What and Where: A Bayesian Inference Theory of Visual Attention

by

Sharat Chikkerur

M.S., University at Buffalo (2005)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author .....
Department of Electrical Engineering and Computer Science
June 10, 2010

Certified by .....
Tomaso Poggio
Eugene McDermott Professor
McGovern Institute
CSAIL, Brain Sciences Department
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Theses

What and Where: A Bayesian Inference Theory of Visual Attention

by

Sharat Chikkerur

Submitted to the Department of Electrical Engineering and Computer Science
on June 10, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

In the theoretical framework described in this thesis, 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 – 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. Finally, we demonstrate that the same model, used to modulate information in an existing feedforward model of the ventral stream, improves its object recognition performance in clutter.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor
McGovern Institute
CSAIL, Brain Sciences Department

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

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 solves routinely 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 and Mishkin, 1982; Ungerleider and Haxby, 1994). 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. 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 work is that visual attention performs this computation (see also (Van Der Velde and De Kamps, 2001; Deco and Rolls, 2004a)).

However, explaining the role of attention is but a small part of understanding visual attention in the brain. The past four decades of research in visual neuroscience have generated a large and disparate body of literature on attention. Several theoretical proposals and computational models have been described to try to explain the main functional and computational role of visual attention. One important proposal by Tsotsos (1997) is 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 system 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. Duncan (1995) suggested that the goal of attention is to bias the choice between competing stimuli within the visual field. These proposals however remain agnostic about how attention should be implemented in the visual cortex and do not yield any prediction about the various behavioral and physiological effects of attention.

On the other hand, computational models attempt to model specific behavioral and physiological effects of attention. Behavioral effects include pop-out of salient objects (Itti *et al.*, 1998; Zhang *et al.*, 2008; Rosenholtz and Mansfield, 2005), top-down bias of target features (Wolfe, 2007; Navalpakkam and Itti, 2006), influence from scene context (Torralba, 2003b), serial vs. 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). A unifying framework that provides a computational goal for attention and at the same time accounts for the disparate effects listed above is missing.

1.2 Our contributions

We present a theory where attention is part of the inference process that determines "what" is "where" in the visual scene. Based on the theory and a few assumptions, we derive a computational model that can be mapped to the anatomy of attentional processing in the brain. Finally, we validate the model experimentally and show that it is consistent with physiological effects and human behavior. In the following, we list our specific contributions.

1.2.1 Theory

- Recently, it has been suggested that visual perception can be interpreted as a bayesian inference process where top-down signals are used to disambiguate noisy bottom-up

sensory input signals (Mumford, 1992; Dayan *et al.*, 1995; Knill and Richards, 1996; Dayan and Zemel, 1999; Weiss *et al.*, 2002; Rao *et al.*, 2002a; Rao, 2004; Lee and Mumford, 2003; Kersten and Yuille, 2003a; Kersten *et al.*, 2004; Friston, 2003; George and Hawkins, 2005; Dean, 2005; Murray and Kreutz-Delgado, 2007; Hinton, 2007; Epshtein *et al.*, 2008). Extending this idea, we propose that attention can also be regarded as an inference process that disambiguates form and location information (Yu and Dayan, 2005; Rao, 2005).

- In the theoretical framework proposed here, we suggest that attention is part of the visual inference process that solves the problem of *what is where*. Spatial attention emerges as a strategy to reduce the uncertainty in shape information while feature-based attention reduces the uncertainty in spatial information. Feature-based 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. The theory explains attention not as a primary mechanism (or a visual routine (Ullman, 1984)), but as an effect of interaction between the 'what' and 'where' streams within this inference framework.

- Conceptual models or theories of attention proposed so far explain the role of attention, but not how it is implemented in the brain. Computational models, on the other hand are far removed from how attention works in the brain. Phenomenological models explain individual physiological phenomena in the brain but do not explain the role of attention in the broader scheme of visual perception. The proposed work bridges the gap between conceptual, computational and phenomenological models of attention. 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.

1.2.2 Computational model of attention

- Starting from a bayesian view of visual perception and assumptions motivated by sample complexity and biology, we derive a specific computational model and relate it to the known anatomy of attention. In the model, visual perception is viewed as infer-

ence of object location and identity given an image. Top-down spatial and feature-based attention can be interpreted as priors/biases that are *set* before the image is presented. The posterior probability of location serves as a *saliency* map that predicts where the attentional spotlight is shifted next.

- The model mimics attentional processing in the brain both in terms of structure as well as behaviour. Within this generative model, the object's identity and location are modelled as being marginally independent. This mimics the separation of 'what' (ventral) and 'where' (dorsal) streams in the human visual system. This segregation raises the question of how object location and identity are tied together during perception. In the proposed model, the interaction between location and identity information occurs through hidden variables that encode positions of individual features. This is similar to regions in the brain (such as V4) that mediate interactions between ventral and dorsal streams.

- The computational model is closely related to the bayesian model of spatial attention proposed by Rao (2005). Here, we significantly extend the model to include feature-based attention in addition to spatial attention. Prior work modelled a single feature dimension with mutually exclusive features. On the other hand, we model conjunction of features that share a common spatial modulation allowing us to model complex search tasks. In addition to reproducing classical results on simple stimuli, we show applications of the model on complex real-world stimuli.

- We show that the bayesian model is also consistent (and in fact equivalent) with phenomenological models such as the normalization model of attention (Reynolds and Heeger, 2009) that can explain several (even conflicting) physiological effects of attention. In contrast to phenomenological models that are *designed* to mimic biological phenomena, we show that the same phenomena emerge naturally as 'predictions' of the model.

- We show that the model exhibits properties consistent with experimental evidence from physiology such as:

    1. Spatial invariance: The response of the model units that encode features are invariant to position of the feature within the image. This property is consistent with

position invariant neurons found in region IT of the ventral stream.

2. Multiplicative modulation: The response of the model units that jointly encode position and feature are modulated under spatial and feature-based attention. This is consistent with region V4 where neuron activities are modulated under attention (McAdams and Maunsell, 1999; Reynolds *et al.*, 1999, 2000; Bichot *et al.*, 2005).

3. Shift in contrast response: It was shown that attention can shift the contrast response curve of neurons (Martınez-Trujillo and Treue, 2002; Reynolds *et al.*, 2000). Other experiments (McAdams and Maunsell, 1999; Treue and Trujillo, 1999a) reported a mulitiplicative gain. Reynolds and Heeger (2009) reconciled these contradictory observations using a normalization model of attention. We show that our model is consistent with both of these experiments.

4. Visual pop-out effect: Psychophysics has shown that parts of an image that are distinct from rest of the image in any feature dimension 'pops-out' and attracts attention independent of the search task. Computational models proposed so far compute an explicit measure of this difference to simulate pop-out (Itti *et al.*, 1998; Gao and Vasconcelos, 2007; Zhang *et al.*, 2008; Rosenholtz, 1985). In contrast, pop-out effect emerges as a natural consequence ('prediction') of our model.

- The model we propose is general in the choice of features used. Using oriented features, the model can reproduce attentional effects found in area V4 as discussed before. When extended to motion features, the model can reproduce attentional effects found in area MT (a region in the dorsal stream that processes motion information) (Treue and Trujillo, 1999b; Beauchamp *et al.*, 1997). Using a combination of color and motion features, we can also explain the interaction between these features (Womelsdorf *et al.*, 2006) that has not been modelled before. Finally, when we use complex shape-based features, we show that the model can predict human eye movements on complex real world images.

1.2.3 Experimental validation of the model

- We test the model on complex real world images through two specific experiments: (i) We show that the model predicts well human eye fixations in natural scenes. (ii) We

demonstrate that the same model, used to modulate information in an existing feedforward model of the ventral stream, improves its object recognition performance in clutter.

- **Modeling eye movements:** Human eye movements can be considered as a proxy for shifts of attention. Also, modeling eye movements has been shown to be useful in priming object detection (Navalpakkam and Itti, 2006; Torralba, 2003a), pruning interest points (Rutishauser *et al.*, 2004) and quantifying visual clutter (Rosenholtz and Mansfield, 2005). Previous work in attention and eye movements has focused on free viewing conditions (Itti and Koch, 2001a; Parkhurst *et al.*, 2002; Peters *et al.*, 2005; Bruce and Tsotsos, 2006) where attention is driven by purely bottom-up information. In reality, top-down effects from the search task can heavily influence attention and eye movements(Yarbus, 1967). In this work, we outline a visual attention model where spatial priors imposed by the scene and the feature priors imposed by the target object are combined in a Bayesian framework to generate a task-dependent *saliency* map. In the absence of task-dependent priors, the model operates in a purely bottom-up fashion.

- **Object recognition in clutter:** The human visual system can recognize several thousand object categories irrespective of their position and size (over some finite range). This combination of selectivity and invariance is achieved by pooling responses from afferents in the previous stage. The cost of this tolerance to position and scale transformations is susceptibility to crowding and clutter. When multiple objects or background clutter are present simultaneously within the receptive field of a neuron, the stimuli compete with each other for representation at a higher layer. This effect has been observed in all stages of the visual processing (Reynolds *et al.*, 1999; Zoccolan *et al.*, 2007), human psychophysics as well as computational models (Serre *et al.*, 2007a). A natural hypothesis – that we adopt here – is that an attentional spotlight may be used to suppress responses from distracting stimulus while enhancing those of the target stimulus.

1.3 Outline

The rest of this thesis is organized as follows. In chapter 2, we provide details about our computational model and describe a tentative mapping to biology. In addition, we illustrate the properties of the model, both designed and emergent. In chapter 3, we show that the model can ”predict” physiological effects of attention. In chapter 4, we use the model to predict human eye movements on complex real world images. In chapter 5, we show that feature-based and spatial attention can be used in conjunction to recognize objects in cluttered visual scenes. We demonstrate that the model can explain improvement in object recognition performance under the influence of attention. Finally in chapter 6, we discuss the limitations of the model and future directions of research.

Chapter 2

Computational framework

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, objects present etc.). 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):

$$P(S, I) = P(I|S)P(S). \tag{2.1}$$

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

Thus, $S = \{O_1, O_2, \cdots, O_n, L_1, L_2, \cdots, L_n\}$. 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. 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 $P(S, I) = P(O_1, L_1, O_2, L_2, \cdots, O_n, L_n, I)$. We can then integrate out the confounding variables (*i.e.*, all objects except one – labeled, without loss in generality, O_1):

$$P(O_1, L_1, I) = \sum_{O_2 \cdots O_n, L_2 \cdots L_n} P(O_1, L_1, O_2 \cdots, O_n, L_2, \cdots, L_n, I). \tag{2.2}$$

We further assume that

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

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

$$P(O, L, I) = P(O)P(L)P(I|L, O). \tag{2.3}$$

In Eq. 2.3 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. 2.3. In a sense, our framework with the key assumption (a), "predicts" attention and – with the approximations to Eq. 2.3 described in the rest of the section – several of its properties.

- Bayesian models of object recognition – but, emphatically, not of attention – assume different (wrt Eq.2.3) factorizations of $P(S, I)$, such as $P(S, I) = P(O_1, L_1, \cdots, O_n, L_n, I)$ (Sudderth *et al.*, 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. 2.3, I corresponds to an entire array of measurements (every feature at every location). Eq. 2.3, 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 (Tsotsos, 1997; Bruce and Tsotsos, 2006; Itti *et al.*, 1998) 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.

Let us assume that ϵ is the error with which the non-parameteric 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 where 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 $\epsilon = 0.1$)!

- Eq. 2.3 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 2-5). 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'. Notice 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 and Itti, 2001; Rutishauser *et al.*, 2004; Walther and Koch, 2007) and also motivates mechanisms such as "inhibiton of return" (Itti and Koch, 2001a). 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 (Monahan, 1982; Smallwood and Sondik, 1973; Lovejoy, 1991).

2.1.1 Model

Consider the generative model specified in Eq. 2.3. 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 *et al.*, 2005; Felzenszwalb and Huttenlocher, 2005; Fergus *et al.*, 2003).

We use intermediate latent variables $\{X^1, X^2, \cdots, X^N\}$ 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 $\{X^i, i = 1, 2, \cdots, N\}$ and the image I. Eq. 2.3 takes the form

$$\begin{aligned} & P(O, L, X^1, \cdots, X^N, I) \\ = P(O)P(L)P(X^1, \cdots, X^N|L, O)P(I|X^1, \cdots, X^N) & \end{aligned} \tag{2.4}$$

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 $P(X^1, \cdots, X^N|L, O)$ is given by the following factorization:

$$P(X^1, \cdots, X^N|L, O) = \prod_{i=1}^{i=N} \left\{P(X^i|L, O)\right\} \tag{2.5}$$

Applying Eq. 2.5, Eq. 2.4 leads to our final probabilistic model

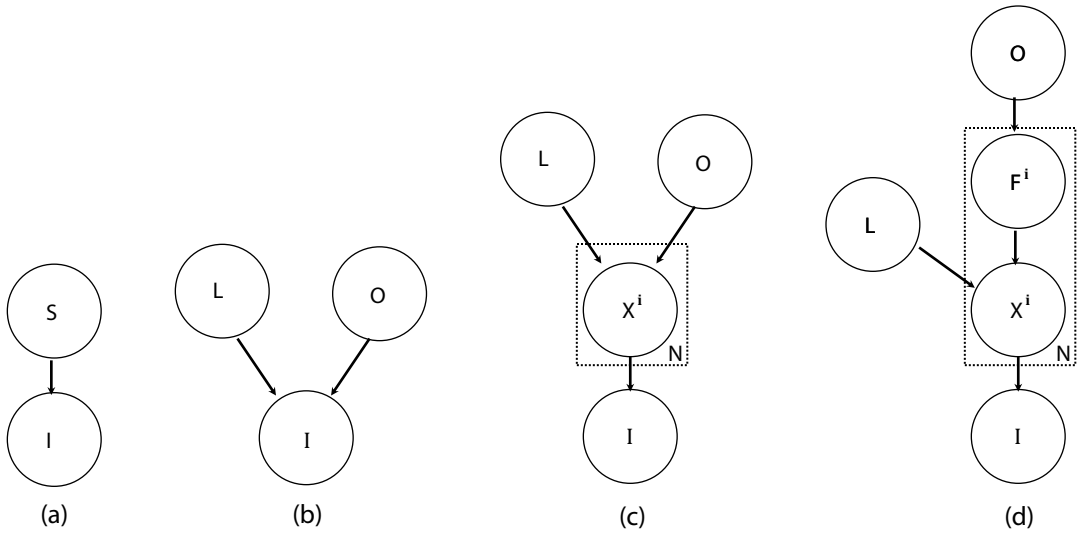


Figure 2-1: The figure illustrates the progression of graphical models corresponding to the sequence of factorizations given in Eq. 2.1 to Eq. 2.5 induced by the three main assumptions.

$$P(O, L, X^1, \cdots, X^N, I)$$
$$= P(O)P(L)\left\{\prod_{i=1}^{i=N}\left\{P(X^i|L,O)\right\}\right\}P(I|X^1, \cdots, X^N) \tag{2.6}$$

The model consists of a location encoding variable L, object encoding variable O, and feature-map variables $\{X^i, i = 1, \cdots, N\}$, 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 $(0, 1, \cdots, L)$.

As we discuss later (Sec. 2.3), it is easier to map the model onto the functional cortical anatomy (see Fig. 2-4) 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 defintion 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. 2.6 can be rewritten as

$$P(O, L, X^1, \cdots, X^N, F^1, \cdots, F^N, I)$$
$$= P(O)P(L)\left\{\prod_{i=1}^{i=N}\left\{P(X^i|L, F^i)P(F^i|O)\right\}\right\}P(I|X^1, \cdots, X^N) \quad (2.7)$$

The model assumes the following generative process

1. Choose $O \sim$ multinomial($P(O)$)

2. Choose $L \sim$ multinomial($P(L)$)

3. For each of the N features

   - Choose $F^i \sim$ Bernoulli($P(F^i|O)$)
   - Choose $X^i \sim$ multinomial ($P(X^i|L, F)$) (see text for details).

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 $X^i = l^*$ 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 [2].The conditional probabilities are specified in Table A1. 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 and 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: The probabilistic model of Eq. 2.7 encodes several constraints resulting from our three assumptions:

[2] $P(I|X^1 \cdots X^N)$ 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.

- *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 mutually exclusive as well.

- *Objects can be represented in terms of a single set of universal features ($F^1, \cdots, F^n$).* Although some objects may have diagnostic features, a large variety of objects can be represented using a shared set of primitive shape features (Ranzato *et al.*, 2007; Mutch and Lowe, 2006; Serre *et al.*, 2007c; Torralba *et al.*, 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.

2.1.2 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 4).

2.2 Model properties

2.2.1 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:

$$P(F^i|I) \quad \propto \quad P(F^i) \sum_{L,X^i} P(X^i|F^i,L)P(L)P(I|X^i) \tag{2.8}$$

Note that the factor $P(F^i)$ is obtained by marginalizing over all objects ($P(F^i) = \sum_O P(F^i|O)P(O)$). Spatial invariance is achieved by marginalizing (summing over) the L variables (see Fig. 2-2).

2.2.2 Spatial attention

As sketched in Fig. 2-5 (panel b), a key limitation of the feedforward (max) pooling mechanisms arises when performing recognition in the presence of clutter (see (Serre *et al.*, 2007b)), because of integration of visual information over relatively large receptive fields that makes the models prone to interference from distractors in the background. A natural way to solve this problem is to rely on spatial attentional mechanisms whereby priors $P(L)$ are concentrated around a region of interest. This may solve the problem of clutter by suppressing the background responses outside the spotlight (see 2-5, panel c).

In our model, spatial attention follows from setting a prior $P(L)$ concentrated around the location/scale of interest (see Fig. 2-2b). Consider the posterior estimate of the feature unit F^i. Ignoring the feature prior, the estimate is given by:

$$P(F^i|I) \quad \propto \quad \sum_{L,X^i} P(X^i|F^i,L)P(L)P(I|X^i). \tag{2.9}$$

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*.

