Detecting natural occlusion boundaries using local cues

Christopher DiMattina
Department of Psychology & Neuroscience Concentration, Grinnell College, Grinnell, IA, USA

Sean A. Fox
Natural Perception Laboratory
Department of Electrical Engineering & Computer Science
Case Western Reserve University, Cleveland, OH, USA

Michael S. Lewicki
Natural Perception Laboratory
Department of Electrical Engineering & Computer Science
Case Western Reserve University, Cleveland, OH, USA

Occlusion boundaries and junctions provide important cues for inferring three-dimensional scene organization from two-dimensional images. Although several investigators in machine vision have developed algorithms for detecting occlusions and other edges in natural images, relatively few psychophysics or neurophysiology studies have investigated what features are used by the visual system to detect natural occlusions. In this study, we addressed this question using a psychophysical experiment where subjects discriminated image patches containing occlusions from patches containing surfaces. Image patches were drawn from a novel occlusion database containing labeled occlusion boundaries and textured surfaces in a variety of natural scenes. Consistent with related previous work, we found that relatively large image patches were needed to attain reliable performance, suggesting that human subjects integrate complex information over a large spatial region to detect natural occlusions. By defining machine observers using a set of previously studied features measured from natural occlusions and surfaces, we demonstrate that simple features defined at the spatial scale of the image patch are insufficient to account for human performance in the task. To define machine observers using a more biologically plausible multiscale feature set, we trained standard linear and neural network classifiers on the rectified outputs of a Gabor filter bank applied to the image patches. We found that simple linear classifiers could not match human performance, while a neural network classifier combining filter information across location and spatial scale compared well. These results demonstrate the importance of combining a variety of cues defined at multiple spatial scales for detecting natural occlusions.

Keywords: occlusions, natural images, psychophysics, edge detection, neural networks

Citation: DiMattina, C., Fox, S. A., & Lewicki, M. S. (2012). Detecting natural occlusion boundaries using local cues. *Journal of Vision*, *12*(13):15, 1–21, http://www.journalofvision.org/content/12/13/15, doi:10.1167/12.13.15.

Introduction

One useful set of two-dimensional cues for inferring three-dimensional scene organization are the boundaries and junctions formed by the occlusions of distinct surfaces (Guzman, 1969; Nakayama, He, & Shimojo, 1995; Todd, 2004), as illustrated in Figure 1. In natural images, occlusion boundaries are defined by multiple cues, including local texture, color, and luminance differences, all of which are integrated perceptually (McGraw, Whitaker, Badcock, & Skillen, 2003; Rivest & Cavanagh, 1996). Although numerous machine vision studies have developed algorithms for detecting occlusions and junctions in natural images (Hoiem, Efros, & Hebert, 2011; Konishi, Yuille, Coughlin, & Zhu, 2003; Martin, Fowlkes, & Malik, 2004; Perona, 1992), relatively little work in visual psychophysics has directly studied natural occlusion detection (McDermott, 2004) or used natural occlusions in perceptual tasks (Fowlkes, Martin, & Malik, 2007).

In this study, we investigate the question of what locally available cues are used by human subjects to detect occlusion boundaries in natural scenes. We approach this problem by developing a novel database of natural occlusion boundaries taken from a set of uncompressed calibrated images used in previous research (Arsenault, Yoonessi, & Baker, 2011; Kingdom, Field, & Olmos, 2007; Olmos & Kingdom, 2004). We demonstrate that our database exhibits strong intersubject agreement in the locations of the labeled occlusions, particularly when compared with edges derived from image segmentation databases. In addition, we find that a variety of simple visual features

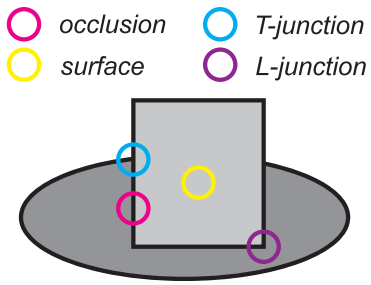

Figure 1. Occlusion of one surface by another in depth gives rise to image patches containing occlusion edges (magenta circle) and junctions (cyan and purple circles).

characterized in previous studies (Balboa & Grzywacz, 2000; Fine, MacLeod, & Boynton, 2003; Geisler, Perry, Super, & Gallogly, 2001; Ing, Wilson, & Geisler, 2010; Rajashekar, van der Linde, Bovik, & Cormack, 2007) can be used to distinguish occlusion and surface patches.

Using a simple two-alternative forced choice experiment, we test the ability of human subjects to discriminate local image regions containing either occlusions or single surfaces. In agreement with related work on junction detection (McDermott, 2004) and image classification (Torralba, 2009), we find that subjects require a fairly large image region (32×32 pixels) in order to make reliable judgments. Using a quadratic classifier analysis, we find that simple visual features defined on the scale of the whole image patch (i.e., luminance gradients) are insufficient to account for human performance, suggesting that human subjects integrate complex spatial information existing at multiple scales.

We investigated this possibility further by training standard linear and neural network classifiers on the rectified outputs of a set of Gabor filters applied to the occlusion and surface patches. We found that a linear classifier cannot fully account for subject performance since this classifier simply detects low spatial frequency luminance edges. However, a neural network having a moderate number of hidden units compared much better to human performance by combining information from filters across multiple locations and spatial scales. Our analysis demonstrates that only one layer of processing beyond the initial filtering and rectification is needed for reliably detecting natural occlusions. Interpreting the hidden units as implementing "second-order" filters, our results are consistent with previous demonstrations that filter-rectify-filter (FRF) models can detect edges defined by cues other than luminance differences (Baker & Mareschal, 2001; Bergen & Landy, 1991; Graham, 1991; Landy, 1991).

This study complements and extends previous work by quantitatively demonstrating the importance of integrating complex, multiscale spatial information when detecting natural occlusion edges (McDermott, 2004). Furthermore, this work provides the larger vision science community with a novel database of occlusion edges as well as a benchmark dataset of human performance on a standard edge-detection problem studied in machine vision (Konishi et al., 2003; Martin et al., 2004; Zhou & Mel, 2008). Finally, we discuss possible mechanisms for natural occlusion detection and suggest directions for future research.

Methods

Image databases

A set of 100 images containing few or no manmade objects were selected from a set of over 850 calibrated uncompressed color images from the McGill Calibrated Color Image Database (Olmos & Kingdom, 2004). Images were selected to have a clear figure-ground organization and plenty of discernible occlusion boundaries. Some representative images are shown in Figure 2 (left column). A group of five paid undergraduate research assistants were instructed to label all of the clearly discernible continuous occlusion boundaries using Adobe Photoshop layers. They were given the following instructions:

"Your task is to label the occlusion contours in the given set of 100 images. An occlusion contour is an edge or boundary where one object occludes or blocks the view of another object or region behind it. Label as many contours as you can, but you do not need to label contours that you are unsure of. Make each distinct contour a unique color to help with future analysis. Each contour must be continuous (i.e., one connected piece). Start by labeling contours on the largest and most prominent objects, and work your way down to smaller and less prominent objects. Do not label extremely small contours like blades of grass."

Students worked independently so their labeling reflected their independent judgment. The lead author (CD) hand-labeled all images as well, so there were six subjects total.

In order to compare the statistics of occlusions with image regions *not* containing occlusions, a database of "surface" image patches was selected from the same images by the same subjects. "Surfaces" in this context were broadly defined as uniform image regions which do not contain any occlusions, and subjects were not given any explicit guidelines beyond the constraint that the regions they select should be relatively uniform and could not contain any occlusions (which was prevented by our custom-authored software). No constraints were imposed with respect to lighting, curvature, material,

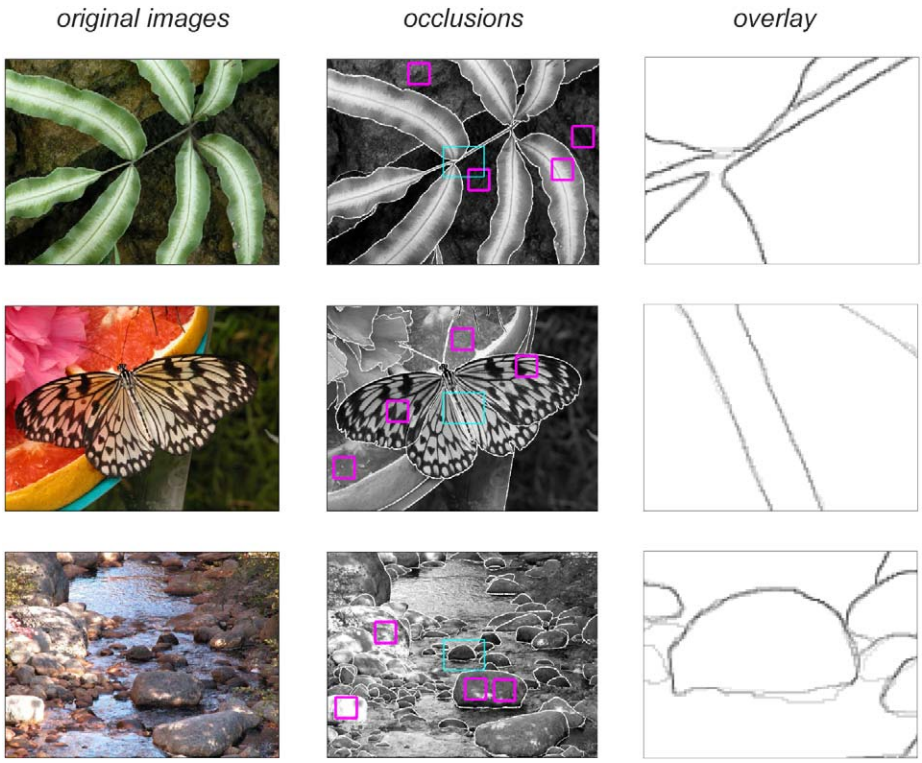

Figure 2. Representative images from the occlusion boundary database, together with subject occlusion labelings. *Left:* Original color images. *Middle:* Grayscale images with overlaid pixels labeled as occlusions (white lines) and examples of surface regions (magenta squares). *Right:* Overlaid plots of subject occlusion labelings taken from an 87 × 115 pixel central region of images (indicated by cyan squares in middle column). Darker pixels were labeled by more subjects, lighter pixels by fewer subjects.

shadows or luminance gradients. Therefore, some surface patches contained substantial luminance gradients, for instance patches of zebra skin (Figure 3). Each subject selected 10 surface regions (60 × 60) from each of the 100 images, and for our analyses we extracted image patches of various sizes (8 × 8, 16 × 16, 32 × 32) at random locations from these larger 60 × 60 regions. Example 32 × 32 surface patches are shown in Figure 2 (middle panel), and examples of both surface and occlusion patches are shown in Figure 3.

