Does spatial invariance result from insensitivity to change?

Frederick A. A. Kingdom McGill Vision Research, Montréal, Québec, Canada

David J. Field Psychology Department, Cornell University, Ithaca, NY, USA

Adriana Olmos McGill Vision Research, Montréal, Québec, Canada

One of the fundamental unanswered questions in visual science regards how the visual system attains a high degree of invariance (e.g., position invariance, size invariance, etc.) while maintaining high selectivity. Although a variety of theories have been proposed, most are distinguished by the degree to which information is maintained or discarded. To test whether information is maintained or discarded, we have compared the ability of the human visual system to detect a variety of wide-field changes to natural images. The changes range from simple affine transforms and intensity changes common to our visual experience to random changes as represented by the addition of white noise. When sensitivity was measured in terms of the Euclidean distance (L_2 norm) between image pairs, we found that observers were an order of magnitude less sensitive to the geometric transformations than to added noise. A control experiment ruled out that the sensitivity difference was caused by the statistical properties of the image difference created by this transformation. We argue that the remarkable difference in sensitivity relates to the processes used by the visual system to build invariant relationships and leads to the unusual result that observers are least sensitive to those transformations most commonly experienced in the natural world.

Keywords: invariant transformations, natural scenes, geometric transformations, photometric transformations

Citation: 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://journalofvision.org/7/14/11/, doi:10.1167/7.14.11.

Introduction

Successful visual recognition in a dynamic environment requires that the processes of recognition show some degree of visual invariance. From drosophila to primates, a wide range of studies have demonstrated that recognition can occur over multiple instances despite changes in the particular position, viewing angle, size, and lighting of the scene (e.g., Brainard, 2004; Busey, Brady, & Cutting, 1990; Cutting, 1987; Desimone, 1991; Ito, Tamura, Fujita, & Tanaka, 1995; Jacobsen & Gilchrist, 1988; Lau, Rensink, & Munzner, 2004; Logothetis, Pauls, & Poggio, 1995; Rensink, 2004; Rutherford & Brainard, 2002; Shepard & Metzler, 1971; Tang, Wol, Xu, & Heisenberg, 2004; Tarr & Pinker, 1989; Wallis & Rolls, 1997). However, although there exist a variety of models that demonstrate various degrees of invariant recognition (Fukushima, 1988; Olshausen, Anderson, & Van Essen, 1995; Wang & Simoncelli, 2005; Wiskott, 2004), no dominant theory has emerged.

Within the vertebrate visual pathway, it is common to find low selectivity and low invariance at early levels (e. g., simple cells), with neurons at higher levels showing both high selectivity (e.g., face selective neurons in IT) with relatively high invariance (invariant to moderate changes in position, size or lighting) (Desimone, 1991; Ito et al., 1995; Wallis & Rolls, 1997). How does the visual system achieve this combination of selectivity and invariance? One approach argues that information regarding "what" and "where" are processed through different pathways, but the information is maintained. Others have argued that information regarding these transformations is simply discarded (Cutting, 1987).

In this paper, we use a simple Euclidean metric, for reasons given below, to determine the extent to which different types of information are retained by the visual system. We consider a variety of transformations performed on images of natural scenes and determine the sensitivity of human observers to those transformations. The number of possible natural image transformations is, of course, very large. Figure 1 shows examples from two broad classes of transformation, termed here *geometric* and *photometric*. The former refers to changes in the positions of image pixels, while the latter refers to changes in the intensive and/or spectral content of image pixels. The geometric transformations in Figure 1 are affine transformations on the two-plane (Watt, 2000). Many of these are quite common in our visual experience. Image "translation" occurs every time we move our eyes, and an image

doi: 10.1167/7.14.11 Received August 7, 2006; published December 13, 2007 ISSN 1534-7362

Figure 1. Example transformations for an image of the Ackee fruit. The middle baseline image has been transformed into the images arranged around it. Rotate, stretch, contract, shear, and translate are affine geometric transformations, and of these, stretch and shear can also be considered distortions. Flatten, brighten, and divide are photometric transformations applied uniformly across the image. Gaussian and fractal are added noise transformations. The multiplicative noise condition (not shown) looks similar to the added Gaussian noise condition.

