Inattentional Blindness in Visual Search

Inattentional Blindness in Visual Search

Matt Chapman-Rounds (m.rounds@ed.ac.uk)
Christopher G. Lucas (c.lucas@ed.ac.uk)
Frank Keller (keller@inf.ed.ac.uk)
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Models of visual saliency normally belong to one of two camps: models such as Experience Guided Search (E-GS), which emphasize top-down guidance based on task features, and models such as Attention as Information Maximisation (AIM), which emphasize the role of bottom-up saliency. In this paper, we show that E-GS and AIM are structurally similar and can be unified to create a general model of visual search which includes a generic prior over potential non-task related objects. We demonstrate that this model displays inattentional blindness, and that blindness can be modulated by adjusting the relative precisions of several terms within the model. At the same time, our model correctly accounts for a series of classical visual search results.

Keywords: Inattentional Blindness; Conjunction Search; Visual Attention; Bayesian Modelling; Predictive Processing

Introduction

Visual search, where agents search for a target amongst distractors, is an important paradigm in the study of human attention (Wolfe, 1994) (see Figure 1 for an example trial). Inattentional blindness, where unexpected objects fail to capture attention, provides a useful insight into how constraints of processing and access lead to failures in the visual system (Simons, 2000). The literature on the two domains is distinct; in this paper we show that extending a model of visual search by adding an environmental prior produces a model that can reproduce empirical results from both domains.

The motivation for our extension hinges on the idea that the brain, due to the pressures of an ever changing environment, never *solely* models a task; it must always additionally maintain what are effectively generic, non-task-specific prior expectations about possible interesting states of the world. For example, in conjunction search (Nakayama & Silverman, 1986), where participants search for a target amongst distractors, a simple model of the search environment should include both "targets" and "distractors" (the statistics of which are learned during training), and "non-task entities" (which are unrelated to the task), as possible kinds. Ignoring non-task entities allows the brain to attend to (and successfully perform) a task, at the expense of potentially missing useful information about the world.

The contributions of this work are threefold. We demonstrate a successful joint model of visual search and inattentional blindness in which search is driven by saliency, generated using precision-weighted error terms. We show the structural equality of two distinct models of saliency, one top-down, and the other bottom-up. Finally, by constructing a model where both task relevant and task irrelevant stimuli contribute to saliency, we shed light on what it means to perform a task – namely, for an agent to have high confidence in its model of those stimuli that constitute the task, compared with its model of other possible stimuli.

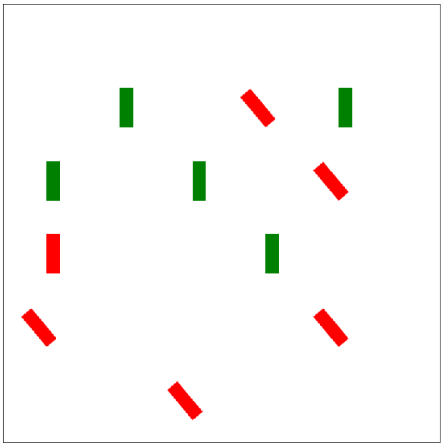

Figure 1: Example trial taken from Task 5 (see Results, below). Task is to find red vertical target amongst green vertical and red titled distractors.

Related Work

Conjunction Search

Empirically, we can distinguish between five forms of guidance in visual search (Wolfe & Horowitz, 2017). The two of interest to this work are bottom-up (where visual properties of aspects of a scene attract more attention than others, Koehler, Guo, Zhang, & Eckstein, 2014), and top-down (where executive control drives attention towards desired targets, Maunsell & Treue, 2006).

The majority of the many models of top-down visual search (Itti, Koch, & Niebur, 1998; Torralba, Oliva, Castelhano, & Henderson, 2006; Navalpakkam & Itti, 2006; Cave, 1999; Choi, Torralba, & Willsky, 2012)[1] share the basic structure of

[1] We cite Itti and Koch's work here, as well as in the section on bottom-up drivers of saliency, because whilst their work focusses on

Guided Search (GS; Wolfe, 1994): primitive visual features are detected across the retina by feature maps, which represent features via a coarse (i.e., highly overlapping) encoding. These feature maps are then passed through a local differencing operator, which enhances local contrasts, and the feature maps are combined using a top-down, task specific weighting to produce a saliency map. A Bayesian treatment of GS called Experience-Guided Search has been proposed by Mozer and Baldwin (2007).