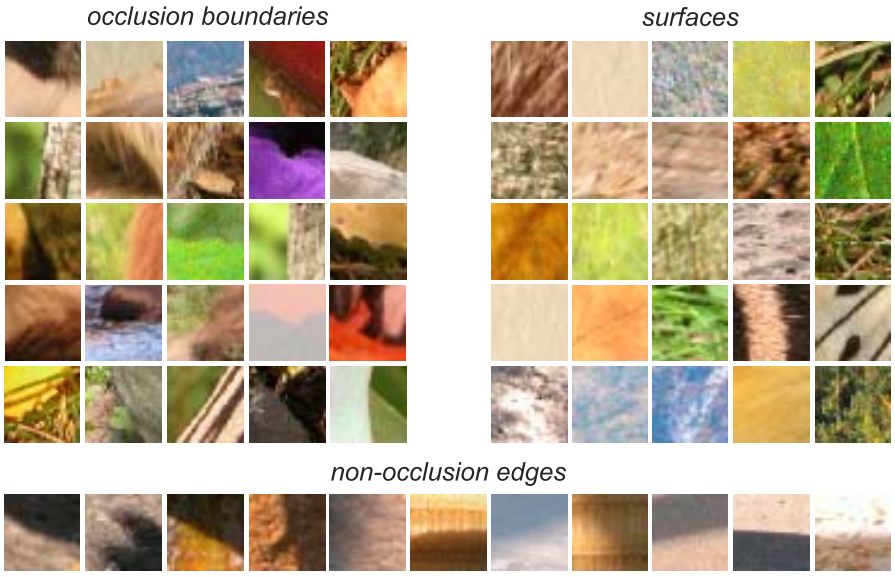

Figure 3. Examples of 32 × 32 occlusions (left), surfaces (right), and shadow edges not defined by occlusions (bottom).

Quantifying subject consistency

In order to quantify intersubject consistency in the locations of the pixels labeled as occlusions we applied *precision-recall* analysis commonly used in machine vision (Abdou & Pratt, 1979; Martin et al., 2004). In addition, we also developed a novel analysis method which we call the *most-conservative subject* (MCS) analysis which controls for the fact that disagreement between subjects in the location of labeled occlusions often arises simply because some subjects are more exhaustive in their labeling than others.

Precision-recall analysis is often used in machine vision studies of edge detection in order to quantify the trade-off between correctly detecting all edges (*recall*) and not incorrectly labeling non-edges as edges (*precision*). Mathematically, precision (P) and recall (R) are defined:

$$P = \frac{tp}{tp + fp}, \tag{1}$$

$$R = \frac{tp}{tp + fn}, \tag{2}$$

where tp, fp, tn, fn are the true and false positives and true and false negatives, respectively. Typically, these quantities are determined by comparing a machine generated test edgemap E to a ground-truth reference edgemap G derived from hand-annotated images (Martin et al., 2004). Since all of our edgemaps were human generated, we performed a "leave-one-out" analysis where we compared a test edgemap from each subject to a reference "ground truth" edgemap defined by combining edgemaps from all of the other remaining subjects. Since our goal for this analysis was simply to compare human performance on two different tasks (occlusion labeling and region labeling), we did not make use of sophisticated boundary-matching procedures (Goldberg & Kennedy, 1995) used in previous studies to optimize the comparisons between human data and machine performance (Martin et al., 2004). We quantify the overall agreement between the test and reference edgemaps using the weighed harmonic mean of P and R defined by:

$$F = \frac{PR}{\alpha P + (1 - \alpha)R}. \tag{3}$$

This quantity F is known as an *F-measure* and was originally developed to quantify the accuracy of document retrieval methods (Rijsbergen, 1979), but is also applied in machine vision (Abdou & Pratt, 1979; Martin et al., 2004). The parameter α determines the relative weight of precision and recall, and we used $\alpha = 0.5$ in our analysis.

In addition to the precision-recall analysis, we developed a novel method for quantifying intersubject consistency, which minimizes problems of intersubject disagreement arising from the fact that certain subjects are simply more exhaustive in labeling all possible occlusions than other subjects. We defined the *most conservative subject* (MCS) for a given image as the subject who had labeled the fewest pixels. Using the MCS labeling, we generate a binary image mask F, which is 1 for any pixel within R pixels of an occlusion labeled by the MCS, and 0 for all other pixels. Applying this mask to the labeling of each subject yields a "reduced" labeling, which is valid for inter-subject comparison since it only includes the most prominent occlusions labeled by all of the subjects. To calculate the comparison between two subjects, we randomly assigned one binary edgemap as the "reference" (I_{ref}) and the other binary edge-map as the "test" (I_{test}). We used the reference map to define a weighting function $f_\gamma(r)$, which was applied to all of the pixels in the test map that quantified how close each pixel in the test map was to a pixel in the reference map. Mathematically, our index is given by

$$\zeta = \frac{1}{N_t} \sum_{(x,y) \in I_{\text{test}}} f_\gamma\Big(r\Big((x,y), I_{\text{ref}}\Big)\Big) I_{\text{test}}(x,y), \tag{4}$$

where $r((x,y), I_{\text{ref}})$ is the distance between (x,y) and the closest pixel in I_{ref}, N_t is the number of pixels in the test set and the function $f_\gamma(r)$ is defined for $0 \leq \gamma < \infty$ by

$$f_\gamma(r) = \begin{cases} e^{-\gamma r} & \text{if } 0 \leq r \leq R \\ 0 & \text{if } r > R \end{cases}, \tag{5}$$

and for $\gamma = \infty$ by

$$f_\infty(r) = \begin{cases} 1 & \text{if } r = 0 \\ 0 & \text{if } r > 0 \end{cases}, \tag{6}$$

where R is the radius of the mask, which we set to $R = 10$ in our analysis. The parameter γ in our weighting function sets the sensitivity of f_γ to the distance between the reference labeling and the test labeling. Setting $\gamma = 0$ counts the fraction of pixels in the test edgemap, which lie inside the mask generated by the reference edgemap, and setting $\gamma = \infty$ measures the fraction of pixels in complete agreement between the two edgemaps.

Statistical measurements

Patch extraction and region labeling

Occlusion patches of varying sizes centered on an occlusion boundary were extracted automatically from the database by choosing random pixels labeled as occlusions by a single subject, cycling through the six subjects. Since we only wanted patches containing a single occlusion separating two regions (figure and ground) of roughly equal size, we only accepted a candidate patch when:

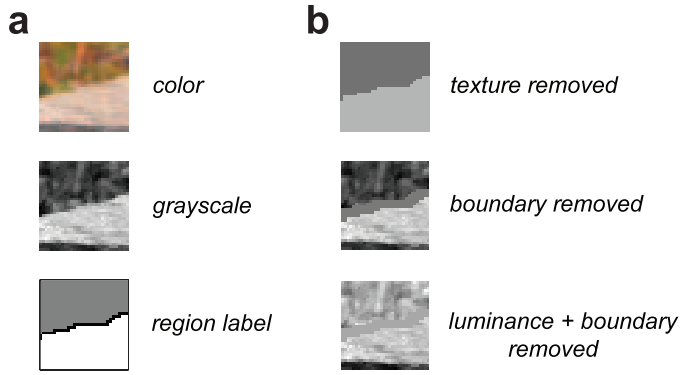

Figure 4. Illustration of stimuli used in the psychophysical experiments. (a) Original color and grayscale occlusion edge (top, middle) and its region label (bottom). The region label divides the patch into regions corresponding to the two surfaces (white, gray) as well as the boundary (black). (b) *Top:* Grayscale image patch with texture information removed. *Middle:* Occlusion patch with boundary removed. *Bottom:* Boundary and luminance information removed.

1. The composite occlusion edge map from all subjects consisted of a single, connected piece in the analysis window.
2. The occlusion contacted the sides of the window at two distinct points and divided the patch into two regions.
3. Each region comprised at least 35% of the pixels.

Each occlusion patch consisted of two regions of roughly equal size separated by a single boundary, with the central pixel ($w/2$, $w/2$) of a patch of size w always being an occlusion (Figure 4a). Note that this procedure actually yields a subset of all possible occlusions, since it excludes T-junctions or occlusions formed by highly convex boundaries. Since the selection of occlusion patches was automated, there were no constraints on the properties of the surfaces on either side of the occlusion with respect to factors like lighting, shadows, reflectance or material. In addition to the occlusion patches, we extracted surface patches at random locations from the set of 60 × 60 surface regions chosen by the subjects. We used 8 × 8, 16 × 16, and 32 × 32 patches for our analyses and psychophysical experiments.

Grayscale scalar measurements

To obtain grayscale images, we converted the raw images into gamma-corrected RGB images using software available online at: *http://tabby.vision.mcgill.ca* (Olmos & Kingdom, 2004). We then mapped the RGB color space to the NTSC color space, obtaining the grayscale luminance $I = 0.2989 \cdot R + 0.5870 \cdot G + 0.1140 \cdot B$ (Acharya & Ray, 2005). From these patches, we measured a variety of visual features, which can be used to distinguish occlusion from surface patches. Some of these features (for instance luminance difference) depended on there being a region labeling of the image patch, which separates it into regions corresponding to two different surfaces (Figure 4a). However, measuring these same features from surface patches is impossible since surface patches only contain a single surface. Therefore, in order to measure region labeling-dependent features from the uniform surface patches, we assigned to each surface patch a set of 25 "dummy" region labelings from our database (spanning all possible orientations). The measured value of the feature was then taken as the maximum value over all 25 dummy labelings, which is sensible since all of the visual features were on average larger for occlusions than uniform surface patches.

Given a grayscale image patch and a region labeling $R = \{R_1, R_2, B\}$ partitioning the patch into regions corresponding to the two surfaces (R_1,R_2) as well as the set of boundary (B) pixels (Figure 4a), we measured the following visual features taken from the computational vision literature:

G1. Luminance difference $\Delta\mu$:

$$\Delta\mu = |\mu_1 - \mu_2|, \tag{7}$$

where μ_1, μ_2 are the mean luminance in regions R_1, R_2, respectively.

G2. Contrast difference $\Delta\sigma$:

$$\Delta\sigma = |\sigma_1 - \sigma_2|, \tag{8}$$

where σ_1, σ_2 are the contrasts (standard deviation) in regions R_1, R_2, respectively. Features G1 and G2 were both measured in previous studies on surface segmentation (Fine et al., 2003; Ing et al., 2010).

G3. Boundary luminance gradient G_B:

$$G_B = \frac{||\nabla I||}{I}, \tag{9}$$

where $\nabla I(x, y) = [\partial I(x, y)/\partial x, \partial I(x, y)/\partial y]^{\mathrm{T}}$ is the gradient of the image patch evaluated at the central pixel, and I is the average intensity of the image patch (Balboa & Grzywacz, 2000).

G4. Oriented energy E_θ:

$$E_\theta = \sum_{i=1}^{N_\theta} (\mathbf{w}_i^{\mathrm{even}} \cdot \mathbf{x})^2 + (\mathbf{w}_i^{\mathrm{odd}} \cdot \mathbf{x})^2, \tag{10}$$

where $\mathbf{x}$ is the image patch in vector form and $\mathbf{w}_i^{\mathrm{even}}$,$\mathbf{w}_i^{\mathrm{odd}}$ are a quadrature-phase pair of Gabor filters of $N_\theta = 8$ evenly spaced orientations θ_i. For patch size w, the filters had means $w/2$ and standard derivations of $w/4$. Oriented energy has been used in several previous studies as a means of detecting edges (Geisler et al.,

2001; Lee & Choe, 2003; Sigman, Cecchi, Gilbert, & Magnasco, 2001).