"contraction" every time we move away from an object. Others, such as "stretch" and "shear," constitute *distortions* that may be less common but occur as the observer moves through an environment. The photometric transformations in Figure 1 are of two classes. Uniform photometric transformations impose the same change to all pixel values: "flatten," "brighten," and "divide" are the examples. Random photometric or "noise" transformations are random perturbations applied either independently to every pixel, as in the "Gaussian noise" example, or independently at different image scales, as in the "fractal noise" example. Of the uniform photometric transformations, "divide" is probably the most commonly experienced as it occurs every time there is a reduction in the ambient light level, as when going from day to night. With the exception of significant levels of photon noise seen under low light conditions, the transformations that involve added noise would normally never occur in our visual experience, and therefore also constitute distortions.

We use a conventional metric of image difference that is intuitively appealing due its simplicity. This is the Euclidean distance E, or L_2 norm. If the images are tri-plane, RGB colored images, as in Figure 1, E can be calculated using the following formula:

$$E = \sqrt{\frac{\sum_{n=1}^{N}\sum_{i=1}^{3}(p_{ni} - q_{ni})^2}{3N}}, \quad (1)$$

where p_{ni} and q_{ni} are the intensities of the corresponding pixels in the two images, with i the image plane ($i = 1{:}3 \mid R, G, B$), n the pixel (i.e., with unique x, y coordinate), and N the number of pixels per image. Euclidean distance has the important property that it defines a straightforward measure of the distance between two images and provides the same answer irrespective of the orthonormal basis used to represent the images, e.g., pixels, Fourier, Haar, etc. (Horn & Johnson, 1990). We are certainly not arguing that the Euclidean distance is the proper *perceptual* metric. Rather, we argue that E is a relatively neutral metric, providing a useful measure for comparing the relative sensitivities to the different types of image transformation shown in Figure 1. It is widely believed that simple visual discrimination tasks are mediated by filters in the early stages of the visual cortex, for example primate area V1, that are tuned to various orientations and spatial frequencies (DeValois & DeValois, 1991). Under the most simplistic model where we assume that the visual system calculates the differences between images from the differences between the magnitudes of m linear, orthonormal filter responses, the Euclidean distance calculated from the filter responses produces similar answers to that calculated from pixel intensities. We should also emphasize that Euclidean distance is a somewhat unusual metric for describing affine transforms. In a Euclidean pixel space, most affine transforms represent a curved trajectory through the space. Although a monotonic increase in the affine transformation (e.g., a shift to the left) will typically result in a monotonic increase in the Euclidean distance, it is not a simple linear relationship. Therefore, although Euclidean distance is a valid metric of physical distance between two images and is easily calculated, we do not expect it to be an accurate perceptual metric. Indeed, it is the failure of this physical metric which is the core of this study.

In this paper, we consider three possible hypotheses. First, if the Euclidean distance provided a good account of visible differences, then we might expect this metric to provide an accurate account of thresholds. However, there is abundant evidence that the Euclidean metric is a poor predictor of perceived image distortions (e.g., Teo & Heeger, 1994). Therefore, although this may be the simplest hypothesis, we do not expect this to produce accurate predictions.

An abundance of evidence suggests that the visual system has evolved neural mechanisms that optimally

