Statistical templates for visual search

John F. Ackermann
Department of Psychology, New York University, New York, NY, USA

Michael S. Landy
Department of Psychology and Center for Neural Science, New York University, New York, NY, USA

How do we find a target embedded in a scene? Within the framework of signal detection theory, this task is carried out by comparing each region of the scene with a "template," i.e., an internal representation of the search target. Here we ask what form this representation takes when the search target is a complex image with uncertain orientation. We examine three possible representations. The first is the matched filter. Such a representation cannot account for the ease with which humans can find a complex search target that is rotated relative to the template. A second representation attempts to deal with this by estimating the relative orientation of target and match and rotating the intensity-based template. No intensity-based template, however, can account for the ability to easily locate targets that are defined categorically and not in terms of a specific arrangement of pixels. Thus, we define a third template that represents the target in terms of image statistics rather than pixel intensities. Subjects performed a two-alternative, forced-choice search task in which they had to localize an image that matched a previously viewed target. Target images were texture patches. In one condition, match images were the same image as the target and distractors were a different image of the same textured material. In the second condition, the match image was of the same texture as the target (but different pixels) and the distractor was an image of a different texture. Match and distractor stimuli were randomly rotated relative to the target. We compared human performance to pixel-based, pixel-based with rotation, and statistic-based search models. The statistic-based search model was most successful at matching human performance. We conclude that humans use summary statistics to search for complex visual targets.

Introduction

How do we find a target embedded in a scene? A long-standing school of thought proposes that we do so by comparing local regions of the scene with a template, i.e., an internal representation of the object we're looking for (DeValois & DeValois, 1990; Graham, 1989; Green & Swets, 1966; Marr, 1982; Verghese, 2001). For a fixed target image in white noise, the optimal form of this internal representation is a matched filter. The response of a matched filter is a linear combination of visual inputs across a localized region of the scene. The largest responses occur when the input is similar to the search target. The template response is a measure of similarity between the thing we're looking at and the thing we're looking for, at each possible location where that thing might be.

Physiological instantiations of such filters are to be found throughout the visual system. For example, the visual system is composed of multiple spatial frequency and orientation channels (Blakemore & Campbell, 1969; Campbell & Robson, 1968; DeValois & DeValois, 1990). A simple cell in primary visual cortex (V1) represents such a channel, tuned for a particular spatial frequency and orientation (Hubel & Wiesel, 1968). It thus acts as a matched filter, i.e., a template for a rudimentary stimulus such as an edge, a line, a patch of grating, etc., at that frequency and orientation.

Linear matched filters have been used extensively to simulate human visual processing in psychophysical tasks. They have been shown to accurately predict performance in search tasks for geometric shapes (Burgess, 1985; Burgess & Colbourne, 1988; Burgess & Ghandeharian, 1984; Rajashekar, Bovik, & Cormack, 2006), gratings. (Burgess, 1981; Najemnik & Geisler, 2005, 2008, 2009), and blob-like targets (Abbey & Eckstein, 2009; Barrett, Yao, Rolland, & Myers, 1993; Eckstein, Abbey, & Bochud, 2009). Below we suggest that a template of a different form, i.e., one that employs various statistics derived from the responses of nonlinear filters, may provide a more valid and biologically plausible representation of the template used in tasks involving more complex stimuli.

Citation: Ackermann, J. F., & Landy, M. S. (2014). Statistical templates for visual search. *Journal of Vision*, *14*(3):18, 1–17, http://www.journalofvision.org/contents/14/3/18, doi:10.1167/14.3.18.

doi: 10.1167/14.3.18 Received May 23, 2013; published March 13, 2014 ISSN 1534-7362

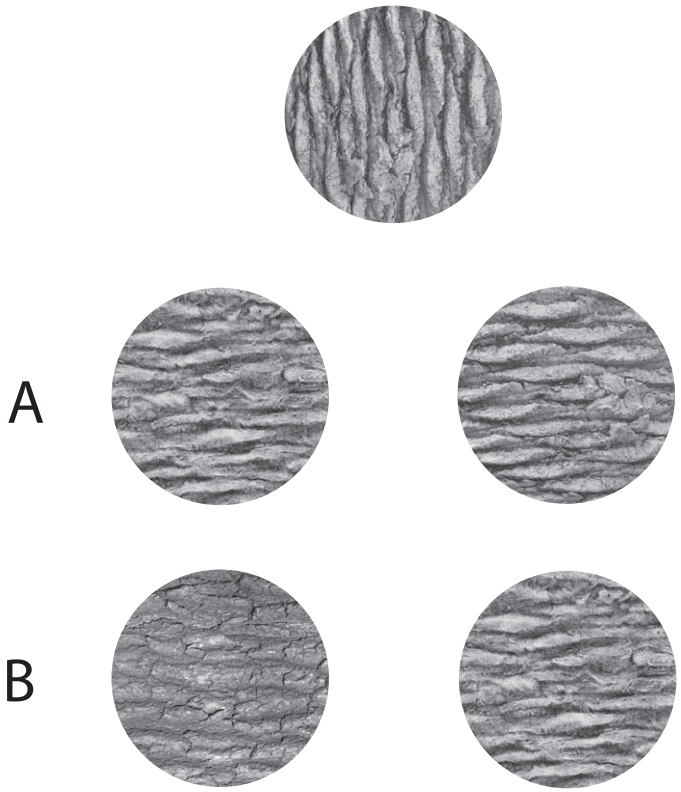

Figure 1. Two search tasks. Most observers can rapidly identify the image in row A that is identical to the top image, in spite of the difference in orientation. Identifying the image in row B that consists of the same type of tree bark as the top image is trivially easy, in spite of the fact that the two images are entirely different, pixel-for-pixel.

Common experience tells us that the thing we're looking for will rarely appear in the exact form that we expect. An example is shown in Figure 1. The search target is the texture patch in the top row. The task is to locate the image in row A that is the same pixel-for-pixel as the target. Most observers will find it relatively easy, given a glance at the two images of reasonable duration, to locate the matching image in spite of the difference in orientation. How can a template-matching model account for our ability to find a search target with uncertain orientation with so little effort?

If the observer's task is to determine the presence of a line of a particular frequency and unknown orientation, signal detection theory suggests that the observer monitors the outputs of multiple templates tuned to all possible orientations. The log-likelihood of the stimulus given each template response is summed across all templates. The observers responds, "The line is present," if the likelihood ratio exceeds a threshold (Green & Swets, 1966). Primary visual cortex contains a representation of edges varying in size and orientation. It is unlikely, however, that this type of matched-filter representation operates in search for stimuli more complex than an edge, since it would require the availability of a cell tuned to every possible image at every possible orientation—an implausibly large number of cells. Thus, template matching does not provide a good account of our ability to detect a complex search target at unpredictable orientation.

Although multiple templates are unlikely to play a part in search for complex images, it is possible that a single, learned template may be mentally rotated and thereby provide a rotation-invariant representation of the target. It is been suggested, for example, that mental rotation is used to match an abstract shape to a rotated version of that shape (Cooper, 1976; Shepard & Metzler, 1971). We propose that a human observer might mentally rotate a learned image template for visual search. When observers are presented with an image that may be a rotated version of the target, they estimate the angle of rotation as the difference in the filter orientations of the subbands (representing the responses of multiple spatial frequency- and orientation-tuned channels) with maximum power in the decompositions of the target and the image being matched. The template is then rotated by the estimated rotation angle, and the template match proceeds as usual.

While a rotated template may account for the relative ease with which we search for a complex target of unknown orientation, it does not explain the ease of search for targets defined statistically rather than by a specific configuration of pixels. Figure 1 shows an example. The search target is the texture patch in the top row. The task is to locate the image in row B that consists of the same type of tree bark as that in the target image. For most observers, this task is trivially easy in spite of the fact that the correct image (on the right) is entirely different from the target image in terms of the arrangement of pixel intensities. A pixel-based template-matching model cannot account for how easy it is to identify such textural "stuff." Search seems to use a template tuned to the textural qualities of the image.

How might a textural template manifest in the visual system? Portilla and Simoncelli (2000) provided a texture analysis algorithm that might also serve as a biologically plausible model of how textures are represented visually. In it, an image is decomposed using localized linear filters at a range of orientations and spatial scales. The resulting oriented subbands at each spatial frequency are analogous to the outputs of a population of simple cell-like filters covering the spatial extent of the image. The image is decomposed twice in this way using two sets of filters that have identical orientation and spatial frequency but orthogonal phase. The filter pairs are used to derive local energy and phase. The linear filter outputs are correlated across scales, and local energy and phase are correlated within and across subbands, resulting in a representation of the image in terms of a vector of correlation coefficients. General pixel statistics (mean, variance,

skew, kurtosis, and range) are also derived and contribute to the representation.

How might this statistical representation be computed in the brain? A cascade of cells, with the outputs from V1 feeding into similar orientation- and frequency-tuned cells in V2, with interposed nonlinearities, can account for our ability to localize more complex stimuli such as edges formed from abutting texture elements rather than changes in luminance (Landy & Graham, 2003; Landy & Oruç, 2002). A further cascade with outputs from the previous level converging onto cells in subsequent levels results in increasingly sophisticated response properties, such as cells in area V4 that appear to respond selectively to complex shape (Tanaka, 2003; Wang, Tanifuli, & Tanaka, 1998). These cascades of linear filters and nonlinearities could form the substrate for computing a texture representation like that described by Portilla and Simoncelli (2000).

There are several lines of evidence to suggest that the Portilla and Simoncelli (2000) model provides an effective model of human representation of visual stimuli. Portilla and Simoncelli's algorithm allows for the synthesis of new texture images using the coefficients derived from the analysis of a texture image. The synthesized textures can be discriminated from the original image with foveal scrutiny, but are generally indistinguishable from the original when viewed briefly or in the periphery. Selectively leaving out sets of coefficients from the synthesized textures leads to systematic changes in texture appearance and increases observers' ability to discriminate them from the originals (Balas, 2006).