2.2.3 Feature-based attention

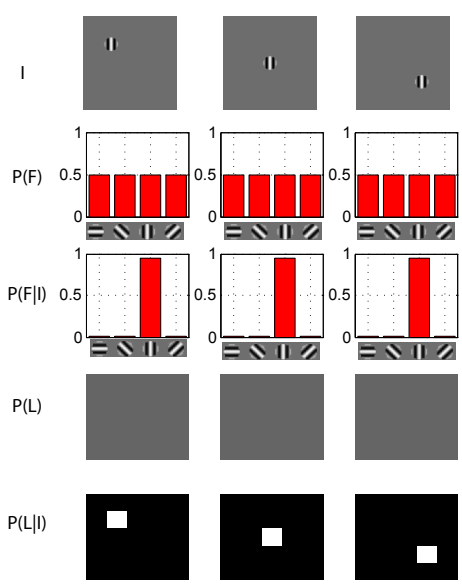
As illustrated in Fig. 2-5 (panel d), when a single isolated object/feature is present, it is possible to read out its location from the posterior probability $P(L|I)$. However when multiple objects/features are present (see Fig. 2-5, panel e), it is no longer possible to readout this information. To 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:

$$P(L|I) \propto P(L) \prod_i \left\{ \sum_{F^i, X^i} P(X^i|F^i, L) P(F^i) P(I|X^i) \right\} \tag{2.10}$$

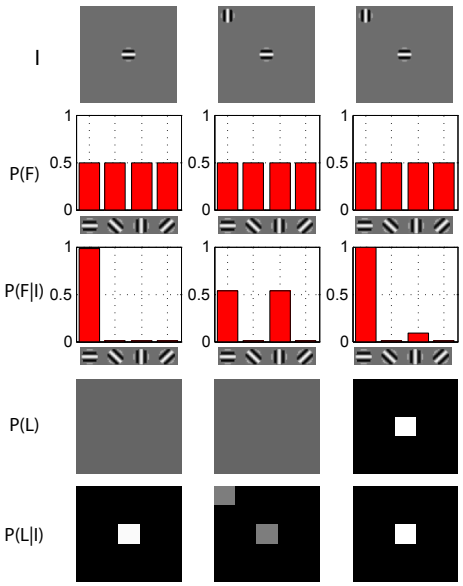
Increasing the concentration of the prior around the target feature F^i enhances the preferred feature at *all* locations while low priors on other features suppress activity from distracting objects. After this, the location of the preferred feature can be read out from the posterior probability $P(L|I)$, which can be interpreted as a saliency map.

2.2.4 Feature pop-out

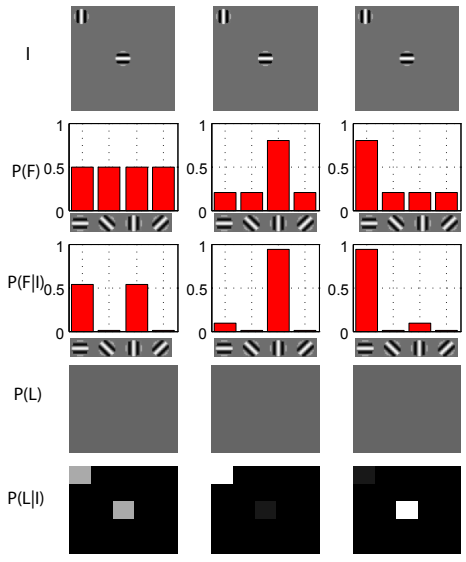
Since the X^i units are mutually exclusive ($\forall i, \sum_{X^i} P(X^i|F^i, L) = 1$), increasing the activity (probability) at one location in an image typically reduces the likelihood of the stimulus being present at other locations (see Fig. 2-2d). In a sense, this is similar to the extra-classical receptive field effects observed throughout the visual cortex (see (Carandini *et al.*, 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



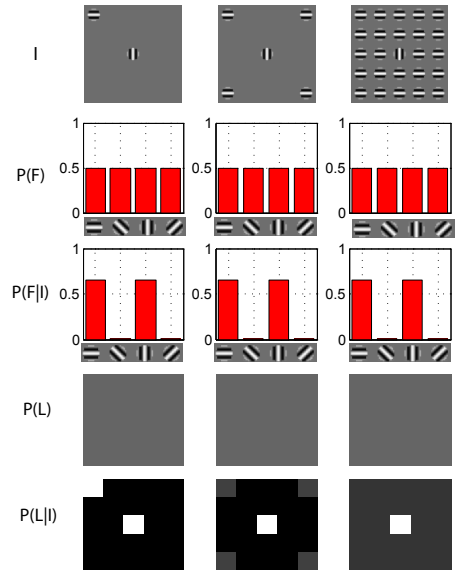
(a) Spatial invariance.



(b) Spatial attention.



(c) Feature attention.



(d) Feature popout.

Figure 2-2: 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.

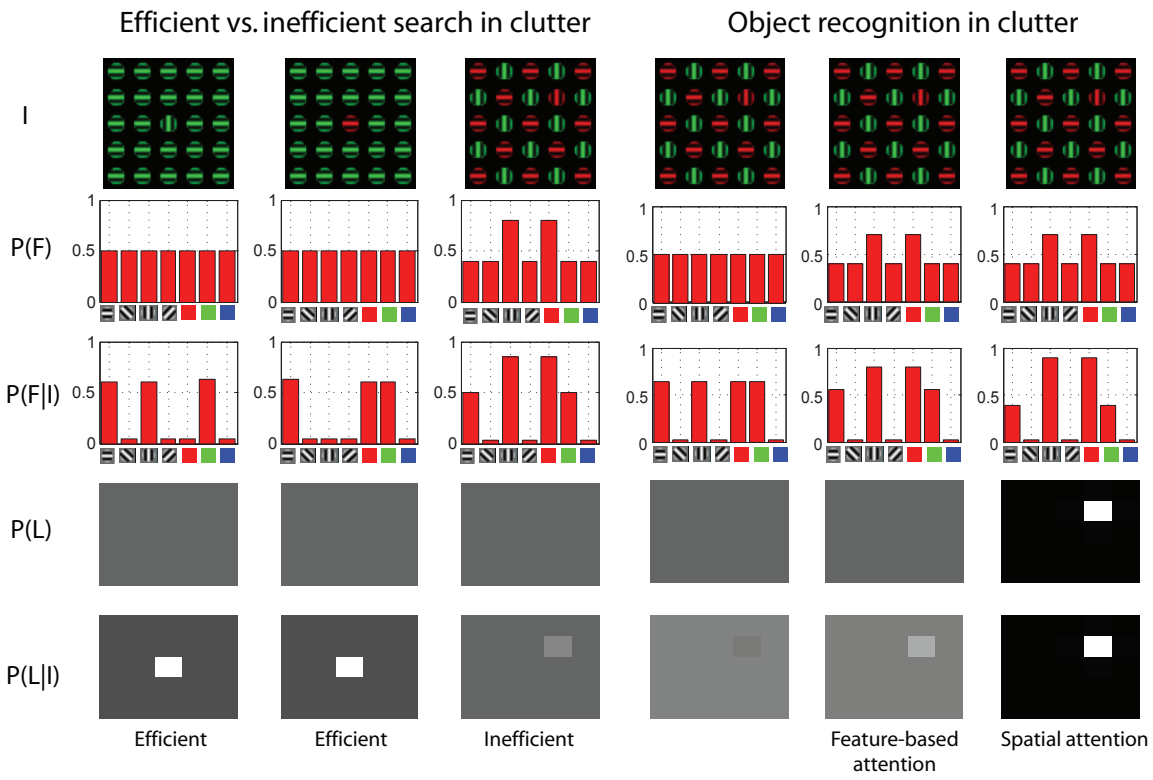


Figure 2-3: Left: Efficient vs. inefficient searches: When a stimulus differs from the background in a single dimension (first and second column) search is easy (indicated by a high contrast saliency map) and independent of the number of distractors. However, when features are shared between the target and distractors (third column), search is more difficult (indicated by a low contrast saliency map). Right: Object recognition in clutter consists of feature-based attention followed by spatial attention. The most likely location of the target is found by feature-based attention by setting appropriately the feature priors (middle column). The hypothesis is then verified by deploying spatial attention around the location of the highest saliency (the spatial priors are changed in the right column). The value of the feature units $P(F|I)$ indicate the presence or absence of an object feature.

of the saliency map towards the location of the salient or 'surprising' item (see Fig. 2-2d).

In contrast to our model, traditional approach to pop-out has been based on image saliency. In (Itti *et al.*, 1998), center-surround difference across color, intensity and orientation dimensions is used as measure of saliency. In (Gao and Vasconcelos, 2007), self information of the stimuli ($-\log(P(I))$) is used as measure of distinctiveness (Zhang *et al.*, 2008). In (Rosenholtz, 1985), the normalized deviation from mean response is used instead.These models, however, cannot account for the task-dependency of eye movements (Yarbus, 1967).

Li and Snowden (2006) proposed a computational model based on V1-like orientation fea-

tures 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. However, recent evidence seems to show V1 to be relatively unaffected by top-down attentional modulation (Hegde and Felleman, 2003), thus moving the locus of attention away from V1 and towards higher regions such as V4. Experiments on spatial attention (McAdams and Maunsell, 1999) and feature-based attention (Bichot *et al.*, 2005) have shown attentional modulation in V4.

2.2.5 Efficient vs. inefficient search tasks

The maximum value of the saliency map (or the posterior probability of location $P(L|I)$) can be viewed as a proxy for the efficiency of search. From this point of view, the bayesian model can predict the relative difficulty of a search task, at least in artificial arrays (see Fig. 2-3a).

2.3 Neural interpretation

Prior work has shown that perception under uncertainty can be modeled well using Bayesian inference (Knill and Richards, 1996; Rao *et al.*, 2002a; Kersten *et al.*, 2004). 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 *et al.*, 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).

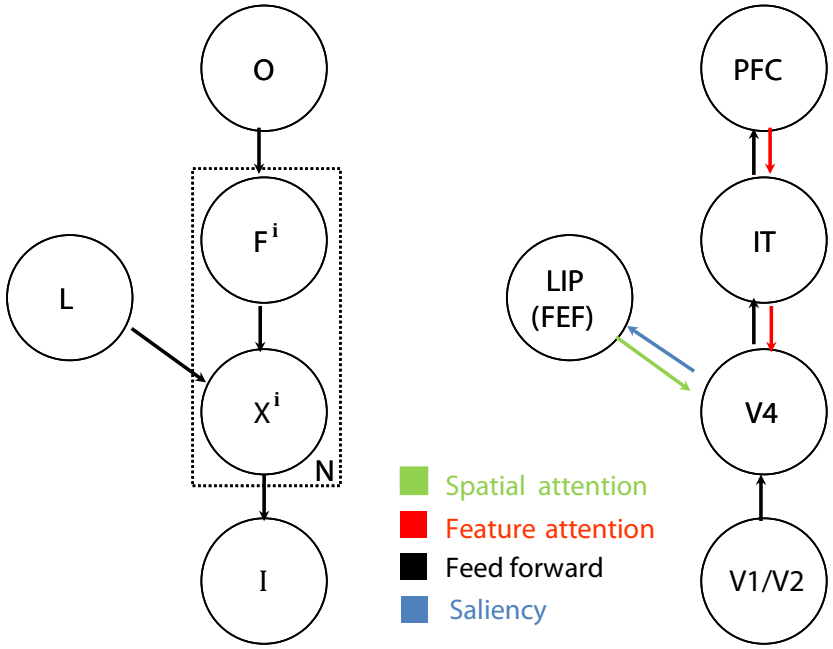


Figure 2-4: Left: Proposed bayesian model. Right: A model illustrating the interaction between the parietal and ventral streams mediated by feedforward and feedback connections. The main addition to the original feedforward model (Serre *et al.*, 2005b) (see also *Supplementary Online Information*) is (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 and Ullman, 1985).

2.3.1 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 2.1, Fig. 2-4). Thus, following the organization of the visual system (Ungerleider and 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 and Tanaka, 1994; Tanaka, 1996; Logothetis and Sheinberg, 1996) and

higher visual areas such as AIT where retinotopy is almost completely lost (Oram and Perrett, 1992; Logothetis *et al.*, 1995) (modeled as F^i units). Prior (non-bayesian) attempts to model this interaction can be found in (Grossberg, 1999; 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 state 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, $P(X^i = 1|I), P(X^i = 2|I) \cdots P(X^i = |L||I)$ can be mapped to the firing rates of the neuron encoding feature F^i at location $1, 2..|L|$ 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 and Poggio, 1999b; Serre *et al.*, 2005b), 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.

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 and 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 and 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 and Wiesel, 1959) to combinations of orientations and features of moderate complexity in intermediate visual areas V2 (Hegde and Van Essen, 2000; Ito and Komatsu, 2004) and V4 (Pasupathy and Connor, 2001; Desimone and Schein, 1987; Gallant *et al.*, 1996), to parts and objects in area IT (Tanaka, 1996; Logothetis and Sheinberg, 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 *et al.*, 2005b) 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 over 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 $(0, 1 \cdots L)$. 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 and 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 stimulus (Hegde and Van Essen, 2000; Ito and Komatsu, 2004). We use the computational model of the ventral stream (Serre *et al.*, 2007c) to derive V2-like features from the image.

Table 2.1: Bayesian model units and tentative mapping to brain areas.

<table>
<tr><th>Conditional Probability</th><th>Modeling</th></tr>
<tr><td>$P(L)$</td><td>Each scene, with its associated view-point, 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).</td></tr>
<tr><td>$P(F^i|O)$</td><td>The probability of each feature being present or absent given the object; it is learned from the training data.</td></tr>
<tr><td>$P(X^i|F^i, L)$</td><td>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
<table>
<tr><td></td><td>$F^i = 1, L = l$</td><td>$F^i = 0, L = l$</td></tr>
<tr><td>$X^i = 0$</td><td>$P(X^i|F^i, L) = \delta_1$</td><td>$P(X^i|F^i, L) = 1 - \delta_2$</td></tr>
<tr><td>$X^i \neq 0$</td><td>$P(X^i|F^i, L) \sim$ Gaussian centered around $L = l$</td><td>$P(X^i|F^i, L) = \delta_2$</td></tr>
</table>
δ_1 and δ_2 are small values (~ 0.01), chosen to ensure that $\sum P(X^i|F^i, L) = 1$.</td></tr>
<tr><td>$P(I|X^i)$</td><td>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 et al., 2007c)).</td></tr>
</table>

Table 2.2: Description of the model conditional probabilities.

2.3.2 Inference using belief propagation

Within the bayesian network, inference can be performed using any of several inference algorithms such as junction tree, variable elimination, MCMC (Markov-chain Monte carlo) and belief propagation (Wainwright and Jordan, 2008; Gilks and Spiegelhalter, 1996). 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.

$$
\begin{aligned}
m_{O\to F^i} &= P(O) && (2.11)\\
m_{F^i\to X^i} &= \sum_{O} P(F^i|O)P(O) && (2.12)\\
m_{L\to X^i} &= P(L) && (2.13)\\
m_{I\to X^i} &= P(I|X^i) && (2.14)\\
m_{X^i\to F^i} &= \sum_{L}\sum_{X^i} P(X^i|F^i,L)(m_{L\to X^i})(m_{I\to X^i}) && (2.15)\\
m_{X^i\to L} &= \sum_{F^i}\sum_{X^i} P(X^i|F^i,L)(m_{F^i\to X^i})(m_{I\to X^i}) && (2.16)
\end{aligned}
$$

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

$$
P(L|I) \propto (m_{L\to X^i})(m_{X^i\to L}) \qquad (2.17)
$$

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, $\{1, 2 \cdots |L|\}$, F^i is summed over $\{0, 1\}$ and X^i, over states $\{0, 1, \cdots |L|\}$. Notice 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:

$$m_{L \to X^i} = P(L) \prod_{j \neq i} m_{X^j \to L} \tag{2.18}$$

$$\tag{2.19}$$

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 $y_1, y_2..y_S$. A possible mapping is to S individual binary indicator variables $I_1, I_2 \cdots I_S$, 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 the 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.

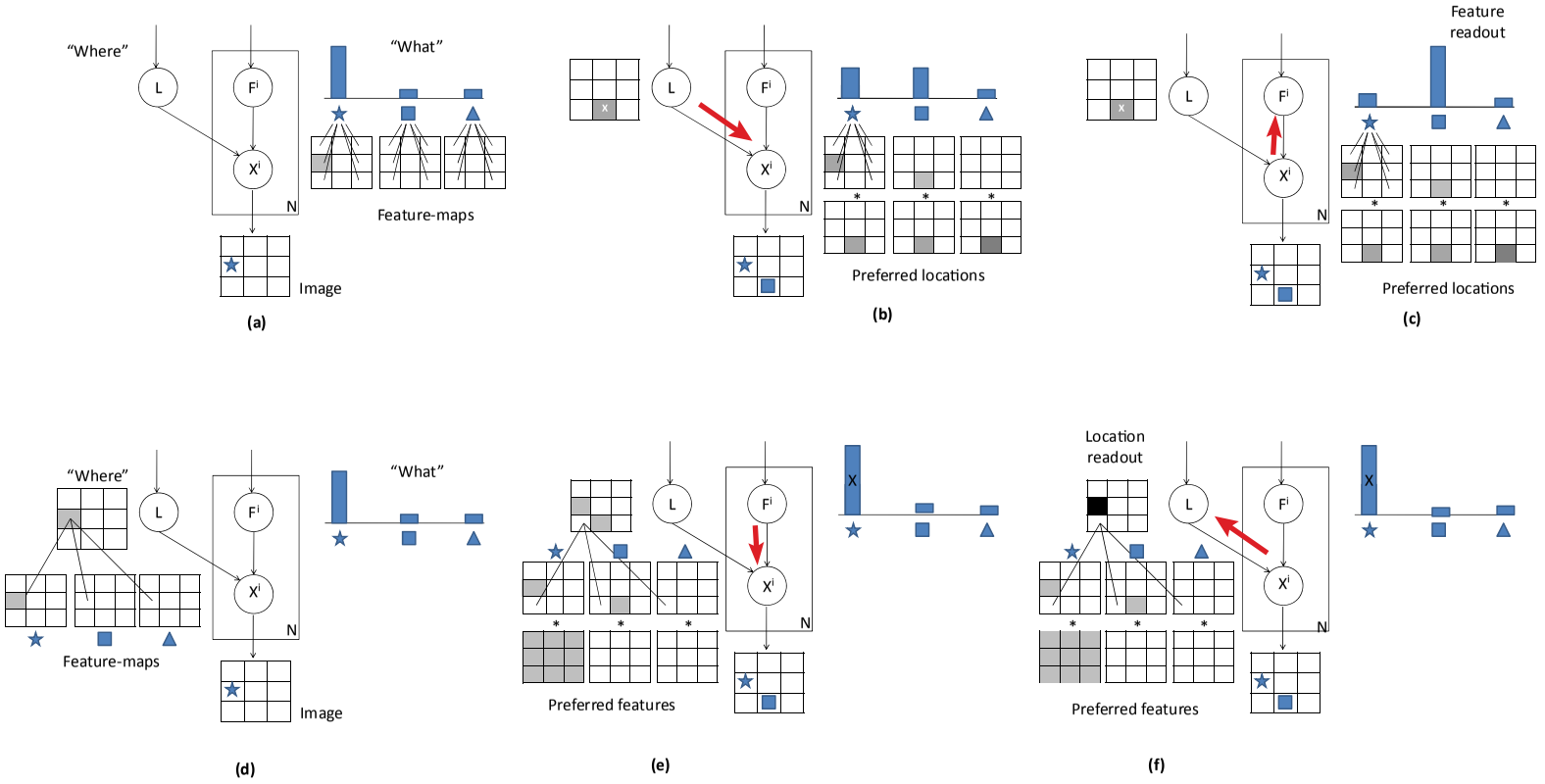


Figure 2-5: Spatial and feature attention re-interpreted using message passing within the model. Spatial attention: *(a)* Each feature unit F^i pools across all locations from the corresponding X^i unit. *(b)* Spatial attention here solves the 'clutter' problem by concentrating the prior $P(L)$ around 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 $P(L|I)$.