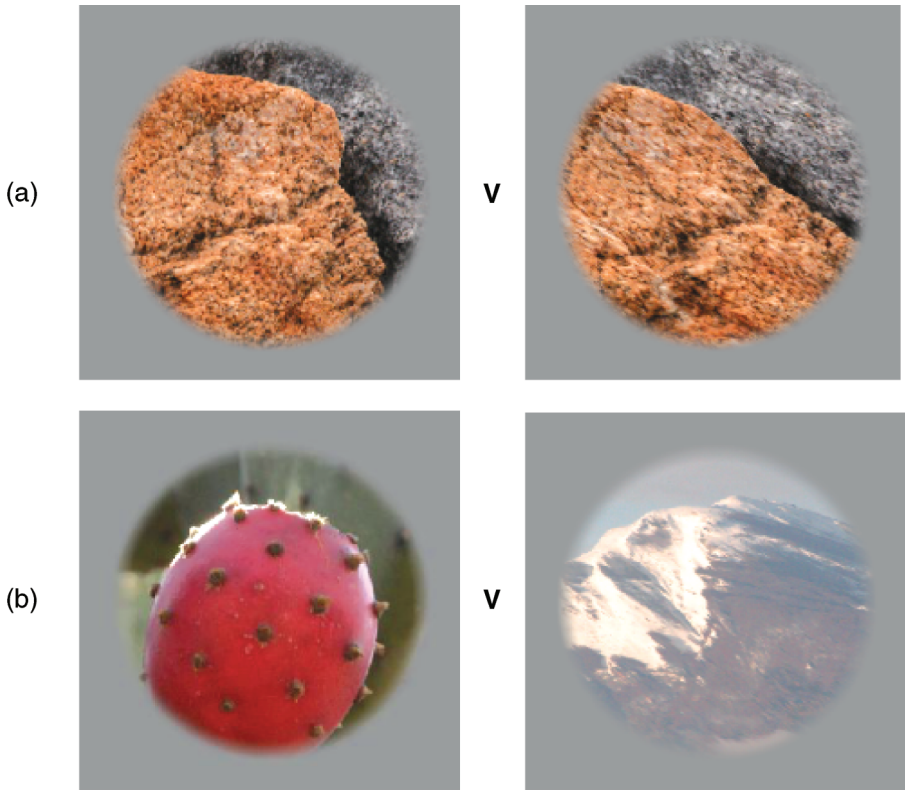

Figure 2. Example forced-choice pairs for the horizontal shear condition in the two experiments. In panel a (Experiment 1), the two images are of the same scene, in panel b (Experiment 2) different scenes. The task for the subject in both experiments was to decide which stimulus was sheared.

represent the statistical properties of natural scenes (Barlow, 1972, 2001; Buchsbaum & Gottschalk, 1983; Field, 1987; Johnson & Baker, 2004; Olshausen & Field, 2004; Ruderman, Cronin, & Chiao, 1998; Thomson, 1999). Our second hypothesis is that human observers would be particularly sensitive to the sorts of transformations that commonly occur in the natural world and therefore relatively sensitive to affine transformations. However, our third hypothesis leads us to the opposite conclusion. If the processes involved in perceptual invariance result from a loss of sensitivity along those common dimensions (the common transformations), then we might expect human observers to be relatively insensitive to these affine transformations.

To test between these predictions, we measured sensitivity to the class of transformations of images of natural scenes shown in Figure 1, and in two ways that are illustrated in Figure 2. In both experiments, the task was to identify which of two images had undergone a pre-specified transformation, i.e., the task was not to detect an unspecified transformation. This minimized the uncertainty of the task and, we assume, maximized sensitivity, in spite of subjects having to pre-learn each transformation. In the first experiment (Figure 2a), subjects were presented on each trial with two images of the *same* natural scene (different scenes on each trial) and were required to indicate which of the pair conformed to a particular transformation. In the second experiment (Figure 2b), subjects were presented on each trial with two images of *different* scenes (different scene-pairs on each trial) and were required to indicate which of the pair conformed to a particular transformation. For the second experiment, only those transformations that could be considered distortions are applicable, and therefore we only tested "stretch," "shear," and "added noise." Importantly, the distortion class of image transformation is the only class uniquely applicable to natural scenes, since knowledge of what is "normal" in a scene is pre-requisite. For the second experiment, we still measured the magnitude of the transformation in terms of E, even though the baseline image was not presented with its transformed version on the same trial.

Methods

Equipment and calibration

The scenes were photographed with a Nikon CoolPix-7500 digital camera. The digital images were first

corrected for the camera's gamma-nonlinearity, as detailed elsewhere (Olmos & Kingdom, 2004). The display monitor was a Sony FD Trinitron 17″, GDM F-500. The RGB phosphors were gamma-corrected after calibration using a photometer (OptiCal, Cambridge Research Systems). Images were displayed using the VSG graphics board (Cambridge Research Systems) housed in a 1800-MHz PC computer. Refresh rate was 120 MHz. Matlab version 7 was used for all image processing tasks.