The phenomenon of "crowding" (Bouma, 1970; Pelli & Tillman, 2008) occurs when identification of a stimulus is impaired in the presence of nearby flanking stimuli. It has been suggested that crowding results from the pooling of the stimulus and flankers for the computation of such a texture representation, where the pooling region size increases with retinal eccentricity (Balas, Nakano, & Rosenholtz, 2009). In this theory, the portions of an image that fall within a pooling region are represented in terms of a single set of summary statistics. As a result, the representation hopelessly entangles the features of discrete objects within the pooling region. A pooled summary-statistic representation has also been shown to predict performance in peripheral search for a target among distractors (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012). The size of the pooling region as a function of retinal eccentricity has been shown to correspond to the size of receptive fields of cells in area V2 of the visual cortex (Freeman & Simoncelli, 2011).

In this viewpoint, at some stage of cortical processing the retinal image is represented as a collection of texture statistics for each of a large collection of pooling regions extending across the visual field (Freeman & Simoncelli, 2011; Rosenholtz, Huang, & Ehinger, 2012). The increase in size of the pooling regions with eccentricity reflects the loss of visual information in the periphery (Balas, 2006; Rosenholtz, Huang, Raj et al., 2012). Objects falling within multiple, tiny pooling regions near the center of the visual field appear distinct while the features of those in the periphery appear jumbled.

Just as a rotated matched filter may be used to match rotated images, the statistical representation of a texture may be "rotated" to match rotated textures. In fact, with the texture representation, the "rotation" of the representation is computationally trivial. Suppose one wants to use the set of texture statistics for an upright texture as a template to match against an image one suspects has been rotated by 45°. The list of texture statistics consists of correlations of energy and phase values of various oriented filters. For example, one statistic might be the correlation between the energy in the outputs of 0° and a 45° filters of a certain spatial scale. After rotation of the template, this statistic need only be relabeled as the correlation between 45° and 90° filters. Thus, a rotation merely relabels (permutes) the list of texture statistics.

Further, the texture statistics themselves may be used to estimate the rotation angle. Suppose that across the multiscale and multi-orientation decomposition of the target, the subband with greatest power happens to have a 90° orientation, and that in the decomposition of the matching image the peak power subband has a 135° orientation. Using this information, one may estimate the rotation as $135° - 90° = 45°$. Using this estimate, we can permute the statistical representation of the target to form a "rotated" template that can then be compared to the statistical representation of the match image.

In the current study, we ask what computation humans use to search for a complex target with unknown orientation. We measured observers' performance in the two search tasks pictured in Figure 1. In one condition, observers were presented with an image of a texture that served as the search target for that trial. They were then presented with two potential matches to the target. One, the match, was the same image as the target. The second, the distractor, was a different image of the same textural material as the target [i.e., a resynthesis using the Portilla & Simoncelli (2000) model]. Both the match and the distractor were randomly rotated relative to the target (Figure 1A). The observers' task was to indicate which stimulus was the match. In the second condition, the search target was also a texture image. But the match was a different resynthesized image of the same textural material as the target and the distractor was a randomly selected resynthesized image of a different texture (Figure 1B). Thus, our task requires the observer to perform a pixel-based match in the first condition and a statistic-based

match in the second. We compared observers' performance to that of three model searchers that utilize either a pixel-based or statistic-based strategy. We describe the models and predictions for their performance below.

Observers' performance was compared to the predictions of three model searchers. The first model, the pixel-based searcher (PBS), uses the target as a matched filter, correlating it with the match and distractor images, and ignoring the possibility that the test images may have been rotated. Note that although the correlation of the target and the match or distractor is performed in image space (i.e., as pixels), it would be equivalent to do so in the space of responses of oriented linear filters at multiple scales as long as that multiscale, multi-orientation representation is a "tight frame," as is the case for the subband representation used by Portilla and Simoncelli (2000). Needless to say, this model proves not to be robust to image rotation.

The second model, pixel-based search with rotation (PBSr), uses the target as a matched filter, but rotates the image before correlating it with the test images based on an estimate of the relative orientation of the two images. The estimate of relative orientation is based on the subbands with maximum power in a multiscale, multi-orientation image decomposition as described above.

The third model is a statistic-based searcher (SBS). It uses the same multiscale, multi-orientation decomposition to derive local energy and phase and computes a template based on correlation statistics of these values [a portion of the texture descriptor described by Portilla & Simoncelli 2000)]. This template is then "rotated" (permuted) based on the same estimate of the relative orientation of the target and test image as in PBSr, and is correlated with an identically computed statistical representation of each test image.

As noted above, our task involves a pixel-defined match in the first condition and a statistically defined match in the second. We predict that the PBS model will perform well in the pixel-defined match when the match is not rotated relative to the target. Its performance should be at chance when the match and distractor stimuli are rotated. The PBSr strategy should perform well at the pixel-defined match independent of the orientation of the match and distractor stimuli. Both pixel-based models should fail at the statistic-based task. On the other hand, we expect the SBS model to perform poorly at the pixel-based task and well at the statistic-based task, independent of the orientation of the statistically defined match and distractor images.

We fit each model to the human data in both conditions (each model has two parameters we describe in the Methods). We found that the SBS model was more predictive of human performance than the pixel-based models (PBS and PBSr).

Methods

Participants

An author (JFA) and three additional male subjects (NYU graduate students and post-docs) participated in four sessions completed on separate days. All subjects other than the first author were compensated in the amount of $10 per session and were naive to the purpose and background of the study. All had normal or corrected-to-normal vision. Subjects signed a consent form approved by the NYU University Committee on Activities Involving Human Subjects.

Apparatus

Stimuli were presented using Psychtoolbox for Matlab (Brainard, 1997; Pelli, 1997) on a gamma-corrected, 36 × 27 cm, Sony Multiscan G400 monitor (Sony, Tokyo, Japan) with a resolution of 1600 × 1200 pixels, a refresh rate of 75 Hz, and a mean luminance of 40 cd/m^2. The monitor was powered by a Dell Precision T3400 PC (Dell, Round Rock, TX) using an Nvidia GeForce 9800 GT video card (Nvidia, Santa Clara, CA). Eye position was monitored using an SR Research Eyelink1000 desktop eyetracker with a sampling rate of 1000 Hz, controlled using the Eyelink Toolbox Matlab interface (Cornelissen, Peters, & Palmer, 2002; Eyelink, Ottawa, Ontario, Canada).

Stimuli

All 486 stimulus images were selected from a database of 256 × 256 pixel, grayscale textures. The database consists of Brodatz (1996) images and photographs compiled by the Laboratory for Computational Vision at NYU.

We expected that performance on these discrimination tasks would depend on the degree to which the texture patches have content at one or many dominant orientations. In both conditions, the potential matches to the target were randomly rotated relative to the template image. In the SBS model, the correlation between the local energy of oriented subbands of the image decompositions for the template and for the rotated potential matches were computed. Rotated images with oriented content that spans the full range of orientations should produce higher template responses than those with oriented content confined to a narrow range. For the former, rotating the image changes the local energy in each oriented subband less than for the latter. Thus, we would like to ensure that matches and distractors have the same degree of

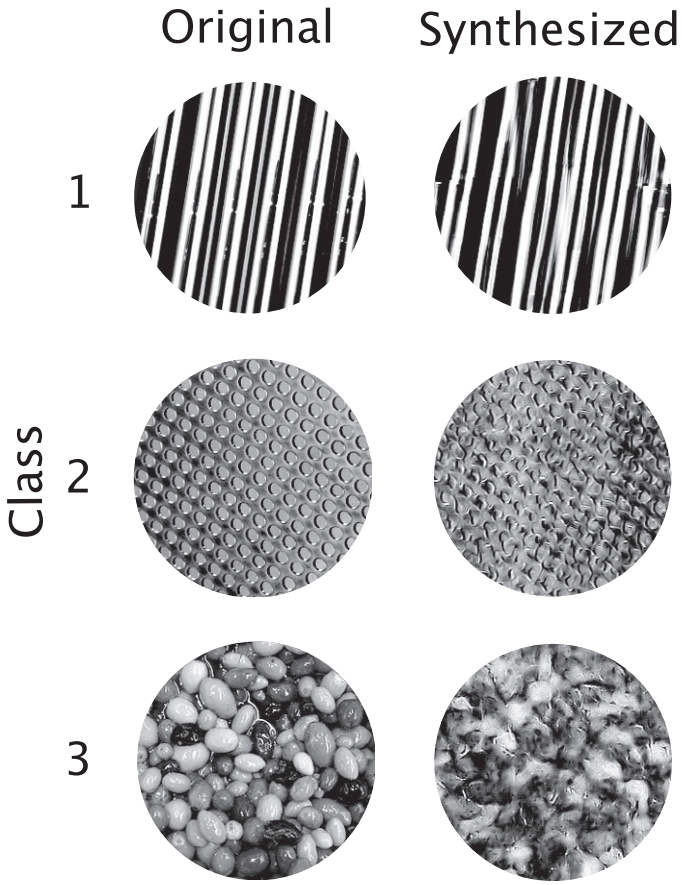

Figure 2. Example images. We defined three classes according to the degree of orientedness (see text). The original images served as target stimuli in Condition 2. Synthesized images served as target stimuli in Condition 1 and as all match and distractor stimuli.

"orientedness," i.e., a similar distribution of power across oriented subbands. We predicted performance to be more rotation-invariant for images that have content spanning a wide range of orientations but performance may be impaired when the content is concentrated at a single orientation. Thus, we classified each image in the database with regard to its relative amount of oriented content.