There is also a body of work that focuses specifically on the bottom-up drivers of saliency (Koehler et al., 2014; Itti & Koch, 2001), an example of which is Attention as Information Maximisation (AIM), proposed by Bruce and Tsotsos (2009). These authors argue that the self-information of a location in an image, estimated on its surrounding context, is a good measure of its visual saliency.

Bottom-up models can be thought of as mapping the saliency of a task-neutral environment; but they provide no account of the relationship between this base saliency and the task at hand. Top-down models provide a qualitative account of various phenomena in visual search (see Results for a detailed discussion of relevant phenomena). However, a limitation of these models is that they do not attempt to model tasks as situated in a wider environment containing task-irrelevant stimuli, or competing tasks. A paradigm where modelling the relative saliency of task-relevant and task-irrelevant items becomes important is that of inattentional blindness (IB; Mack & Rock, 1998).

Inattentional Blindness

To model IB in the context of visual search, we use an "additional singleton" approach (see, e.g., Simons, 2000), where an unexpected single item has a distinctive unique feature, and that item is never the target item. There are several factors which have been shown to affect the rate of unexpected object detection when performing a task: increased cognitive load increases blindness (Kreitz, Furley, Memmert, & Simons, 2016), the similarity of the unexpected object to task-relevant objects increases the probability that the unexpected object will capture attention (Most et al., 2001; Simons & Chabris, 1999), as do shared features between task relevant objects and unexpected objects (Koivisto & Revonsuo, 2009).

Models of the causes of inattentional blindness range from claims of inattentional amnesia (we see the object, but fail to report it after the trial, Wolfe, 1999), to arguments that we are blind to objects we do not expect to see (Braun, 2001). More recent accounts have focused on the relationship between bottom-up saliency (which drives transient attentional capture) and a top-down attentional set, which governs whether transient attention is sufficient to generate sustained attention, and subsequent awareness (Most, Scholl, Clifford, & Simons, 2005). In spirit, our approach falls under this latter umbrella, but we show that blindness can be explicitly thought of as a result of a ratio of precisions in a mathematical model that extends the conjunction search literature (and is also able to replicate standard results in that domain).

saliency maps, they assume that these maps are combined according to top-down attentional drivers, which makes them less purely bottom-up than AIM, for example. See the section on bottom-up visual attention, below.

Model

Our starting point is Experience-Guided Search (E-GS; Mozer & Baldwin, 2007), a Bayesian treatment of GS developed to overcome a shortcomings of GS (Wolfe, 1994; Wolfe & Horowitz, 2017), namely that GS produces better than human performance without the addition of noise or regularising constraints on the top-down weighting of features. Mozer and Baldwin's premise is that a location in the visual field is salient if a target is likely to be at that location. They define $P(T_x = 1|\boldsymbol{F}_x)$ as a measure of saliency computed using statistics obtained from recent experience performing the task:

$$P(T_x|\boldsymbol{F}_x,\boldsymbol{\rho}) = \frac{P(T_x)\prod_i P(F_{xi}|T_x,\boldsymbol{\rho}_i)}{\sum_{t=0}^{1} P(T_x = t)\prod_i P(F_{xi}|T_x = t,\boldsymbol{\rho}_i)} \quad (1)$$

Here, $\boldsymbol{F}_x$ is the feature activity vector at retinal location x, T_x is the binary indicator of targethood, $\boldsymbol{\rho}$ parameterises the stimulus environment, and F_{xi} is the feature vector corresponding to feature i.

Whilst we lack the space to give a full treatment here, by assuming a generative model with $F_{xi}|\{T_x = t,\boldsymbol{\rho}\} \sim \text{Binomial}(n,\rho_{it})$, where ρ_{it} is the parameterising spike rate associated with feature i for target and non-target items ($t = 1, t = 0$), in the limit of reasonably large n we can approximate $P(F_{xi}|\ldots)$ as Gaussian, with mean $n\rho_{it}$ and variance $n\rho_{it}(1-\rho_{it})$. This allows the authors to derive a measure of saliency, S_{EGS}, as:

$$S_{EGS} = \sum_i \left[\Lambda_{\rho i0}(f_{xi} - n\rho_{i0})^2 - \Lambda_{\rho i1}(f_{xi} - n\rho_{i1})^2\right] \quad (2)$$

Where $\Lambda_{\rho it}$ denotes the precision (inverse variance) of the model's current estimate of ρ_{it}.[2]

This is a sum of terms, two for each feature i, which capture how surprising the activation corresponding to that feature, f_{xi}, is with respect to the target or not-target cases, the model's beliefs about which are parameterised by ρ_{i1} and ρ_{i0} respectively. The saliency of feature i increases if the observed activation is distant from the mean activity observed in the past in the absence of a target. It decreases if the observed behaviour is distant from the mean activity observed its presence. This surprisal is weighted by observed precisions: high variance features contribute less to saliency.