Chapter 3

”Predicting” physiological effects

3.1 "Predicting" the physiological effects of attention

Here we show that the proposed model is consistent with neurophysiology experiments about the effects of feature-based and spatial attention (McAdams and Maunsell, 1999; Bichot *et al.*, 2005; Reynolds and Heeger, 2009). We also find that, surprisingly, several key attentional phenomena such 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.

3.1.1 Attentional effects in V4

Within our model, V4 neurons are represented with variables $\{X^1, X^2 \cdots X^N\}$. 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

$$P(X^i|I) \;=\; \frac{P(I|X^i)\ \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}} \tag{3.1}$$

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. 3.1 may be re-written 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 $A(L,F^i) = P(L)P(F^i)$; (iii) a divisive normalization factor $S(L,F^i)$. With this

notation, Eq. 3.1 can be rewritten as:

$$P(X^i|I) = \frac{E(X^i)A(F^i, L)}{S(F^i, L)} \tag{3.2}$$

Equation 3.2 turns out to be closely related to a phenomenological model of attention recently proposed by Reynolds and Heeger (2009). In this model, the response of a neuron at location x and tuned to orientation θ is given by:

$$R(x, \theta) \quad = \quad \frac{A(x, \theta)E(x, \theta)}{S(x, \theta) + \sigma} \tag{3.3}$$

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. $A(x, \theta)$ represents the attentional modulation that enhances specific orientations and locations, based on the search task. Reynolds & Heeger showed that the model of Eq.3.3 can reproduce key physiological effects of attention such as contrast gain behavior under different stimulus conditions. A comparison of Eq. 3.2 with Eq. 3.3 suggests that the normalization model of Reynolds & Heeger model is a special case of our model e.g. Eq. 3.2. Normalization in our model emerges directly from the bayesian formulation, instead of being an ad hoc assumption[1].

3.1.2 Multiplicative modulation

Spatial attention In (McAdams and 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

$$P(X^i = x|I) \quad \propto \quad \sum_{F^i, L} P(X^i = x|F^i, L)P(I|X^i)P(F^i)P(L). \tag{3.4}$$

[1]It is to be noted that the normalization model (as described in (Reynolds and Heeger, 2009)) explains the individual phenomenon of change in contrast response due to attention. The bayesian model described here can explain and predict several additional phenomena discussed in this section that are not addressed by the normalization model.

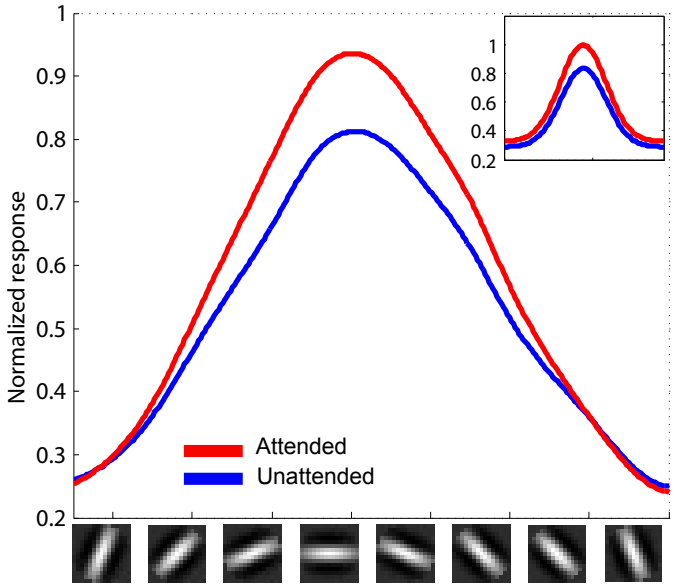


Figure 3-1: Effect of spatial attention on tuning response. The tuning curve shows a multiplicative modulation under attention. The inset shows the replotted data from (McAdams and Maunsell, 1999).

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 enhance the response, even under the same stimulus condition (see Fig. 3-3).

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. 2-2b). 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 $i = 1 \ldots n$ at locations $l_1 \ldots l_m$ that overlap with the attentional spotlight $P(L)$ and in suppression everywhere else. The message passing is initiated at the level of the L units assumed to be in parietal cortex) and should manifest itself after a short delay in the F^i units (in the ventral stream), in agreement with physiological data (Buschman and Miller, 2007).

Feature based attention Recent findings in physiology (Bichot *et al.*, 2005) show multiplicative modulation of neuronal response under attention. Units in the PFC and higher

areas seem to modulate arrays of "feature detectors" 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. 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 ($P(I|X^i)$ is high).

In terms of message passing, objects priors are first concentrated around the object(s) of interest (*e.g.*, (see Fig. 2-2c). '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. 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 $j = 1 \ldots n$ 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. 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 & Miller (Buschman and Miller, 2007)(see Fig.2-2).

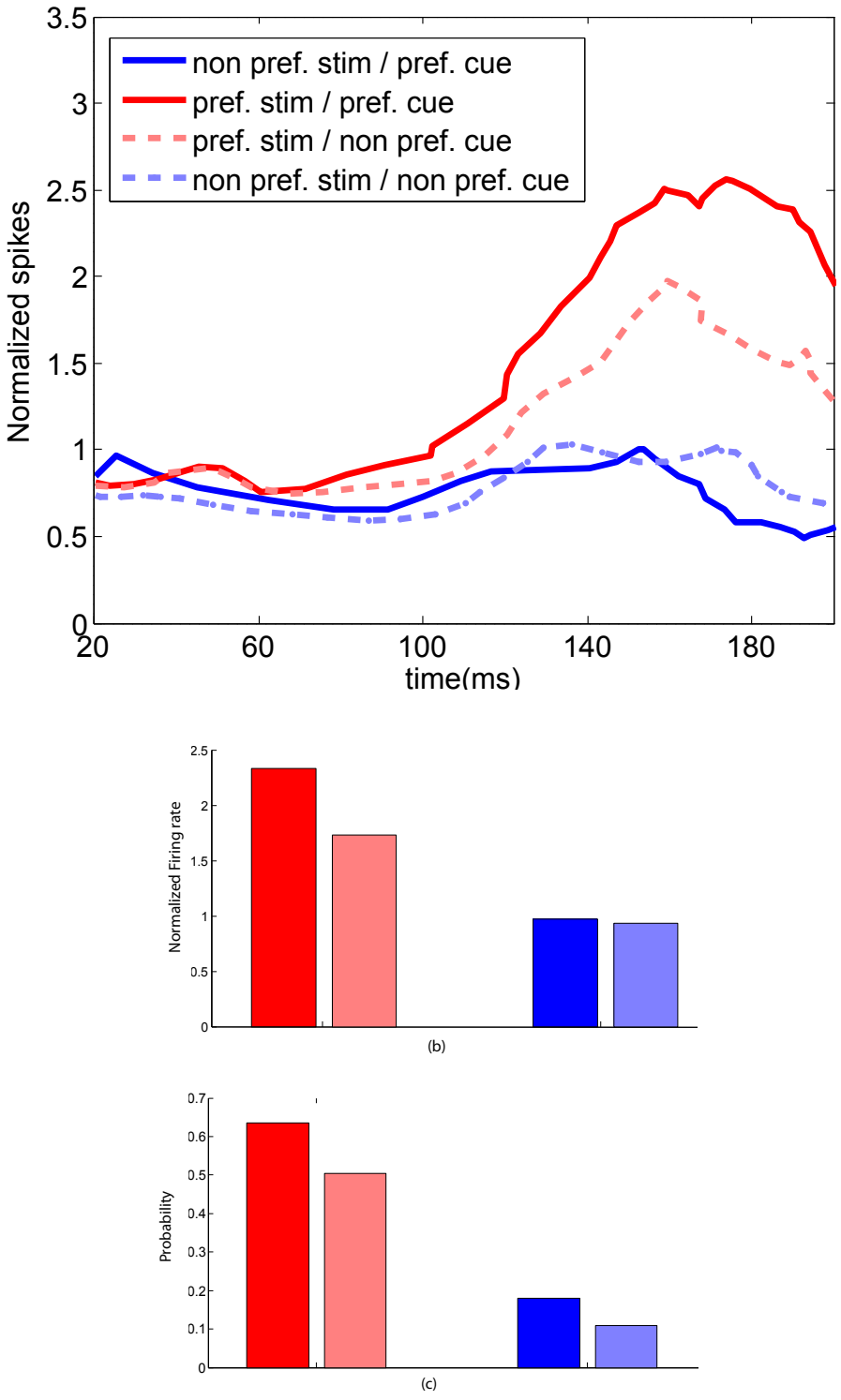
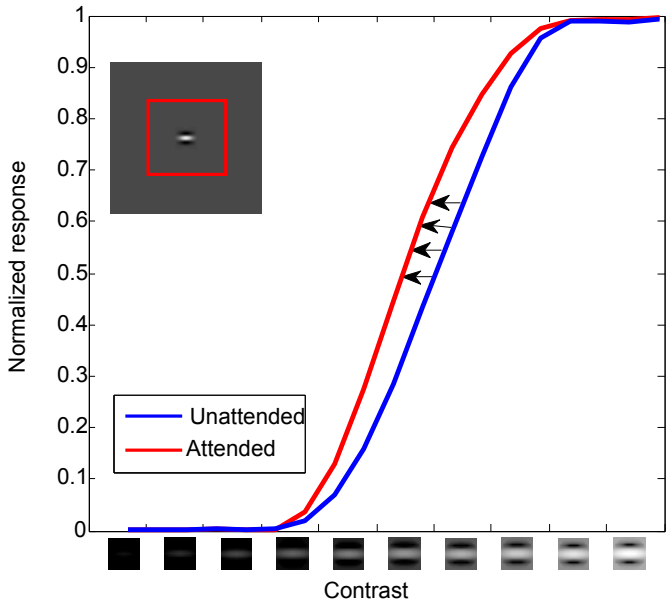


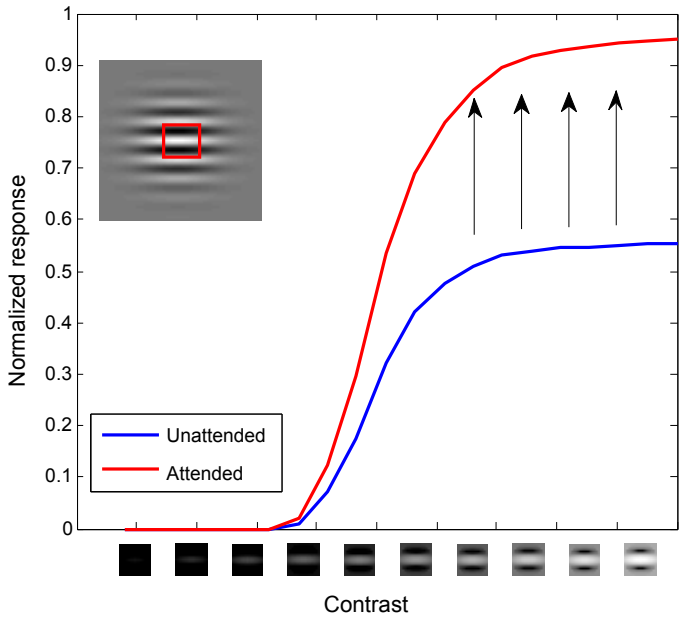
Figure 3-2: *(a)* Effect of feature attention on neuron response (Replotted from (Bichot *et al.*, 2005)). *(b)* The time course of the neuron response is sampled at 150ms. *(c)* The model predicts multiplicative modulation of the response of X^i units under attention.

3.1.3 Contrast response

The influence of spatial attention on contrast response of V4 neurons have been studied extensively. Prior work showed two major, apparently contradictory, effects: in (Martınez-Trujillo and Treue, 2002; Reynolds *et al.*, 2000) attention was shown to shift the contrast response of neurons, while in (McAdams and Maunsell, 1999; Treue and Trujillo, 1999a) 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 a contrast gain was observed, the spotlight of attention was larger than the stimulus. In the experiments in which a response gain was seen, the stimulus was observed to be larger than the spotlight of attention. Reynolds & Heeger showed that their normalization model of attention(see Sec.2.3) produces the observed effects under these different conditions.
In Fig.3-3c and Fig.3-3d we show that our bayesian model, as expected given its close relation with Reynolds & 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.



(a) Shift in contrast response under attention.



(b) Response gain under attention.

Figure 3-3: 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.

3.2 Attentional effects in MT

3.2.1 Multiplicative modulation

So far we have restricted our analysis and to interaction of the ventral stream and parietal regions (and consequently to static images alone). However, motion serves as a powerful cue for visual attention. In the brain, the dorsal stream comprising of regions V1, MT, MST, STS (Ungerleider and Pasternak, 2004) processes motion information. Similar to properties of the hierarchical ventral stream, the specificity of stimulus is found to increase as we progress higher in the hierarchy. Physiological studies have shown attention can modulate neurons in region MT in the dorsal stream, similar to the modulation found in V4 region in the ventral stream (Treue and Trujillo, 1999b; Beauchamp *et al.*, 1997; Womelsdorf *et al.*, 2006). We propose to incorporate motion based features in addition to existing shape and color features within the attentional framework. Retaining its parallel to biology, we proposed to use motion features derived from the computational model of the dorsal stream (Jhuang *et al.*, 2007) and specifically that of region MT (Simoncelli and Heeger, 1998; Rust *et al.*, 2006). The extended model comprising shape and motion features will be demonstrated using real world video sequences.

Preliminary results: The Bayesian model of attention is agnostic to the origin of the features and only requires a spatially organized feature map for its functioning. We use the computational model of the area MT proposed by Simoncelli (Simoncelli and Heeger, 1998; Rust *et al.*, 2006) to compute the motion features. Thus, in addition to the shape-based features we introduce motion based feature maps (corresponding to four orientations and two speeds). The message passing equations remain unaltered. In the following, we shall present preliminary results illustrating spatial and feature-based attention operating on moving stimuli.

The stimulus consists of three horizontal sections of moving dot patterns (see Fig.3-4). Each section can have a separate direction and speed of motion in addition to its color. The arrows indicate the direction of motion of the dot pattern and is not a part of the stimulus. $P(L)$ and $P(F)$ correspond to the spatial and feature prior respectively. These

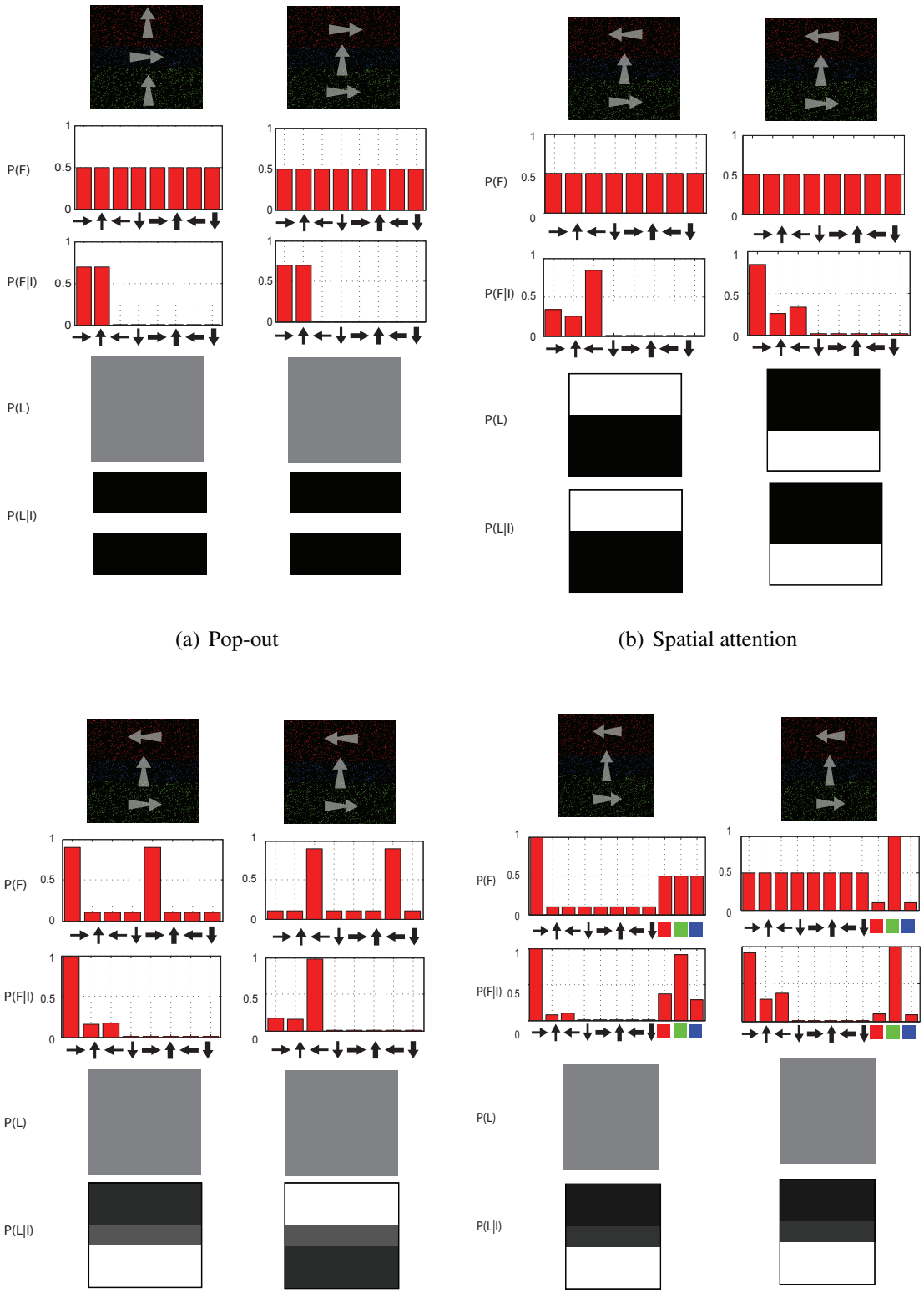
indicate the location and features of interest before the stimulus is presented. $P(L|I)$ and $P(F|I)$ are the posterior probabilities obtained after the sitmulus is observed. Thus, these values account for the bottom-up evidence from the stimulus as well as top-down task-based priors.

Pop-out: The region of the stimulus where the direction differs from rest of the stimuls has the highest saliency. This may be attributed to lateral inhibition within each feature dimension (implied by the mutual exclusion of the different states of variable X^i). The absolute direction of motion does not affect the result. In the example (see Fig.3-4a), the central band region has the highest saliency because its direction differs from rest of the stimulus.

Spatial attention: A specific region within the stimulus can be attended while ignoring the rest by altering the spatial prior $P(L)$. In the example, the 'read-out' from the feature units indicate the speed and direction of the dots in the attended region. We observed that the feature in the attended region is enhanced while the others are supressed (see Fig.3-4b).

Feature-based attention: Instead of attending to a location, attention can be directed to a specific feature. The 'read-out' from the location unit indicates the area where the attended feature is likely to be found. In the example, the feature to be attended is specified by altering $P(F)$. The most likely location can then be 'read-out' from $P(L|I)$ (see Fig.3-4c).