We classified each image by first performing a pyramid decomposition using the steerable, Fourier-domain filters described by Portilla and Simoncelli (2000). Matlab code for implementing the steerable pyramid is publicly available at: http://www.cns.nyu.edu/~eero/STEERPYR/. The decomposition starts by generating high- and low-pass filtered versions of the original image. It then samples the image at increasingly coarse spatial scales by recursively downsampling (by a factor of two) and low-pass filtering the low-pass images. Each low-pass image is split into oriented subbands. We used four spatial scales and 16 orientations per scale for the classification.

We calculated the power spectral density, P, for the subband at each spatial scale, S, and orientation, θ: $P_{S,\theta} = (1/N)\sum_{i=1}^{N} |F_{S,\theta}(x_i, y_i)|^2$, where N is the number of samples at a given spatial scale and the Fs are the complex-valued filter responses; the real and imaginary parts correspond to the responses at each location to two filters having orthogonal phase.

Images in which the content is concentrated at a single orientation will have values of $P_{S,\theta}$ that peak at a single θ for a given value of S. Images with more broadly distributed oriented content will have multiple peaks. To quantify this, we found the scale, S_{max}, that contains the maximum power over θ: $S_{max} = \arg\max_S (\max_\theta P_{S,\theta})$, $\theta = \{0, (\pi/16), \ldots, \text{etc. } (15/16)\pi\}$. We then defined the orientation index, ω, to be the coefficient of variation (SD/mean) of $P_{S_{max},\theta}$ over θ.

The images were coarsely divided into three orientation classes using this criterion. Class 1 contains images with values of $\omega > 0.6$ and content concentrated at a single orientation. Class 2 contains images with $0.2 < \omega < 0.4$ and content at two or three orientations. Class 3 contains images with $\omega < 0.1$ and content distributed across all orientations. We then visually inspected the images in each class and selected 16 from each that fall unequivocally into three qualitative classes. The images selected from Class 1 have a single dominant orientation (and a single peak within $P_{S_{max},\theta}$). Those from Class 2 are "plaid-" or "grid-like" (and have two peaks within $P_{S_{max},\theta}$). Images from Class 3 are "blob-like" (and have essentially flat profiles within $P_{S_{max},\theta}$). Example stimuli from each class are shown in Figure 2.

The 16 images from each orientation class served as stimuli in the experiment. In Condition 1, the target/match stimuli were generated by synthesizing eight new images from each of the original 48 (i.e., 16 images $\times$ 3 orientation classes) using Portilla and Simoncelli's (2000) texture analysis and synthesis algorithms (examples are shown in Figure 2). Target/match stimuli were the same image pixel-for-pixel. Distractors in Condition 1 were a different set of 8 $\times$ 48 images synthesized from the same originals as the target/match stimuli. Distractors in Condition 1 were paired, on each trial, with a target/match stimulus that was synthesized from the same original.

For Condition 2, target stimuli were selected from the original 48 (unsynthesized) images. Match and distractor stimuli were selected from the 8 $\times$ 48 set of match and 8 $\times$ 48 set of distractor stimuli, respectively, used in Condition 1. Thus, match stimuli in Condition 2 had the same statistical content as the target image but a different arrangement of pixels. Match stimuli in Condition 2 were paired, on each trial, with a distractor that was synthesized from a different original (and hence with different statistical content than the target/match).

All stimulus images were windowed within a 256-pixel–diameter disc. The contrast of the image content

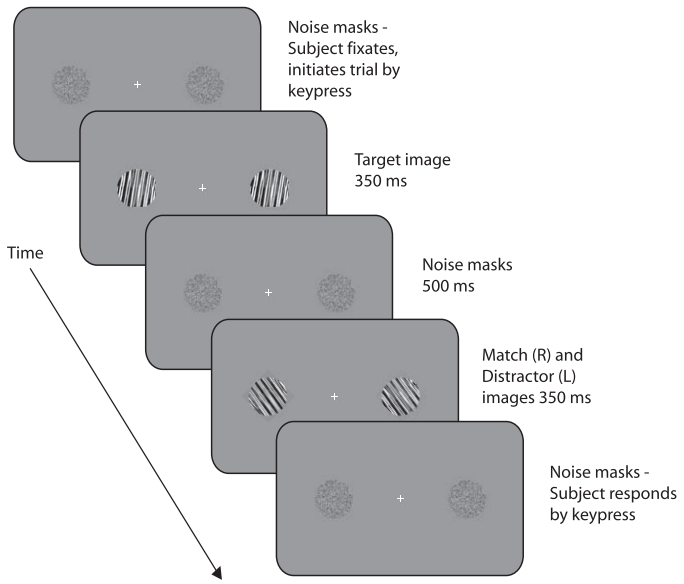

Figure 3. Trial sequence for the experiment.

appearing within the window was normalized to a mean luminance of 40 cd/m^2 and a *SD* of 13 cd/m^2. Match and distractor stimuli were randomly rotated by 0°, 45°, 90°, 135°, or 180° on each trial (matches and distractors were rotated by the same amount on each trial) prior to being windowed and normalized. Nearest-neighbor rotation was used throughout. Only these five orientations were included because pilot experiments indicated that performance for orientations between 180° and 360° mirrored performance between 180° and 0°.

Procedure

Subjects viewed the computer monitor from a distance of 57 cm with head position constrained by a chinrest. Each block began with a nine-point calibration of the eyetracker. The trial sequence is shown in Figure 3.

Each trial began with a central fixation cross flanked by two circular, 6°-diameter binary noise masks centered 12° left and right of fixation. The subject fixated the cross and pressed a key. If the subject's eye position was greater than 1° from the cross, a central red "X" appeared briefly, instructing the subject to fixate and try again to initiate the trial. When the trial was successfully initiated, the noise masks were replaced on both sides by the target stimulus for that trial. The target stimuli were displayed for 350 ms after which they were replaced by the noise masks for 500 ms. The noise masks were then replaced by the match and distractor stimuli, which appeared at positions (left or right) that were randomly selected for each trial. After 350 ms, noise masks replaced the stimuli. If the subject's eye position moved more than 1° from fixation between target onset and match/distractor offset, the trial was canceled and rerun later in the session. If the trial was successfully completed, a question mark appeared in place of the fixation cross instructing the subject to indicate by keypress the side on which the match stimulus had appeared. Auditory feedback in the form of high- and low-pitched tones indicated at the end of each trial whether their choice was correct or incorrect.

The size of the stimulus images (6° diameter) and their position (12° eccentricity) were selected in order to place them within a single pooling region according to Bouma's law (Bouma, 1970; Pelli & Tillman, 2008). Given this placement, we can assume that each image is represented, in terms of the SBS model, by a single set of summary statistics.

On each trial, the target, match, and distractor stimuli were selected from the same orientation class. Stimuli from the three classes were evenly and randomly intermixed within each condition. Each of the 1,920 match and distractor images (48 images × 8 synthesized copies × 5 orientations) appeared once in each condition. In Condition 1, each of the 384 (synthesized) target images (48 images × 8 copies) appeared five times. In Condition 2, each of the 48 (unsynthesized) target images appeared 40 times.

The experiment was completed in four sessions on different days. Observers completed each condition in two consecutive sessions of 960 trials each. The order in which the conditions were run was counterbalanced across subjects. In Condition 1, subjects were instructed that the match stimulus was exactly the same as the target and that the distractor was a different stimulus of the "same textural stuff." In Condition 2, subjects were instructed that the match was of the same textural stuff as the target and that the distractor was of a different textural stuff. In each condition, observers were told that the match and distractor stimuli would be randomly rotated (each by the same amount) between 0° and 180° relative to the target. Observers completed 40 practice trials at the beginning of the first session of each condition to ensure that they understood the instructions and stimulus characteristics.

Models

Pixel-based searcher

Our pixel-based searcher model is depicted in Figure 4. It is based on a standard signal-detection-theoretic ideal observer (Abbey & Eckstein, 2009; Geisler, 2010; Green & Swets, 1966; Lu & Dosher, 2008; Palmer, Verghese, & Pavel, 2000). The observer treats the target as a matched-filter template and compares it to both the match and distractor images. Note that the target, match, and distractor images analyzed by each of our models are the same images viewed by the human observers in the

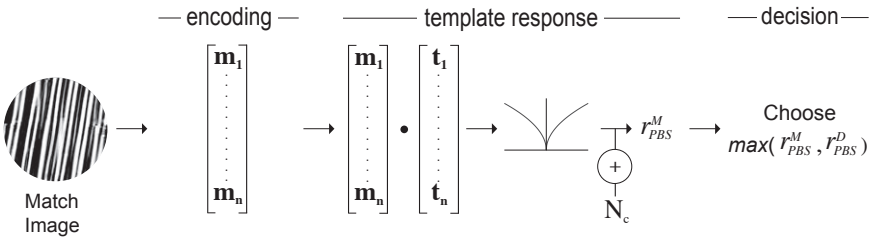

Figure 4. The PBS model. **Encoding phase**: The image (here, the match), is encoded in terms of pixel values contained in vector **m**. **Template response**: **m** is correlated with a similar representation, **t**, of the target image. The result is rectified, passed through a power function, and corrupted by additive noise, N_c, yielding noisy scalar response value, r^M_{PBS}. **Decision phase**: The observer derives a similar response to the distractor, r^D_{PBS}, and compares it to r^M_{PBS}. The stimulus yielding the maximum response is the observer's choice for the match stimulus.

