A theory of visual search for foveated visual systems.

Technical Report by R. Calen Walshe

November 1, 2019

0.1 Abstract

Searching the environment for specific targets is a critical capability for humans and other animals. Here we present a principled theory of visual search in arbitrary backgrounds that is based on the statistical properties of natural images, the retinal falloff in resolution and sampling with eccentricity, the increase in intrinsic location uncertainty with retinal eccentricity, and the prior probabilities of target presence and location within the image. The important theoretical result is that near-optimal fixation selection for arbitrary targets added to natural backgrounds can be obtained with relatively simple, biologically plausible mechanisms. We test the theory here by predicting human covert search behavior in white noise backgrounds. Predictions are generated as follows. First, the effective prior probability distribution on target location is computed from the prior and the intrinsic location uncertainty. Second, the effective amplitude of the target (also dependent on retinal eccentricity) is computed and the target (if present) is added to the background. Third, template responses are computed at each image location by taking the dot product of a template (having the shape of the target) with the image and then adding a random sample of internal noise. Fourth, the responses are correctly normalized by the sum of the internal noise variance and the estimated variance due to external factors (the background statistics). Fifth, the normalized responses are combined with the effective prior on target location to obtain values proportional to the posterior probability. If the maximum of these values exceeds a criterion, the response is that the target is present at the location of the maximum. Human error rates averaged over spatial location correspond to the theoretical predictions. However, humans make a higher proportion of errors near the fovea, possibly due to underestimation of priors near the fovea or to allocating more decoding resources to the periphery.

0.2 Theory of visual search

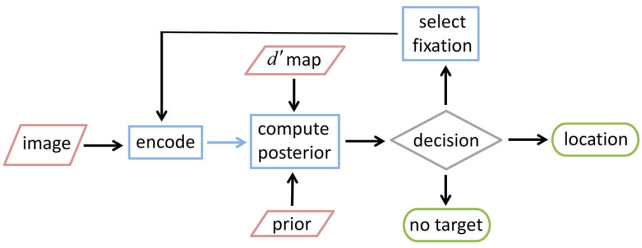

Figure 1

Here we consider search for a specific target object that, if present, is added to the background image at a location randomly sampled from an arbitrary prior probability distribution. For the covert search task, we assume that the stimulus is presented for the duration of a typical fixation in overt search tasks (nominally 250 ms).

It is well known that the optimal procedure for detecting known targets at a fixed position in Gaussian white noise is a template matching observer. Template matching involves applying a template (i.e. receptive field), that has the same shape as the target, to the image. In general, when the target is present the response to the background will be larger. In visual search, the location of the target is by definition unknown so there is uncertainty about the location of the target. A simple, and under many conditions near optimal, strategy for identifying targets under position uncertainty is to compute a normalized template response, $R(\mathbf{x})$, at each location of the image. Recent work in natural image statistics has shown that template response distributions to natural images are Gaussian and the standard deviations that vary as a product of the local luminance, RMS contrast, and similarity (cosine of the angle between the amplitude spectrum of the target and the background). This result implies that a template matching observer, suitably normalized by local image statistics, is approximately optimal for natural scenes.

In general, the template response distribution takes the following form:

$$R(\mathbf{x}) \sim \begin{cases} \mathcal{N}\left[d'(\mathbf{x})\rho_\pi\left(\mathbf{x}-\mathbf{x}_T\right),1\right] \\ \mathcal{N}[0,1] \end{cases} \tag{1}$$

where the template response is distributed as a standard normal with a mean equal to the target detectability at location $\mathbf{x}$ multiplied by the correlation between the template and the target when the target is present and 0 otherwise. The standard deviation is equal to 1 because the template response is normalized. Normalization is taken care of by the model of detectability in natural scenes that is described next.

The model of detectability follows from recent work studying the performance of template matching observers for additive targets in natural scenes. The detectability at a location $\mathbf{x}$ takes the following form:

$$d'(\mathbf{x}) \propto \left[a \frac{\|T_\lambda(\mathbf{x})\|}{L(\mathbf{x})C(\mathbf{x})S(\mathbf{x})} F(\mathbf{x}) \right]^\beta \tag{2}$$

where detectability is a product of the amplitude of the target a, local luminance (L), RMS contrast (C) and pattern similarity (S) of the background, partial masking factor $\|T_\lambda(\mathbf{x})\|$. $F(\mathbf{x})$ is a factor that describes the fall-off in detectability for foveated visual systems.