Multi-modal interaction: In this example, we illustrate the interaction of color and motion features, a question explored ealier in (Wannig *et al.*, 2007). We use a simpler stimulus, where dots moving in different directions are non-overlapping. Furthermore, the color of the dots can vary independently of its direction. In the example (see Fig. 3-4d), attending to a specific color results in the 'read out' of its corresponding direction of motion. Alternatively, attending to a specific direction allows us to 'read out' its color. This effect may be explained as a two stage process in which, attending to a specific motion or color feature biases attention towards the locations where this feature is present. This in turn, enhances other features also present within the attended region (see Fig.3-4d).



(a) Pop-out

(b) Spatial attention

(c) Feature-based attention

(d) Multimodal interaction

Figure 3-4: Attention applied to motion features.

3.3 "Predicting" effects of spatial attention in IT

3.3.1 Experimental evidence

In their seminal study, Moran and Desimone (1985) showed that the response of neurons in the extrastriate cortex is modulated by spatial attention while neurons in the striate cortex are not. Specifically, when multiple objects were presented within the receptive field of a V4/IT neuron, it was found that the response of the neuron depended only on the properties of the attended stimulus. The response of the neuron to the unattended stimulus was reduced even when it was the preferred stimulus of the neuron. The study measured the effects of attention at the level of individual neurons. However, physiological studies have shown that objects are encoded using a population of neurons in IT (Kobatake and Tanaka, 1994; Tanaka, 1996). Thus, in order to study the effect of attention on object perception, it is essential to study the phenomena at the population level. A recent (and ongoing) study in Poggio and Desimone labs attempts to quantify the effect of attention on IT neurons at a population level. Specifically, the study attempts to measure how the information [2] about objects in the receptive field is affected by spatial attention. In the following, we briefly describe the original experiment followed by simulations using the model.

Data The stimuli used in the experiment consists of images composed using one (isolated condition) or three (cluttered condition) out of a pool of 16 objects. The individual objects used are shown in Fig. A3. Each object was placed at one of three possible positions on the contralateral hemifield. Overall the experiment used 912 images consisting of 48 images containing isolated objects (16 objects presented at one of the three positions) and 768 images containing three objects (see Fig 3-5)[3].

Experiment The stimuli was presented to two alert monkeys. The monkey fixated on a spot at the center of the stimulus. The stimulus consisted of one or three objects. In case of a stimulus with a single object, it was always considered as the target object. When the

[2] as measured by neural decoding performance

[3] Note that this is less than the total number($16 \times 15 \times 14$) of possible combinations of 16 objects present at three location

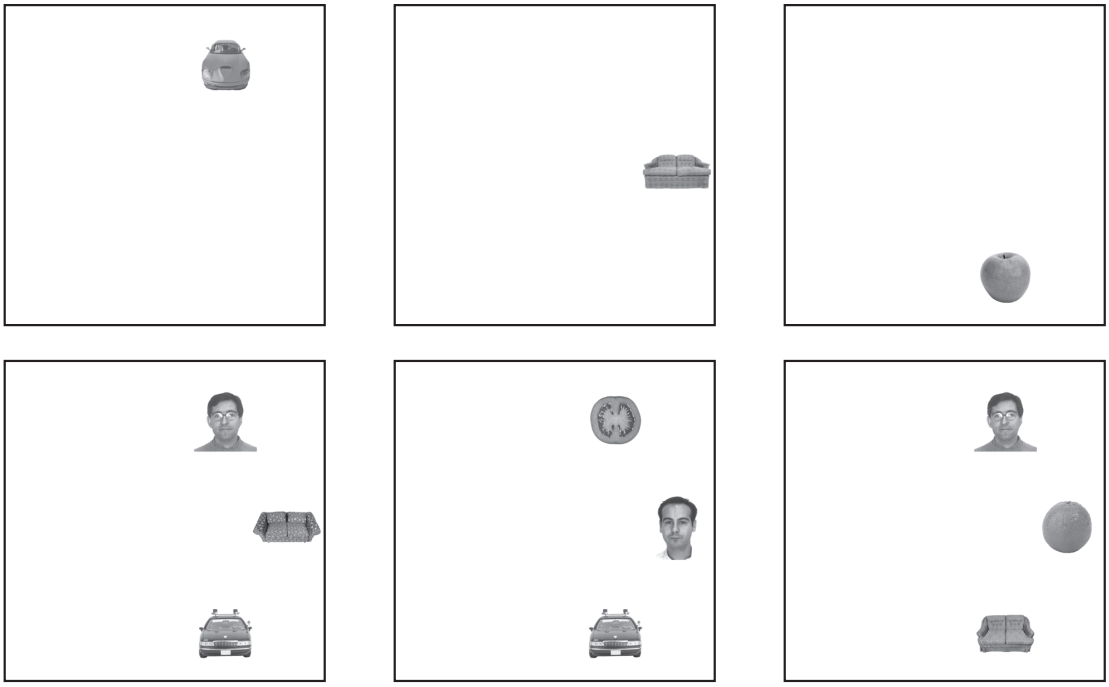


Figure 3-5: Illustration of some of the stimuli presented during the experiment. (Top): Stimuli where a single object was present. (Bottom): Stimuli where three objects were present. In both cases, the fixation point was placed at the center of the image.

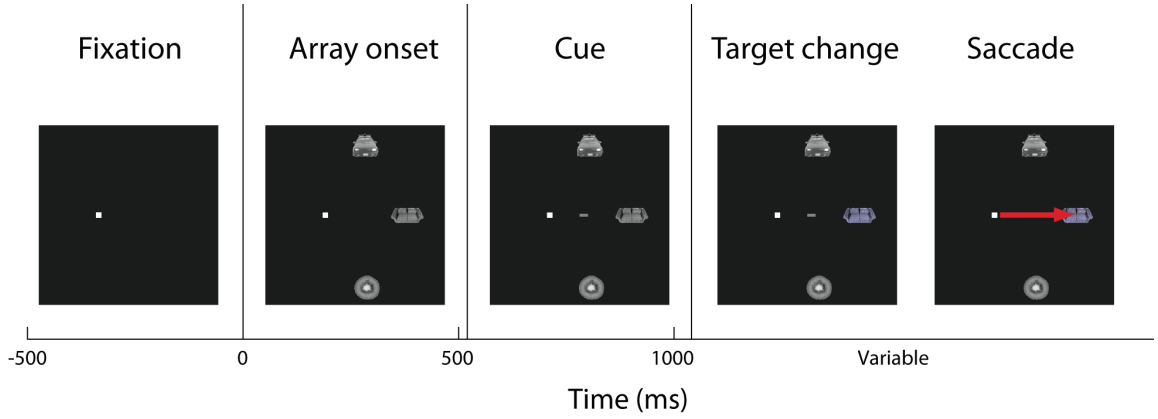


Figure 3-6: Experimental protocol used for recording neurons from IT. Notice the spatial cue in the form of a small bar directed at the target object. (image,courtesy Ethan Meyers)

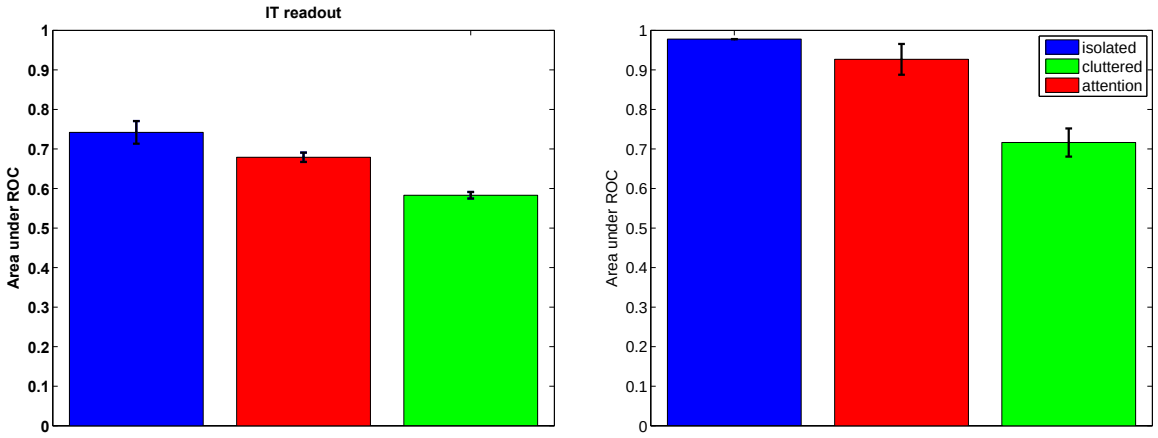
stimulus consisted of three objects, one of the objects was designated as the target and the other two objects were considered as distracters. Between 518-528 ms after the objects appeared, a cue was presented in the form of a short line (see Fig. 3-6). The monkey was rewarded for saccading to the cued object when it changed in color, which happened between 518-2160ms after the objects appeared on screen. A total of 98 and 139 cells were recorded from the first and second monkey respectively.

Population decoding Using the neural responses obtained during isolated object presentation, a statistical classifier is trained to associate the neural responses to the identity of the object. A classifier is trained for each of the 16 objects (one vs. all paradigm). The prediction score obtained from the classifier represents the extent of information present about the target object. The decoding performance is measured in terms of the area under ROC curve. This kind of neural decoding paradigm has been used in interpretation of neural data before (Hung *et al.*, 2005).

Results The study showed that when multiple objects are present within the receptive field of an IT neuron, the response consists of a mixture of information about objects present in the stimuli. However, once an attentional cue is provided, information about the cued object is enhanced relative to the other objects in the display (see Fig. 3-7 a). This effect could be explained as (i) attention restoring activities of neurons similar to that of an isolated object or (ii) attention changing the representation of object by inclusion of additional information. The study showed that the former hypothesis is better supported by the evidence. In the presence of attention, the pattern of neural activities reverts to activity similar to the condition when the cued object was present in isolation. The study also showed that bottom-up cues such as change of color can temporarily override top-down effects of spatial attention.

3.3.2 Bayesian model

Simulation We study the effects of spatial attention in the model and test if the predictions of the proposed model are consistent with the experimental evidence. We presented



(a) Decoding performance using IT neurons

(b) Decoding performance using the model before fitting the parameters

Figure 3-7: Comparison between IT neurons and model simulation before parameter fitting.

the same set of 912 stimuli to the model. The prior on object identity was set to be uniform. In the case where no spatial cue was provided in the original experiment, the spatial prior was set to be uniform. If a cue was provided, spatial prior is set to be a gaussian around the object location. The size of the gaussian is chosen to be such that the probability mass is concentrated within the spatial support of the cued object. The posterior probability of the object provides a quantity that is similar to prediction score given by the classifier in the original experiment. The prediction is assumed to be correct if the object with the highest probability was the designated target. The performance of the model is measured in terms of the area under ROC.

Results The results (see Fig. 3-7 b) show that similar to the original experiment, the decoding performance is the highest in the isolated condition and decreases under clutter when no attentional cue is provided. However, under spatial attention, the performance is restored and is similar to the condition when only isolated objects are presented. The absolute decoding performance of the model is much higher than obtained from neural data. During the simulation, the features are assumed to be noise free. Furthermore, the size of the spotlight of attention is fixed.

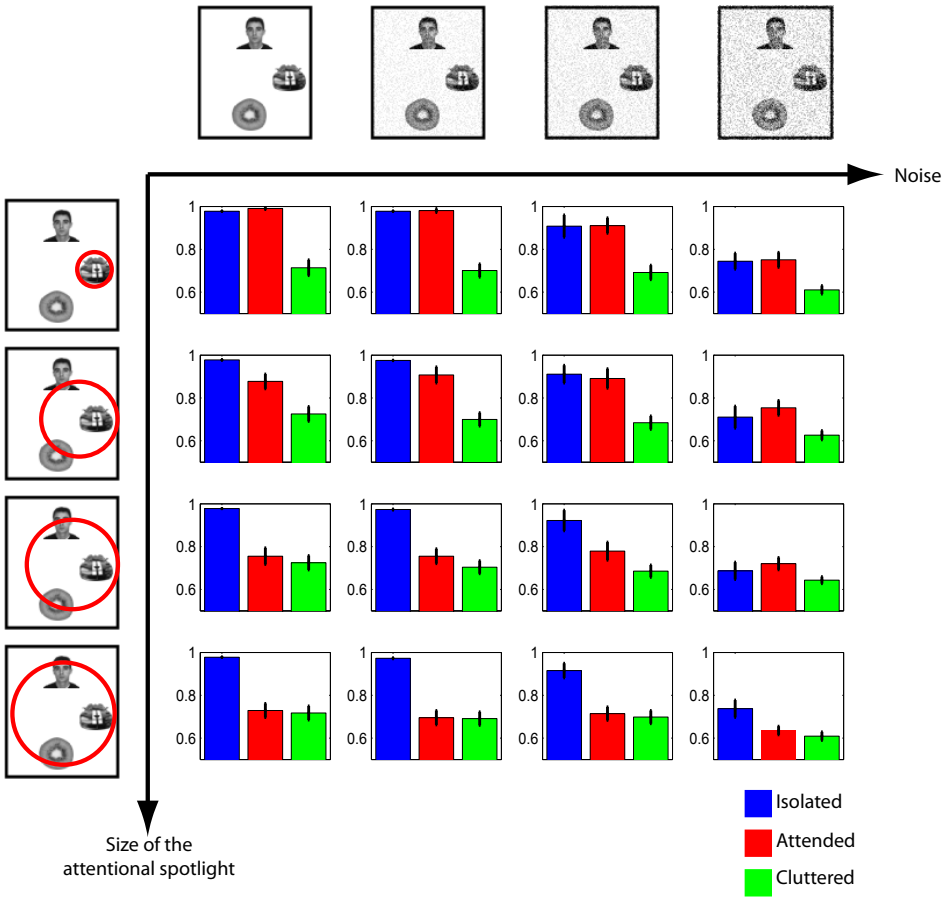
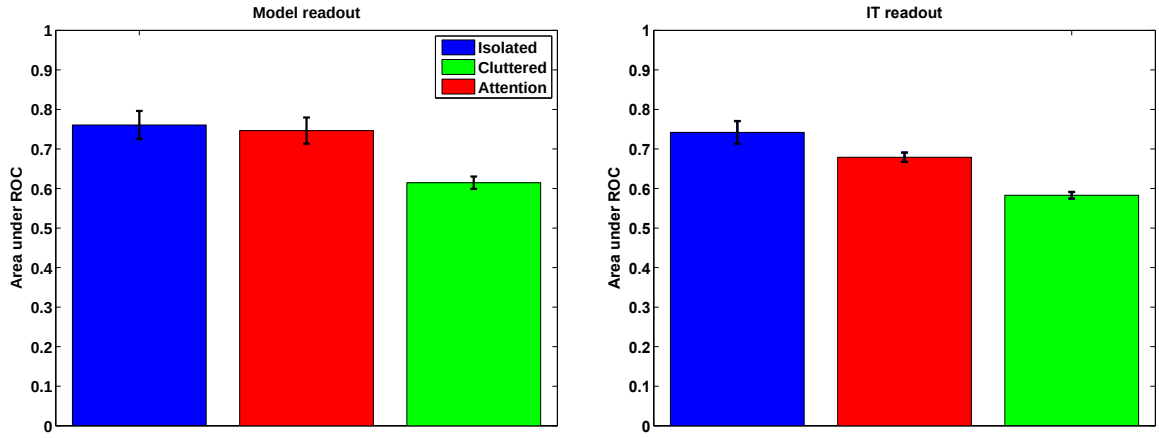


Figure 3-8: Effect of noise and size of the attentional spotlight on decoding performance.

Parameter fitting We studied the effects of these parameters on the decoding performance (see Fig 3-8). We observed that increasing the noise (probability of error) decreases the absolute decoding performance in both the isolated and attended condition. On the other hand, increasing the size of the attentional spotlight decreases the performance for the attended condition while not affecting the decoding performance for the isolated condition. When the size of the attentional spotlight is enlarged, features from the other object interfere causing a drop in performance. With a proper choice of the noise level and the size of the attentional spotlight, the performance of the model can be made to be close to the performance obtained from IT neurons. This can be considered a crude form of fitting model parameters (See Sec. A1.4 for implementation details).



(a) Decoding performance using the model after fitting the parameters

(b) Decoding performance using IT neurons

Figure 3-9: Comparison between IT neurons and model simulation.

3.3.3 Discussion

The effect of attention on IT can be summarized eliminating interference of clutter and restoring information of the target object. This is similar to the attentional spotlight metaphor (Crick and Koch, 1990b,a), where attention ”illuminates” area/features of interest while suppressing the distracters. The prediction of the model is consistent with this explanation. In this study, we fit the model parameters such that the decoding performance was similar to that of IT neurons. Conversely, it can be speculated that the model parameters is predictive of the level of noise and the size of the attentional spotlight in the brain.

Chapter 4

Predicting eye-movements

4.1 Predicting human eye movements

Human and animal studies (see (Wolfe, 2007) for a recent review) have isolated at least three main components used to guide the deployment of an eye movements. First, studies have shown that image-based *bottom-up* cues can capture attention, particularly during free viewing conditions[1]. Second, task dependence also plays a significant role in visual search (Wolfe, 2007; Yarbus, 1967)[2]. 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.

4.1.1 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 and Tsotsos (Bruce and 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 and 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 Sec. A1).

[1] 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 and Koch, 2001a).