experiment. The results of the comparison are passed through a nonlinearity and corrupted by noise to provide a basis for predicting human performance data. To be specific, let **T** be the windowed and normalized target image. We rearrange the pixel values of **T** as a column vector **t**. The match and distractor images are similarly represented as vectors **m** and **d** respectively, and correlated with **t**. The resulting correlation coefficients are fullwave rectified and passed through a power function (i.e., nonlinear transducer function) with exponent ω_{PBS}. This effectively passes the response through a nonlinear transducer function that accounts for typical Weber's Law-like behavior on the part of the observer (Foley & Legge, 1981; Legge, 1981; Lu & Dosher, 2008). Thus, the template response, r, of the PBS model to the match image is given by:

$$r^M_{PBS} = \left| \frac{\mathbf{m}^T \cdot \mathbf{t}}{||\mathbf{m}||\ ||\mathbf{t}||} \right|^{\omega_{PBS}}$$

and analogously for the response r^D_{PBS} to the distractor. The PBS observer chooses the image with the highest corresponding template response. We assume the comparison process is corrupted by Gaussian noise, so that the probability of choosing correctly is:

$$p_{PBS}(c) = \Phi\left(\frac{r^M_{PBS} - r^D_{PBS}}{\sigma_{PBS}} \right),$$

where Φ is the cumulative standard normal distribution. The predicted proportion correct for a condition is the average over all trials of the predicted probability correct for each trial in that condition (i.e., for the triad of target, match and distractor images presented in that trial). The two parameters (ω_{PBS} and σ_{PBS}) were adjusted to fit the data in both conditions by maximum likelihood. The fit was carried out (for all models) using a custom grid-search technique in which the range of parameter values tested was iteratively shrunk to converge on those corresponding to the maximum binomial likelihood. The search was terminated when the maximum likelihood parameters changed with a tolerance of less than 0.01.

Pixel-based searcher with rotation

The PBSr model, shown in Figure 5, behaves identically to the PBS observer except that the target image is rotated before being compared to the match and distractor images by an estimate of the relative orientation of the pair of images being compared. The estimate of relative image orientation is based on a multiscale, multi-orientation representation of the images by determining, for each image, the subband with the highest power. The estimate of relative image orientation is the difference of the orientations of these highest power subbands.

Prior to the pyramid decomposition, each 256-pixel–diameter, windowed, and normalized target image, **T**, distractor image, **D**, and match image, **M**, were centered within a 256 × 256 matrix with the nonimage elements equal to the mean luminance. The resulting images were then decomposed into a 16 subband, multiscale, multi-orientation representation, $T_{\mathrm{S},\theta,}$ with four frequency bands/scales S and four orientations $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$. At each scale, size $m \times m$, the oriented subbands were point-wise multiplied by a $m \times m$ binary mask, with the central m-diameter disk (corresponding to the filtered image) equal to 1, and all other elements equal to 0.

The filters used for the decomposition provide a steerable basis (Freeman & Adelson, 1991; Portilla & Simoncelli, 2000). That is, weighted combinations of the four oriented subbands at each scale can be used to generate filter responses at any orientation at that scale. Using this technique, the PBSr calculates the power, $P_{S,\theta'}$, of each of the masked, oriented subbands at each scale of the decomposed target image, for $\theta' \in \{0, \pi/16, \ldots, (15/16)\pi\}$:

$$P_{S,\theta'} = \frac{1}{n} \sum_{i=1}^{N} |F_{S,\theta'}(x_i, y_i)|^2,$$

where N is the number of samples at a given spatial scale, and n equals the number of nonzero elements of the mask at that scale. F represents the complex-valued filter responses.

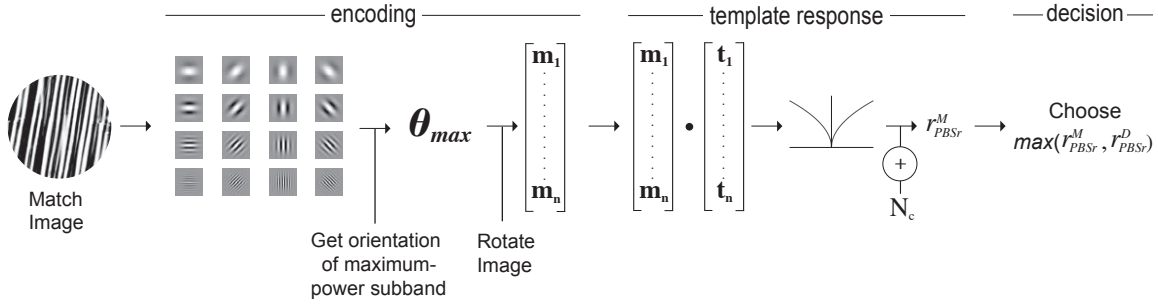

Figure 5. The PBSr model. **Encoding phase**: The image is subjected to a multiscale and multi-orientation decomposition from which it derives an estimate of the orientation of the image, θ_{max}, relative to the target (see text). It rotates the original (nondecomposed) image so that its maximum power subband is aligned with that of the target. The vectorized, rotated, intensity-based representation (of the original, nondecomposed image), **m**, constitutes the search template. **Template response**: An intensity-based representation, **t**, of the target image is formed, as in the PBS model, and correlated with **m**. The result is rectified, passed through a power function, and corrupted by additive noise, N_c, yielding noisy scalar response value, r^M_{PBSr}. **Decision phase**: The observer derives a similar response, r^D_{PBSr}, to the distractor and compares it to r^M. The stimulus yielding the maximum response is the observer's choice for the match stimulus.

It then finds the orientation that contains the maximum power for any value of S: $\theta^T_{max} = \arg\max_{\theta'}(\max_S P_{S,\theta'})$ (and analogously for the match and distractor images, yielding θ^M_{max} and θ^D_{max}). The original, nondecomposed, 256-pixel–diameter match image **M** is rotated by $\theta^T_{max} - \theta^M_{max}$ and, similarly, the distractor image **D**, by $\theta^T_{max} - \theta^D_{max}$.

These rotated images are then represented as vectors (**m** and **d**) and compared to the target image in exactly the same fashion as for PBS. Again, two parameters (now called ω_{PBSr} and σ_{PBSr}) were adjusted to fit the data in both conditions by maximum likelihood.

Statistic-based searcher

The SBS model, depicted in Figure 6, begins by decomposing the target, match, and distractor images into the same multiscale, multi-orientation representation as PBSr, and uses the same technique to estimate the relative orientation of the target versus match or distractor images (e.g., $\Delta\theta_M = \theta^T_{max} - \theta^M_{max}$). Note that, as with PBSr, each oriented subband of the decompositions is masked so that the nonimage elements equal 0. We use the relative orientation to "relabel" the subband decomposition of the match image, forming decomposition $\mathbf{M}'$ where $\mathbf{M}'_{S,\theta} = \boldsymbol{M}_{S,\theta + \Delta\theta_M}$ (and similarly for the distractor image resulting in rotated decomposition $\mathbf{D}'$).

The multiscale, multi-orientation decomposition is a set of complex-valued responses at each location in each subband. The real and imaginary parts of the responses correspond to the outputs of a quadrature pair of linear filters that have the same orientation and spatial frequency tuning, but differ in phase by 90° (Heeger, 1992; Simoncelli, Freeman, Adelson & Heeger, 1992). The local power (magnitude), P, and phase, ϕ, of the linear filter responses are computed:

$$P_{S,\theta}(x,y) = R_{S,\theta}(x,y)^2 + I_{S,\theta}(x,y)^2$$

$$\phi_{S,\theta}(x,y) = \tan^{-1}\left(\frac{I_{S,\theta}(x,y)}{R_{S,\theta}(x,y)}\right),$$

where R and I are the real and imaginary parts, respectively, of the responses within each subband (e.g., $\mathbf{T}_{S,\theta}$).

Next we derive two classes of statistics. The first class includes, (a) the central 7×7 neighborhood of the autocorrelation (computed using the standard frequency-domain calculation) of each subband $P_{S,\theta}$, (b) the cross-correlation of each oriented subband $P_{S,\theta}$ with every other subband within each scale, and (c) the cross-correlation of each $P_{S,\theta}$ with the corresponding subband at the next coarser scale, $P_{S+1,\theta}$. The second class is the cross-correlation of $\phi_{S,\theta}$ with the corresponding subband at the next coarser scale, $\phi_{S+1,\theta}$. These are two of the four classes of statistics in the Portilla and Simoncelli (2000) texture model. They have been shown to correspond to specific visual qualities of natural textures made evident by leaving each out of the texture synthesis algorithm (Balas, 2006). Broadly speaking, the magnitude correlations represent low frequency, repetitive content across the image. The phase correlations represent shading and depth-from-shading qualities of the image. The other two classes (excluded from the model) are (a) first-order pixel statistics (i.e., mean, variance, skew, kurtosis, and range of pixel intensities) and (b) the autocorrelation of the linear filter outputs at each spatial frequency independent of orientation. The pixel statistics represent the absolute luminance and contrast of the image. By normalizing the contrast of all stimulus images, we rendered this class of statistics essentially useless for the purpose of discriminating the images and thus it was excluded from the model. The linear filter autocorre-