This remains a strictly task-based model, however. To expand it, we need to consider how the saliency of a feature changes under a generic, non-task specific prior. To do this, we turn to the literature on bottom-up measures of saliency,

[2] A difference between this presentation and that in Mozer and Baldwin is that we have not ignored the scaling n; whilst in E-GS only the relative magnitude of the terms in (2) is relevant, here we do care about how much data the model has seen.

in particular AIM (Bruce & Tsotsos, 2009). In doing so, we note the structural similarity between AIM and E-GS.

The premise of AIM is that those areas of an image that contain the most Shannon self-information are those that contain content of interest. Hence visual saliency is driven by surprise with respect just to visual input. First, "a sparse spatiochromatic basis" is generated in an unsupervised fashion using ICA, such that every image patch can be expressed as a vector of coefficient contributions (if projected back into image space, the coefficients, not incidentally, look a lot like Gabor filters and colour opposition patches[3]).

For each location x we can characterise the content of the local neighbourhood C_x by a vector $\boldsymbol{\alpha}_x$. For each of the i features, the p.d.f of the surround is estimated by making a histogram of all α_i values for every nearby patch. Then α_{xi}'s likelihood $P(\alpha_{xi})$ can be estimated from the histogram and thus its Shannon information content computed by $\log(1/P(\alpha_{xi}))$. Adding the Shannon information from each coefficient in $\boldsymbol{\alpha}_x$ gives us an estimate of the Shannon information contained in patch x, and hence the saliency of that patch:

$$S_{\text{AIM}} = -\sum_i \log P(\alpha_{xi}) \tag{3}$$

We then approximate the histogram $P(\alpha_{xi})$ as Gaussian distributed with mean $\bar{\alpha}_i$ and variance $\sigma^2_{\alpha i}$, which are the statistical mean and unbiased variance computed from the activations of surrounding patches for feature i.

This means we can rewrite (3) as:

$$S_{\text{AIM}} = \sum_i \left[\frac{1}{\sigma^2_{\alpha i}} (\alpha_{xi} - \bar{\alpha}_i)^2 \right] \tag{4}$$

Comparing to (2), we can see that if we make the same assumptions about the form of the likelihood of our incoming data, the measures of saliency used by E-GS and AIM are both sums of precision-weighted errors. The main difference is that AIM learns its model statistics from surrounding, synchronic activations, whereas E-GS learns its statistics diachronically, and with respect to the pertinent categories of a task oriented model.

Our final step is to argue that true saliency is a combination of many such terms, driven by the pressure to balance attention between task-driven stimuli and the world in which a task takes place. We therefore propose the following measure of saliency of location x, $\boldsymbol{S}_x$:

$$\boldsymbol{S}_x = \sum_i \left[-\Lambda_{i,1}(f_i - \mu_{i,1})^2 + \Lambda_{i,0}(f_i - \mu_{i,0})^2 + \Lambda_{i,\alpha}(f_i - \mu_{i,\alpha})^2 \right] \tag{5}$$

Where for clarity we have simplified the learned means to μ, and the learned precisions to Λ, for target, 1, distractors, 0, and non-task foils, α.

[3] AIM uses ICA to find a roughly orthogonal basis which the authors argue can be usefully compared to sparse coding in early visual cortex. E-GS uses the handcrafted sparse basis from GS. Both can summarised as: response activity is computed **in parallel** for multiple features. Activity which is surprising on a feature is salient.

We might think of the third term in (5) as constant: in the absence of any task, this is likely the term that dominates saliency. However, once I have a particular task, then the other terms will contribute to my estimate of the saliency. Rather nicely, we can also see how expertise might play a role: if a task has been repeated many times, then the precisions associated with those task-relevant features will be high, and so will dominate the saliency computation.