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

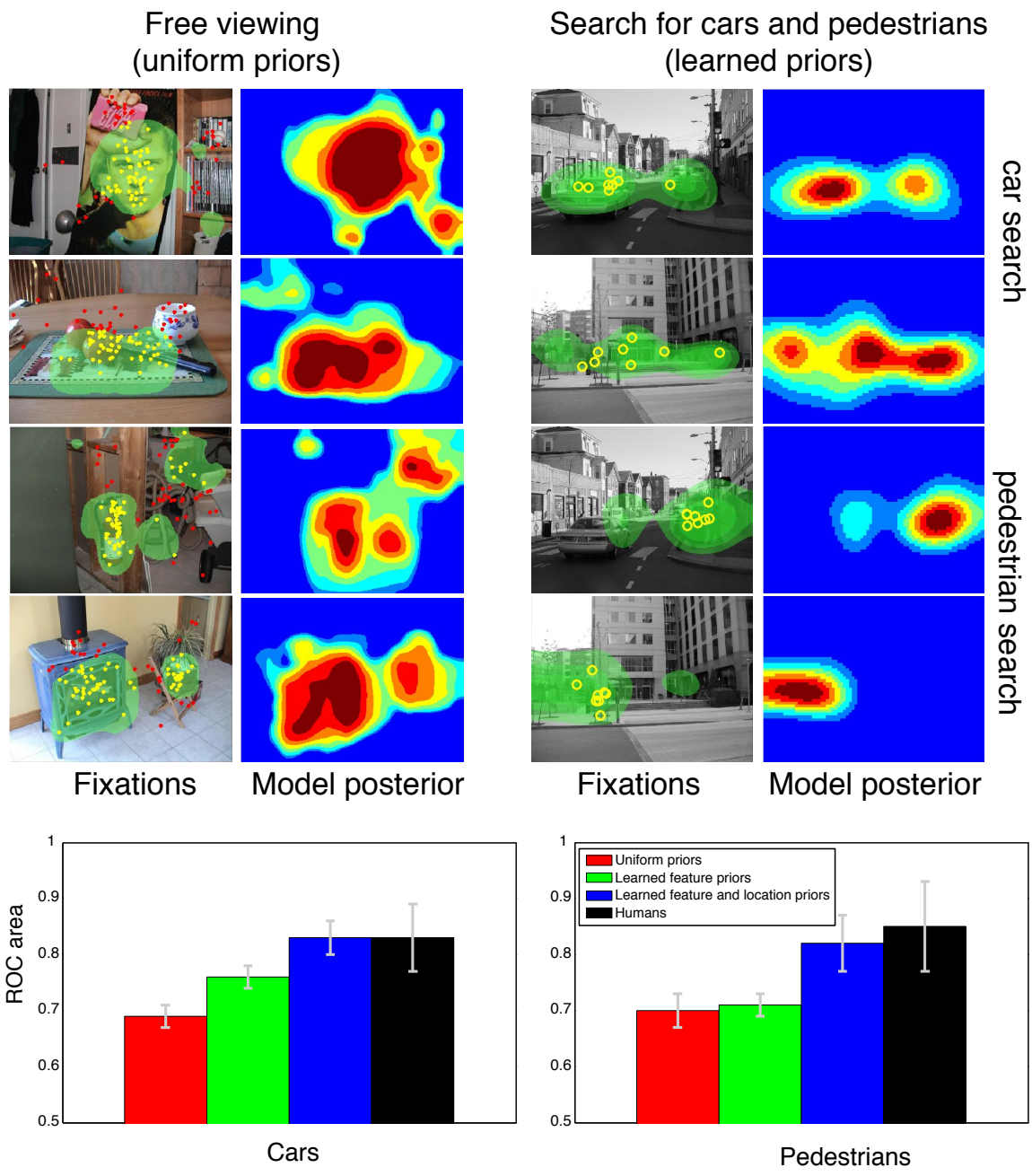


Figure 4-1: Predicting human eye movements: *(a)* Agreement between the model and human eye fixations during free viewing (left) and a complex visual search for either cars or pedestrians. Sample images overlaid with most salient (top 20%) regions predicted by the model (green) along with human eye movements (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.

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; 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. 4-1). 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 4.1).[3] 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.

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

Table 4.1: Comparison of the proposed bayesian model with shape-based features with prior work that relies on low level features.

4.1.2 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 *et al.*, 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 8 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),

[3] 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.

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 seconds, and within this time observers had to count the number of targets (cars or pedestrians) and press a key to indicate completion. 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 *et al.* (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 is provided in Sec. A1.

As assumed in several previous psychophysical studies (Itti and Koch, 2001a; Rao *et al.*, 2002a; Torralba *et al.*, 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. 4-1(b) shows the agreement between the model (and how the location and feature

priors influence performance) and human observers for the first fixation. Table A2 and A3 provide comparisons for additional number of fixations and against other models of eye movements (Itti and Koch, 2001a; 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. 4-1(b)). Beyond the first saccade, the agreement between model and human fixations decreases while the inter subject agreement increases (see Table A2 and A3). 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 and Koch, 2001a) in Table A2 and A3), does worse than chance. Learning either spatial priors or feature priors improve 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 *et al.* (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 *et al.*, 2003). Here we provide a quantitative evaluation of the proposed model of attention for detecting objects in images as opposed to predicting human eye movement. Table 4.2 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 lied 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

	Car	Pedestrian
Bottom up (Itti and Koch, 2001a)	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**

Table 4.2: Comparison between the performance of the various models to *localize* objects. The values indicate the area under the ROC.

the predictive ability of the individual models. The context (gist) representation derived from shape-based units (Serre *et al.*, 2005b) performs better than the representation based on simple oriented features (Torralba *et al.*, 2006). As expected, bottom-up cues derived using shape-based features performs better than bottom-up saliency obtained using simple oriented features (Itti and Koch, 2001a).

Chapter 5

Beyond attention: a non-bayesian extension

5.1 Beyond attention: a non bayesian extension

5.1.1 Recognition in clutter

In previous work, we have shown that feedforward hierarchical models of object recognition work relatively well for the recognition of objects presented in isolation and/or with limited (up to 3-5 objects) background clutter (Serre *et al.*, 2005a, 2007c) but that their level of performance is affected by the presence of clutter (see (Serre *et al.*, 2007a,b)). Other clutter effects have been well documented in the context of human psychophysics for the recognition of artificial letter stimuli (Pelli and Tillman, 2008) as well as objects in natural images (Serre *et al.*, 2007a). The detrimental effect of clutter has also been reported at the single cell electrophysiology level whereby the selectivity of neurons is reduced when multiple stimuli fall within their receptive fields (Zoccolan *et al.*, 2007; Reynolds *et al.*, 1999; Riesenhuber and Poggio, 1999a; Missal *et al.*, 1997). A natural solution to this problem is a spotlight of attention – a mechanism to suppress regions of the image that are unlikely to contain objects to be recognized (Walther and Koch, 2007). We propose that this mechanism – and the algorithms to support its function, in particular the choice of the regions to be suppressed – is in fact visual attention.

During a visual search for a specific feature or object, the top-down feature-based attentional mechanisms first bias the saliency map towards locations that share features with the target. The sequence of messages are identical with feature-based attention ($O \rightarrow F^i \rightarrow X^i \rightarrow L$). The saliency map ($P(L|I)$) then provides the most likely location containing the target. The region around the location with the maximum saliency ($P(L|I)$) can be chosen as the center of the spotlight of attention.

The search now proceeds with the deployment of the spatial attention around the region of interest. It is to be noted that this step is outside the bayesian framework. Thus searching for an object involves several iterations of the core bayesian model and thus several inference cycles.

To locate subsequent objects, the attentional spotlight is shifted (possibly via the PFC and/or FEF onto LIP) to the next location (Posner and Cohen, 1984). In subsequent experiments we have set the spotlight to correspond to a fixed size grid of X^i units (*i.e.,* 3×3

spatial arrangements). It is interesting to note that because of the multi-scale representation of these features in the original model (Serre *et al.*, 2005b) this translates into a spotlight in retinotopic coordinates with a variable size and capable of attending to objects at multiple scales.

In the following, we consider two realistic visual tasks involving object recognition in clutter and provide evidence suggesting that the strategy described above does work for the recognition of objects in complex cluttered visual scenes.

5.1.2 Artificial search arrays

We created search arrays of artificial stimuli and evaluated the performance of the model operating in two distinct modes to detect the presence/absence of an object: (a) a pre-attentive mode characterized by a single feedforward sweep through the system (corresponding to processing by the original feedforward hierarchical model with no attentional mechanisms) and (b) an attentive mode whereby cortical loops are active. Examples of stimulus display used is provided on Fig. 5-1 (see also Fig. A1). Here we used the same set of stimuli as in (Hung *et al.*, 2005) (see Fig. A2). The dataset consists of 76 images belonging to 8 image categories (food, faces, monkey faces, hands, vehicles, lines and toys). Search arrays were generated with 1, 2, 4 and 8 items placed at random locations. We generated 50 images containing the target and 50 images where the target was absent. We repeated this process for each target category and cardinality thus generating 800×4 images in all.

Because the positions of items in these artificial search arrays are selected at random, the location priors $P(L)$ were set to be uniform. Fig. 5-1 shows how switching the identity of the object to be detected influences the computation of the posterior probabilities for the location variable $P(L|I)$. Details about the learning of object and feature priors is described in Sec. A1. Fig. 5-1 shows that the detection performance of the model operating in pre-attentive feedforward mode degrades with increasing number of distractors (performance reaching chance level around eight objects). This is similar to previously reported results (Serre *et al.*, 2007b). Conversely, in the attentive mode, the effect of clutter is much less

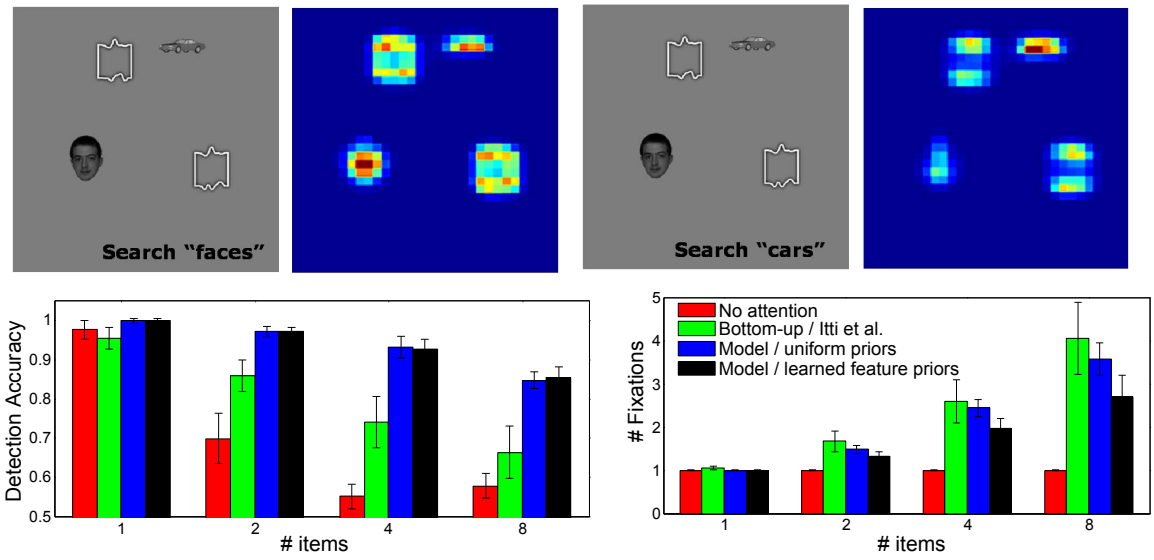


Figure 5-1: Performance measure for feature-based attention during search and recognition in artificial search arrays. *(a)* Top-down modulation and biasing of the image saliency towards the search target with feature-based attentional mechanisms. *(b)* (Left) Object recognition performance in an array of distractors: In the absence of attentional mechanisms, the performance of the feedforward hierarchical model of object recognition (red) degrades with the number of items. Extending the model to incorporate top-down feature-based attention (black) improves recognition performance at a level higher than obtained with an alternative bottom-up saliency model (Itti and Koch, 2001a). (Right) Search efficiency: The average number of attentional shifts required to find the target increases with the number of distractors. For a given number of distractors, top-down feature-based attention using shape features is the most efficient (on average) for locating the target.

pronounced. Fig. 5-1 provides a comparison between the number of shifts of attention required to correctly localize a target image by (a) the top-down feature-based approach, (b) the same model setting the object and feature priors to be uniform (effectively approximating simple bottom-up saliency computations) and a classical bottom-up saliency algorithm by Itti & Koch (Itti and Koch, 2001a). On average, top-down feature based attention can locate objects of interest much faster than the bottom-up saliency version (2.6 shifts of attention vs. 3.6 for 8 item search arrays).

5.1.3 Complex Natural Scenes

In previous work, a feedforward computational model of the ventral stream of the visual cortex (Serre *et al.*, 2005b) was shown to account for the level of performance of human observers during a rapid (masked) animal vs. non-animal recognition task (Serre *et al.*,

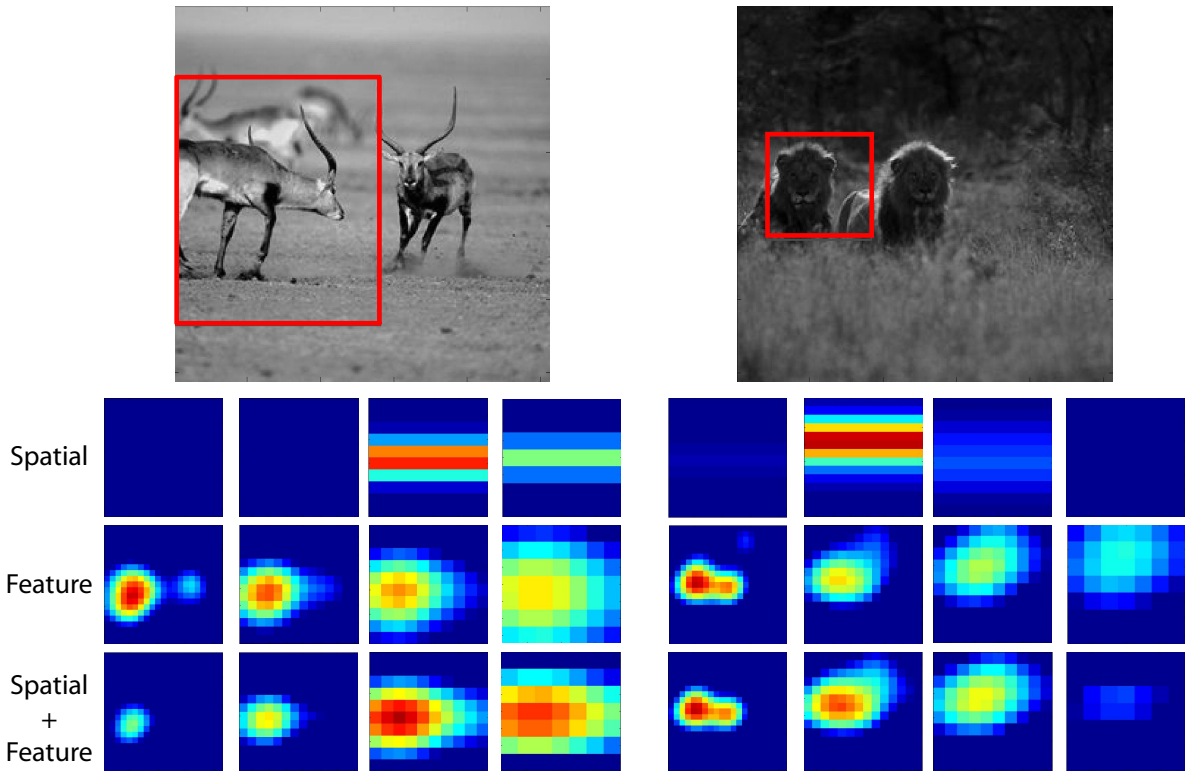
2007a). This finding suggested that under rapid masked presentation conditions, a hierarchical feedforward model may provide a satisfactory description of information processing in the ventral stream of the visual cortex. Additionally, we had also found that even for rapid but unmasked presentation, the agreement between the model and human observers decreased and in particular, human observers started to significantly outperform the model. Here we show the addition of cortical feedback and attentional mechanisms improves the recognition capability of the original hierarchical recognition architecture. We emphasize that the overall model considered here is not bayesian – only the attentional part is.

The dataset consists of 600 images containing one or more animals and 600 distractor images comprising of natural and artificial objects (see Fig. A5 and ref. (Serre *et al.*, 2007a) for examples). Image are organized into four categories based on the amount of clutter and relative size of the animal with respect to background (The categories are "head", "near body", "medium body" and "far-body"). We trained the model of attention to localize animals (details are provided in Sec. A1.5). Fig. 5-2 (a) shows how the combination of top-down spatial priors together with top-down feature-based priors about animals bias the saliency map ($P(L|I)$) towards locations which are likely to contain an animal.

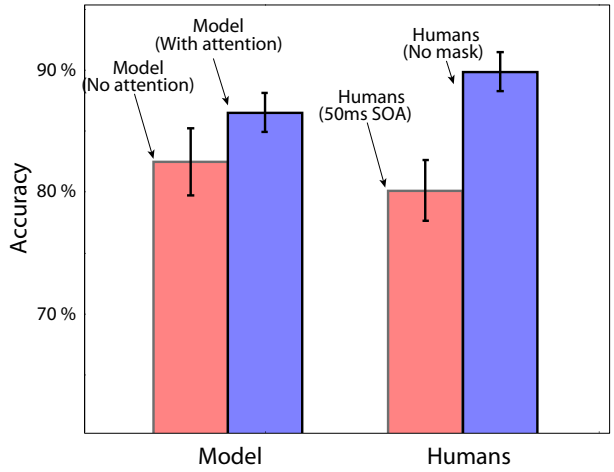
A comparison between the performance of human observers vs. the models is shown in Fig. 5-2 (b). The bayesian model of attention combines top-down feature and spatial priors with evidence from the image to generate a saliency map which is then used to generate spotlights of attention around the likely locations of animals. Examples of attentional spotlights generated by the model are provided in Fig. 5-2(a) (see also Fig. A6).

The detection accuracy of the model increases from 82.5 ± 2.75 under purely feedforward condition to 86.5 ± 1.6 when attentive processing is added to the feedforward network. Human performance with the same stimuli increases from 80.12 ± 2.5 when the SOA (corresponding to the delay between stimuls onset and the mask) is 50ms to 85.375 ± 2.125 when the SOA is 80ms. When the mask is removed entirely, human performance improves further to 89.875 ± 1.625. Our results show that the performance of the extended model increases in a manner similar to the performance of human subjects who were given additional time to process the images. This suggests that attention may play an important role in recognition in clutter. However, the performance for the no-mask condition is not

fully accounted by the attentive processing considered here. The human visual system may rely upon higher level information that is not available to the model. It is to be noted that attention in general is not guaranteed to improve recognition performance. In fact, in an early filtering scheme such as the one described, an object is not detected (no matter how good the classifier is) if it is not selected by the attention mechanism. Our empirical studies show that attention can improve recognition only in situations where the features used for the attentional mechanism are sufficiently decorrelated from those used for recognition. However, attentional mechanisms always improve the speed of recognition.



(a) Combination of spatial and feature-based priors in the model



(b) Human vs. model on the animals vs. non-animal task.

Figure 5-2: Animal vs. non-animal categorization task: comparison of the model with human performance. Typical posterior maps produced by the model for spatial ($P(L|I)$) and feature-based ($\sum_i P(X^i|I)$) attention as well as their combination (organized by scale from left to right: $0.23\times, 0.36\times, 0.55\times$ and $0.85\times$ the size of the image). (b) Performance of the model with and without attention and comparison against human observers with and without mask (see (Serre *et al.*, 2007a) for details).

Chapter 6

Discussion

6.1 Relation to prior work

A few theories and several specific models (see Table A4 and A5 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 Tsotos (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; Zhang *et al.*, 2008; Rosenholtz and Mansfield, 2005), top-down bias of target features (Wolfe, 2007; Navalpakkam and Itti, 2006), influence from scene context (Torralba, 2003b), serial vs. 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.

6.2 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. The first test for the theory is computational, *i.e.,* whether it indeed "solves" the basic recognition problem. For this we checked that the attentional model helps a feedforward model to improve recognition performance in the case of natural, complex images. We also 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 – suprisingly – a number of psychophysical and physiological properties of attention that were so far explained using different, and somewhat *ad hoc* mechanisms.