G5. Global patch contrast ρ:

$$\rho = \mathrm{std}(I). \tag{11}$$

This quantity has been measured in previous studies, which quantify statistical differences between fixated image regions and random image regions for subjects' free-viewing natural images while their eyes are being tracked (Rajashekar et al., 2007; Reinagel & Zador, 1999). Several studies of this kind have suggested that subjects may be preferentially looking at edges (Baddeley & Tatler, 2006).

Note that features G3-G5 are measured globally from the entire patch, whereas features G1, G2 are differences between statistics measured from different regions of the patch.

Color scalar measurements

Images were converted from RGB to LMS color space using a MATLAB program, which accompanies the images in the McGill database (Olmos & Kingdom, 2004). We converted the logarithmically transformed LMS images into an $L\alpha\beta$ color space by performing principal components analysis (PCA) on the set of LMS pixel intensities (Fine et al., 2003; Ing et al., 2010). Projections onto the axes of the $L\alpha\beta$ basis represent a color pixel in terms of its overall luminance (L), blue-yellow opponency (α), and red-green opponency (β).

We measured two additional properties from the LMS color image patches represented in the $L\alpha\beta$ basis:

C1. Blue-Yellow difference $\Delta\alpha$:

$$\Delta\alpha = |\alpha_1 - \alpha_2|, \tag{12}$$

where α_1, α_2 are the mean values of the B-Y opponency component α in regions R_1, R_2, respectively.

C2. Red-Green difference $\Delta\beta$:

$$\Delta\beta = |\beta_1 - \beta_2|, \tag{13}$$

where β_1, β_2 are the mean values of the R-G opponency component β in regions R_1, R_2, respectively.

These color scalar statistics were motivated by previous work studying human perceptual discrimination of different surfaces (Fine et al., 2003; Ing et al., 2010).

Machine classifiers

Quadratic classifier analysis

In order to study how well various feature subsets measured from the image patches could predict human performance, we made use of a quadratic classifier analysis. The quadratic classifier is a natural choice for quantifying the discriminability of two categories defined by features having multivariate Gaussian distributions (Duda, Hart, & Stork, 2000), and has been used in previous work studying the perceptual discriminability of surfaces (Ing et al., 2010). Assume that we have two categories C_1,C_2 of stimuli from which we can measure n features $\mathbf{u} = (u_1,u_2, \ldots, u_n)^{\mathrm{T}}$, and features measured from each category are Gaussian distributed

$$p(\mathbf{u}|C_i) = N(\mathbf{u}|\mu_i,\Sigma_i), \tag{14}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the means and covariances of each category. Given a novel observation with feature vector $\mathbf{u}^*$, assuming that the two categories are equally likely a priori we evaluate the log-likelihood ratio

$$\mathrm{L}_{12}(\mathbf{u}^*) = \ln p(\mathbf{u}^*|C_1) - \ln p(\mathbf{u}^*|C_2), \tag{15}$$

choosing C_1 when $\mathrm{L}_{12} \geq 0$ and C_2 when $\mathrm{L}_{12} < 0$. In the case of Gaussian distributions for each category as in Equation 14, Equation 15 can be rewritten as

$$\mathrm{L}_{12}(\mathbf{u}^*) = \frac{1}{2}\left[|\Sigma_2| - |\Sigma_1|\right] + \frac{1}{2}\left[Q_1(\mathbf{u}^*) - Q_2(\mathbf{u}^*)\right], \tag{16}$$

where

$$Q_i(\mathbf{u}) = (\mathbf{u} - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{u} - \mu_i). \tag{17}$$

The task for applying this formalism is to define a set of n features to measure from the set of image patches, and then use these measurements to define the means and covariances of each category in a supervised manner. New image patches that are unlabeled can then be classified using this quadratic classifier.

In our analyses, category C_1 was occlusion patches and category C_2 was surface patches. We estimated the parameters of the classifiers for each category by taking the means and covariances of the statistics measured from a set of 2,000 image patches (training set) and we applied these classifiers to a different 400 patch subsets of 1,000 image patches, which were presented to the subjects (test set). This analysis was performed for multiple classifiers defined by different subsets of parameters, and for image patches of all sizes (8×8, 16×16, 32×32).

SVM classifier analysis

As an additional control, in addition to the quadratic classifier analysis we trained a Support Vector Machine (SVM) classifier (Cristianini & Shawe-Taylor, 2000) to discriminate occlusions and surfaces using our gray-scale visual feature set (G1–G5). The SVM classifier is a standard and well-studied method in machine learning, which achieves good classification results by learning the separating hyperplane that maximizes the margin

between two categories (Bishop, 2006). We implemented this analysis using the function *svmclassify.m* and *svmtrain.m* in the MATLAB Bioinformatics Toolbox.

Multiscale classifier analyses on Gabor filter outputs

One weakness of defining machine classifiers using our set of visual features is that these features are defined on the scale of the entire image patch. This is problematic because it is well known that occlusion edges exist at multiple scales, and that appropriate scale selection and integration across scale is an essential computation for accurate edge detection (Elder & Zucker, 1998; Marr & Hildreth, 1980). Furthermore, it is well known that the neural code at the earliest stages of cortical processing is reasonably well described by a bank of multi-scale filters resembling Gabor functions (Daugman, 1985; Pollen & Ronner, 1983), and that such a basis forms an efficient code for natural scenes (Olshausen & Field, 1996).

In order to define a multiscale feature set resembling the early visual code, we utilized the rectified outputs of a bank of filters learned using Independent Component Analysis (ICA). These filters closely resemble Gabor filters, but have the additional useful property of constituting a maximally independent set of feature dimensions for encoding natural images (Bell & Sejnowski, 1997). Example filters for 16 × 16 image patches are shown in Figure 5a. The outputs of our filter bank were used as inputs to two different standard classifiers: (a) a linear logistic regression classifier and (b) a three-layer neural network classifier (Bishop, 2006) having 4, 16, 64 hidden units. These classifiers were trained using standard gradient descent methods,

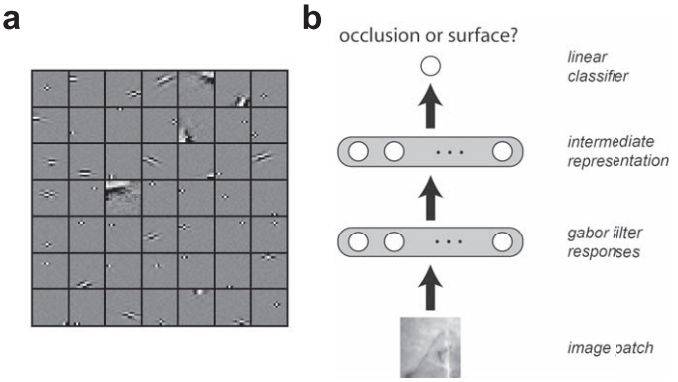

Figure 5. Illustration of the multiscale classifier analysis using the outputs of rectified Gabor filters. (a) Gabor functions learned using ICA form an efficient code for natural images, which maximize statistical independence of filter responses. (b) Grayscale image patches are decomposed by a bank of multiscale Gabor filters resembling simple cells, and their rectified outputs transformed by an intermediate layer of representation whose outputs are passed to a linear classifier.

and their performance was evaluated on a separate set of validation data not used for training. A schematic illustration of these classifiers is shown in Figure 5b.

Experimental paradigm

In the psychophysical experiments, image patches were displayed on a 24-inch Macintosh cinema display (Apple, Inc., Cupertino, CA). Since the RGB images were linearized to correct for camera gamma (Olmos & Kingdom, 2004), we set the display monitor to have gamma of approximately 1 using the Mac Calibration Assistant so that the displayed images would be as natural looking as possible. Subjects were seated in a dim room in front of the monitor and image patches were presented to the center of the screen, scaled to subtend 1.5 degrees of visual angle at the approximately 12-inch (30.5 cm) viewing distance. All stimuli subtended the same visual angle to eliminate confounds between the size of the image on the retina and its pixel dimensions. Previous studies have demonstrated that human performance on a similar task is dependent only on the number of pixels (McDermott, 2004), so by holding the retinal size constant the only variable is the number of pixels.

Our experimental paradigm is illustrated in Figure 6. Surrounding the image patch was a set of flanking crosshairs whose imaginary intersection defines the "center" of the image patch. In the standard task, the subject decides whether an occlusion boundary passes through the image patch center with a binary (1/0) keypress. Half of the patches were taken from the occlusion boundary database (where all occlusions pass through the patch center) and the other half of the patches were taken from the surface database. Therefore, guessing on the task would yield performance of 50 percent correct. The design of this experiment is similar to a previous study on detecting *T*-junctions in natural images (McDermott, 2004). To optimize performance, subjects were given as much time as they needed for each patch. Furthermore, since perceptual learning can improve performance (Ing et al., 2010), we provided positive and negative feedback.

We performed the experiment on two sets of naive subjects having no previous exposure to the images or knowledge of the scientific aims, as well as the lead author (S0), for a total of six subjects (three males, three females). One set of two naive subjects (S1, S2) was allowed to browse grayscale versions of the full-scale images (576 × 768 pixels) prior to the experiment. After 2–3 seconds of viewing, they were shown, superimposed on the images, the union of all subject labelings. This was meant to give the subjects an intuition for what is meant by an occlusion boundary in the context of the full-scale image, and was inspired by previous work on

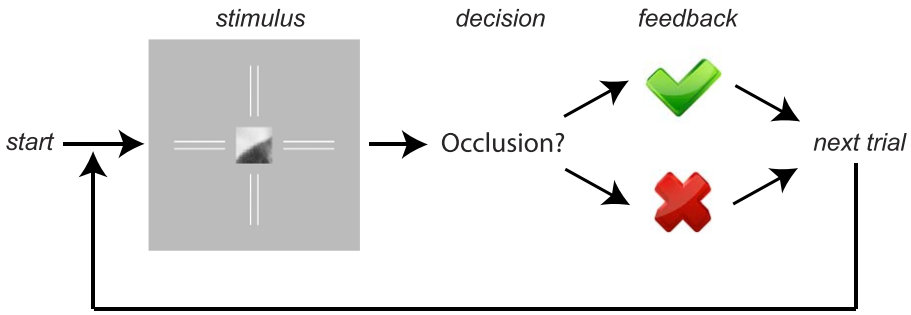

Figure 6. Schematic of the two-alternative forced choice experiment. A patch was presented to the subject, who decided whether an occlusion passes through the imaginary intersection of the crosshairs. After the decision, the subject was given positive or negative feedback.

surface segmentation where subjects were allowed to preview large full-scale images of foliage from which surface patches were drawn (Ing et al., 2010). A second set of three naive subjects (S3, S4, S5) were not shown the full-scale images beforehand, in order to control for the possibility that this pre-exposure may have helped to improve task performance.