It is easy to see how this formulation could give rise to inattentional blindness: if a surprising object is neutral with respect to a task, then whether it affects the overall saliency measure will depend on the relative precision weighting of the first two terms and the third (if it is extremely surprising because it is blue, for example, but our task is clear-cut so the precisions associated with the top-down terms is high, then it still might not be that salient overall – if it is the only blue object we have seen, then its associated precision may be quite low). In a free viewing paradigm, the search task-relevant terms would be absent, the model would collapse to AIM, and unexpected objects would be salient. If the object possesses some task relevant features, then the first and second terms will contribute to its saliency, and it is more likely to be attended.

The leveraging of precision weighted errors to produce different effects can also be related to the predictive coding work of Friston and colleagues (Friston, Adams, Perrint, & Breakspear, 2012), where a free-energy minimising agent passes precision-weighted surprisals up a processing hierarchy, and expectations down. Indeed, Friston has explicitly claimed that (covert) attention can be thought of as precision weighting (Feldman & Friston, 2010), which our simple model certainly aligns with.

Methods

To test our model in a conjunction search paradigm, we simulated image environments of distractor and target objects on a 5×5 grid with a white background (See Figure 1 for an example trial). We represented images both at the object level and the pixel level (see Feature Spaces, below); in either case, at test time a vector valued representation $\boldsymbol{F}_x$ scaled to [0,1] was passed to a learned model, which returned the saliency score for each location x by computing $\mathbb{E}_\rho[\boldsymbol{S}_x]$, where we assume Beta priors over the expected activations, such that $\rho_{i,t} \sim \text{Beta}(\alpha_{i,t},\beta_{i,t})$, $\rho_{i,d} \sim \text{Beta}(\alpha_{i,d},\beta_{i,d})$ and $\rho_{i,nt} \sim \text{Beta}(\alpha_{i,nt},\beta_{i,nt})$. We used a Beta prior as the model assumes f_{xi} are rate activations, and hence fall in [0,1].

$\mathbb{E}_\rho[\boldsymbol{S}_x]$ is the sum of the expected value of each of the terms in (5). For the i^{th} feature of the k^{th} term at location x, this is:

$$\begin{aligned} \mathbb{E}_{\rho k}\left[\Lambda_{i,k}(f_{xi} - \mu_{i,k})^2\right] = n_k \Big[f_{xi}^2 \frac{(\alpha_{ik}+\beta_{ik}-1)(\alpha_{ik}+\beta_{ik}-2)}{(\alpha_{ik}-1)(\beta_{ik}-1)} \\ - f_{xi} \frac{2(\alpha_{ik}+\beta_{ik}-1)}{\beta_{ik}-1} + \frac{\alpha_{ik}}{\beta_{ik}-1} \Big] \end{aligned} \tag{6}$$

In the case of an object-level representation, x corresponds to an object in the image. In the case of the pixel-level representation, we computed saliency for every second pixel, which

gave reasonable results and was less costly than computing for every pixel.

For each task, the posterior beliefs of the model were learned from 100 labelled example trials. The learned model was then used to generate saliency maps for 1000 unlabelled trials, where, for object-level saliency maps, rank order of objects by saliency was taken to be directly proportional to response time (RT).

When pixel-level saliency maps were generated, we explicitly "saccade" through the most salient pixels in order, and introduce inhibition of return, which depresses the saliency S at pixel i at step t according to:

$$\begin{aligned} S_{i,t} &= S_{i,t} - (S_{i,t} \cdot R_{i,t-1}) \\ R_{i,t-1} &= \mathrm{G}(S_{i,t}) + \frac{1}{2} R_{i,t-2} \end{aligned} \tag{7}$$

where $\mathrm{G}(S_{i,t})$ is a Gaussian function, with a standard deviation $1/16$ the size of the image, of the distance of i from the target of the t^{th} saccade. The sequence of response times to any particular object is then taken to be proportional to the value of t when a pixel of that object is first visited.

Feature Spaces

Our saliency measure relies on the assumption that we have access to a sparse, independent feature representation of the visual space; either at the object level or at the pixel level. In principle, both should perform similarly, and so we tested our hypotheses (see Results) against both a variant of Guided Search's handcrafted approach to generating activations from features (Wolfe, 1994), and AIM's unsupervised approach (Bruce & Tsotsos, 2009), which uses ICA to generate a vector of activations from an image patch.

