Attention as a Bayesian inference process

Citation: Chikkerur, Sharat et al. "Attention as a Bayesian Inference Process." 2011. Proc. SPIE 7865, 786511–786511–10. Web. 11 Apr. 2012. © 2011 SPIE - International Society for Optical Engineering

As Published: http://dx.doi.org/10.1117/12.876734

Publisher: SPIE - International Society for Optical Engineering

Persistent URL: http://hdl.handle.net/1721.1/69982

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Attention as a Bayesian inference process

Sharat Chikkerur[a], Thomas Serre[a,b], Cheston Tan[a] and Tomaso Poggio[a]

[a]Massachusetts Institute of Technology, Cambridge, MA, US
[b]Brown University, Providence RI, US

ABSTRACT

David Marr famously defined vision as "knowing what is where by seeing". In the framework described here, attention is 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 performs well in recognition tasks and that predicts some of the main properties of attention at the level of psychophysics and physiology. We propose an algorithmic implementation a Bayesian network that can be mapped into the basic functional anatomy of attention involving the ventral stream and the dorsal stream. This description integrates bottom-up, feature-based as well as spatial (context based) attentional mechanisms. We show that the Bayesian model predicts well human eye fixations (considered as a proxy for shifts of attention) in natural scenes, and can improve accuracy in object recognition tasks involving cluttered real world images. In both cases, we found that the proposed model can predict human performance better than existing bottom-up and top-down computational models.

Keywords: Attention, Bayesian inference, Eye-movements

1. INTRODUCTION

Visual processing in the brain proceeds 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. The two streams underlie the perception of 'what' and 'where' concurrently and relatively independently of each other.[1] This 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[2]). 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.[3,4] On the other hand, computational models attempt to model specific behavioral and physiological effects of attention.[5–8,8–11] A unifying framework that provides a computational goal for attention and at the same time accounts for the disparate effects listed above is missing. We validate the model experimentally and show that it is consistent with physiological effects in IT and human behavior (eye-movements).

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.[12–19] Extending this idea, we propose that attention can also be regarded as an inference process that disambiguates form and location information.[11,20] 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[21]), but as an effect of interaction between the 'what' and 'where' streams within this inference framework. 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.

Human Vision and Electronic Imaging XVI, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7865, 786511 · © 2011 SPIE-IS&T
CCC code: 0277-786X/11/$18 · doi: 10.1117/12.876734

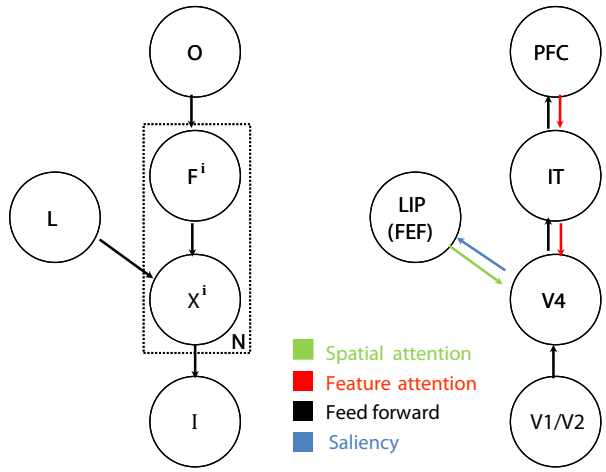


Figure 1. Left: Proposed Bayesian model. Right: A model illustrating the interaction between the parietal and ventral streams mediated by feedforward and feedback connections.

2. MODEL

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. These receive bottom-up evidence from the input variable I. The object variable O is modeled as a multinomial random variable with $|O|$ values corresponding to objects known by the model. The prior $\mathrm{P(O)}$ is set based on the search task. Each feature-encoding unit $\mathrm{F^i}$ is a binary random variable that represents the presence or absence of a feature irrespective of location and scale. 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 $(0, 1, \cdots, L)$. 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 represented by the graphical model can be rewritten as

$$\begin{aligned} &P(O, L, X^1, \cdots, X^N, F^1, \cdots, F^N, I) \\ &= P(O)P(L)\left\{\prod_{i=1}^{i=N}\left\{P(X^i|L, F^i)P(F^i|O)\right\}\right\}P(I|X^1, \cdots, X^N) \end{aligned} \tag{1}$$

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 *.The conditional probabilities are specified in Table 1. 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,[22] 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.

Further author information: (Send correspondence to Sharat Chikkerur)
Sharat Chikkerur: E-mail: sharat@mit.edu

*$P(I|X^1 \cdots X^N)$ 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.