We used both color and grayscale patches of sizes 8 × 8, 16 × 16, and 32 × 32. In addition to the raw image patches, a set of "texture-removed" image patches was created by averaging the pixels in each of the two regions, thus creating synthetic edges where the only cue was luminance or color contrast. For the grayscale patches, we also considered the effects of removing luminance difference cues. However, it is a much harder problem to create luminance-removed patches as was done in previous studies on surface segmentation (Ing et al., 2010), since simply subtracting the mean luminance in each region of an image patch containing an occlusion often yields a high spatial frequency boundary artifact, which provides a strong edge cue (Arsenault et al., 2011). Therefore, we circumvented this problem by setting a 3-pixel thick region around the boundary to the mean luminance of the entire patch, in effect covering up the boundary. Then we could remove luminance cues equalizing the mean luminance on each side without creating a boundary artifact since there was no boundary visible. We called this condition "boundary + luminance removed." One issue, however, with comparing these boundary + luminance removed patches to the normal patches is that now two cues are missing (the boundary and the luminance difference), so in order to better assess the combination of texture and luminance information we also created a "boundary removed" condition, which blocks the boundary but does not modify the mean luminance on each side. Illustrative examples of these stimuli are shown in Figure 4b.

All subjects were shown different sets of 400 image patches sampled randomly from a set of 1,000 patches in every condition, with the exception of two subjects in the luminance-only grayscale condition (S0, S2), who were shown the same set of patches in this condition only. Informed consent was obtained from all subjects, and all experimental procedures were approved beforehand by the Case Western Reserve University IRB (Protocol #20101216).

Tests of significance

In order to determine whether or not the performance of human subjects was significantly different in different conditions of the task, we utilized the standard binomial proportion test (Ott, 1993), which relies on a Gaussian approximation to the binomial distribution. This test is well justified in our case because of the large number of stimuli ($N = 400$) presented in each experimental condition. For proportion estimate $\hat{\pi}$ we compute the $1 - \alpha$ confidence interval as

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \quad (18)$$

We use a significance level of $\alpha = 0.05$ in all of our tests and calculations of confidence intervals.

In order to evaluate the performance of the machine classifiers we performed a Monte Carlo analysis where the classifiers were evaluated on 200 different sets of 400 image patches randomly chosen from a validation set of image patches. This validation set was distinct from the set of patches used to train the classifier, allowing us to study classifier generalization. We plot 95% confidence intervals around the mean performance of our classifiers.

Results

Case occlusion boundary (COB) database

We developed a novel database of occlusion edges and surface regions not containing any occlusions for

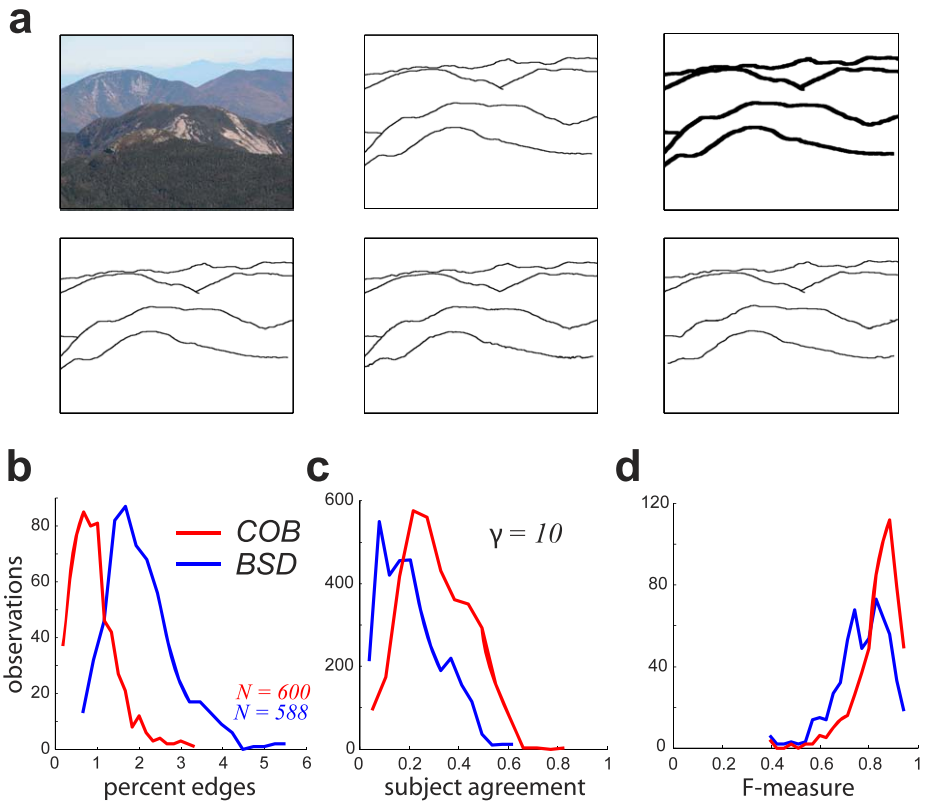

Figure 7. Labeling occlusions directly marks fewer pixels than inferring occlusions from image segmentations and yields greater agreement between subjects. (a) *Top:* An image from our database (left) together with the labeling (middle) by the most conservative subject (MCS). The right panel shows a binary mask of all pixels near the MCS labeling (10 pixel radius). *Bottom:* Product of MCS mask with labelings from three other subjects. (b) Histogram of the fraction of pixels labeled as edges in the COB (red) and BSD (blue) databases across all images and all subjects. (c) Histogram of the subject consistency index for edgemaps obtained from the COB (red) and BSD (blue) databases for $\gamma = 10$. (d) Precision-recall analysis also demonstrates better consistency (*F*-measure) for COB (red) than BSD (blue).

use in our perceptual experiments. Representative images from our database are illustrated in Figure 2 (left column). Note the clear figural objects and many occlusions in these images. In the middle column of Figure 2, we see grayscale versions of these images, with the set of all pixels labeled by any subject (logical OR) overlaid in white. The magenta squares show examples of surface regions labeled by the subjects. Finally, the right column shows an overlay plot of the occlusions marked by all subjects, with darker pixels being labeled by more subjects and the lightest gray pixels being labeled by only a single subject. Note the high degree of consistency between the labelings of the multiple subjects. Figure 3 shows some representative examples of occlusion (left) and surface (right) patches from our database. It is important to note that while occlusions may be a major source of edges in natural images, edges may arise from other cues like cast shadows or changes in material properties (Kersten, 2000). The bottom panel of Figure 3 shows examples of patches containing edges defined by shadows rather than by occlusions.

In other annotated edge databases like the Berkeley Segmentation Dataset (BSD) (Martin, Fowlkes, Tal, & Malik, 2001), occlusion edges were labeled indirectly by segmenting the image into regions and then denoting the boundaries of these regions to be edges. We observed that when occlusions are labeled directly instead of being inferred indirectly from region segmentations that a smaller number of pixels are labeled, as shown in Figure 7b, which plots the distribution of the percentage of edge pixels for all images and subjects in our database (red) and the BSD database (blue) for 98 BSD images segmented by six subjects. We find that averaged across all images and subjects that about 1% of the pixels were labeled as edges in our database, whereas about twice as many pixels were labeled as edges by computing edgemaps using the BSD segmentations (COB median = 0.0084, N = 600; BSD median = 0.0193, N = 588; $p < 10^{-121}$, Wilcox rank-sum).

We observed a higher level of intersubject consistency in the edgemaps obtained from the COB than those obtained from the BSD segmentations, which we

quantified using a novel analysis we developed, as well as a more standard precision-recall analysis (Abdou & Pratt, 1979; Rijsbergen, 1979, Methods, Image database). Figure 7a shows an image from the COB database (top left) together with the edgemap and derived mask for the *most conservative subject* (MCS), which is the subject who labeled the fewest pixels (top middle, right). When the MCS mask is multiplied by the edgemap of the other subjects, we see reasonably good agreement between all subjects (bottom row). In order to quantify this over all images, we computed our novel intersubject similarity index ζ defined in Equation 4 and in Figure 7c we see that on average that ζ is larger for our dataset than for the BSD, where here we plot the histogram for $\gamma = 10$ (COB median $= 0.2890$, $N = 3500$; BSD median $= 0.1882$, $N = 3430$; $p < 10^{-202}$, Wilcox rank-sum). Similar results were obtained over a wide range of values of γ (Supplementary Figure S1). In addition to our novel analysis, we also implemented a precision-recall analysis (Abdou & Pratt, 1979; Rijsbergen, 1979) using a "leave-one-out" procedure where we compared the edges labeled by one subject to a "ground truth" labeling defined by combining the edgemaps of the five other subjects (Methods, Image database). Agreement was quantified using the *F*-measure (Rijsbergen, 1979), which provides a weighted mean of precision (not labeling nonedges as edges) and recall (detecting all edges in the image). We observe in Figure 7d that for this analysis there is a significantly better agreement between edgemaps in the COB database than those obtained indirectly from the BSD segmentations using this analysis ($p < 10^{-28}$).

Visual features measured from occlusion and surface patches

We were interested in determining which kinds of locally available visual features could possibly be utilized by human subjects to distinguish image patches containing occlusions from those containing surfaces. Toward this end, we measured a variety of visual features taken from previous studies of natural image statistics (Balboa & Grzywacz, 2000; Field, 1987; Ing et al., 2010; Lee & Choe, 2003; Rajashekar et al., 2007; Reinagel & Zador, 1999) from both occlusion and surface patches. For patches containing only a single occlusion, which can be divided into two regions of roughly equal size, we obtained a region labeling, illustrated in Figure 4a (bottom). The region labeling consists of sets of pixels corresponding to the two surfaces divided by the boundary (white and gray regions), as well as the boundary (black line). Using this region labeling, we can measure properties of the image on either side of the boundary and compute differences in these properties between regions. By

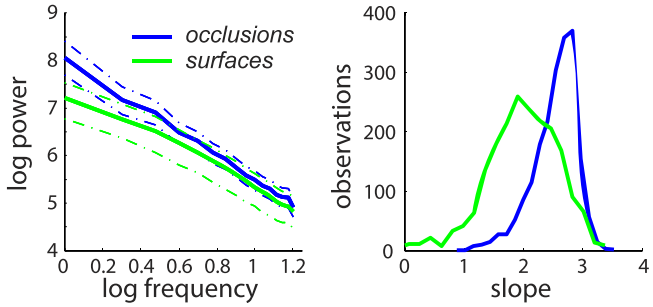

Figure 8. Power spectra of 32 × 32 patches. *Left:* Median power spectrum of occlusions (blue) and surfaces (green). Thin dashed lines show the 25th and 75th percentiles. *Right:* Power-spectrum slopes for occlusions (blue) and surfaces (green).

definition, surface patches are comprised of a single region and therefore it is unclear how we can measure quantities (like luminance differences), which depend on a region labeling. Therefore, in order to measure the same set of features from the surface patches, we assigned to each patch a set of 25 dummy region labelings (spanning all orientations) and for each dummy labeling performed the measurements. Since all features were, on average, larger for occlusions, for surfaces we took as the measured value of the feature the maximum over all 25 dummy labelings. A full description of measured features is given in Methods (Statistical measurements).