Guided Search has an eight-dimensional feature space: four activations correspond to colour, four to orientation. The four orientation dimensions are given by:

Steep: $\cos(2x)^{0.25}$, $-45 < x < 45$
Shallow: $|\cos(2x)|^{0.25}$, $-90 < x < -45$ and $45 < x < 90$
Left: $|\sin(2x)|^{0.25}$, $-90 < x < 0$
Right: $\sin(2x)^{0.25}$, $0 < x < 90$

The four colour receptors are red, yellow, green, and blue, described as the "quite arbitrary … third root of triangluar functions" (Wolfe, 1994) that have peaks at positions evenly spaced at their ordinal positions in the spectrum. These activations are then passed through a local differencing operator to yield a bottom-up activation.

For unsupervised extraction of a sparse basis, we sampled 250,000 image patches of size 21×21 from a dataset of natural images (Hodosh & Hockenmaier, 2013), and used JadeICA (Cardoso, 1999), preserving 90% variance to extract an independent basis (27 dimensions were retained). ICA infers the mixing matrix, $\boldsymbol{B}$, between the independent causes and the perceived data (the patches). We then use $\boldsymbol{B}^{-1}$ to produce a vector of activations for any new patch.

Both approaches are claimed to produce activations corresponding to neuronal activity; Wolfe (1994) chose the eight features of Guided Search accordingly, and Bruce and Tsotsos (2009) argue that the roughly orthogonal basis learned by ICA can be usefully compared to sparse coding in early visual cortex. Hence it should be the case that our model produces similar performance from both forms of preprocessing.

Learning

The posteriors are computed using:

$$\begin{aligned} \rho_k | F_{xk} \sim \mathrm{Beta}\Big(&\lambda\alpha_k^0 + (1-\lambda)\Big[\alpha_k + \sum_{x \in X_k} f_{xk}\Big], \\ &\lambda\beta_k^0 + (1-\lambda)\Big[\beta_k + \sum_{x \in X_k} 1 - f_{xk}\Big]\Big) \end{aligned} \tag{8}$$

where X_k denotes the set of points labelled k in the training examples. This, as in Mozer and Baldwin (2007), interpolates between the prior distribution $\sim \mathrm{Beta}(\alpha_k^0, \beta_k^0)$ and the empirical posterior. This interpolation regularises the model's fit to the data, and improves its performance.

For all experiments, $\alpha_{id}^0 = \beta_{jt}^0 = 10$, $\alpha_{it}^0 = \beta_{jd}^0 = 25$, for all i and j, $\alpha_{int}^0 = \beta_{int}^0 = 10$, and $\lambda = 0.3$. These parameter values are mostly taken from Mozer and Baldwin (2007), as there was no reason to change them.

Results

We tested two hypotheses: that our model would reproduce a range of standard effects in visual search, and that our model could reproduce two standard results from the inattentional blindness literature.

Visual Search

To evaluate the performance of the model in the visual search paradigm, we followed Wolfe (1994) and Mozer and Baldwin (2007), and tested our model against six search tasks used to evaluate the original guided search model. These tasks are as follows. All graphs shown are using the eight simple features of guided search. Standard error bars are included.

1. Vertical target among homogeneous distractors (Figure 2): As the angle of the distractors increases from 0–55 degrees (where 0 is vertical), time to target should become constant with respect to the number of distractors (i.e., pop-out occurs).
2. Categorical search (Figure 3): Target among two types of distractors defined with respect to a single feature (angle of orientation). Distractors are 100 degrees apart, and target is 40/60 degrees from the distractors in two cases, but in the third case it is the only near vertical item, allowing pop-out.
3. Target-distractor similarity (Figure 4): Search efficiency for target among heterogeneous distractors. There are two target orientations, and two degrees of target similarity. For each orientation, search should be more efficient when target and distractors are dissimilar.

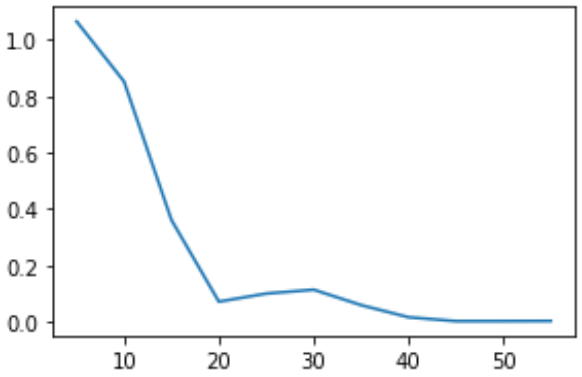

Figure 2: (Task 1) Horizontal; distractor orientation (degrees). Vertical; Gradient of time-to-target against number of distractors. Pop-out clearly occurs at around 20 degrees from the vertical.