Stimuli and procedures

For each experiment, 188 different scenes were employed, and these were taken from the McGill Calibrated Colour Image database (Olmos & Kingdom, 2004). The scenes represented a range of natural (forests, mountains, flowers, and fruits) and urban (buildings, traffic signs, man-made objects) environments photographed under different illumination conditions (sunny and cloudy) and distances (0.5 m–1000 m). The camera's smallest aperture setting (f 7.4) was chosen to capture the images with minimum within-image differences in focus. The images were presented on a mid-gray background of 42 cd/m^2. Intensity resolution was 24 bits (256 levels for each *R, G,* and *B* image). Each image was circular with a diameter of 300 pixels subtending 11 deg at the viewing distance of 100 cm. The stimulus edges were softened using a 0.55 × 0.55 deg Gaussian filter with a standard deviation of 2 deg. Each stimulus was presented for a total of 500 ms with a temporal ramp of 100 ms at stimulus onset and offset.

For each scene, an original plus six levels of transformation were generated. The six levels were chosen to ensure a range of approximately 50–100% correct. The two images on each trial were presented in a two-interval forced-choice (2IFC) procedure, with an inter-stimulus interval of 500 ms. Subjects indicated by key press the image that conformed to the particular transformation. Practice sessions ensured that for each type of transformation subjects understood the task and the key allocation, which was adapted for each type of transformation. For example, in the "rotation" condition, subjects pressed the left or right key depending on whether the second of the 2IFC pair appeared rotated clockwise or anticlockwise relative to the first. For the "translation vertical" task, subjects pressed the top or bottom key depending on whether the second of the 2IFC pair appeared shifted upwards or downwards relative to the first. During each session of 96 trials one type of transformation was tested (the order of transformations was random), with the 6 levels of transformation presented 16 times each in random order. The scenes on each trial were randomly selected from the 188 available, with the constraint that a given scene would be presented once only. There were 5 repeat sessions for each transformation, making a total of 480 trials per transformation.

Image transformations

Geometric

All geometric transformations were affine transformations, achieved using a mapping function that related points in the original image to corresponding points in the transformed image. The procedure involved two steps. In the first step, the pixels were rearranged using a matrix transformation of the general form:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} \\ m_{2,1} & m_{2,3} & m_{2,3} \\ m_{3,1} & m_{3,2} & m_{3,3} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{2}$$

where x, y are the original and x', y' the transformed image pixel coordinates. For the four classes of geometric transformation, the matrix coefficients were

$$\begin{array}{ll} \text{Scale} & \text{Rotate} \\ \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \text{Translate} & \text{Shear} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t_1 & t_2 & 1 \end{bmatrix} & \begin{bmatrix} 1 & h_1 & 0 \\ h_2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \end{array} \tag{3}$$

where s_1 and s_2, t_1 and t_2, and h_1 and h_2 are the transformation levels, with subscripts 1 and 2 for the x (horizontal) and y (vertical) coordinates. θ is orientation in degrees. For the scale transformation, s_1 and s_2 were covaried. For the stretch horizontal transformation, s_2 was set to zero while s_1 was varied, and similarly for the horizontal and vertical versions of the translation and shear transformations.

The second step involved allocating pixel intensity values for the resulting non-integer x's and y's. We employed a bi-cubic interpolation method, in which the new pixel value was the weighted average of the four neighboring pixel values. Although the range of transformations was tailored to each subject to ensure an average performance of about 75% correct, the total range across subjects for the different geometric transformations was 0.2–52% of image width/height for scaling (specifically contraction); 0.1–45 deg for rotation; 0.001–1.3 aspect ratio for shear; and 0.1–5.7% of image height/width (corresponding to 0.011–0.63 deg) for translation. The 6 levels of each transformation were spaced logarithmically.