The behavioral response is determined first by taking the log-likelihood ratio at each location of the template response map:

$$l(\mathbf{x}) = d'(\mathbf{x})R(\mathbf{x}) - 0.5d'^2(\mathbf{x}) \tag{3}$$

Next, this quantity was combined with the effective prior and compared against a criterion. If the location with the maximum posterior probability exceeds a criterion then this location is returned as the response, otherwise location with the maximum posterior probability is returned, otherwise the response is absent:

$$\max_{\mathbf{x}}[l(\mathbf{x}) + \ln \text{prior}(\mathbf{x})] > \gamma \tag{4}$$

0.3 Covert search in white noise: A test of the theory

A test of the covert search theory was conducted in a search for a 6 cpd raised cosine windowed sine wave added to a white noise background. Subjects fixated the center of the search display which was a circular area of white noise subtending 18° of visual angle. Target positions were uniformly sampled from the search display subtending a radius of 8° . The stimulus was presented for 250 ms and subjects gave their response with a mouse click on the location of the target if present, or clicking a different button to decide absent. Preliminary psychophysical experiments were conducted to measure subject specific detectability maps (d'). The human results and the theoretical predictions are shown in Figure 2. The response type is plotted as a function of the binned distance from the fovea (center of display). The number of responses in each distance bin is normalized by the total number of targets presented in that bin. Key results obtained are: i) humans make fewer hits and more misses in the near fovea than is predicted from the ideal and ii) humans make more false alarms in the periphery than is predicted by the ideal. An alternative, sub-optimal, model was tested by specifying an alternative prior distribution to the one that was used to sample the target positions in the image. A cartoon of the priors in 1D profile are shown in the insets of Figure 2. Although preliminary, this is suggestive that the divergence between the ideal and humans may arise due to underestimation of prior probability in the foveal region.

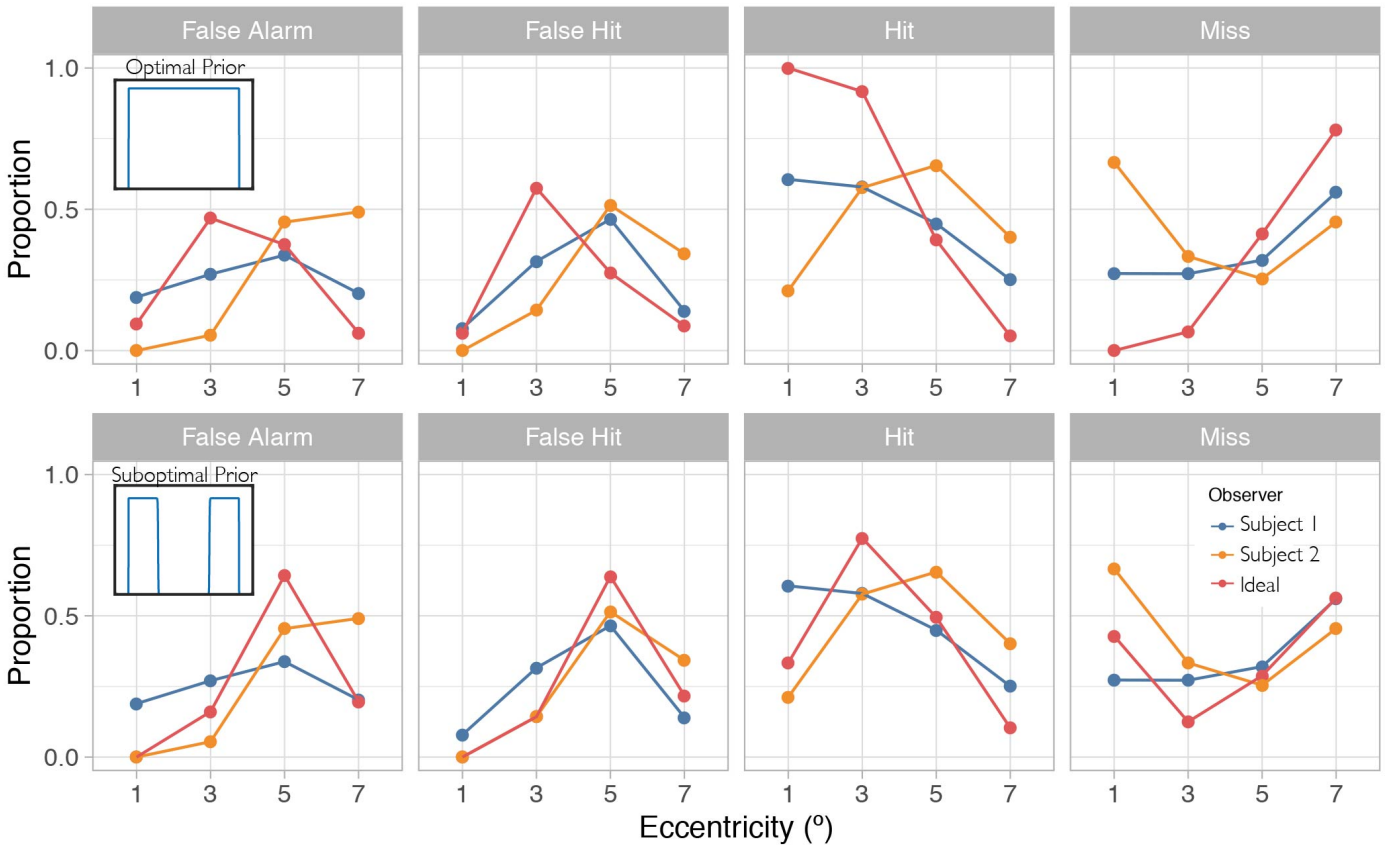

Figure 2: Search Results