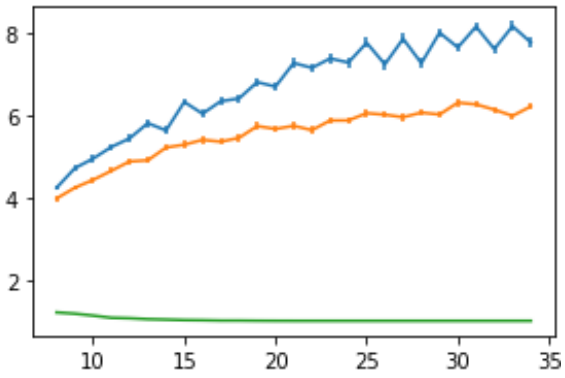

Figure 3: (Task 2) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Blue; target at 10, distractors at -30 and 70 degrees. Orange; target at 20, distractors at -20 and 80 degrees. Green; corresponds to case where distractor is the only near-vertical item. Target at 10, distractors at -50 and 50 degrees.

4. Feature search asymmetry (Figure 5): It is more efficient to find a tilted bar among verticals than a vertical among tilted. This is because tilted items activate features that make them more discriminable; for example in the 8 dimensional feature space described above, one feature activates when presented with vertical objects, but two activate when presented with objects at 20 degrees.

5. Conjunction search – distractor confusability (Figure 6): Red vertical target among green vertical and red tilted distractors. Red tilt can be 90 or 40 degrees: both are inefficient, but should vary in relative difficulty.

6. Distractor ratio effect (Figure 7): Response times for red vertical target amongst red tilted and yellow vertical distractors, as a function of ratio of distractor types. Search should be most efficient in the extremes, where there are a minimum of distractors of one particular type.

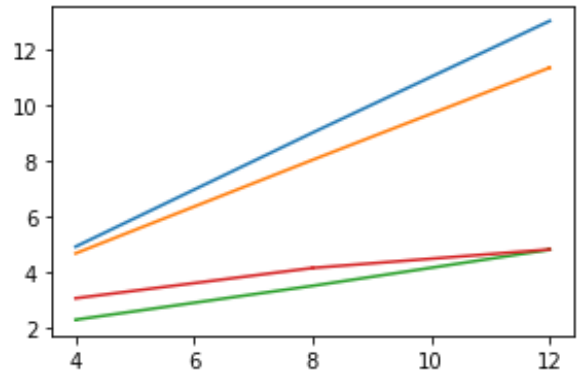

Figure 4: (Task 3) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Blue and Orange; target at 0, distractors at -20 and 20 degrees, and -40 and 40 degrees respectively. Green and Red; target at 20, distractors at 0 and 40 degrees, and -20 and 60 degrees, respectively.

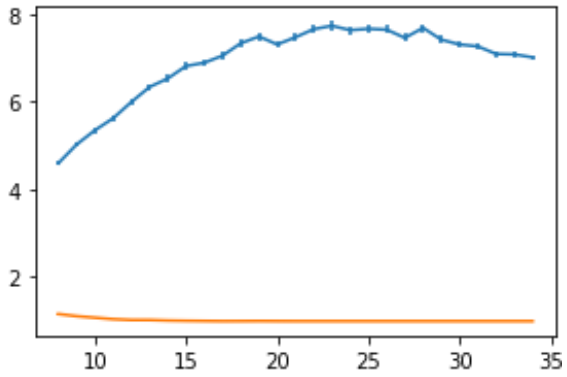

Figure 5: (Task 4) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Blue; Target at 0, distractors at 20, orange; target at 20, distractors at 0.

Inattentional Blindness

We aimed to test two basic results in the inattention blindness literature. First, that performing a task reduces the probability of fixating or reporting unexpected objects, when compared to a task-free control (Simons & Chabris, 1999).

With reference to Equation (5), we assume that $\Lambda_{i,1} \approx \Lambda_{i,0} = \Lambda_T, \forall i$ (i.e., the two task-specific confidences are similar), the relative magnitude of Λ_T to Λ_α should be central to the relationship between performing a task, and corresponding inattentional blindness. This is because if Λ_T is much larger than Λ_α then the task-specific terms dominate the saliency score, and objects which are surprising in features that are not task specific have lower probability of detection.