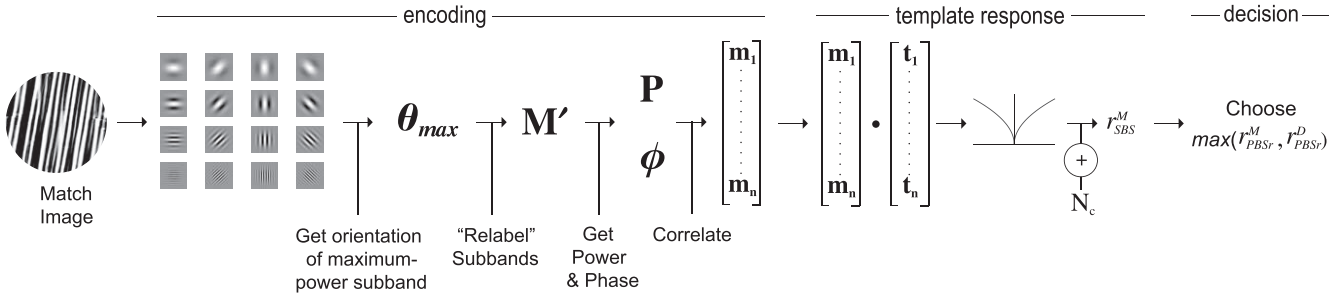

Figure 6. The SBS model. **Encoding phase**: The image is subjected to a multiscale and multi-orientation decomposition. The orientation of the highest power subband, θ_{max}, is derived as in the PBSr model. The subbands are "relabeled" and a new decomposed version of the image, **M**′, is constructed in which the maximum power subband of the image is aligned with that of the target. (Note that the actual image is not rotated.) The local energy, **P**, and phase, ϕ, are derived and correlated within and across subbands (see text) resulting in statistic vector, **m**. **Template response**: **m** is correlated with a similar representation, **t**, of the target image. The result is rectified, passed through a power function, and corrupted by additive noise, N_c, yielding scalar response value, r^M_{SBS}. **Decision phase:** The observer derives a similar response, r^D_{SBS}, to the distractor and compares it to r^M_{SBS}. The stimulus yielding the maximum response is the observer's choice for the match stimulus.

lations represent high-frequency repetitive content in the image. By simulating the performance of the SBS model in our task, given arbitrary values of the nonlinearity (ω_{SBS}) and noise (σ_{SBS}) (see below), we found that leaving this class of statistics out of the model did not qualitatively change its predictions and thus it is excluded for the sake of parsimony. Excluding the magnitude and phase correlations did substantially change the predictions of the model within the context of our task.

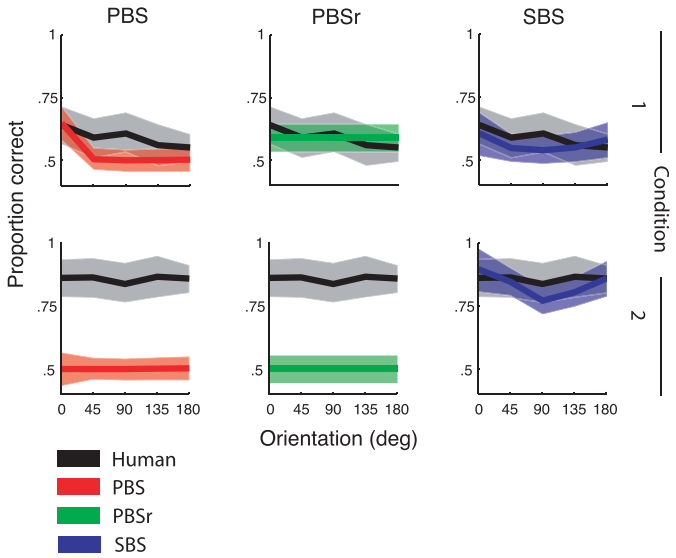

Figure 7. The predicted proportion of correct responses of the PBS, PBSr, and SBS models averaged across subjects and orientation classes plotted as a function of match and distractor orientation relative to the target. The across-subjects averages of the human data for each condition are plotted with each model prediction for comparison. Error regions show the *SE* derived by bootstrapping.

The resulting vector of correlation coefficients is used as the search template. The statistics computed for the two classes are combined into a single vector **t** for the statistical representation of the target. The rotated match and distractor representations (**M**′ and **D**′) are similarly processed to yield statistical representations **m** and **d**. These vectors are processed in the same manner as the image vectors in PBS and PBSr to model performance. Again, two parameters (now called ω_{SBS} and σ_{SBS}) were adjusted to fit the data in both conditions by maximum likelihood. Note that the "rotation" of the target representation does not rotate the image data within each subband, but only relabels the subbands so that appropriate orientation subbands of the target and match or distractor can be compared. This should affect only the autocorrelation statistics, which are not rotated to align with those of the match or distractor's autocorrelation.

In all three models, fullwave rectification of template correlation values was used, rather than halfwave rectification, because it provides a better fit to the human observers' performance (see Results). The power function applied to template correlations leads to model behavior analogous to adding signal-dependent noise to the correlation values.

Results

We calculated the proportion of correct responses for the human observers and compared them to the best-fit values of the PBS, PBSr, and SBS models. Figure 7 shows human and model proportion correct averaged across subjects and orientation classes (data and model fits for individual subjects and for all orientation classes are shown in Figure 1 of the

Subject	ω_{PBS}	σ_{PBS}	ω_{PBSr}	σ_{PBSr}	ω_{SBS}	σ_{SBS}
1	0.82	0.11	0.80	0.17	0.45	0.03
2	0.81	0.20	0.50	0.25	0.57	0.17
3	0.64	0.25	0.44	0.25	0.43	0.15
4	0.79	0.15	0.82	0.24	0.64	0.13

Table 1. The best-fit parameters for each subject and model.

Supplement). The best fitting values of ω and σ for each model are shown in Table 1.

Several patterns are apparent. First, the PBS model (Figure 7, column 1) predicts a high proportion correct in Condition 1 when a pixel-based match is called for and the match/distractor images are presented at the original orientation. It predicts chance performance when the match/distractor images are rotated relative to the target. It also predicts chance performance in Condition 2 when a statistic-based match is required. Second, the PBSr model (Figure 7, column 2), as expected, predicts no effect of match/distractor orientation on proportion correct. It predicts above-chance performance in Condition 1 and at-chance performance in Condition 2. Third, the SBS model (Figure 7, column 3) predicts relatively poor performance across orientations in Condition 1 and better performance in Condition 2. Human observers show a pattern qualitatively similar to that of the SBS model with poor performance in Condition 1 and better in Condition 2 across orientation classes and match/distractor orientations.

As we predicted, there is a slight improvement in performance in Condition 1 as the amount of oriented content in the images increases (i.e., going from orientation Class 1 to Class 3) for the SBS model (Figure 8). Human observers in Condition 1 show an

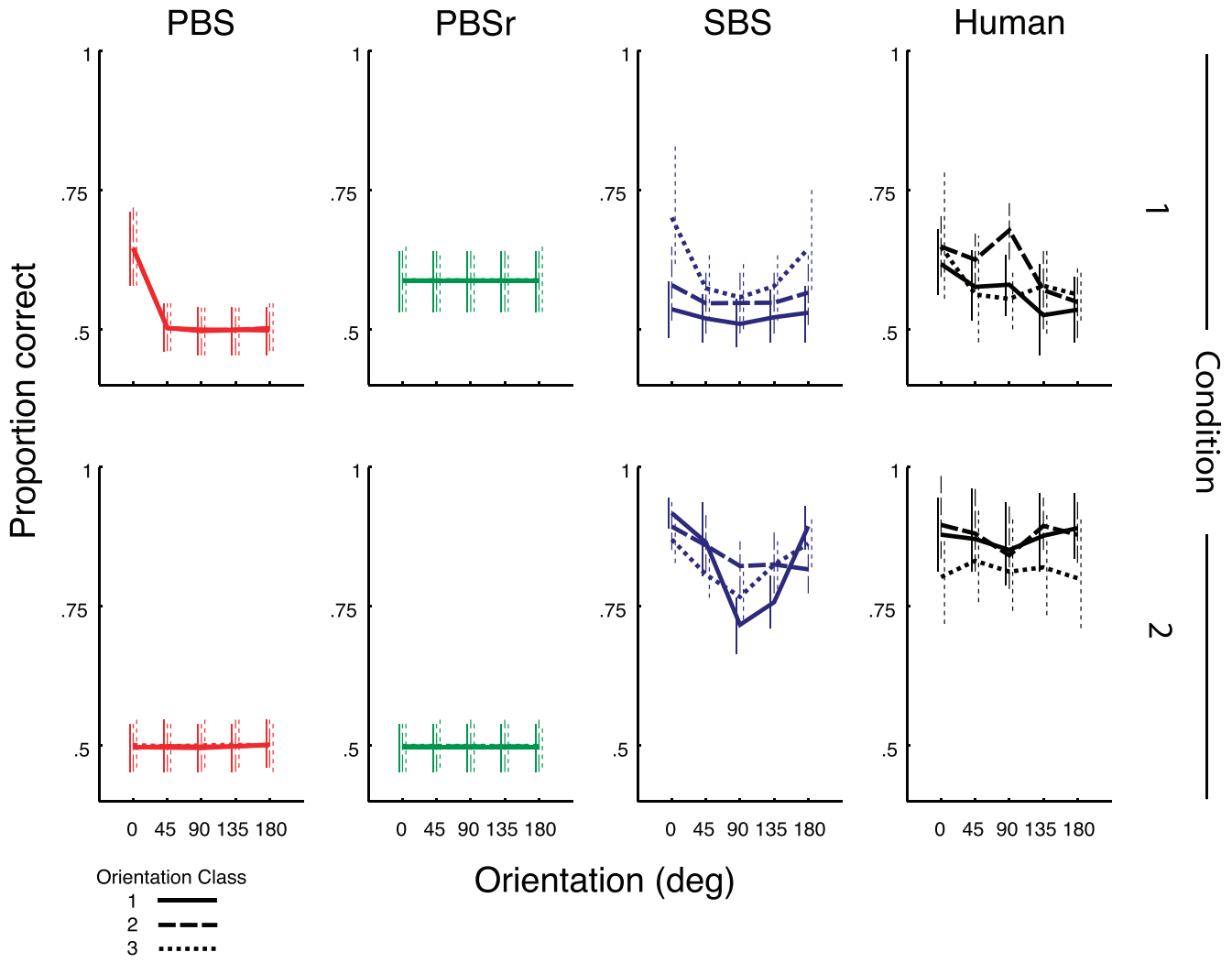

Figure 8. Proportion of correct responses for each orientation class in each condition averaged across subjects plotted as a function of match and distractor orientation relative to the target. As predicted, there is an increment in proportion correct for the SBS model in Condition 1 as the orientation content of the images increases (from Class 1 to Class 3). Error bars show the *SE* derived by bootstrapping.