Appendix A

Implementation details

A1 Methods

In this work, we extend the feedforward model of the ventral stream to include position information and attentional feedback. The main addition to the original feedforward model (Serre *et al.*, 2005b) is (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 serve as a 'saliency map' encoding the visual relevance of individual locations (Koch and Ullman, 1985). The feedforward and feedback information are combined using a bayesian framework. In the following, we describe the model of the ventral stream as well as the bayesian extension.

A1.1 Ventral ('what') stream model

The 'what' stream corresponds to the model described in (Serre *et al.*, 2005b). This model is representative of the class of feedforward hierarchical models of object recognition. It builds on previous models (Wallis and Rolls, 1997; Mel, 1997; Riesenhuber and Poggio, 1999b; Ullman *et al.*, 2002; Thorpe, 2002; Amit and Mascaro, 2003; Wersing and Koerner, 2003), conceptual proposals (Hubel and Wiesel, 1968; Perrett and Oram, 1993; Hochstein and Ahissar, 2002; Biederman, 1987) and computer vision systems (Fukushima, 1975; LeCun *et al.*, 1998). It was showed (Serre *et al.*, 2005b) to be consistent with experimental data in V4 (Cadieu *et al.*, 2007) and IT (Hung *et al.*, 2005) (see (Serre *et al.*, 2005b) for reviews). The model was also shown to be able to fit the performance of human observers during a rapid animal vs. non-animal categorization task (Serre *et al.*, 2007a), a task which may likely to rely on bottom-up feedforward processing (VanRullen and Koch, 2003).

Here for simplicity we focused on the part of the model called the 'bypass route', modeling the ventral stream in a 4-layer hierarchy corresponding to V1/V2 $\rightarrow$ V4/PIT $\rightarrow$ AIT (layers $S_1/C_1 \rightarrow S_{2b} \rightarrow C_{2b}$). Details about the model can be found elsewhere (Serre *et al.*, 2005b). In the following we provide a short description of its implementation. The ventral stream model is built up using a hierarchy of simple (S) and complex (C) units (here the term *unit* is used instead of *cell* to distinguish the entities in the model from their

counterpart in biology). The S units provide selectivity to specific stimulus while the C units provide invariance by aggregating information from the afferent S units. Both the selectivity and invariance exhibited by individual units increases as we go higher up in the hierarchy. At the highest level we find units that respond to individual object categories at any location or scale. In this paper, we make use of layers S1, C1, S2 and C2 as outlined in Serre *et al.* (2007c). Several other variations of this model exists (Mutch and Lowe (2006); Ranzato *et al.* (2007)).

S1 Units: The S1 units are tuned to bars of specific size and orientation. S1 unit response $S1_{(s,d)}(x,y)$ at scale s and direction d is computed using a normalized convolution operation

$$S1_{(s,d)}(x,y) = \frac{\hat{G}(x,y,\theta) * I(x,y)}{\sqrt{I^2(x,y) * H(x,y)}} \tag{A1}$$

Here $\hat{G}(x,y,\theta)$ is the zero mean,unit normal Gabor filters tuned to orientation θ and $H(x,y)$ a filter with all ones and the same size as the Gabor filter. The Gabor filters are given by

$$G(x,y) = \exp\left(-\frac{u^2+\gamma v^2}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}\right) \tag{A2}$$

$$\text{and} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{A3}$$

Here γ represents the eccentricity/anisotropy of the filter, θ is the orientation and λ is the wavelength of the filter. In the current implementation, we compute the S1 response at sixteen different scales and four orientations.

C1 Units: The C1 units pool afferents from several S1 units within its receptive field and also across adjacent scales using a max operation. This operation done by the C units is the key to the invariance exhibited by the model, since the C unit responds to the afferent pattern occurring anywhere within its receptive field and at any scale.

S2 Units: The S2 units respond to simple shapes and have a gaussian-like tuning function

to their preferred stimulus. The response of an individual S2 unit is given by

$$S2_i = \exp\left(-\frac{||C_1 - P_i||^2}{2\sigma_i^2}.\right) \tag{A4}$$

Here, P_i represents the preferred stimulus (dictionary elements) that was learnt during training. The sharpness of the tuning is determined by σ_i and is designed to provide low response for blank input.

C2 Units: The C2 cells again pool responses S2 units across 3×3 locations and adjacent scales. The C2 features are used as the bottom-up input data in our attentional model.

A1.2 Bayesian model

We implemented the bayesian network using Kevin Murphy's bayesian toolbox available at `http://bnt.sourceforge.net`.

Integrated model of attention and recognition The bayesian model takes as input, the V4-like ($C2$) feature maps of the feedforward model of the ventral stream (Serre *et al.*, 2005b). These are mapped to the X^i units in the bayesian model. The $\mathrm{X^i}$ units receives bottom-up evidence $P(I|X^i)$. For instance, if features F^i correspond to orientations, the feature map is computed using oriented Gabor filters (Daugman, 1980). Given the image I, for each orientation and location, $P(I|X^i)$ is set proportional to the output of the filter. The response can be passed through a sigmoid or even discretized without affecting the model. The original ventral stream (Serre *et al.*, 2005b) model used several thousands of continuous valued shape-based features. In practice, in order to minimize the computational complexity of the attentional component, a subset of the shape-based features that are discriminative for the target object category was selected using a mutual information driven process (Fleuret, 2004). Given the desired number of features, the feature selection algorithm provides the most discriminatory and mutually independent features to use and also the corresponding thresholds that maximize the mutual information between the features and the objects. Using these thresholds, we discretize all the feature maps (Thus, $P(I|X^i)$ can be 0 or 1). The next stage in the computation proceeds using message passing

algorithm and is agnostic to the origin of the individual features.

<table>
<tr><th>Conditional Probability</th><th>Modeling</th></tr>
<tr><td>$P(L)$</td><td>Each scene or view-point places constraints on the location and sizes of objects that can be encountered in the image. Such constraints can be specified explicitly (e.g. during spatial attention) or learned using a set of training examples (Torralba, 2003b).</td></tr>
<tr><td>$P(F^i|O)$</td><td>The probability of each feature being present or absent given the object and is directly learned from the training data.</td></tr>
<tr><td>$P(X^i|F^i, L)$</td><td>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 can be captured succinctly by the following table
<table>
<tr><td></td><td>$F^i = 1, L = l$</td><td>$F^i = 0, L = l$</td></tr>
<tr><td>$X^i = 0$</td><td>$P(X^i|F^i, L) = \delta_1$</td><td>$P(X^i|F^i, L) = 1 - \delta_2$</td></tr>
<tr><td>$X^i \neq 0$</td><td>$P(X^i|F^i, L) \sim$ Gaussian centered around $L = l$</td><td>$P(X^i|F^i, L) = \delta_2$</td></tr>
</table>
δ_1 and δ_2 are small values. They are chosen to ensure that $\sum P(X^i|F^i, L) = 1$.</td></tr>
<tr><td>$P(I|X^i)$</td><td>For each location within the feature map, $P(I|X^i)$ provides the likelihood that X^i is active. In the model, this bottom-up evidence or likelihood is set proportional to the activations of the shape-based units (see (Serre et al., 2007c)).</td></tr>
</table>

Table A1: Description of the model conditional probabilities.

A1.3 Search and recognition of objects in artificial search arrays

The feedforward recognition component was trained with images of isolated objects. Additional (virtual) training examples were generated using translation and scaling of the original images. Here we used ≈ 100 shape-tuned features and a linear SVM for readout analysis as done in (Hung *et al.*, 2005). A separate classifier was trained for each category (using the 'one vs. all' strategy).

In the pre-attentive mode, *i.e.,* for recognition in the absence of attention, the entire image was analyzed by the hierarchical feedforward model. A target was considered detected if the classifier output for the corresponding object category was positive. To evaluate the

performance of the integrated model of recognition and attention, the attentional component was first used to generate attentional spotlights based on local maxima in the saliency map (corresponding to $P(L|I)$, the posterior probability of location). Processing by the feedforward model was limited to this region of interest. The corresponding region (and its immediate surrounding) was then inhibited and the process was repeated until the target was found or the maximal image saliency fell below a fixed threshold (10% of the maximum value).

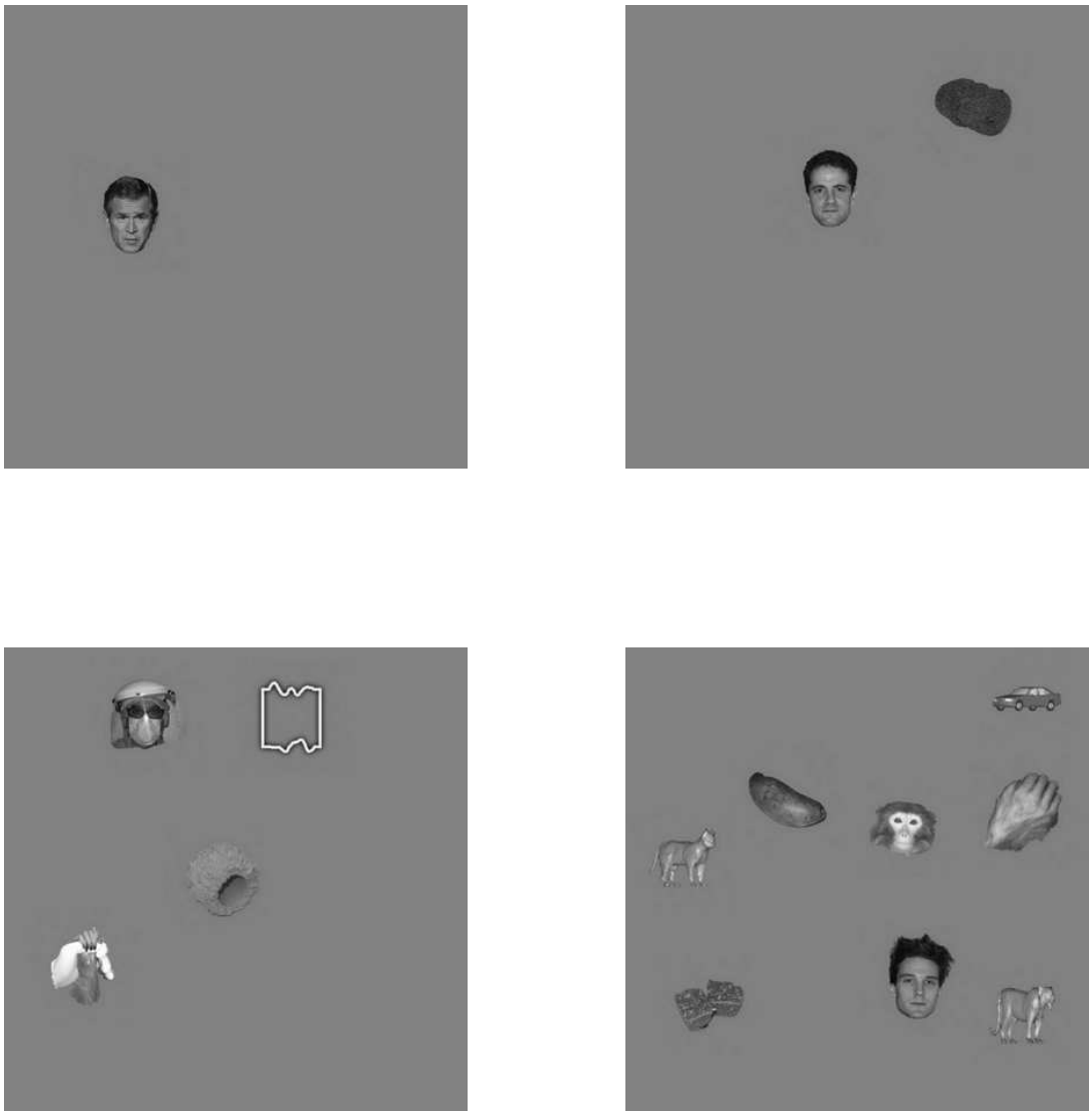


Figure A1: Example stimuli with one, two, four and eight items.

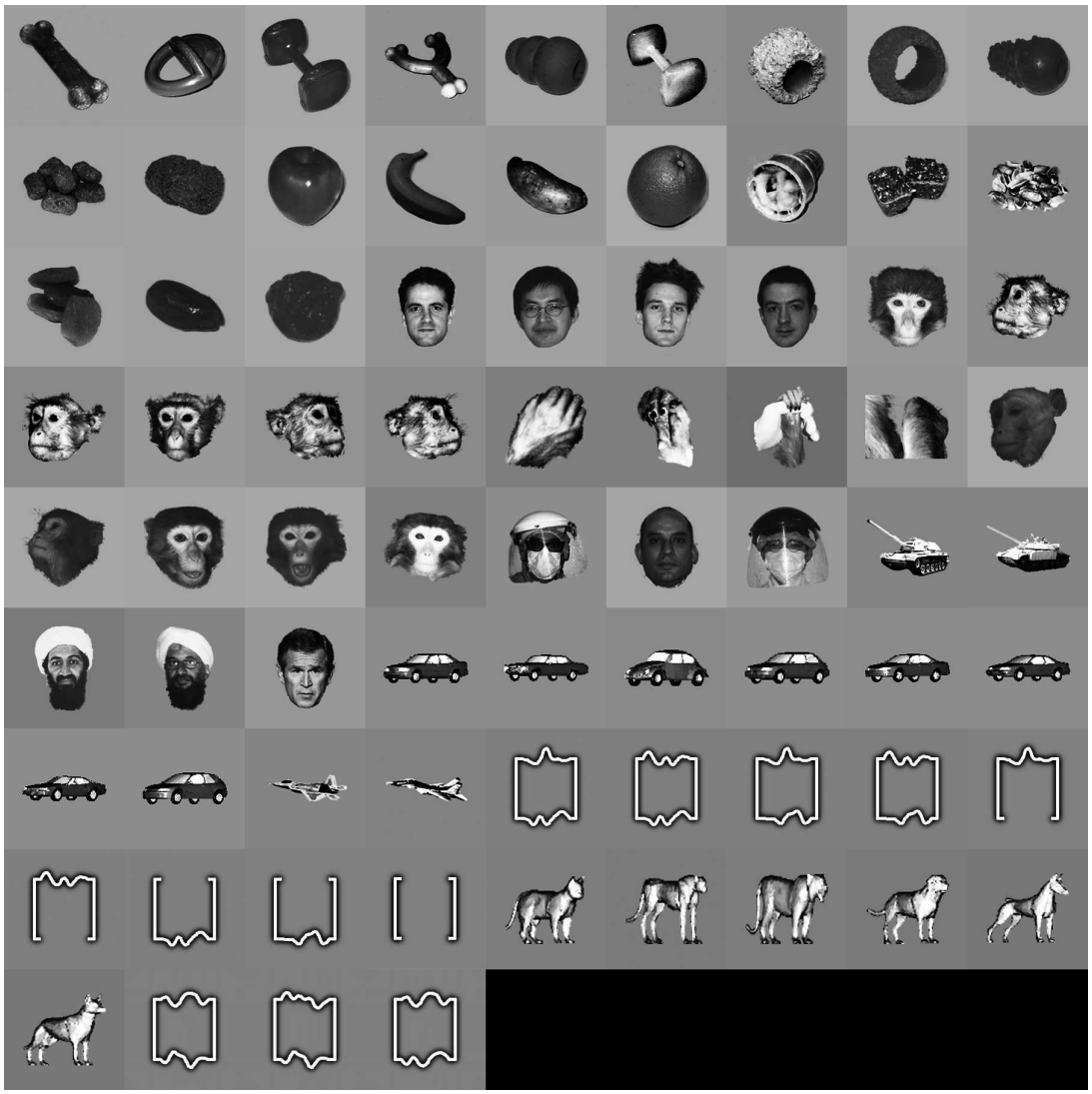


Figure A2: 77 stimuli used to create the artificial search array.

A1.4 Predicting effects of attention in IT

Data The 16 objects used to create the stimuli are shown in Fig. A3. The stimuli was generated by placing one or three of these objects in the image. The objects can occupy only three possible positions in the image that subtended an angle $0°$,$-60°$ or $60°$ to the horizontal. The objects appear at a distance of $5.5°$ of visual angle from the center of the image.

Features The objects consist of fruits, couches, face and cars. To extract features that are responsive to these objects, we used the CSAIL Labelme (Russell *et al.*, 2008) dataset (855 cars,63 couch,120 face, 51 fruit images) to derive the feature dictionary. Starting from a large pool of randomly sampled features, a mutual information driven feature selection (Fleuret, 2004) was performed to select 60 most informative features. The selection al-

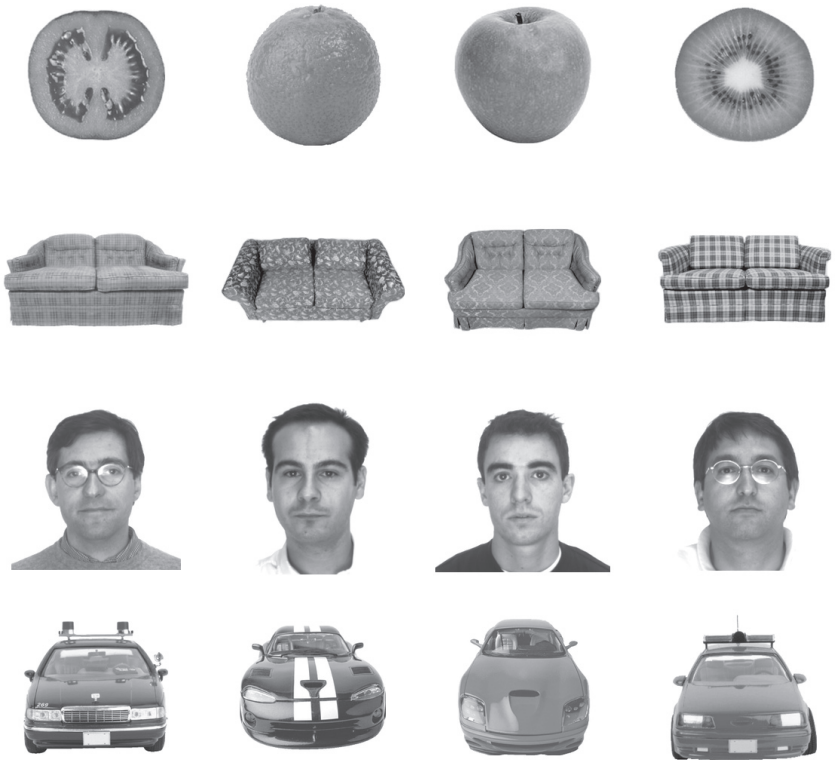


Figure A3: The set of 16 objects used to create the stimulus.