In free viewing, however, where Λ_α is larger than, or equal to Λ_T (the task does not dominate attention), the context-dependent surprisals should contribute to the overall saliency, and generically unexpected objects (persons in gorilla suits, for example), are more likely to capture attention.

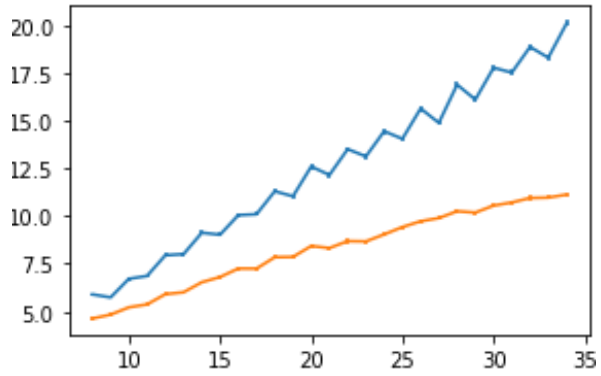

Figure 6: (Task 5) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Red targets at 0 degrees, one set of green distractors at 0 degrees. Blue; second set of red distractors at 40 degrees. Orange; second set of red distractors at 90 degrees.

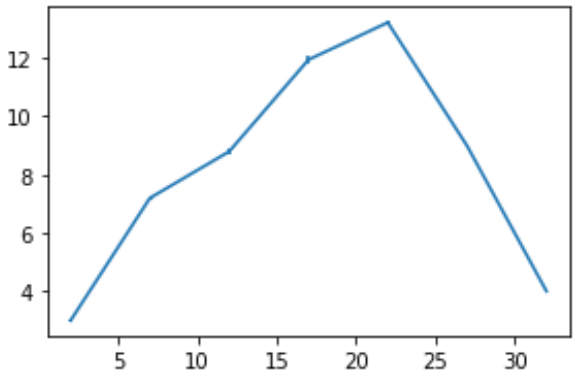

Figure 7: (Task 6) Horizontal; number of red tilted distractors (of a total 35 distractors). Vertical; response time/time-to-target (in fixations, t). Distractors were yellow vertical, and red 60 degrees. Target was red vertical.

To test this we introduced critical trials into our normal experiment. On a critical trial an unexpected (blue, left-leaning) object (see Figure 8a for an example) is also present along with the normal distractors and target. We varied the ratio n_T/n_α between 0.005 and 20. Figures 8b and 8c show a clear transition between the scenario in which the $\boldsymbol{\alpha}$ term dominates the saliency computation – in which the unexpected item pops out amidst the red task-relevant objects – and that in which the $\boldsymbol{T}$ terms dominate – where the same object does not pop out of the target and distractors, even though it is clearly surprising to an outside observer.

Second, we checked that if an unexpected object possesses features that are also task relevant, it is more likely to be fixated or reported (Most et al., 2001). We modified Task 2 (see Visual Search, above), as here the target and distractors are defined with respect to only one feature dimension. For a critical trial with a red target at 10 degrees, and red distractors at 30 and 70 degrees, we added an unexpected blue singleton at −70 or 15 degrees. Average number of fixations to target for the singleton at a task-relevant angle (70 degrees) was 13.99 ± 0.002. For the singleton at a task-irrelevant angle it was 2.0 ± 0.001. This was for a constant 12 distractors, and the ratio n_T/n_α was set to 100. This is quite a substantial difference (probably because the experimental set-up was as simple as possible), but it bears out our hypothesis.

Conclusion and Future Work

A weakness of this work is that as it is intended as a theoretical starting point, our analysis is primarily qualitative, and we have not compared the original predictions of our model to data from human participants. We will focus on these deficiencies in upcoming work via two main avenues.

The first approach is to test human participants to show that modulating the relative model precisions of (i.e., confidences in) targets specifically affects the probability that unexpected objects might be detected. If, for example, participants were initially provided only with a verbal descriptions of a visual target, we would expect probability of inattention to a non-target singleton to increase over the course of several trials, as participants became more confident in the target of their search task. We would also expect probability of inattention to be greater for a comparable task where participants are provided with a visual example of their target.

Our second approach, which lies solely in the conjunction search paradigm, would be to include distractors in a conjunction search task that shared no features with the target. We hypothesise that both overt indications of attention (fixations) and covert indications of attention (average time to target) to these non task-relevant objects would decrease over the course of several trials.