	Condition 1						Condition 2					
	PBS		PBSr		SBS		PBS		PBSr		SBS	
Subject	χ^2	p	χ^2	p	χ^2	p	χ^2	p	χ^2	p	χ^2	p
1	52.9	**<0.001**	14.1	0.37	20.1	0.10	753.5	**<0.001**	756.9	**<0.001**	18.1	0.15
2	15.5	0.28	10.6	0.65	14.5	0.34	411	**<0.001**	411.9	**<0.001**	9.9	0.70
3	19.2	0.12	11.4	0.58	14.5	0.34	391.7	**<0.001**	392.7	**<0.001**	11.3	0.59
4	29.9	**0.005**	6.9	0.90	21.3	0.07	461.7	**<0.001**	464.2	**<0.001**	8.8	0.79

Table 2. The χ^2 statistics and p values for the goodness-of-fit tests for each model in each condition. p values for model predictions that differ significantly from the human data are shown in bold.

improvement going from Class 1 to 2, but no improvement for images in Class 3. The pattern of improved performance is not apparent for the pixel-based model observers in Condition 1 or for human and model observers in Condition 2.

We quantified the effect of condition, orientation class, and match/distractor orientation on the human observers' proportion correct using a repeated-measures ANOVA (2 conditions × 3 orientation classes × 5 match/distractor orientations). We find a significant main effect of condition on proportion correct, $F(1, 3) = 425$, $p < 0.001$, reflecting the observers' overall improvement in performance in Condition 2 over Condition 1. We find a significant main effect of orientation class, $F(2, 6) = 22.7$, $p = 0.02$, and a significant main effect of match/distractor orientation, $F(4, 12) = 9.74$, $p = 0.02$. There were no significant interactions.

Next, we compare human performance (across conditions, orientation classes, and match/distractor orientations) to that of each model by way of Pearson's χ^2 tests. Tests comparing each subject's performance to SBS model predictions did not achieve significance, $\chi^2(28) = 38, 24, 26, 30$, all ps > 0.05. Tests for all subjects showed a significant difference between human performance and PBS model predictions, $\chi^2(28) = 806, 426, 411, 491$, all ps < 0.001, and between human performance and the PBSr model predictions, $\chi^2(28) = 770, 422, 404, 471$, all ps < 0.001.

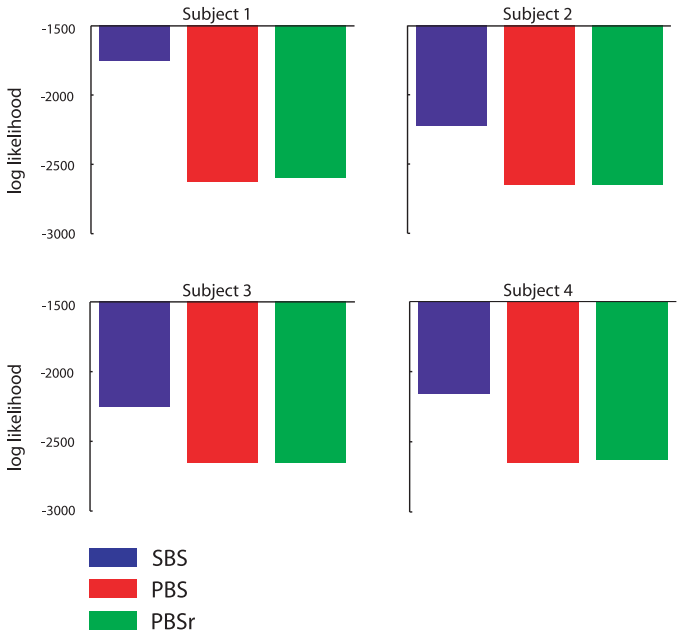

Figure 9. Log likelihood of the fits of the SBS, PBS, and PBSr models to each subject's data. Greater values indicate a better fit.

We hypothesize that a pixel-based model should do well at predicting human performance in Condition 1 when a pixel-based match is called for, and provide less accurate predictions in Condition 2 when a statistic-based match is required. Conversely, a statistic-based model should do poorly at predicting performance in Condition 1 and better in Condition 2. We quantified the comparison of model predictions and human performance in each condition using a χ^2 test. The results are shown in Table 2. The PBS model predictions in Condition 1 are not significantly different from human performance for Subjects 2 and 3. The PBSr model's predictions are not significantly different from all four subjects' performance in Condition 1. Both pixel-based models fail to predict performance in Condition 2, as expected. Contrary to our hypothesis, the SBS model's predictions are not significantly different from human performance in Condition 1, as well as Condition 2, for all subjects.

We compare the three models in their ability to predict human performance. Since each model has the same number of parameters, we can simply compare the likelihood of their respective fits to the data. Figure 9 shows the log-likelihood of each model for each subject (summed across stimulus sets, conditions, and orientations). The SBS model outperforms the PBS and PBSr models by a substantial margin for all subjects.

Evidence for a dual pixel- and statistic-based search strategy

The two tasks our observers performed had substantially different requirements. In Condition 1, a pixel-by-pixel comparison was required, whereas Con-

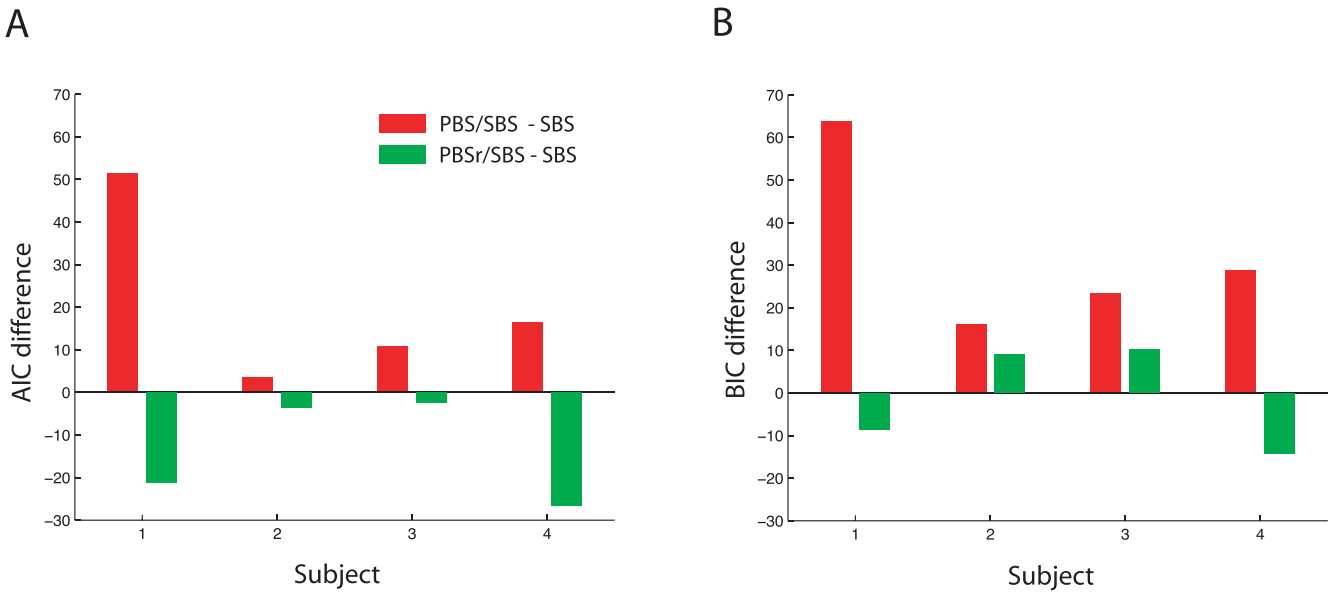

Figure 10. (A) Evidence for a dual strategy. Graphed here are differences in the AIC for dual models and AIC for the SBS model. The dual models use either the PBS or PBSr model in Condition 1 and the SBS model in Condition 2. Negative values in green indicate that the dual strategy PBSr/SBS provides a better fit to the data than the SBS model alone. (B) Differences in the BIC between the dual models and SBS model. Positive values in green indicate lack of support for a dual strategy (see text).

dition 2 required a qualitative comparison of textures, so that one might predict that subjects would use a pixel-based strategy in Condition 1 and a statistic-based strategy in Condition 2. Thus, we next ask whether the search strategy employed by human observers was task-dependent. Did they use a pixel-based template when the match was defined by pixels in Condition 1 and a statistic-based one when the match was defined by statistics in Condition 2? Or, did they use a statistic-based template for both?

We fit each of our pixel-based search models, PBS and PBSr, to each subject's data in Condition 1 (across all three orientation classes) by maximum likelihood, each with one value of ω and one value of σ. We fit the SBS model to the data in Condition 2 with one value of ω_{SBS} and one value of σ_{SBS}. We then compared the likelihood of these two four-parameter fits (PBS/SBS and PBSr/SBS) to that of the two-parameter SBS model fit to both conditions, using the Akaike information criterion (AIC; Akaike, 1974). (Note that a likelihood ratio test is not called for here since the models are not nested.) The AIC essentially compares the negative log likelihoods of each model fit but penalizes each model depending on the number of parameters in the fit. Smaller values of the AIC indicate a better fit. Figure 10A shows the difference in AIC values for the four- and two-parameter models for each subject. The negative values for the comparison PBSr/SBS to the SBS model are evidence for a dual strategy. However, we also calculated the less conservative Bayesian information criterion (BIC), which is similar to the AIC except that it considers the total number of trials being fit. We get mixed results using the BIC (Figure 10B), obtaining negative differences for the model comparisons for Subjects 1 and 4 (evidence for a dual strategy) and positive differences for Subjects 2 and 3 (indicating a lack of evidence for a dual strategy). Note also that the PBSr model predicts no effect of orientation on human observers' performance—a prediction that is contradicted by our finding above of a significant effect of orientation. The PBSr model can be rejected on the basis of that result alone. We conclude that the present evidence is not sufficient to indicate the use of a dual strategy.