The graphical model can be tentatively mapped 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 Fig. 1). Thus, following the organization of the visual system,[1] 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[23] and higher visual areas such as AIT where retinotopy is almost completely lost[24,25] (modeled as F^i units). Prior attempts to model this interaction have been non-Bayesian.[2,26]

2.1 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. In the simulations of this paper, the inference mechanism used is the 'belief propagation' algorithm,[28] 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,[11] where the top-down messages, $\pi()$ and bottom-up messages $\lambda()$ are replaced by a uniform $m()$ term.

$$m_{O\to F^i} = P(O) \tag{2}$$

$$m_{F^i\to X^i} = \sum_O P(F^i|O)P(O) \tag{3}$$

$$m_{L\to X^i} = P(L) \tag{4}$$

$$m_{I\to X^i} = P(I|X^i) \tag{5}$$

$$m_{X^i\to F^i} = \sum_L \sum_{X^i} P(X^i|F^i,L)(m_{L\to X^i})(m_{I\to X^i}) \tag{6}$$

$$m_{X^i\to L} = \sum_{F^i} \sum_{X^i} P(X^i|F^i,L)(m_{F^i\to X^i})(m_{I\to X^i}) \tag{7}$$

<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.[10]</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[27]).</td></tr>
</table>

Table 1. Description of the model conditional probabilities.

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

$$P(L|I) \quad \propto \quad (m_{L\to X^i})(m_{X^i\to L}) \tag{8}$$

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. When considering multiple features, the Bayesian inference proceeds as in a general polytree.[28] 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:

$$m_{L\to X^i} \quad = \quad P(L)\prod_{j\neq i} m_{X^j\to L} \tag{9}$$

3. EXPERIMENTAL VALIDATION

3.1 "Predicting" effects of spatial attention in IT

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,[29,30] human psychophysics as well as computational models.[27] 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.

In their seminal study, Moran and Desimone[31] 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.[23,32] 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 [†] 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. 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 2)[‡].

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 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). 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.

[†] as measured by neural decoding performance

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

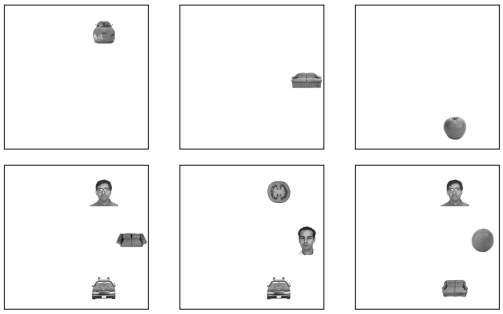


Figure 2. Illustration of some of the stimuli presented during the experiment. (Top): Stimuli where a single object was present. (Bottom): Stimuli where three objects were present. In both cases, the fixation point was placed at the center of the image.

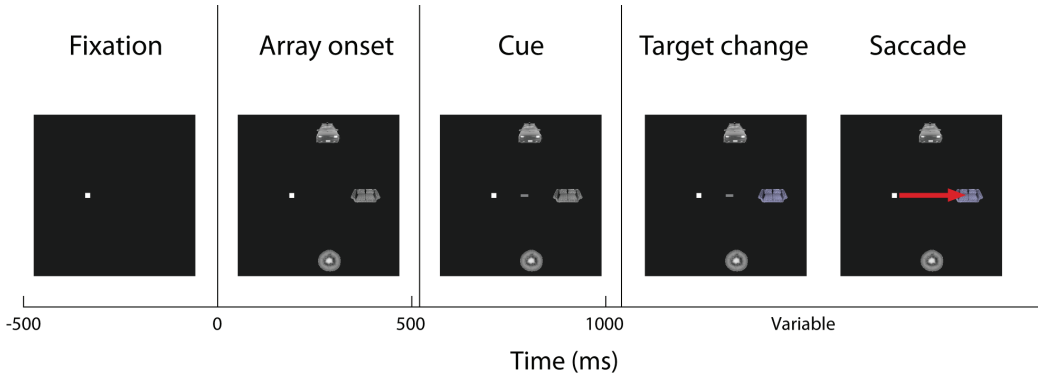
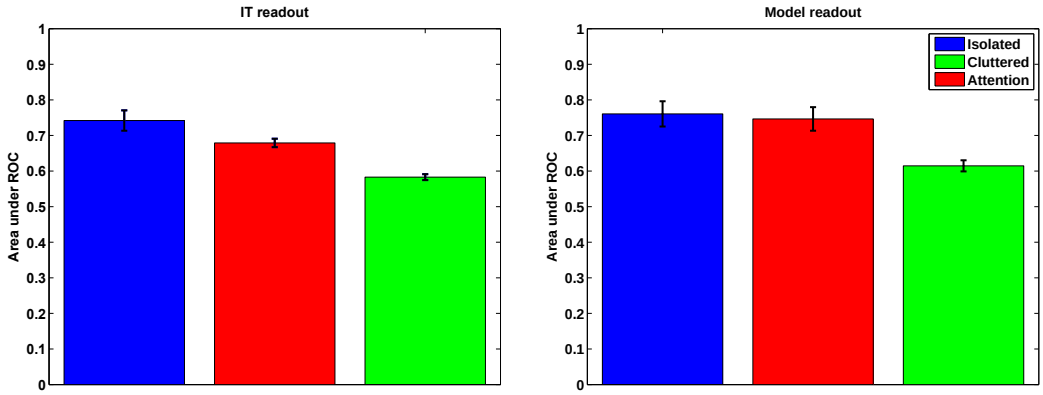