We compared the power spectra of 32 × 32 occlusion and texture patches (Figure 8, left). On average there is significantly less energy in low spatial frequencies for textures than for occlusions, which is intuitive since many occlusions contain a low spatial frequency luminance edge (Figure 3, left panel). Analysis of the exponent by fitting a line to the power spectra of individual images found a different distribution of exponents for the occlusions and the textures (Figure 8, right). For textures the median exponent is close to 2, consistent with previous observations (Field, 1987), whereas for occlusions the median exponent value is slightly higher (≈2.6). This is consistent with previous work analyzing the spectral content of different scene categories, which show that landscape scenes with a prominent horizon (like an ocean view) tend to have lower spatial frequency content (Oliva & Torralba, 2001).

Figure 9 shows that all of the grayscale features (G1-G5, Methods, Statistical measurements) measured from surfaces (green) and occlusions (blue) are well approximated by Gaussians when plotted in logarithmic coordinates. Supplementary Table 1 lists the means and standard deviations of each of these features. We see that on average these features are larger for occlusions than for textures, as one might expect intuitively since these features explicitly or implicitly measure differences or variability in an image patch,

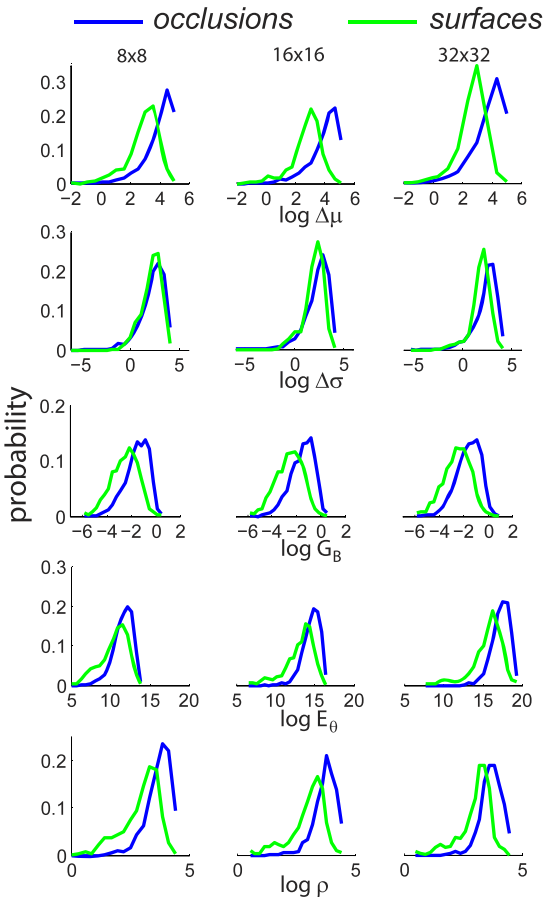

Figure 9. Univariate distributions of grayscale visual features G1-G5 (see Methods) for occlusions (blue) and textures (green). When plotted on a log scale, these distributions are well described by Gaussians and exhibit separation for occlusions and textures.

which will tend to be larger for occlusion patches. Analyzing the correlation structure of the grayscale features reveals that they are all significantly positively correlated (Supplementary Table 2). Although some of these correlations are unsurprising (like that between luminance difference log $\Delta\mu$ and boundary gradient log G_B), many pairs of these positively correlated features can be manipulated independently of each other in artificial images (for instance global contrast log ρ and contrast difference log $\Delta\sigma$). These results are consistent with previous work, which demonstrates that in natural images there are often strong conditional dependencies between supposedly independent visual feature dimensions (Fine et al., 2003; Karklin & Lewicki, 2003; Schwartz & Simoncelli, 2001; Zetzsche & Rohrbein, 2001). Supplementary Figure S2 plots all possible bivariate distributions of the logarithm of the grayscale features for 32 × 32 textures and surfaces.

In addition to grayscale features, we measured color features (C1–C2, Methods, Statistical measurements) as well by transforming the images into the $L\alpha\beta$ color space Fine et al. (2003); Ing et al. (2010). Using our region labelings (Figure 4a, bottom), we measured the color parameters log $\Delta\alpha$ and log $\Delta\beta$ from our occlusions. Supplementary Table 3 lists the means and standard deviations of each of these features, and we observe positive correlations ($r = 0.43$) between their values, similar to previous observations of positive correlations between these same features for two nearby patches taken from the same surface (Fine et al., 2003). This finding is interesting because occlusion boundaries by definition separate two or more different surfaces, and the color features we measure constitute a set of independent feature dimensions. From the two-dimensional scatterplots in Supplementary Figure S3, we see that there is separation between the multivariate distributions defined by these color parameters, suggesting that color contrast provides a potential cue for occlusion edge detection, much as it does for surface segmentation (Fine et al., 2003; Ing et al., 2010).

Human performance on occlusion boundary detection task

Effects of patch size, color and pre-exposure

In order to determine the local visual features used to discriminate occlusions from surfaces, we designed a simple two-alternative forced choice task, illustrated schematically in Figure 6. An image patch subtending roughly 1.5 degrees of visual angle (30.5 cm viewing distance) was presented to the subject, who had to decide with a binary (1/0) keypress whether an image patch contains an occlusion. Patches were chosen from a pre-extracted set of 1,000 occlusions and 1,000 surfaces, with equal probability of each type so guessing would yield 50% correct. Subjects were allowed to view each patch as long as they needed (1–2 seconds typical) and following previous work, auditory feedback was provided to optimize performance (Ing et al., 2010). The task was performed on six subjects having varying degrees of exposure to the image database (Methods, Experimental paradigm).

Figure 10a illustrates the effect of patch size on the task for grayscale image patches for all subjects. Thick lines indicate the mean subject performance and thin dashed lines denote 95% confidence intervals. We see from this that performance is significantly different for the different size image patches which we tested (8 × 8, 16 × 16, 32 × 32). For a subset of subjects (S0, S1, S2), we also tested color image patches, and the results for color image patches are shown in Figure 10b (red line) together with the grayscale data from these same subjects (black dashed line). We see from this plot that for all patch sizes tested performance is significantly better for color patches, which is sensible because color is a potentially informative cue for distinguishing different surfaces, as has been shown in previous work (Fine et al., 2003; Ing et al., 2010).

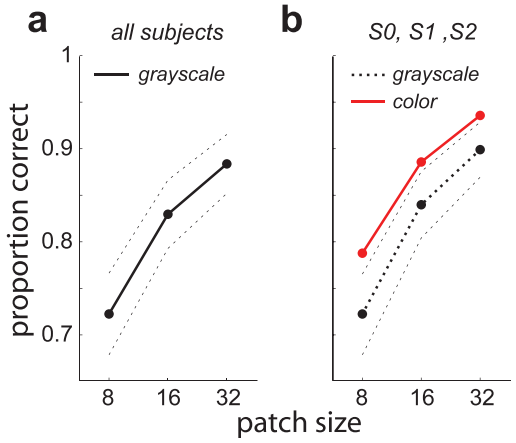

Figure 10. Performance of human subjects at the occlusion detection task for 8 × 8, 16 × 16, and 32 × 32 image patches. (a) Subject performance for grayscale image patches. Thin dashed lines denote 95% confidence intervals. Note how performance significantly improves with increasing patch size. (b) Subject performance for color image patches is significantly better than for grayscale at all patch sizes.

One concern with the interpretation of our results is that the brief pre-exposure (3 seconds) of subjects S1, S2 to the full-scale (576 × 768) images prior to the first task session (Methods, Experimental paradigm) may have unfairly improved their performance. In order to control for this possibility, we ran three additional subjects (S3, S4, S5) who did not have this pre-exposure to the full-scale images, and we found no significant difference in the performance of the two groups of subjects (Supplementary Figure S4a). We also found that the lead author (S0) was not significantly better at the task than the two subjects (S1, S2) who were only briefly pre-exposed to the images (Supplementary Figure S4b). Therefore, we conclude that pre-exposure to the full-scale images makes little if any difference in our task.

Effects of luminance, boundary and texture cues

In order to better understand what cues subjects are making use of in the task, we tested subjects on modified image patches with various cues removed (Methods, Experimental paradigm). In order to determine the importance of texture, we removed all texture cues from the patches by averaging all pixels in each region ("texture removed"), as illustrated in Figure 4b (top). We see from Figure 11a that subject performance in the task is substantially impaired without texture cues for the 16 × 16 and 32 × 32 patch sizes. Note how performance is roughly flat with increasing patch size when all cues but luminance are removed, suggesting that luminance gradients are a fairly "local" cue, which is very useful even at very small scales. Similar results were obtained for color patches (Figure 12), where the only cues available were luminance and color contrast.

Removing luminance cues while keeping texture cues intact is slightly more tricky, since simply equalizing the luminance in the two image regions can lead to problems of boundary artifact (a high spatial frequency edge along the boundary), as has been noted by others (Arsenault et al., 2011). In order to circumvent this problem, we removed the boundary artifact by covering a 3-pixel wide region including the boundary by setting all pixels in this strip to a single uniform value (Figure 4b, bottom). We then were able to equalize the luminance on each side of the patch without creating boundary artifact (luminance + boundary removed). Since the boundary pixels may could potentially contain long-range spatial correlation information (Hess & Field, 1999) useful for identifying occlusions,

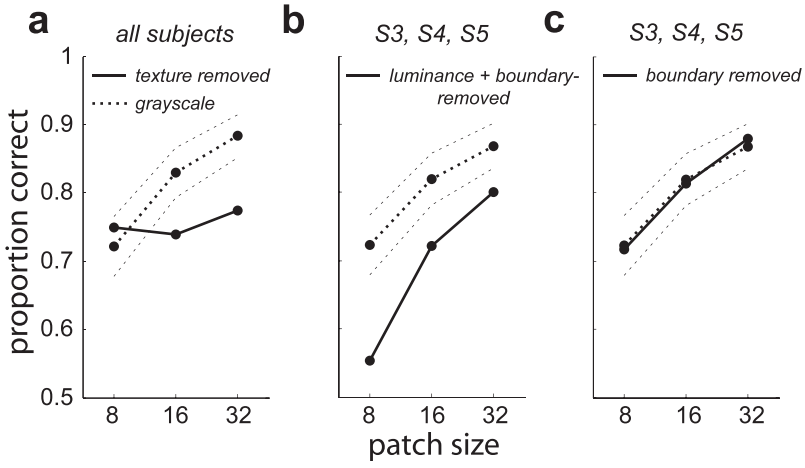

Figure 11. Subject performance for grayscale image patches with various cues removed. Dashed lines indicate the average performance for unaltered image patches, solid lines performance in the cue-removed case. (a) Removal of texture cues significantly impairs subject performance for larger image patches. (b) Removal of the boundary and luminance cues substantially impairs subject performance at all patch sizes. (c) Removal of the boundary cue alone without altering texture cues or luminance cues does not affect subject performance.