Evidence for use of only a subregion of the stimulus images

In Condition 1 of our experiment, the observer's task was to locate the image that was identical, pixel-for-pixel, to the learned template. A pixel-based searcher need not use all pixels in the images to carry out the task. Textures, and natural images in general, are redundant (Attneave, 1954; Barlow, 1961; Hyvarinen, Hurri, & Hoyer, 2009). The observer need not consider a region of the image in which the pixel values change very little when a single pixel from the region may suffice to summarize its content. Pixel-based search in our task could proceed by comparing a small, minimally redundant region of the image to a similar region of the learned template. In fact, all subjects remarked that they used such a strategy, selectively attending to a small region of the target stimulus and

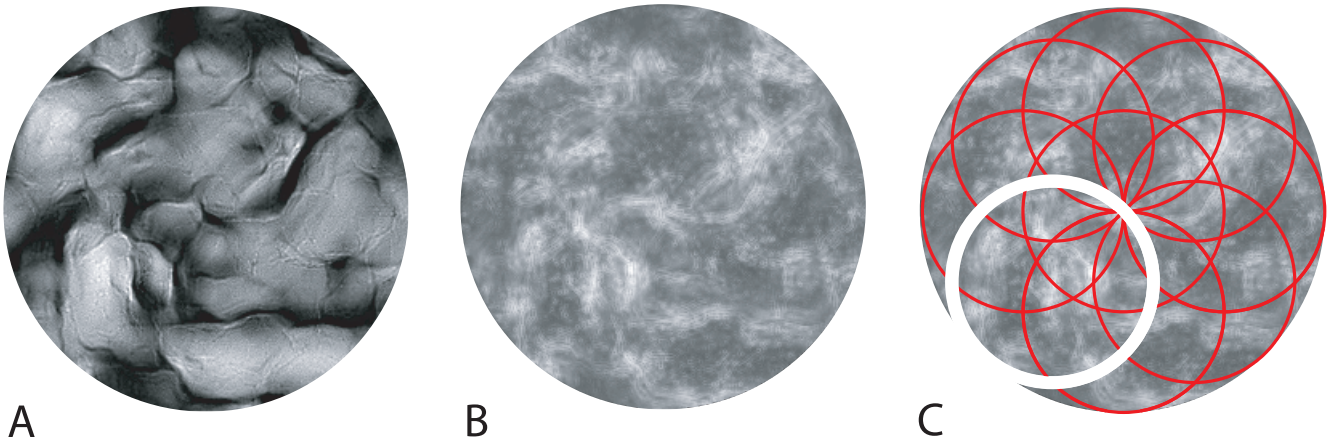

Figure 11. Example of the method for selecting the most salient region of target, match, and distractor images. (A) Sample 256-pixel–diameter image. (B) The image's saliency map (see text). (C) The nine 128-pixel–diameter subregions of the saliency map in red with the region containing the highest average salience outlined in white.

then searching the match/distractor stimuli for the same content.

Adopting such a strategy of using only the least redundant region of the images in Condition 1 may improve the performance of the pixel-based searcher. To investigate this, we simulated pixel-based searchers that carry out the task in Condition 1 using the most

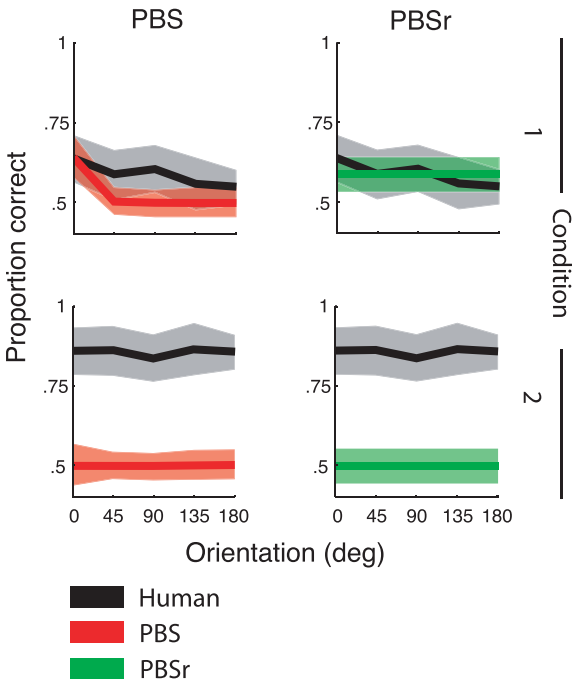

Figure 12. The predicted proportion of correct responses of the PBS and PBSr models averaged across subjects and orientation classes plotted as a function of match and distractor orientation relative to the target. The across-subjects averages of the human data for each condition are replotted with each model prediction for comparison. The PBS and PBSr model predictions assume use of only the most salient region of the images in Condition 1 and the entire image in Condition 2.

"salient" region of the target, match, and distractor stimuli. We assume here that the most salient, least redundant region of the image is the one in which the pixel values vary the most with respect to the mean pixel value of the image. The most salient region is thus that with the highest average local contrast. Our simulation employed an information-theoretic saliency algorithm (Bruce & Tsotsos, 2009; Itti, Koch, & Niebur, 1998) that effectively calculates the local contrast at each pixel in the image and then averages it across small regions. Each windowed and normalized target, match, and distractor image was decomposed into a 16-band multiscale, multi-orientation representation $\mathbf{T}_{\mathbf{S},\theta}$. A histogram of the filter responses was derived for each subband. We calculated the probability, $p(I)$, of sampling each intensity value, I, in a given subband from that subband's respective histogram. We then derived the self-information, $-ln(p[I])$, for each intensity value within each subband. The resulting self-information matrices (with size equal to that of the respective subband of $\mathbf{T}_{\mathbf{S},\theta}$) were upsampled and summed resulting in a 256-pixel–diameter saliency map for that image. We computed the average self-information within each of nine 128-pixel–diameter (3° diameter) regions of the saliency map, placed so that they tile the image (Figure 11). The 128-pixel–diameter circular region of the image with the highest average self-information was extracted and used to simulate the search tasks for each condition (in place of the original 256-pixel–diameter image).

Figure 12 shows model fits in each condition, averaged across subjects and stimulus classes. Average proportion correct for the human subjects is plotted for reference (data for individual subjects and orientation classes are shown in Figure 2 of the Supplement). The PBS model (red) and the PBSr model (green) are the predictions for pixel-based searchers that use the most salient region of the target, match, and distractor

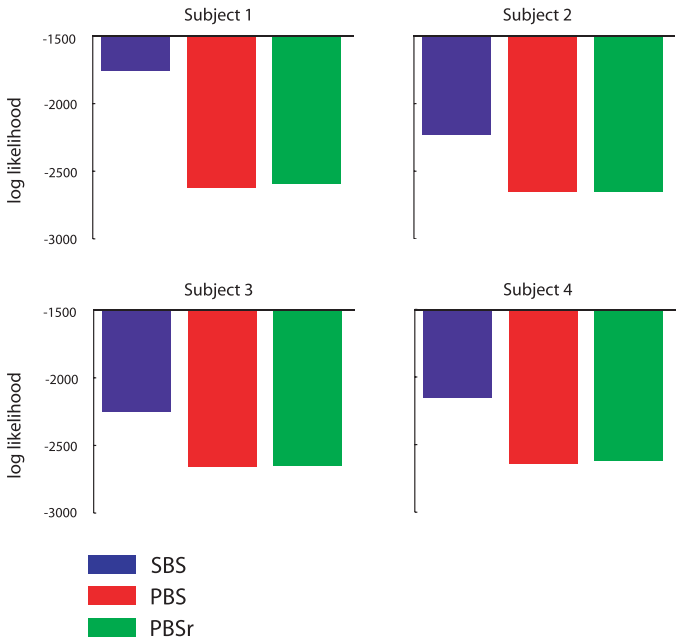

Figure 13. Log likelihood of the model fits in Figure 12. Greater values indicate a better fit.

images in Condition 1 and the entire image in Condition 2.

We compared the likelihood of each model's fit to the data. The log-likelihoods are shown in Figure 13. As without use of the salient region, the SBS model provides a better fit to the data.

We also used the AIC, as above, to compare the likelihoods of fits assuming the use of a dual strategy in which pixel-based search, using the most salient regions of the images, is used in Condition 1 and a statistic-based strategy, using the entire image, is used in Condition 2. The results are shown in Figure 14A. We again obtain negative values for the AIC differences comparing the PBSr/SBS to the SBS model and mixed results using the BIC (Figure 14B). As before, the PBSr/SBS model using the most salient region makes a qualitatively incorrect prediction in comparison to the human data in that it predicts no effect of match/distractor orientation on performance. The PBSr/SBS model can still be rejected on those grounds.

Finally, we compared each pixel-based model using the entire image to the same model using the most salient region in Condition 1. We obtain positive values for AIC difference comparing the PBS model using the salient regions to the PBS model using the entire image and comparing the PBS/SBS model using the salient region to the PBS/SBS model using the entire image. Use of the salient region in both cases provides a better fit to the data in Condition 1 when the match/distractor stimuli are not rotated relative the target. There is a slight decrement in performance assuming use of the salient region, particularly apparent in the PBS model, when the match/distractor stimuli are rotated relative to the target.