Figure 3. Experimental protocol used for recording neurons from IT. Notice the spatial cue in the form of a small bar directed at the target object. (image,courtesy Ethan Meyers)

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.[33]

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. 4 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.

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 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.



(a) Decoding performance using IT neurons. (b) Decoding performance using the model.

Figure 4. Comparison between IT neurons and model simulation before parameter fitting.

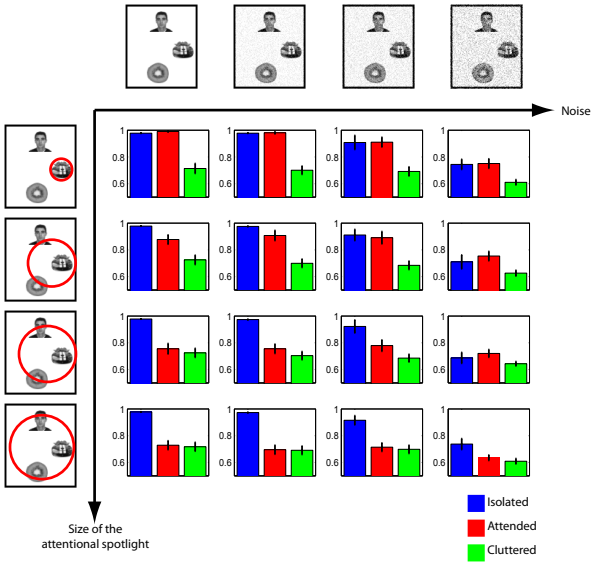


Figure 5. Effect of noise and size of the attentional spotlight on decoding performance.

Results The results (see Fig. 4 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.

Parameter fitting We studied the effects of these parameters on the decoding performance (see Fig 5). 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.

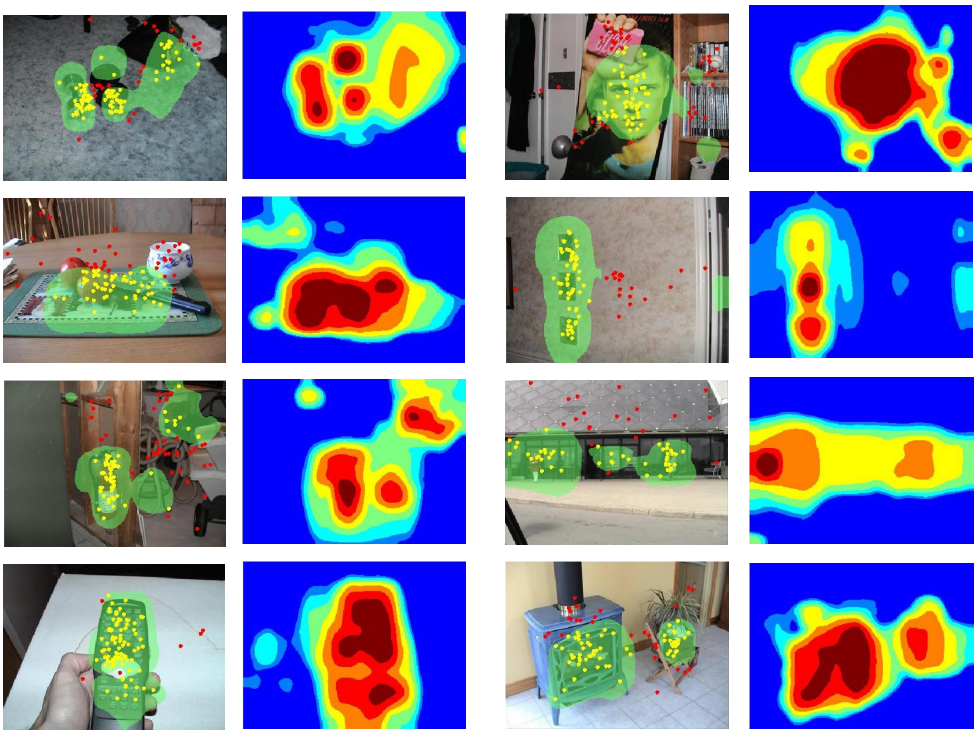


Figure 6. Predicting human eye movements: *(a)* Agreement between the model and human eye fixations during free viewing (left) and a complex visual search for either cars or pedestrians. Sample images overlaid with most salient (top 20%) regions predicted by the model (green) along with human eye movements (yellow: agree with prediction, red: not predicted by model) and corresponding model posteriors (*i.e.,* predicted image saliency).

