., Olshausen, B., & Lewicki, M. (2002). *Probabilistic models of the brain: Perception and neural function*. The MIT Press.
Reynolds, J., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in Macaque Areas V2 and V4. *Journal of Neuroscience, 19*(5), 1736.
Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron Review, 61*, 168–184.
Reynolds, J., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26*(3), 703–714.
Riesenhuber, M., & Poggio, T. (1999b). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.
Rosenholtz, R. (1985). A simple saliency model predicts a number of motion poput phenomena. *Human Neurobiology, 39*(19), 3157–3163.
Rosenholtz, R., & Mansfield, J. (2005). Feature congestion: a measure of display clutter. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 761–770). New York, NY, USA: ACM.
Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1), 157–173.
Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? In *Proc. IEEE computer vision and pattern recognition* (Vol. 2).
Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., & Poggio, T. (2005b). A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. MIT AI Memo 2005-036/CBCL Memo 259.
Serre, T., Wolf, L., Bileschi, S., Reisenhuber, M., & Poggio, T. (2007c). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
Simoncelli, E. P., & Schwartz, O. (1999). Modeling surround suppression in V1 neurons with a statistically derived normalization model. *Advances in Neural Information Processing Systems*, 153–159.
Smallwood, R., & Sondik, E. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 1071–1088.
Steimer, A., Maass, W., & Douglas, R. (2009). Belief propagation in networks of spiking neurons. *Neural Computation, 21*, 2502–2523.
Sudderth, E., Torralba, A., Freeman, W., & Willsky, A. (2005). Learning hierarchical models of scenes, objects, and parts. In *Proc. IEEE computer vision and pattern recognition* (Vol. 2, pp. 1331–1338).
Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109–139.
Thorpe, S. (2002). Ultra-rapid scene categorisation with a wave of spikes. *Proceedings of Biologically Motivated Computer Vision*, 1–15.
Torralba, A. (2003a). Contextual priming for object detection. *International Journal of Computer Vision, 53*(2), 169–191.
Torralba, A. (2003b). Modeling global scene factors in attention. *Journal of Optical Society of America, 20*(7), 1407–1418.
Torralba, A., Murphy, K., & Freeman, W. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. IEEE Computer Vision and Pattern Recognition* (Vol. 2).
Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013

scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786.

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

Treue, S., & Trujillo, J. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature, 399*, 575–579.

Tsotsos, J. (1997). Limited capacity of any realizable perceptual system is a sufficient reason for attentive behavior. *Consciousness and cognition, 6*(2–3), 429–436.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687.

Ungerleider, L., & Haxby, J. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology, 4*(2), 157–165.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. *Analysis of Visual Behavior, 549*, 586.

Van Der Velde, F., & De Kamps, M. (2001). From knowing what to knowing where: Modeling object-based attention with feedback disinhibition of activation. *Journal of Cognitive Neuroscience, 13*(4), 479–491.

Wainwright, M., & Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning, 1*(1–2), 1–305.

Wallis, G., & Rolls, E. T. (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology, 51*, 167–194.

Walther, D., & Koch, C. (2007). *Computational Neuroscience: Theoretical insights into brain function* (Progress in Brain Research).

Weiss, Y., Simoncelli, E., & Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*(6), 598–604.

Wersing, H., & Koerner, E. (2003). Learning optimized features for hierarchical models of invariant recognition. *Neural Computation, 15*(7), 1559–1588.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. *Integrated Models of Cognitive System*, 99–119.

Yarbus, A. L. (1967). *Eye movements and vision*. Plenum press.

Yu, A., & Dayan, P. (2005). Inference, attention, and decision in a Bayesian neural architecture. *Advances in Neural Information Processing Systems, 17*, 1577–1584.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation, 10*(2), 403–430.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 1–20.

Please cite this article in press as: Chikkerur, S., et al. What and where: A Bayesian inference theory of attention. *Vision Research* (2010), doi:10.1016/j.visres.2010.05.013