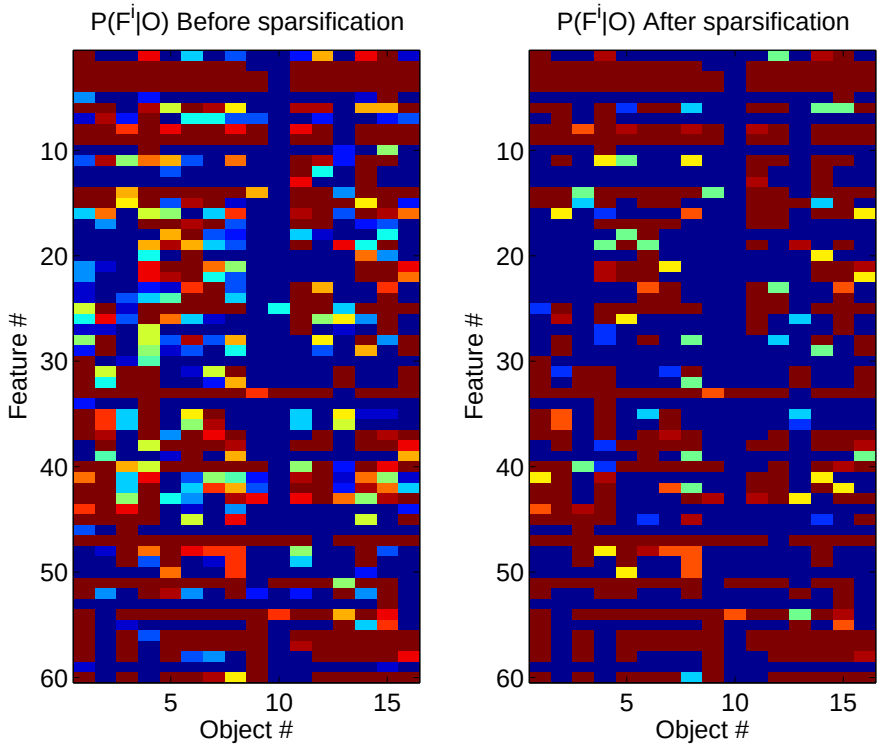


Figure A4: The arrays represents the conditional probability table $P(F^i|O)$ before and after the sparsification procedure. Each pixel in the array represents the value of $P(F^i = 1|O = o)$, for a specific $\{i, o\}$ combination.

gorithm also provides the thresholds to optimally detect these features. Although this is far fewer than the number of IT neurons recorded in the original experiment, they respond more selectively to the set of 16 objects.

Training To estimate the probabilities $P(F^i|O)$ for each object, we used scaled versions of the 16 objects (corresponding to 16 logarithmically spaced scales between 0.8 and 1.25 times the original size). A feature was assumed to be presented if the value of the feature map at any location exceeded the threshold obtained during feature selection. $P(F^i|O)$ was then estimated using the counts. An undesirable effect of having to share all the features among the objects is that the number of features whose presence indicates the object is far fewer than the number of features that are usually absent in the object. The absence of these features provide evidence for the presence of objects. Furthermore, when the target object is present among other distracters, the desirable features are still present. However,

the presence of distracter features artificially decrease the likelihood of the object being present. In order to prevent this, we set $P(F^i|O) = min(0.5, P(F^i|O))$ such that when features are likely to be absent in an objects, the probability is modified such that they are non-informative. The conditional probability matrices before and after this procedure is shown in Fig. A4.

A1.5 Attention helps object recognition in complex natural scenes

Feature priors $(P(F^i|O))$ Here we used a dictionary of ≈ 100 shape-tuned features. The features are computed at four different scales corresponding to $0.85\times$, $0.55\times$, $0.36\times$ and $0.23\times$ the size of the image. Accordingly, the location units span multiple scales (corresponding directly with that of X^i). Having multi-scale features allows us to determine the location as well as the size of the animals. Using a set of 600 training images, the occurrence probability of these individual features ($P(F^i|O)$) within animal and non-animal images was learned.

Location priors $(P(L))$ Computer vision systems that are based on scanning (Bileschi, 2006; Dalal *et al.*, 2006) do not make any assumption about the scene in which the objects are found. However, in the natural world, the scene and the objects within it share *contextual* relationships. Based on this observation, Torralba et al. (Torralba, 2003b) proposed that a computational model of scene gist could similarly be used to reason about object probability and position prior to object detection in computer vision applications. Previous approaches have used gist representations based on spatial distribution of oriented filter responses (Torralba, 2003b; Itti *et al.*, 2005). In this work, we use biologically inspired shape descriptors (Serre *et al.*, 2007c) to describe the 'shape' of the scene.

The association between the image and the location (and scale) of animals was learned using a mixture of regressors (Murphy *et al.*, 2003). Contextual priors were learned in a way similar to described by Torralba (Torralba and Oliva, 2003). Given a vectorial scene-gist representation G, the probability of x-location, y-location and scale ($X = x, y, \sigma$) is

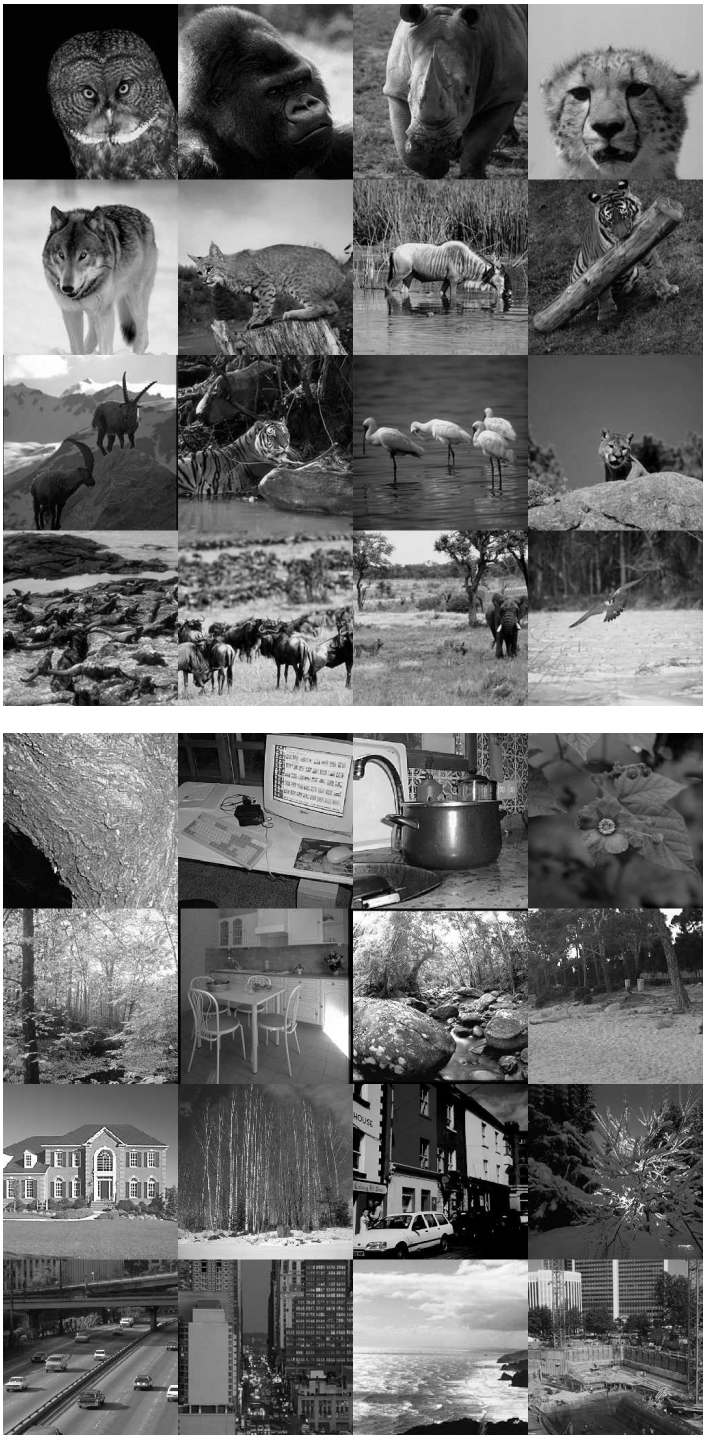


Figure A5: The animals vs. non-animals dataset used in (Serre *et al.*, 2007a).The images in the dataset are divided into four categories with the depth of view and the amount of clutter increasing along the rows. The distractor non-animal images are matched for depth.

given by

$$P(X|G) = \sum_k P(K|G)P(X|K,G) \quad \text{(A5)}$$

$$P(X|K,G) \sim N(\mu_K + A_k^T G, \Sigma_K) \quad \text{(A6)}$$

$$P(K|G) \sim \text{softmax}(K; W_k^T G) \quad \text{(A7)}$$

K represents the canonical view, each of which imposes a different distribution on the object location and size. A_k, μ_K represent the parameters of the individual regressors. The weights W_k and the softmax function provide a smooth transition between different constraints. $P(X|K)$ specifies the individual regressor for view K. To decrease the learning time and to avoid over fitting, we reduce the dimension of the individual representations using PCA. We retain only the top 32 principal components. The number of regressors was fixed at $K = 5$ for all representations. Our informal study on the selection of K did not show any difference for higher values of K. Once $P(X|G)$ is obtained, it is transformed to its discrete counterpart $P(L)$. Admittedly, this procedure is done outside of the bayesian framework before inference is done.

Recognition Given a test image, a multi-scale saliency map (corresponding to the posterior probability of location, $P(L|I)$) was generated using the bayesian model of attention (see Fig. A6). The most salient location and scale was then used to generate a spotlight of attention (the size of the spotlight being determined by the scale of the saliency map for the corresponding local max). The region within the spotlight was isolated and further processed by the feedforward hierarchical model for final animal vs. non-animal categorization. Here we only allowed the model to make one shift of attention.

A1.6 Predicting eye movements

Free viewing

Here we used set of 500 randomly chosen images of outdoor and indoor scenes from the *CSAIL LabelMe* dataset (Russell *et al.*, 2008) for training the attention model. These im-

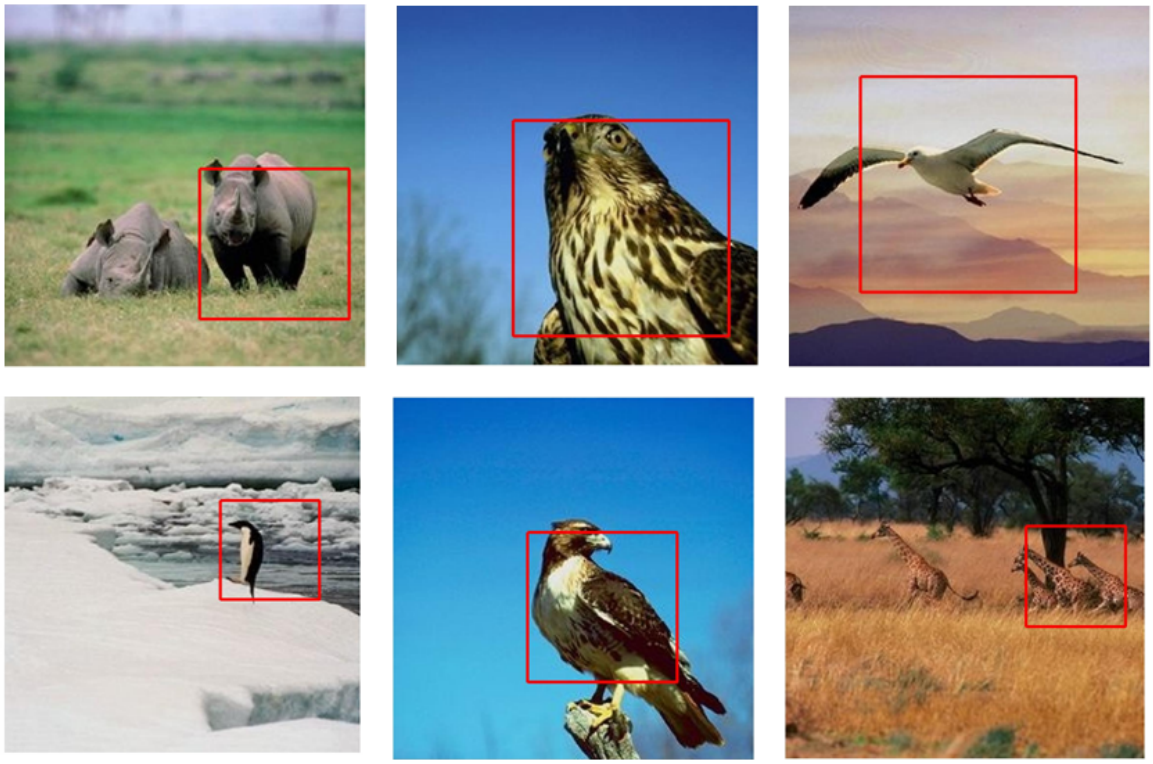


Figure A6: The figure illustrates the attentional windows computed using the bayesian model on several images. Note that the size of the attention is not fixed. Instead it is determined by the scale corresponding to the most salient location.

ages are visually similar to the images used in the actual psychophysics experiment. Several hundred shape-tuned features were first extracted via random sampling in these images (see (Serre *et al.*, 2005b) for details). To speed up the system at run-time, we selected a small subset of 32 features with the largest variance (which we assume to be the most informative). Here for a fair comparison with the approach by Bruce & Tsotsos, in addition to the shape features, we included 6 color features corresponding to normalized (z-score) and half rectified intensities in the LAB color space. Training consisted in determining the threshold for both color and shape features. We estimated the distribution of each feature by exhaustively sampling the responses from all the training images. The threshold was then chosen such that only 20% of the responses would be active (*i.e.,* above threshold) on average across images. This is equivalent to the 80th percentile value in the estimated distribution. Choosing a higher threshold produces a sparser saliency map and does not seriously affect the results.

Search in natural images

Feature priors $(P(F^i|O))$ To train the model, we used part of the CBCL Street scene database (Bileschi, 2006) and part of *LabelMe* (Russell *et al.*, 2008). We used about $32,000$ training examples (crops) total, which included both cars and pedestrians (about $3,000$ examples of each). The negative examples were randomly extracted from the database, and then pruned to exclude regions that overlapped with cars or pedestrians. To train the model, we started by extracting $1,000$ shape-tuned features randomly sampled from training data (see (Serre *et al.*, 2005b). Using this data, 200 features were selected using a feature selection process based on mutual information (Fleuret, 2004). Probabilities $P(\mathbf{F^i}|\mathbf{O})$ were obtained via maximum-likelihood estimation.

Location priors $(P(L))$ Here, we consider ~ 500 shape-based units to represent the scene gist. These units have a larger receptive field compared to the ones used for representing objects, but are derived using the same computation (Serre *et al.*, 2007c). The responses are pooled in a 3×3 overlapping grid (each grid corresponding half the width of the original image) using a *max* operation. This permits the detection of local image configurations in a translation invariant manner (Serre *et al.*, 2007c). The resulting 4500 dimensional vector is further reduced to 32 dimensions using PCA. This 32 dimensional vector represents the "context" of the objects in the scene that can be used to determine likely locations of objects in the scene. We use a mixture-of-regressors as in (Murphy *et al.*, 2003) to learn the mapping between the context features and location/scale priors for each object.

Fig. A7 shows the result of a manifold-learning analysis (Roweis and Saul, 2000) directly on the output of the $|\mathrm{K}| = 5$ experts trained on the street scenes (each image point is assigned to an expert – one color for each expert – by soft-assignment on the corresponding probability distributions). Visual inspection suggests that the mixture of experts learned canonical views of the street scenes (e.g. dark blue centers being "side views with building" while light blue centers represent far views). Overall the analysis reveals that the shape-based features are able to capture the smooth variations going from one canonical view to another, thus providing a good representation for the scene.

Figure A7: Visual inspection suggests that centers in the mixture of experts correspond to canonical street scenes (see text for details).

	First	First two	First three
Bottom up (Itti and Koch, 2001a)	0.44 ± 0.03	0.43 ± 0.03	0.42 ± 0.02
Context (Torralba *et al.*, 2006)	0.80 ± 0.04	0.80 ± 0.05	0.79 ± 0.05
Model / uniform priors	0.69 ± 0.02	0.68 ± 0.01	0.68 ± 0.01
Model / learned spatial priors	0.82 ± 0.05	0.81 ± 0.05	0.80 ± 0.04
Model / learned feature priors	0.76 ± 0.02	0.74 ± 0.03	0.73 ± 0.03
Model / full	$\mathbf{0.83 \pm 0.03}$	$\mathbf{0.81 \pm 0.03}$	$\mathbf{0.80 \pm 0.03}$
Humans	0.83 ± 0.06	0.88 ± 0.04	0.88 ± 0.03

Table A2: Search for cars in street scene images. Values indicate the area under the ROC. For each object, the ability of the models to predict the first, two and three fixations is indicated. The saliency model by Itti & Koch (Itti and Koch, 2001a) corresponds to the implementation available at (http://saliencytoolbox.net).

	First	First two	First three
Bottom up (Itti and Koch, 2001a)	0.44 ± 0.03	0.42 ± 0.03	0.42 ± 0.02
Context (Torralba *et al.*, 2006)	0.780 ± 0.07	0.79 ± 0.07	0.77 ± 0.07
Model / uniform priors	0.70 ± 0.026	0.71 ± 0.02	0.70 ± 0.01
Model / learned spatial priors	0.80 ± 0.08	0.79 ± 0.08	0.78 ± 0.07
Model / learned feature priors	0.71 ± 0.02	0.71 ± 0.02	0.69 ± 0.02
Model / full	$\mathbf{0.82 \pm 0.05}$	$\mathbf{0.81 \pm 0.05}$	$\mathbf{0.80 \pm 0.05}$
Humans	0.85 ± 0.08	0.85 ± 0.08	0.87 ± 0.03

Table A3: Search for pedestrians in street scene images. Values indicate the area under the ROC. For each object, the ability of the models to predict the first, two and three fixations is indicated.

A2 Discussion

A2.1 The neuroscience of visual attention

In this section, we identify the primary regions of the brain involved in attention and their interconnections. Koch and Ullman (1985) postulated the existence of a *saliency map* in the visual system. The *saliency map* combines information from several abstract feature maps (*e.g.*, local contrast, orientations, color) into a global saliency measure that indicates the relevance of each position in the image. Consistent with this hypothesis, models of attention have assumed that there exist two stages of visual processing. In a pre-attentive parallel processing mode, the entire visual field is processed at once to generate a saliency map which is then used to guide a slow serial attentive processing stage, in which a region of interest (attentional spotlight) is selected for "specialized" analysis. The attentional spotlight may be guided in several ways. It can be driven by a spatial cue (e.g. "what object is at the center of the image?") or can be feature/object based (e.g. "where is the red square?"). Further, attention can be classified as being bottom-up (stimulus driven) or top-down (task driven). In bottom-up attention, the attentional shifts are purely image driven and is independent of any task. (e.g. a bright red sign on the street attracts our attention irrespective of whether we were looking for it). However, the neural correlate of the saliency map remains to be found.