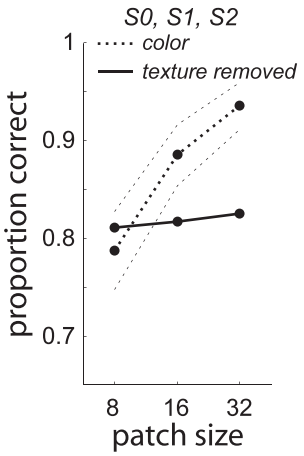

Figure 12. Subject performance for unaltered color image patches (dashed line) and with texture cues removed (solid line).

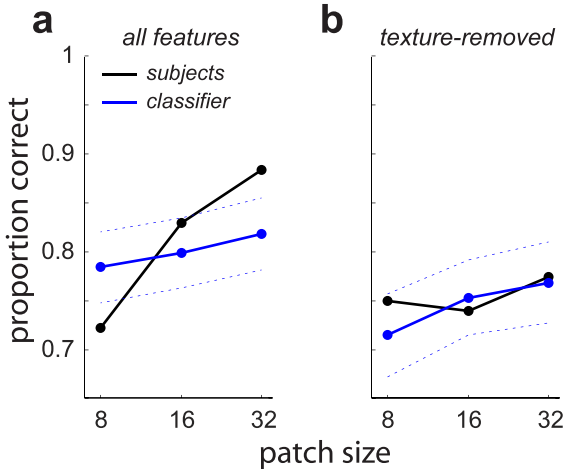

Figure 13. Comparison of subject and quadratic classifier performance for grayscale image patches. (a) Classifier defined using all grayscale features (thick blue line) does not accurately model human performance on the task with unmodified patches (black line). (b) Classifier defined using only luminance cues accurately models human performance in the texture-removed condition where luminance differences are the only available cue.

as an additional control we also considered performance on patches where we only removed the boundary (boundary removed).

We see from Figure 11c that simply removing the boundary pixels has no effect on subject performance, suggesting that subjects do not need to see the boundary in order to perform the task, so long as other cues like luminance and texture differences are present. However, we see from Figure 11b that there are profound effects of removing the luminance cues while keeping the texture cues intact. We see that subject performance is impaired at every patch size, although in contrast to removing texture cues, we see a substantial improvement in performance as patch size is increased. This suggests that texture cues are complementary to luminance cues in the sense that texture information becomes more useful for larger patches, while luminance is useful even for very small patches.

Quadratic classifier analysis

Comparison with human performance

Our analysis of image patches taken from the database suggests that relatively simple visual features measured in previous related studies can potentially be used to distinguish occlusions and surfaces (Figure 9). Therefore, it was of interest to determine whether machine classifiers defined using these simple visual features could potentially account for human performance on the task. Since the logarithmically transformed features are approximately Gaussian distributed (Figure 9 and Supplementary Figures S2, S3), following previous work (Ing et al., 2010) we defined a quadratic classifier on class-conditional distributions which we modeled as multivariate Gaussians (Methods, Machine classifiers). When data from both classes is exactly Gaussian this classifier yields Bayes-optimal performance (Duda et al., 2000). This classifier was then tested on validation sets of 400 image patches drawn from the same set shown to humans subjects (which were not used to train the classifier), using 200 Monte Carlo resamplings.

In Figure 13, we compare the performance of human subjects (black lines) on the grayscale version of the task with that of quadratic classifiers (blue lines). Figure 13a shows task performance with the full grayscale image patches. We see that the quadratic classifier defined on all grayscale parameters (G1-G5, Methods, Statistical measurements) does not capture the experimentally observed improvement in subject performance with increasing patch size. This suggests that in addition to the scalar grayscale cues used to define the classifier, subjects must be making use of more complex spatial cues like differences in texture (Bergen & Landy, 1991) or long-range spatial correlations (Hess & Field, 1999), which are not available at low resolutions but become available as the image patch gets larger. We see from Figure 13b that a quadratic classifier defined using only the luminance difference is in much better agreement with human performance when this is the only available cue for performing the task. Similar results were obtained for color image patches (Figure 14).

Ranking of feature dimensions

In order to understand which feature dimensions are most important for the quadratic classifier detecting

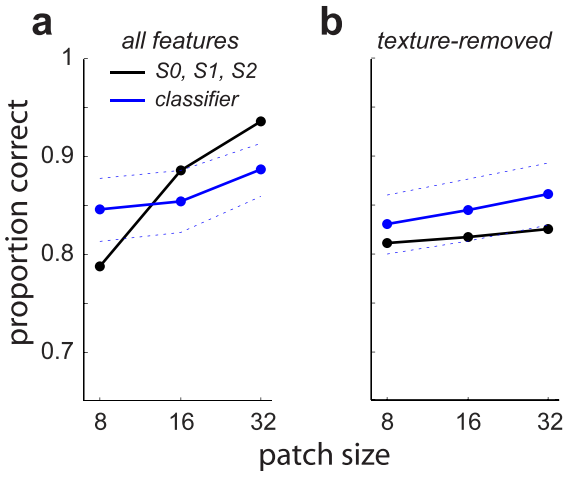

Figure 14. Comparison of subject and quadratic classifier performance for color image patches. (a) Classifier defined using all color features (thick blue line) does not accurately model human performance (black line) on the task with unmodified patches. (b) Classifier defined using only luminance and color contrast more accurately models human performance in the texture-removed condition.

occlusions, we computed the d' value for each of the feature dimensions, with the results of this analysis being summarized in Supplementary Table 4. We see from this table that the best single grayscale parameters for discriminating occlusion and surface patches are the luminance difference ($\Delta\mu$) and overall patch variance (ρ). We also see that contrast difference ($\Delta\sigma$) by itself is a fairly weak parameter for all patch sizes, while for the larger patch sizes (16 × 16, 32 × 32) color cues are very informative.

Our correlation analyses presented in Supplementary Table 2 demonstrated that many of these features were correlated with each other, so therefore it was of interest to determine which features contributed the most independent information to the classifier. We did this by adding feature dimensions to the classifier one by one in order of decreasing d' for 16 × 16 and 32 × 32 patches, and applying each classifier to the task for 200 Monte Carlo simulations. The results of this analysis performed on 16 × 16 and 32 × 32 image patches are shown in Figure 15. We see from Figure 15 that for grayscale classifiers containing ρ and $\Delta\mu$ that no performance increase is seen when one adds the features G_B (boundary gradient) and E_θ (oriented energy), most likely because these features are strongly correlated with ρ, $\Delta\mu$ (Supplementary Table 2). However, we do see a jump in performance when the contrast difference $\Delta\sigma$ is added, most likely because this feature dimension is only weakly correlated with ρ and $\Delta\mu$ (Supplementary Table 2). For our analysis of the color classifier, we see that there are substantial improvements in performance when the color parameters $\Delta\alpha$ and $\Delta\beta$ are added to the classifier, which makes sense because color cues are largely independent from luminance cues.

Comparison with SVM classifier

There are two logical possibilities for the poor match between human performance in our task and the performance of the quadratic classifier. The first possibility is that the feature sets used to define the classifier do not adequately describe the features actually used by human subjects to perform the task. A second possibility is that the feature set is adequate, but the quadratic classifier is somehow suboptimal. Although the quadratic classifier is Bayes optimal for Gaussian distributed class-conditional densities (Duda

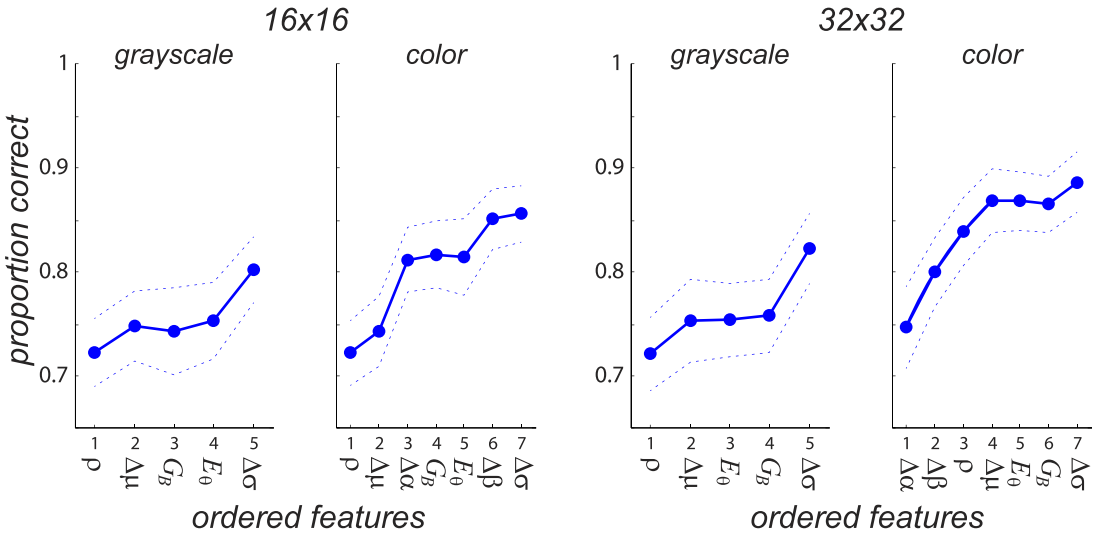

Figure 15. Quadratic classifier performance on 32 × 32 patches as a function of number of features. Feature sets of varying lengths are defined by ranking individual dimensions by their d' measures and adding them in order from highest to lowest.

et al., 2000), and our feature sets are reasonably well described by Gaussians (see Figure 9 and Supplementary Figures S2, S3), this description is far from exact. Therefore, it is possible that some other, more sophisticated classifier may yield superior performance to the quadratic classifier, and hence provide a better match with our perceptual data.

In order to control for this possibility, we compared the performance of the quadratic classifier to a Support Vector Machine (SVM) classifier (Bishop, 2006; Cristianini & Shawe-Taylor, 2000) on the task of discriminating grayscale occlusion patches from surfaces. We found that the SVM classifier performed worse than the quadratic classifier (Supplementary Figure S5). This failure of the SVM classifier strengthens the argument that the reason for the poor match of the quadratic classifier with human performance is the inadequacy of the measured feature set used to define the classifier, rather than the inadequacy our classifier method.

Multiscale classifier analysis

Gabor feature set and classifiers

One weakness of the set of simple visual features quantified in this study is that they are defined on the scale of the entire image patch, and therefore do not integrate information across scale. It is well known that occlusions in natural images can exist on multiple scales, and that accurate edge detection requires an appropriate choice of scale (Elder & Zucker, 1998; Marr & Hildreth, 1980). Furthermore, our feature set does not include any features which quantify texture differences, which is problematic because our psychophysical results suggests that texture is a very important cue, and previous work in machine vision has demonstrated that texture cues are highly informative for detecting edges and segmenting image regions (Bergen & Landy, 1991; Malik & Perona, 1990; Martin et al., 2004; Voorhees & Poggio, 1988).