Evidence for full- versus halfwave rectification

As noted above, the use of fullwave rectification in our models provided a better fit to the human data. We verified this by generating predictions from each model using halfwave rectification, i.e., negative values of the

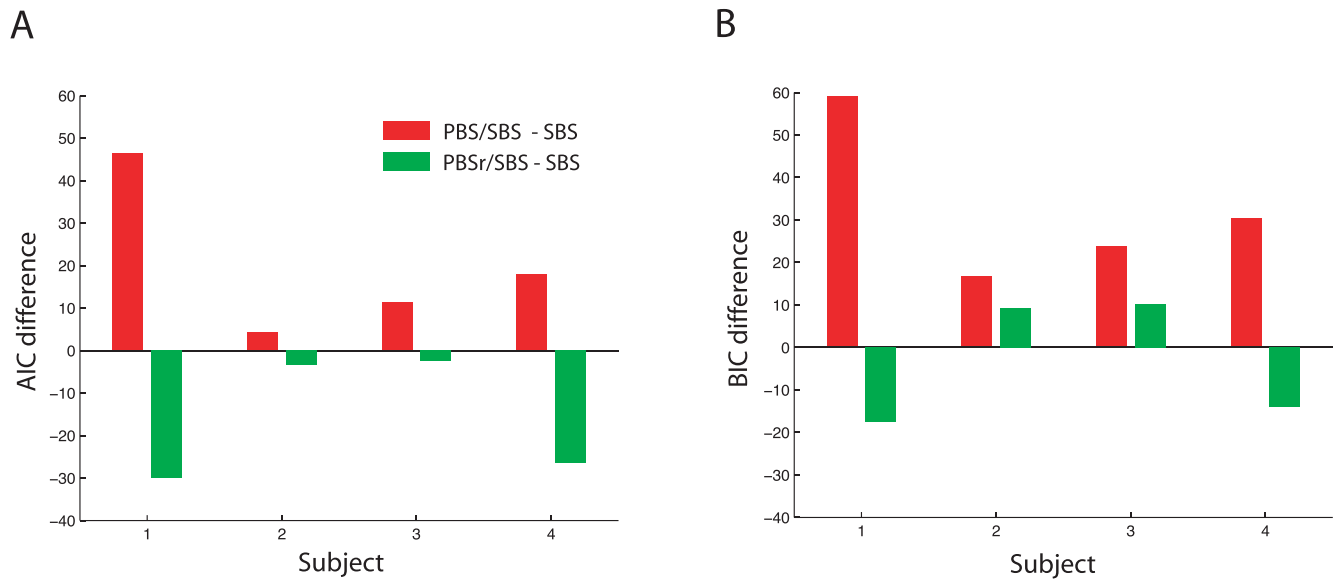

Figure 14. (A) The AIC values derived by fitting the PBS and PBSr models to Condition 1, using the most salient regions, and SBS model to Condition 2 minus the AIC for the fit of the SBS model (using the entire image) to both conditions. Negative values in green indicate evidence for a dual strategy. (B) Differences in the BIC values between the dual models and SBS model. Positive values in green indicate lack of support for a dual strategy (see text).

correlations, r, were set to 0. The log likelihoods of the resulting predictions for each model were compared to those utilizing fullwave rectification (as shown in Figure 9). Note that negative values of r did not occur in the PBS or PBSr model; thus halfwave versus fullwave rectification had no bearing on their predictions. We compared the log likelihoods of the SBS model predictions by taking the difference between that derived using fullwave and that derived using halfwave rectification. Positive values of the difference indicate a better fit utilizing fullwave rectification. The values for Subjects 1–4 are: −8.6, 2.1, 3.4, and 3.7, respectively. Thus, fullwave rectification provides a better overall fit to the data for three of the four subjects.

Discussion

We compared human performance to that of three models in a search task with complex images. Our statistic-based searcher uses linear filter responses to derive the local magnitude and phase. It uses correlations between neighboring magnitudes and phases as the search template. Our pixel-based searchers use the target image as the search template. The statistic-based searcher was shown to more accurately reflect human performance. It predicted the relative ease with which humans are able to identify a target based on its textural content, as compared to when the target is defined by a specific arrangement of pixels. It also predicted human observers' above-chance performance in locating a randomly rotated, pixel-defined target as well as the relative ease of locating a pixel-defined target when it is at or near the original orientation. Whereas the PBS model can predict human observers' performance at detecting the pixel-defined target at the original orientation, it predicted at-chance performance for targets that did not appear at the original orientation, and the PBSr model predicted equal performance across orientations.

Thus, a model of visual search that uses nonlinear filter responses and their correlations does well at predicting human performance in a search task in which the precise form of the search target is unknown to the observer, either because it appears at an unpredictable orientation or is specified only in terms of its general textural content. Our statistic-based model yielded good predictions of human performance regardless of whether the search target was defined as a pixel-based or statistic-based match suggesting that observers use a statistical template independent of stimulus characteristics or task demands. A statistic-based search model offers a viable, biologically plausible model of human search for complex visual targets.

Keywords: texture, visual search, image statistics

Acknowledgments

We would like to acknowledge both code and helpful comments from Jeremy Freeman and Eero Simoncelli. This work was supported by NIH grant EY08266.

Commercial relationships: none.
Corresponding author: Michael Landy.
Email: landy@nyu.edu.
Address: Department of Psychology, New York University, New York, NY, USA.

References

Abbey, C. K., & Eckstein, M. P. (2009). Frequency tuning of perceptual templates changes with noise magnitude. *Journal of the Optical Society of America*, *26*, 72–83.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*, 183–193.

Balas, B. J. (2006). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research*, *46*, 299–309.

Balas, B. J., Nakano, L., & Rosenholtz, R. (2009). A summary statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12):13, 1–18, http://www.journalofvision.org/content/9/12/13, doi:10.1167/9.12.13. [PubMed] [Article]

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge: MIT Press.

Barrett, H., Yao, J., Rolland, J., & Myers, K. (1993). Model observers for image quality assessment. *Proceedings of the National Academy of Sciences*, *90*, 9758–9765.

Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology*, *203*, 237–260.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*, 177–178.

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Brodatz, P. (1996). *Textures: A photographic album for artists and designers*. New York: Dover.

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi: 10.1167/9.3.5. [PubMed] [Article]

Burgess, A. E. (1981). Efficiency of human visual signal discrimination. *Science*, *214*, 93–94.

Burgess, A. E. (1985). Visual signal detection III: On Bayesian use of prior knowledge and cross correlation. *Journal of the Optical Society of America*, *2*(9), 1498–1507.

Burgess, A. E., & Colbourne, B. (1988). Visual signal detection IV: Observer inconsistency. *Journal of the Optical Society of America*, *5*(4), 617–627.

Burgess, A. E., & Ghandeharian, H. (1984). Visual signal detection II: Signal-location identification. *Journal of the Optical Society of America*, *1*(8), 906–910.

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology*, *197*, 551–566.

Cooper, L. A. (1976). Demonstration of a mental analog of an external rotation. *Perception & Psychophysics*, *19*, 296–302.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with Matlab and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*, 613–617.

DeValois, R. L., & DeValois, K. K. (1990). *Spatial vision*. New York: Oxford University Press.

Eckstein, M., Abbey, C., and Bochud, F. (2009). A practical guide to model observers for visual detection in synthetic and natural noisy images. In J. Beutel, H. Kundel, & R. Van Metter (Eds.), *Handbook of medical imaging, Volume 1: Physics & psychophysics* (pp. 593–628). Bellingham, WA: SPIE Press.

Foley, J. M., & Legge, G. E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research*, *21*, 1041–1053.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*, 1195–1201.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*, 891–906.

Geisler, W. S. (2010). Contributions of ideal observer theory to vision research. *Vision Research*, *51*, 771–781.

Graham, N. V. S. (1989). *Visual pattern analyzers*. New York: Oxford University Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Kreiger Publishing Co.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197.

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

Hyvarinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics: A probabilistic approach to early computational vision*. London: Springer.

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*, 1254–1259.

Landy, M. S., & Graham, N. (2003). Visual perception of texture. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1106–1118). Cambridge: MIT Press.

Landy, M. S., & Oruç, İ. (2002). Properties of second-order spatial frequency channels. *Vision Research*, *42*, 2311–2329.

Legge, G. E. (1981). A power law for contrast discrimination. *Vision Research*, *21*, 457–467.

Lu, Z.-L., & Dosher, B. A. (2008). Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review*, *115*, 44–82.

Marr, D. (1982). *Vision*. Cambridge: MIT Press.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*(3):4, 1–14, http://www.journalofvision.org/content/8/3/4, doi: 10.1167/8.3.4. [PubMed] [Article]

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387–391.

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision Research*, *49*, 1286–1294.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, *40*, 1227–1268.

Pelli, D. G. (1997). The Videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded

window of object recognition. *Nature Neuroscience*, *11*, 1129–1135.

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*, 49–71.

Rajashekar, U., Bovik, A., & Cormack, L. (2006). Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. *Journal of Vision*, *6*(4):7, 379–386, http://www.journalofvision.org/content/6/4/7, doi: 10.1167/6.4.7. [PubMed] [Article]

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology*, *3*(13), 1–15, doi:10.3389/fpsyg.2012.00013.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4):14, 1–17, http://www.journalofvision.org/content/12/4/14, doi:10.1167/12.4.14. [PubMed] [Article]

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *IIEE Transactions on Information Theory*, *38*, 587–607.

Tanaka, K. (2003). Inferotemporal response properties. In L. S. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1151–1164). Cambridge: MIT Press.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron*, *31*, 523–535.

Wang, G., Tanifuli, M., & Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research*, *32*, 33–46.