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,[34,35] 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.

3.2 Predicting human 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,[9,36] pruning interest points[37] and quantifying visual clutter.[7] Previous work in attention and eye movements has focused on free viewing conditions[38–41] 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.[42] 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. This work has been described in detail elsewhere.[43] Here, we describe a summary of the results.

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.[41] 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.

Models	Agreement with humans (ROC area)
Bruce and Tsotsos[41]	0.728
Itti and Koch[5]	0.727
Proposed model	**0.779**

Table 2. Comparison of the proposed Bayesian model with shape-based features with prior work that relies on low level features.

There are at least two measures that have been used to compare models of attention to human fixations: normalized scan path saliency (NSS) from[44] and fixations in the most salient region ($FMSR$) from.[41,45] 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. 6). 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 2).[§] 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.

4. DISCUSSION

4.1 Relation to prior work

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[3] is that attention reflects evolution's attempt to fix the processing bottleneck in the visual system[46] by directing the finite computational capacity of the visual system preferentially to relevant stimuli within the visual field while ignoring everything else.[4] suggested that attention is used to *bind* different features (e.g. color and form) of an object during visual perception.[47] 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,[5–7] top-down bias of target features,[8,9] influence from scene context,[10] serial vs. parallel-search effect[8] etc. Physiological effects include multiplicative modulation of neuron response under spatial attention[11] and feature based attention.[48] 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.

4.2 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. 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.

[§] 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.

REFERENCES

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

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

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

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

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

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

[7] Rosenholtz, R. and Mansfield, J., "Feature congestion: a measure of display clutter," in [*Proceedings of the SIGCHI conference on Human factors in computing systems*], 761–770, ACM New York, NY, USA (2005).

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

[9] Navalpakkam, V. and Itti, L., "An integrated model of top-down and bottom-up attention for optimizing detection speed," in [*Proc. IEEE Computer Vision and Pattern Recognition*], (2006).

[10] Torralba, A., "Modeling global scene factors in attention," *Journal of Optical Society of America* **20**(7), 1407–1418 (2003).

[11] Rao, R., "Bayesian inference and attentional modulation in the visual cortex," *NeuroReport* **16**(16), 1843–1848 (2005).

[12] Mumford, D., "On the computational architecture of the neocortex – II: The role of cortico-cortical loops," *Biological Cybernetics* **66**, 241–251 (1992).

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

[14] Rao, R., Olshausen, B., and Lewicki, M., [*Probabilistic models of the brain: Perception and neural function*], The MIT Press (2002).

[15] Rao, R., "Bayesian computation in recurrent neural circuits," *Neural Computation* **16**(1), 1–38 (2004).

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

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

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

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

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

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

[22] Koch, C. and Ullman, S., "Shifts in selective visual attention: towards the underlying neural circuitry.," *Human Neurobiology* **4**(4), 219–27 (1985).

[23] Tanaka, K., "Inferotemporal cortex and object vision," *Annual Review of Neuroscience* **19**, 109–139 (1996).

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

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

[26] Grossberg, S., "How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex," *Spatial Vision* **12**(2), 163–185 (1999).

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

[28] Pearl, J., [*Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*], Morgan Kaufmann Publishers (1988).

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

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

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

[32] Kobatake, E. and Tanaka, K., "Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex," *Journal of Neurophysiology* **71**, 856–867 (1994).

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

[34] Crick, F. and Koch, C., "Towards a neurobiological theory of consciousness," in [*Seminars in the Neurosciences*], **2**(263-275), 201 (1990).

[35] Crick, F. and Koch, C., "Some reflections on visual awareness," in [*Cold Spring Harbor Symposium on Quantitative Biology*], **55**, 953–962 (1990).

[36] Torralba, A., "Contextual priming for object detection.," *International Journal of Computer Vision* **53**(2), 169–191 (2003).

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

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

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

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

[41] Bruce, N. and Tsotsos, J., "Saliency based on information maximization," *Advances in Neural Information Processing Systems* **18**, 155 (2006).

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

[43] Chikkerur, S., Serre, T., Tan, C., and T., P., "What and where: A Bayesian inference theory of visual attention," *Vision Research* **229**(4715), 782 (2010).

[44] Peters, R. J. and Itti, L., "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in [*Proc. IEEE Computer Vision and Pattern Recognition*], (2007).

[45] Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M., "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 (2006).

[46] Broadbent, D. E., [*Perception and communication*] (1958).

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

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