Prior studies (Colby and Goldberg, 1999) have shown that the parietal cortex maintains a spatial map of the visual environment and in fact maintains several frames of reference (eye-centered, head-centered etc) making it a likely candidate for the saliency computation. Studies show that response of LIP neurons within the parietal are correlated with likelihood ratio of the target object (Bisley and Goldberg, 2003). In this paper, our hypothesis – which is not critical for the theory and is mainly dictated by simplicity – is that the saliency map is represented in LIP. In addition to computing saliency, circuits are also needed to plan the shifts of attention, that is, to plan and serialize the search by prioritizing candidate shifts of attention and holding them in working-memory until the saccade has been initiated. Because of its overlap with the prefrontal cortex (PFC), the frontal eye field (FEF) is a good candidate for shifting the focus of attention. Recent evidence (Buschman and Miller, 2007)

further supports the role of FEF in spatial and feature based attention. We speculate that the FEF stimulation effect reported by Moore and Armstrong (2003) (i.e., an enhancement observed in V4 receptive field locations that match the region of the visual field represented at the FEF stimulation site) is indirect, mediated through LIP.

In addition to the parietal region, the ventral stream is also intimately involved in attention. Li and Snowden (2006); Itti *et al.* (1998) have proposed computational models based on V1-like features showing that they are sufficient to reproduce attentional effects such as pop-out and search asymmetries. However, recent evidence shows V1 to be relatively unaffected by top-down attentional modulation (Hegde and Felleman, 2003), thus moving the locus of attention away from V1 and towards higher regions such as V4. Experiments on spatial attention (McAdams and Maunsell, 1999) and feature-based attention (Bichot *et al.*, 2005) have shown attentional modulation in V4. In particular, feature-based attention is found to modulate the response of V4 neurons at all locations–the activities are increased if the preferred stimulus of the neurons is the same as the target stimulus and suppressed otherwise. Under spatial attention, V4 neurons that have receptive fields overlapping with the locus of attention are found to be enhanced. Thus V4 neurons are involved in feature-based attention as well as spatial attention suggesting that V4 serves as the area of interaction between ventral and parietal cortices.

In this work, we explicitly model the interaction between the ventral and parietal cortical regions (Rao, 2005; Van Der Velde and De Kamps, 2001) and integrate these interactions within a feedforward model of the ventral stream (Serre *et al.*, 2007c). The main addition to the feedforward model is (1) the inclusion of cortical feedback within the ventral stream (providing feature-based attention) and (2) from areas of the parietal cortex onto areas of the ventral stream (providing spatial attention) and, (3) feedforward connections to the parietal cortex that serve as a 'saliency map' encoding the visual relevance of individual locations (Koch and Ullman, 1985). The model is directly inspired by the physiology of attention and extends a bayesian model of spatial attention proposed by Rao (2005).

A2.2 Computational models of attention

Several theoretical proposals and computational models have been described to try to explain the main functional and computational role of visual attention. One important proposal by Tsotsos (1997) is 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 system 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. Duncan (1995) suggested that the goal of attention is to bias the choice between competing stimuli within the visual field. These proposals however remain agnostic about how attention should be implemented in the visual cortex and do not yield any prediction about the various behavioral and physiological effects of attention.

On the other hand, several computational models have attempted to account for specific behavioral and physiological effects of attention. Behavioral effects include pop-out of salient objects (Itti *et al.*, 1998; Zhang *et al.*, 2008; Rosenholtz and Mansfield, 2005), top-down bias of target features (Wolfe, 2007; Navalpakkam and Itti, 2006), influence from scene context (Torralba, 2003b), serial vs. 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). Table A4 provides a comparison of our approach with existing work in literature.

A2.3 Other approaches for modeling human eye movements

Our work builds on a number of computational (Tsotsos *et al.*, 1995; Itti and Koch, 2001a; Rao *et al.*, 2002b; Torralba *et al.*, 2006; Walther and Koch, 2007) and conceptual proposals (Wolfe, 2007) that have been suggested over the years to explain visual search tasks (see (Walther and Koch, 2007) for a recent review). Work on modeling visual attention most related to our approach can be characterized based on the type of cues that are used and how they are combined

Studies have shown that image-based *bottom-up* cues can capture attention, particularly

	Proposed	(Bruce and Tsotsos, 2006)	(Zhang *et al.*, 2008)	(Deco and Rolls, 2004b)	(Ehinger *et al.*, 2009)	(Fukushima, 1986)	(Hou and Zhang, 2007)	(Harel *et al.*, 2007)	(Itti and Koch, 2001b)	(Rao, 2005)	(Torralba, 2003b)	(Walther and Koch, 2007)	(Wolfe, 2007)	(Yu and Dayan, 2005)
Biologically plausible	✓	✓	✓	✓	×	✓	×	×	✓	✓	✓	✓	✓	✓
Real world stimuli	✓	✓	✓	×	✓	×	✓	✓	✓	×	✓	✓	×	×
Pop-out	✓	✓	✓	×	✓	×	✓	✓	✓	×	✓	✓	×	×
Feature-based attention	✓	×	×	✓	✓	✓	×	×	×	×	✓	×	✓	✓
Spatial attention	✓	×	×	×	✓	×	×	×	×	✓	✓	✓	×	✓
Parallel vs. serial search	✓	×	×	×	×	×	×	×	×	×	×	×	✓	×
Models ventral/parietal	✓	×	×	✓	×	×	×	×	×	✓	×	×	×	×

Table A4: The matrix compares the features of prior computational models.

during free viewing conditions. Locations where stimulus differs significantly from rest of the image is said to 'pop-out'. In (Itti *et al.*, 1998), center-surround difference across color, intensity and orientation dimensions is used as measure of saliency. In (Gao and Vasconcelos, 2007), self information of the stimuli ($-\log(P(I))$) is used as measure of distinctiveness (Zhang *et al.*, 2008). In (Rosenholtz, 1985), the normalized deviation from mean response is used instead. Spectral methods for computing bottom-up saliency have also been proposed (Hou and Zhang, 2007). These models, however, cannot account for the task-dependency of eye movements (Yarbus, 1967). Depending on the search tasks, human eye movements may differ substantially–even when the stimuli are identical

A seminal proposal to explain how top-down visual search may operate is the *Guided Search* model proposed by Wolfe (Wolfe, 2007) according to which the various feature maps are weighted according to their relevance for the task at hand to compute a solitary saliency map. Building on Wolfe's model, several approaches have been suggested (Navalpakkam and Itti, 2006; Gao and Vasconcelos, 2005; Zhang *et al.*, 2008). Computational models that use feature-based cues have relied upon low-level features such as color, contrast, orientation (Peters and Itti, 2007; Navalpakkam and Itti, 2006) that are too simple for real-world object-based visual searches. These models also ignore the role of spatial attention. In situations where the location of the target is explicitly cued, the role of spatial attention cannot be overlooked. In (Desimone, 1998), it was shown that activity of V4

	BU	Loc	Sc	Feat	R-W	Comb
Fukushima et al. (Fukushima, 1986)	×	×	×	✓	×	N/A
(Itti and Koch, 2001a)	✓	×	×	×	✓	N/A
(Zhang *et al.*, 2008)	✓	×	×	×	✓	Bayes
(Gao and Vasconcelos, 2007)	✓	×	×	×	✓	N/A
(Hou and Zhang, 2007)	✓	×	×	×	✓	N/A
(Navalpakkam and Itti, 2006)	×	×	×	✓	✓	Lin
(Gao and Vasconcelos, 2005)	×	×	×	✓	✓	N/A
(Torralba, 2003b)	✓	✓	✓	×	✓	Bayes
(Bruce and Tsotsos, 2006)	✓	×	×	×	✓	N/A
(Walther and Koch, 2007)	✓	✓	×	✓	✓	N/A
Proposed	✓	✓	✓	✓	✓	Bayes

Table A5: A summary of the differences between different approaches to model attention and eye movements. The various approaches are compared based on the type of cues that are used to derive a saliency map, how those cues are combined and whether the work was evaluated on real-world images. 'BU' column indicates if bottom-up cues are used, 'Loc' (location) and 'Sc' (scale) columns indicate if contextual cues are used to predict object location and scale respectively. The 'Feat' (feature) column indicates if the model relies on **top-down** feature cues. 'RW' (real-world) shows if the model has been evaluated on real world images. In cases where multiple cues are combined, 'Comb' (combination) indicates if the combination is bayesian ('Bayes') or linear ('Lin').

neurons are reduced when multiple stimuli are present within its receptive field. However, when a specific location is cued and subsequently attended, the neurons at the attended locations are selectively enhanced. The neuron responds as if there is a single stimulus within the receptive field. In (Rao, 2005), a bayesian model of spatial attention is proposed that reproduces this effect. Our work can be viewed as an extension of this approach. In addition to direct cueing, spatial cues may also be derived indirectly, by context, in natural scenes. Spatial relations between objects and their locations within a scene have been shown to play a significant role in visual search and object recognition (Biederman *et al.*, 1982). In (Oliva *et al.*, 2003), Oliva, Torralba and colleagues showed that a combination of spatial context and bottom-up attention could predict a large fraction of human eye movements during real-world visual search tasks in complex natural images. With the exception of (Ehinger *et al.*, 2009), computational models have not considered the interaction between spatial, bottom-up and top-down attentional effects. Table A5 provides a succinct comparison of our approach with existing work in literature.

Bibliography

Amit, Y. and Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vision Research*, **43**(19), 2073–2088.

Beauchamp, M. S., Cox, R. W., and Deyoe, E. A. (1997). Graded effects of spatial and featural attention on human area MT and associated motion processing areas.

Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, **7**(6), 1129–1159.

Bichot, N., Rossi, A., and Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, **308**(5721), 529–534.

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

Biederman, I., Mezzanotte, R., and Rabinowitz, J. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, **14**(2), 143.

Bileschi, S. M. (2006). *StreetScenes: Towards scene understanding in still images*. Ph.D. thesis, MIT.

Bisley, J. and 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. and Tsotsos, J. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems*, **18**, 155.

Buschman, T. and Miller, E. (2007). Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. *Science*, **315**(5820), 1860.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., and Poggio, T. (2007). A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, **98**(3), 1733.

Carandini, M., Heeger, D., and Movshon, J. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, **17**(21), 8621–8644.

Colby, C. and Goldberg, M. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, **22**(1), 319–349.

Crandall, D., Felzenszwalb, P., and Huttenlocher, D. (2005). Spatial priors for part-based recognition using statistical models. In *Proc. IEEE Computer Vision and Pattern Recognition*, volume 1.

Crick, F. and Koch, C. (1990a). Some reflections on visual awareness. In *Cold Spring Harbor Symposium on Quantitative Biology*, volume 55, pages 953–962.

Crick, F. and Koch, C. (1990b). Towards a neurobiological theory of consciousness. In *Seminars in the Neurosciences*, volume 2, page 201.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, pages 428–441.

Daugman, J. (1980). Two-dimensional spectral analysis of cortical receptive field profile. *Vision Research*, **20**, 847–856.

Dayan, P. and Zemel, R. (1999). Statistical models and sensory attention. *Proceedings of the International Conference on Artificial Neural Networks*, page 2.

Dayan, P., Hinton, G. E., and 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*, volume 20, page 938.

Deco, G. and Rolls, E. (2004a). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, **44**(6), 621–642.

Deco, G. and Rolls, E. (2004b). 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. and 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.

Duncan, J. (1995). Target and nontarget grouping in visual search [comment]. *Percept. Psychophys.*, **57**(1), 117–20.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., and Oliva, A. (2009). Modelling search for people in 900 scenes: a combined source model of eye guidance. *Visual Cognition*.

Epshtein, B., Lifshitz, I., and Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proc. of the National Academcy of Sciences*.

Felzenszwalb, P. and Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, **61**(1), 55–79.

Fergus, R., Perona, P., and 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., and 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. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, **20**(3-4), 121–136.

Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics*, **55**(1), 5–15.

Gallant, J. L., Connor, C. E., Rakshit, S., Lewis, J. W., and 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. and Vasconcelos, N. (2005). Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *Proc. IEEE Computer Vision and Pattern Recognition*.

Gao, D. and Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In *Proc. of the International Conference on Computer Vision*.

George, D. and Hawkins, J. (2005). A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In *International Joint Conference on Neural Networks*, volume 3.

Gilks, W. and 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.

Harel, J., Koch, C., and Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, **19**, 545.

Hegde, J. and Felleman, D. J. (2003). How selective are V1 cells for pop-out stimuli? *Journal of Neuroscience*, **23**(31).

Hegde, J. and Van Essen, D. (2000). Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience*, **20**(5), 61–61.

Hinton, G. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, **11**(10), 428–434.

Hochstein, S. and Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, **36**, 791–804.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Proc. IEEE Computer Vision and Pattern Recognition*.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, **148**, 574–91.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, **195**, 215–243.

Hung, C., Kreiman, G., Poggio, T., and DiCarlo, J. (2005). Fast read-out of object identity from macaque inferior temporal cortex. *Science*, **310**, 863–866.

Ito, M. and 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. and Koch, C. (2001a). Computational modelling of visual attention. *Nature Reviews on Neuroscience*, **2**(3), 194–203.

Itti, L. and Koch, C. (2001b). Computational modelling of visual attention. *Nature Reviews on Neuroscience*, **2**(3), 194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(11).

Itti, L., Rees, G., and Tsotsos, J. (2005). *Neurobiology of attention*. Academic Press.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A Biologically Inspired System for Action Recognition. In *Proc. of the International Conference on Computer Vision*, pages 1–8.

Kanwisher, N. and Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nature Reviews on Neuroscience*, **1**(2), 91–100.

Kersten, D. and Yuille, A. (2003a). Bayesian models of object perception. *Current Opinion in Neurobiology*, **13**(2), 150–158.

Kersten, D. and Yuille, A. (2003b). Bayesian models of object perception. *Current Opinion in Neurobiology*, **13**(2), 150–158.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference.

Knill, D. and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge Univ Pr.

Kobatake, E. and 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. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, **4**(4), 219–27.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. of the IEEE*, **86**(11), 2278–2324.

Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*.

Li, Z. and 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.

Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, **19**, 577–621.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, **5**, 552–563.

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*. Henry Holt and Co., Inc. New York, NY, USA.

Martınez-Trujillo, J. and Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron*, **35**(2), 365–370.

Maunsell, J. H. and Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neuroscience*, **29**(6), 317–22.

McAdams, C. and 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.

Meyers, E., Zhang, Y., Bichot, N., Serre, T., and Poggio, T. (????). Attentional clutter suppression in it population activity.

Miau, F. and 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*, pages 789–792.

Missal, M., Vogels, R., and Orban, G. (1997). Responses of macaque inferior temporal neurons to overlapping shapes. *Cerebral Cortex*, **7**(8), 758.

Monahan, G. (1982). A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, pages 1–16.

Moore, T. and Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, **421**(6921), 370–3.

Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, **229**(4715), 782.

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., and 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. and Kreutz-Delgado, K. (2007). Visual recognition and inference using dynamic overcomplete sparse learning. *Neural Computation*, **19**(9), 2301–2352.

Mutch, J. and Lowe, D. (2006). Multiclass object recognition with sparse, localized features. In *Proc. IEEE Computer Vision and Pattern Recognition*.

Navalpakkam, V. and 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*.

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

Oram, M. and Perrett, D. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology*, **68**, 70–84.

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, **42**(1), 107–23.

Pasupathy, A. and 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.

Pelli, D. and Tillman, K. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, **11**(10), 1129–1135.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, **11**, 317–333.

Peters, R. J. and 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.

Peters, R. J., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, **45**(18), 2397–416.

Posner, M. and Cohen, Y. (1984). Components of visual orienting. *Attention and performance*, pages 531–556.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews on Neuroscience*, **1**(2), 125–132.

Ranzato, M., Huang, F., Boureau, Y., and 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., and Lewicki, M. (2002a). *Probabilistic models of the brain: Perception and neural function*. The MIT Press.

Rao, R., Olshausen, B., and Lewicki, M. (2002b). *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, Cambridge, MA.

Reynolds, J., Chelazzi, L., and Desimone, R. (1999). Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4. *Journal of Neuroscience*, **19**(5), 1736.

Reynolds, J., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, **26**(3), 703–714.

Reynolds, J. H. and Heeger, D. J. (2009). The normalization model of attention. *Neuron Review*, **61**, 168–184.

Riesenhuber, M. and Poggio, T. (1999a). Are Cortical Models Really Bound by the Binding Problem. *Neuron*, **24**(1), 87–93.

Riesenhuber, M. and 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. and Mansfield, J. (2005). Feature congestion: a measure of display clutter. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 761–770. ACM New York, NY, USA.

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500), 2323–2326.

Russell, B., Torralba, A., Murphy, K., and Freeman, W. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, **77**(1), 157–173.

Rust, N. C., Mante, V., Simoncelli, E. P., and Movshon, J. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, **9**(11), 1421–1431.

Rutishauser, U., Walther, D., Koch, C., and Perona, P. (2004). Is bottom-up attention useful for object recognition? In *Proc. IEEE Computer Vision and Pattern Recognition*, volume 2.

Serre, T., Wolf, L., and Poggio, T. (2005a). Object recognition with features inspired by visual cortex. In I. C. S. Press, editor, *Proc. IEEE Computer Vision and Pattern Recognition*, volume 2, pages 994–1000, San Diego.

Serre, T., Kouh., M., Cadieu, C., Knoblich, U., Kreiman, G., and 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., Oliva, A., and Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proc. of the National Academcy of Sciences*, **104**(15), 6424.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007b). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, **165**, 33–56.

Serre, T., L., W., Bileschi, S., Reisenhuber, M., and Poggio, T. (2007c). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Simoncelli, E. P. and Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, **38**(5), 743–761.

Smallwood, R. and Sondik, E. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, pages 1071–1088.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2005). Learning hierarchical models of scenes, objects, and parts. In *Proc. IEEE Computer Vision and Pattern Recognition*, volume 2, pages 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. *Proc. Biologically Motivated Computer Vision*, pages 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. and Oliva, A. (2003). Statistics of natural image categories. *Network: computation in neural systems*, **14**, 391–412.

Torralba, A., Murphy, K., and Freeman, W. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. IEEE Computer Vision and Pattern Recognition*, volume 2.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, **113**(4), 766–86.

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

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

Treue, S. and Trujillo, J. (1999b). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, **399**(6736), 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.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial intelligence*, **78**(1-2), 507–545.

Ullman, S. (1984). Visual routines. *Cognition*, **18**(1-3), 97–159.

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

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

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

Ungerleider, L. G. and Pasternak, T. (2004). Ventral and dorsal cortical processing streams. *The visual neurosciences*, **1**(34), 541–562.

Van Der Velde, F. and 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.

VanRullen, R. and Koch, C. (2003). Visual selective behavior can be triggered by a feed-forward process. **15**, 209–217.

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

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

Walther, D. and Koch, C. (2007). *Computational Neuroscience: Theoretical insights into brain function, Progress in Brain Research*, chapter Attention in Hierarchical Models of Object Recognition.

Wannig, A., Rodríguez, V., and Freiwald, W. (2007). Attention to surfaces modulates motion processing in extrastriate area MT. *Neuron*, **54**(4), 639–651.

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

Wersing, H. and 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*, pages 99–119.

Womelsdorf, T., Anton-Erxleben, K., Pieper, F., and Treue, S. (2006). Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature Neuroscience*, **9**(9), 1156–1160.

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

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

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

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J. (2007). Trade-Off between Object Selectivity and Tolerance in Monkey Inferotemporal Cortex. *Journal of Neuroscience*, **27**(45), 12292.