In order to choose a set of biologically motivated, multiscale features to measure from the image patches, we obtained a Gabor filter bank by applying the Independent Components Analysis (Bell & Sejnowski, 1997) to natural image patches taken from our dataset (Figure 5a), and this filter bank was applied to the image patches. The independent component analysis (ICA) algorithm learns a set of linear features that are maximally independent, which is ideal for classifier analyses since correlated features add less information than independent features. The rectified outputs of this Gabor filter bank were then used as inputs to two supervised classifiers: (a) A linear binary logistic classifier, and (b) a neural network classifier (Bishop, 2006) having 4, 16, and 64 hidden units. The logistic classifier can be considered as a special case of the neural network classifier illustrated in Figure 5b where instead of learning the best hidden layer representation for solving the classification task, the hidden layer representation is always identical to the input representation and only the output weights are modified. After training, both classifiers were tested on a separate validation set not used for training to evaluate their performance.

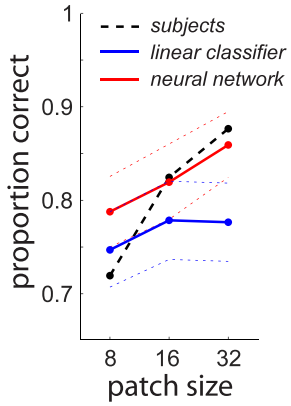

Figure 16. Performance of humans and classifiers defined on multiscale Gabor feature set. We see that a simple linear classifier (blue line) does not account for human performance (black dashed line) in the task, a neural network classifier having an additional hidden layer of processing (red line) compares well with human performance.

Classifier comparison

Figure 16 shows the performance of these two classifiers together with subject data for the grayscale occlusion detection task. We see that the linear classifier (blue line) does not accurately describe human performance (black line), while the neural network (red line) having 16 hidden units compares much more favorably to the human subjects. Similar results were obtained for a neural network with 64 hidden units, but worse performance was seen with four hidden units (not shown). We can see from Figure 16 that the linear classifier does not increase its performance with increasing patch size, and seems to perform similarly to the quadratic classifier defined using only luminance differences. This suggests that this classifier may simply be learning to detecting low spatial frequency edges at various orientations. Indeed, plotting the connection strengths of the output unit with each of the Gabor filters as in Figure 17 demonstrates that there are strong positive weights for filters at low spatial frequency located in the center of the image patch, thereby detecting luminance differences on the scale of the entire image patch.

In contrast, the neural network classifier learns to spatially pool information across spatial scale and location (Supplementary Figure S6), although the hidden unit "receptive fields" do not seem to resemble

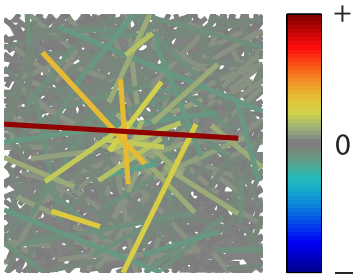

Figure 17. Schematic illustration of weights learned by the linear classifier. We see that the linear classifier learns strong positive weights (hot colors) with Gabor filters having low spatial frequencies (long lines) located in the center of the image patch.

the hypothetical texture-edge detectors illustrated in Figure 18, which could, in principle, encode edges defined by second-order cues like differences in orientation or spatial frequency. Interestingly, such units tuned for texture-edges have been learned in studies seeking to find a sparse code for patterns of correlation of Gabor filter activities in natural images (Cadieu & Olshausen, 2012; Karklin & Lewicki, 2003, 2009), and it is of interest for future work to examine whether a population of such units can be decoded in order to reliably detect occlusion edges. Nevertheless, this analysis provides a simple existence proof that a multiscale feature set coupled with a second layer of representation, which pools across this set of multiscale features may be adequate to explain natural occlusion edge detection, and a similar computational scheme has been proposed by various filter-rectify-filter models of texture edge detection (Baker & Mareschal, 2001; Landy & Graham, 2003). Interestingly, training neural networks directly on pixel representations did not yield models, which could match human performance (Supplementary Figure S7, demonstrating the importance of the multiscale Gabor decomposition as a preprocessing stage. Of course, we cannot suggest from our simple analysis the detailed form, which this integration takes or what the receptive fields of neurons performing this operation may be like, but this is certainly a question of great interest for future work in neurophysiology and computational modeling.

Discussion

Summary of results

In this study, we performed psychophysical experiments to better understand what locally available cues the human visual system utilizes when detecting natural occlusion edges. In agreement with previous work (McDermott, 2004), we found that fairly large regions (32×32 pixels) were required for subjects to attain reliable performance on our occlusion detection task (Figure 10). Studying the grayscale version of the task in detail, we found that texture and luminance cues were both utilized, although the relative importance of these cues varied with the patch size. In particular, we found that luminance cues were equally useful at all patch sizes tested (Figure 11a). This suggests that luminance differences are a local cue, being equally useful for small as well as large patches. In contrast, texture cues were only useful for larger image patches (Figure 11b). This is sensible because texture is defined by information occurring at a variety of scales (Portilla & Simoncelli, 2000; Zhu, Wu, & Mumford, 1997), and making patches larger makes more spatial scales visible.

The importance of textural cues was further underscored by comparing human performance on the task to the performance of a quadratic classifier defined using a set of simple visual features taken from previous work which did not include texture cues. Although many of the visual features provided independent sources of information (Figure 15) and permitted some discrimination of the stimulus categories, this feature set could not account for human performance in the task, overpredicting performance for small (8×8) patches, and underpredicting

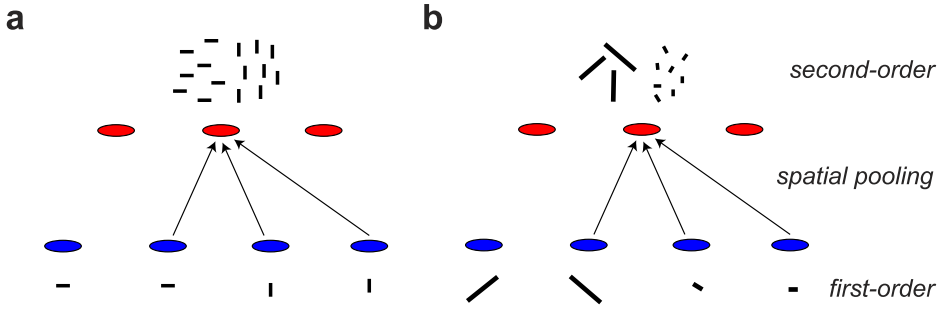

Figure 18. Hypothetical neural models for detecting occlusion edges defined by textural differences. (a) Hypothetical unit which detects texture edges defined by differences in orientation energy. (b) Hypothetical unit that detects texture edges defined by spatial frequency differences.

performance for large (32 × 32) patches (Figure 13). The overprediction of human performance for small image patches is particularly interesting since it demonstrates that subjects are not simply relying on luminance cues, but are integrating texture cues as well, which we found to be quite poor for small image patches (Figure 11b).

One limitation of the visual feature set used to define the quadratic classifier analysis is that the features are defined on the scale of the entire image patch. This is problematic since it is well-known that the earliest stages of human visual processing perform a multiscale analysis using receptive fields resembling Gabor filters (Daugman, 1985; Jones & Palmer, 1987; Pollen & Ronner, 1983), and that such a representation forms an efficient code for natural images (Field, 1987; Olshausen & Field, 1996). Furthermore, many computational methods for texture segmentation and edge detection use of a multiscale Gabor filter decomposition (Baker & Mareschal, 2001; Landy & Graham, 2003). Therefore, we defined a new set of classifiers taking as inputs the rectified responses of a set of Gabor filters applied to the image patches (Figure 5). We found that the multiscale feature set was adequate to explain human performance, but only provided that the classifier making use of this feature set was sufficiently powerful. In particular, a linear classifier was unable to account for human performance since this classifier only learned to detect luminance-edges at low spatial frequencies (Figure 17). However, a neural network classifier having an additional hidden layer of processing was able to learn to reliably detect occlusions nearly as well as human subjects (Figure 16). These results suggest a multiscale feature decomposition together with relatively simple computations performed in an additional layer of processing may be adequate to explain human performance in an important natural vision task.

Comparison with previous work

Relatively few studies in visual psychophysics have directly considered the problem of what cues are useful for detecting natural occlusion edges, with the closest study to our own focusing on the more specialized problem of detecting *T*-junctions in natural grayscale images (McDermott, 2004). This study found that rather large image regions were needed for subjects to attain reliable performance for junction detection, consistent with our results for occlusion detection. In this work, it was found that many *T*-junctions easily visible on the scale of the whole image were not detectable using local cues, suggesting that they may only be detectable by mechanisms making use of feedback of global scene information (Lee & Mumford, 2003). Our task is somewhat complementary to that studied by McDermott (2004) since we exclude *T*-junctions and only consider nonjunction occlusions. This gave us a much larger stimulus set and permitted quantitative comparisons of human task performance with a variety of models based on features measured from occlusions and non-occlusions. However, we arrive at the very similar conclusion for our task that a substantial amount of spatial context is needed for accurate occlusion detection (Figure 10), and that many occlusions cannot be reliably detected using local cues alone.

A related set of studies has considered the problem of judging when two image patches belong to the same or different surfaces, and comparing the performance of human subjects with the predictions of Bayesian models defined using sets of measured visual features similar to those used to define our quadratic classifiers. In contrast to one such surface segmentation study (Ing et al., 2010), we found texture cues to be very important for our task. One possible reason for the discrepancy between our results and those of Ing et al. (2010) is that their study only considered the problem of discriminating a very limited class of images, in particular the surfaces of different leaves. Since the leaves typically belonged to the same kind of plant, they were fairly smooth and quite similar in terms of visual texture. Therefore, for this particular choice of stimuli texture is not a particularly informative cue, since leaves of the same kind tend to have very similar textures. For our task however, where very different surfaces were juxtaposed to define occlusions, we found removal of texture cues to be highly detrimental to subject performance (Figure 11).

Previous work on edge localization reveals that luminance, texture and color cues are all combined to determine the location of edges (McGraw et al., 2003; Rivest & Cavanagh, 1996), and similarly we find that in our task these multiple sources of information are combined as well. However, in our task we did not consider quantitative models for how disparate cues may be combined optimally, as has been done in a variety of perceptual tasks (Knill & Saunders, 2003; Landy & Kojima, 2001). One challenge with studying cue combination in our task is that one of the major cues used by the subjects are texture differences, and unlike luminance or color, natural textures require very high-dimensional models to fully specify (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000; Zhu et al., 1997). The problem of optimal cue combination for natural occlusion detection has been mostly considered in the computer vision literature (Konishi et al., 2003; Zhou & Mel, 2008), but models of cue combination for occlusion detection have not been directly compared with human data in a controlled psychophysical experiment.