We have made three distinct contributions; we have presented a model of visual search that exhibits inattentional blindness, we have shown the equivalency of AIM and E-GS under certain assumptions, and we have argued that an interpretation of what it is to "perform a task" should be grounded on the relative precisions of parts of the brain's generative model.

We conclude that modelling task-based behaviour as explicitly located in a wider context can bear explanatory fruit.

Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Braun, J. (2001). It's great but not necessarily about attention. *Psyche*, *7*, 6.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 5–5.

Cardoso, J. (1999). High-order contrasts for independent component analysis. *Neural Computation*, *11(1)*, 157-192.

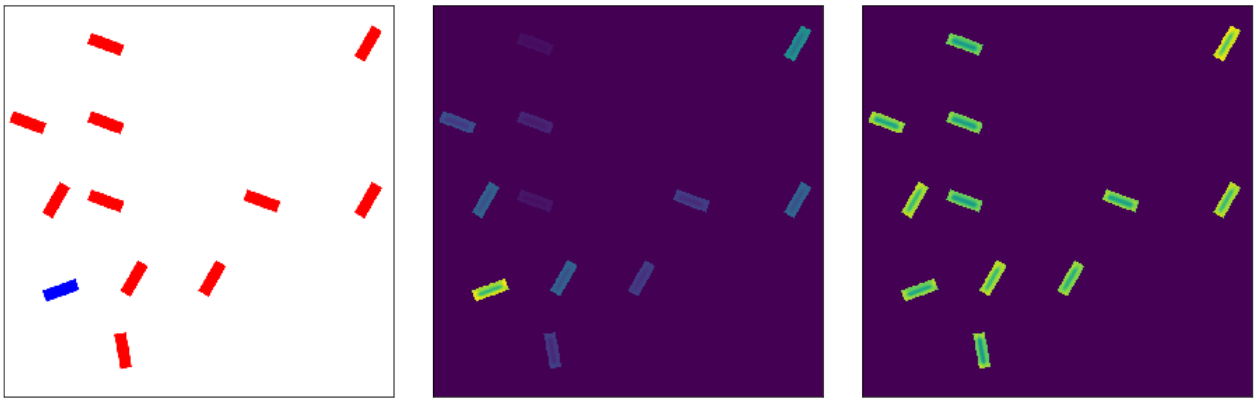

Figure 8: Maps of relative saliency in critical trial in search for red near-vertical target amongst red distractors. Left (a): trial image, Center (b): saliency map when $\boldsymbol{\alpha}$ term dominates the saliency computation, in which the unexpected item pops out amidst the red targets and distractors, Right (c): saliency map when $\boldsymbol{T}$ terms dominate, where the same object does not pop out.

Cave, K. R. (1999). The featuregate model of visual selection. *Psychological Research*, *62*(2), 182–194.

Choi, M. J., Torralba, A., & Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recognition Letters*, *33(7)*, 853–862.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.

Friston, K., Adams, R. A., Perrint, L., & Breakspear, M. (2012). Perceptions as hypothesis, saccades as experiments. *Frontiers in Psychology*, *151(3)*, 1-20.

Hodosh, M., & Hockenmaier, J. (2013). Sentence-based image description with scalable, explicit models. *CVPR Workshops*, 294-300.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3), 14–14.

Koivisto, M., & Revonsuo, A. (2009). The effects of perceptual load on semantic processing under inattention. *Psychonomic Bulletin & Review*, *16*(5), 864–868.

Kreitz, C., Furley, P., Memmert, D., & Simons, D. J. (2016). The influence of attention set, working memory capacity, and expectations on inattentional blindness. *Perception*, *45*(4), 386-399.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322.

Most, S. B., Scholl, B. J., Clifford, E. R., & Simons, D. J. (2005). What you see is what you set: sustained inattentional blindness and the capture of awareness. *Psychological Review*, *112*(1), 217.

Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., & Chabris, C. F. (2001). How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological Science*, *12*(1), 9-17.

Mozer, M., & Baldwin, D. (2007). Experience-guided search: A theory of attentional control. In *NIPS* (p. 1033-1040).

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature*, *320*, 264–265.

Navalpakkam, V., & Itti, L. (2006). Top–down attention selection is fine grained. *Journal of Vision*, *6*(11), 4–4.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, *4*(4), 147–155.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074.

Torralba, A., Oliva, A., Castelhano, M. S., & 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.

Wolfe, J. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. (1999). Inattentional amnesia. *Fleeting Memories*, *17(5)*, 71-94.

Wolfe, J., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*, 0058.