Our computational analysis demonstrates the necessity of multiscale image analysis for modeling human performance in our occlusion detection task. Furthermore, our work and that of others demonstrates the importance of textural cues for occlusion detection, as a wide variety of computational models, which explain human performance at detecting texture-defined edges utilize a multiscale Gabor decomposition (Landy & Graham, 2003). One broad class of computational models of texture segmentation and texture edge detection can be described "Filter-Rectify-Filter" (FRF) models, sometimes also referred to as "back-pocket" models, where the rectified outputs of a Gabor filter bank are then subject to a second stage of filtering, and the outputs of these filters are then processed further to obtain a decision about whether or not a texture edge is present (Baker & Mareschal, 2001; Bergen & Landy, 1991; Graham & Sutter, 1998; Landy, 1991; Landy & Graham, 2003). Our simple three-layer neural network model (Figure 5) constitutes a very simple model of the FRF type, with the hidden units performing the role of the second-order filters. The ability of our simple FRF model to solve this task as well as human subjects provides additional evidence in favor of the FRF computational framework.

Resources for future studies

Our database of hand-labeled occlusions (*http://case.edu*) avoids many limitations of previously collected datasets of hand-labeled edges. Perhaps the greatest improvement is that unlike many related datasets (Collins, Wright, & Greenway, 1999; Martin et al., 2001), we make use of uncompressed calibrated color images taken from a larger image set used in previous psychophysical research (Arsenault et al., 2011; Kingdom et al., 2007; Olmos & Kingdom, 2004). JPEG compression creates artificial statistical regularities, which complicates feature measurements (Zhou & Mel, 2008) and causes artifact when training statistical image models (Caywood, Willmore, & Tollhurst, 2004).

Another improvement of this database over previous work is that we explicitly label occlusion edges rather than inferring them indirectly from labels of image regions or segments. Although de facto finding region boundaries does end up mainly labeling occlusions, the notion of an image region is much vaguer than that of an occlusion. However, due to the specificity of our task, our database most likely neglects quite a few edges which are not occlusions, since many edges are defined by properties unrelated to ordinal depth like shadows, specularities and material properties (Kersten, 2000). Therefore, our database only represents a subset (albeit an important one) of possible natural edges.

Some databases only consider a restricted class of images like close-up foliage (Ing et al., 2010) or lack clear figural objects (Doi, Inui, Lee, Wachtler, & Sejnowski, 2003). Our database contains a wide variety of images at various spatial scales containing a large number of occlusions. We do not claim that our images are fully representative of natural images, but represent a modest effort to obtain a diverse sample, which is important since different categories of natural images have different statistics (Torralba & Oliva, 2003), and training statistical image models on different image databases often yields different linear filter properties (Ziegaus & Lang, 2003).

Acknowledgments

This work was supported by NIH grant 0705677 to M.S.L. We thank the five Case Western Reserve University undergraduate research assistants who helped to collect the databases. Parts of this work have been presented previously in poster form at the 2011 Society for Neuroscience Meeting.

Commercial relationships: none.
Corresponding author: Christopher DiMattina.
Email: dimattina@grinnell.edu.
Address: Department of Psychology & Neuroscience Concentration, Grinnell College, Grinnell, IA, USA.

References

Abdou, I., & Pratt, W. (1979). Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proceedings of the IEEE, 67*, 753–763.

Acharya, T., & Ray, A. K. (2005). *Image processing: Principles and applications*. Hoboken, NJ: Wiley-Interscience.

Arsenault, E., Yoonessi, A., & Baker, C. L. (2011). Higher-order texture statistics impair contrast boundary segmentation. *Journal of Vision, 11*(10): 14, 1–15, http://www.journalofvision.org/content/11/10/14, doi:10.1167/11.10.14. [PubMed] [Article]

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46*, 2824–2833.

Baker, C. L., & Mareschal, I. (2001). Processing of second-order stimuli in the visual cortex. *Progress in Brain Research, 134*, 1–21.

Balboa, R. M., & Grzywacz, N. M. (2000). Occlusions

and their relationship with the distribution of contrasts in natural images. *Vision Research, 40*, 2661–2669.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research, 37*, 3327–3338.

Bergen, J. R., & Landy, M. S. (1991). Computational modeling of visual texture segregation. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing*, pp. 253–271. Cambridge, MA: MIT Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Cadieu, C., & Olshausen, B. A. (2012). Learning intermediate-level representations of form and motion from natural movies. *Neural Computation, 24*, 827–866.

Caywood, M. S., Willmore, B., & Tolhurst, D. J. (2004). Independent components of color natural scenes resemble v1 neurons in their spatial and color tuning. *Journal of Neurophysiology, 91*, 2859–2873.

Collins, D., Wright, W. A., & Greenway, P. (1999). The Sowerby image database. In *Image processing and its applications (Seventh International Conference)*, (pp. 306–310). London: Institution of Electrical Engineers.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel based learning methods*. Cambridge, UK: Cambridge University Press.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A, 2*, 1160–1169.

Doi, E., Inui, T., Lee, T. W., Wachtler, T., & Sejnowski, T. J. (2003). Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Computation, 15*, 397–417.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: Wiley-Interscience.

Elder, J. H., & Zucker, S. W. (1998). Local scale control for edge detection and blur estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 699–716.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, 4*, 2379–2394.

Fine, I., MacLeod, D. I. A., & Boynton, G. M. (2003). Surface segmentation based on the luminance and color statistics of natural scenes. *Journal of the Optical Society of America A, 20*, 1283–1291.

Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure-ground cues are valid for natural images. *Journal of Vision, 7*(8):2, 1–9, http://www.journalofvision.org/content/7/8/2, doi:10.1167/7.8.2. [PubMed] [Article]

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41*, 711–724.

Goldberg, A. V., & Kennedy, R. (1995). An efficient cost scaling algorithm for the assignment problem. *Mathematical Programming*, *71*, 153–177.

Graham, N. (1991). Complex channels, early local nonlinearities and normalization in texture segregation. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing*, pp. 274–290. Cambridge, MA: MIT Press.

Graham, N., & Sutter, A. (1998). Spatial summation in simple (Fourier) and complex (non-Fourier) texture channels. *Vision Research, 38*, 231–257.

Guzman, A. (1969). Decomposition of a visual scene into three-dimensional bodies. In A. Griselli (Ed.), *Automatic interpretation and classification of images*, pp. 291–304. New York: Academic Press.

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. In *Proceedings of ACM SIGGRAPH*, 229–238.

Hess, R. F., & Field, D. J. (1999). Integration of contours: New insights. *Trends in Cognitive Sciences, 3,* 480–486.

Hoiem, D., Efros, A. A., & Hebert, M. (2011). Recovering occlusion boundaries from an image. *International Journal of Computer Vision, 91*(3): 328–346.

Ing, A. D., Wilson, J. A., & Geisler, W. S. (2010). Region grouping in natural foliage scenes: Image statistics and human performance. *Journal of Vision, 10*(4):10, 1–19, http://www.journalofvision.org/content/10/4/10, doi:10.1167/10.4.10. [PubMed] [Article]

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*, 1187–1211.

Karklin, Y., & Lewicki, M. S. (2003). Learning higher-

order structures in natural images. *Network, 14*, 483–499.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature, 457*, 83–86.

Kersten, D. (2000). High-level vision as statistical inference. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences*, pp. 353–363. Cambridge, MA: MIT Press.

Kingdom, F. A. A., Field, D. J., & Olmos, A. (2007). Does spatial invariance result from insensitivity to change? *Journal of Vision, 7*(14):11, 1–13, http://www.journalofvision.org/content/7/14/11, doi:10.1167/7.14.11. [PubMed] [Article]

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of slant? *Vision Research, 43*, 2539–2558.

Konishi, S., Yuille, A. L., Coughlin, J. M., & Zhu, S. C. (2003). Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*, 57–74.

Landy, M. S. (1991). Texture segregation and orientation gradient. *Vision Research, 31*, 679–691.

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, MA: MIT Press.

Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A, 18*, 2307–2320.

Lee, H.-C., & Choe, Y. (2003). Detecting salient contours using orientation energy distribution. In *International Joint Conference on Neural Networks* (pp. 206–211). IEEE Conference Publications.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*, 1434–1448.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A, 7*, 923–932.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London, Series B, 29*, 187–217.

Martin, D., Fowlkes, C. C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings of the 8th International Conference on Computer Vision*, *2*, 416–423.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*, 530–549.

McDermott, J. (2004). Psychophysics with junctions in real images. *Perception, 33*, 1101–1127.

McGraw, P. V., Whitaker, D., Badcock, D. R., & Skillen, J. (2003). Neither here nor there: Localizing conflicting visual attributes. *Journal of Vision, 3*(4):2, 265–273, http://www.journalofvision.org/content/3/4/2, doi:10.1167/3.4.2. [PubMed] [Article]

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In D. N. Osherson, S. M. Kosslyn, & L. R. Gleitman (Eds.), *An invitation to cognitive science: Visual cognition*, pp. 1–70. Cambridge, MA: MIT Press.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*, 145–175.

Olmos, A., & Kingdom, F. A. A. (2004). McGill calibrated colour image database. Internet site: http://tabby.vision.mcgill.ca (Accessed November 20, 2012).

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

Ott, R. L. (1993). *An introduction to statistical methods and data analysis* (4th ed.). Belmont, CA: Wadsworth, Inc.

Perona, P. (1992). Steerable-scalable kernels for edge detection and junction analysis. *Image and Vision Computing, 10*, 663–672.

Pollen, D. A., & Ronner, S. F. (1983). Visual cortical neurons as localized spatial frequency filters. *IEEE Transactions on Systems, Man and Cybernetics, 13*, 907–916.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*, 4971.

Rajashekar, U., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2007). Foveated analysis of image features at fixations. *Vision Research, 47*, 3160–3172.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10*, 341–350.

Rijsbergen, C. V. (1979). *Information retrieval*. London: Butterworths.

Rivest, J., & Cavanagh, P. (1996). Localizing contours defined by more than one attribute. *Vision Research, 36*, 53–66.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*, 819–825.

Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and gestalt rules. *Proceedings of the National Academy of Sciences, 98*, 1935–1940.

Todd, J. T. (2004). The visual perception of 3D shape. *Trends in Cognitive Science, 8*, 115–121.

Torralba, A. (2009). How many pixels make an image? *Visual Neuroscience, 26*, 123–131.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems, 14*, 391–412.

Voorhees, H., & Poggio, T. (1988). Computing texture boundaries from images. *Nature, 333*, 364–367.

Zetzsche, C., & Rohrbein, F. (2001). Nonlinear and extra-classical receptive field properties and the statistics of natural scenes. *Network, 12*, 331–350.

Zhou, C., & Mel, B. W. (2008). Cue combination and color edge detection in natural scenes. *Journal of Vision, 8*(4):4, 1–25, http://www.journalofvision.org/content/8/4/4, doi:10.1167/8.4.4. [PubMed] [Article]

Zhu, S. C., Wu, Y. N., & Mumford, D. (1997). Minimax entropy principle and its application to texture modeling. *Neural Computation, 9*, 1627–1660.

Ziegaus, C., & Lang, E. W. (2003). Independent component analysis of natural and urban image ensembles. *Neural Information Processing Letters, 1*, 89–95.