Photometric—uniform

To *flatten* or reduce the contrast of the image, we decreased the range of pixel intensities (0–255) in each *R, G,* and *B* plane according to the formula

$$\begin{aligned} &\text{If } I(x,y) > M \text{ then} \\ &I'(x,y) = [I(x,y) - I(x,y)] * (1-k) + I(x,y) \\ &\text{If } I(x,y) < M \text{ then} \\ &I'(x,y) = [I(x,y) - I(x,y)] * (1-k) + I(x,y), \end{aligned} \tag{4}$$

where $I(x, y)$ is the original, $I'(x, y)$ the transformed image plane, and M the average value of the original image plane. k determined the degree of flattening. The k values spanned 0.05–0.45. To *brighten* the image, all pixel RGB values were incremented by a specified amount, ranging from 1 to 38. To *divide* the image, all pixel values were divided by an amount ranging from 1.01 to 1.31. As with the geometric transformations, the 6 levels of each transformation were spaced logarithmically.

Photometric—random noise

For the *added Gaussian noise* condition, each *R, G,* and *B* pixel value (0–255) was perturbed by an amount randomly drawn from a Gaussian probability distribution with mean zero and standard deviation σ equal to k, where k ranged from 1 to 15. *Multiplicative noise* was achieved by setting σ proportional to the pixel value, i.e., $\sigma = k \cdot I(x, y)$, where k ranged from 0.0002 to 0.0019. The *added fractal noise* images were generated by adding to each *RGB* image plane a fractal noise mask (Simoncelli, 2003) whose power spectral density fell with spatial frequency f according to $1/f^n$, with n set to 3 and the image variance normalized to 1. The choice of exponent $n = 3$ may seem odd because natural scenes have an average exponent of 2 (Field, 1987). However, we measured the spectra of our actual test images and found they had an average exponent of 3.2, and so took 3 rather than 2 for our fractal noise. The steeper-than-normal power spectra of our images is likely caused by the fact that they contained a more than average number of close-ups of objects. The different levels of fractal noise were achieved by multiplying the noise mask by a constant k that varied in logarithmic intervals from 1 to 12.

Data analysis

For each trial, the Euclidean distance E (between the original and transformed image) was recorded along with the response "correct" or "incorrect." Although there were 6 discreet levels for each transformation, the computed values of E for each level of a given transformation varied according to the image. In order to fit psychometric functions, the Es were divided into 6 "bins" for each transformation. The first bin was set to have a minimum of zero, while the last, sixth bin was set to have a maximum equal to the maximum E for that transformation. The first bin "divider" was determined iteratively to be the value such that when the remaining bin dividers were logarithmically spaced, the between-bin variance in the number of trials was minimized. This method ensured that the trials were distributed as evenly as possible between bins under the constraint that all except the first bin were logarithmically spaced (because the first bin began at zero). After the Es were binned, the mean log E, proportion correct, and number of trials were calculated for each bin. The psychometric functions relating proportion correct to log E were fitted using the logistic function: $0.5 + 0.5 \cdot \exp[(\log E - a)/b]/\{1+\exp[(\log E - a)/b]\}$, where a is the threshold at the 75% correct level and b is the slope. The fitting procedure used a weighting function given by the reciprocal of the binomial standard deviation $\sigma_i = \sqrt{(p_i(1-p_i)/N_i}$, where p_i and N_i are the proportion correct and number of trials for the ith log E level.

Results

Two female student observers (KW, SG) participated in the first two experiments and were unaware of their purpose. Figure 3 shows example psychometric functions for KW's added Gaussian noise and vertical translation conditions from Experiment 1. The psychometric functions give the mean proportion of correct trials as a function of log E. The threshold was calculated as the value of log E giving 75% correct (see Methods for details). Threshold Es (note: not log Es) for all transformations are shown in Figure 4a for Experiment 1 and Figure 4b for Experiment 2.

Our results show that for a range of different types of natural image transformation measured in terms of Euclidean distance E, what the eye sees best is added noise. How much more sensitive our subjects are to added noise can be gleaned from a comparison of the Gaussian noise condition, which had the lowest thresholds, with the average of the geometric transformations, which had the highest thresholds.

For the first experiment, the Gaussian noise thresholds were approximately 11 times, and for the second experiment approximately 14 times lower than the mean of the geometric transformation thresholds. In all instances, the noise conditions produced lower thresholds than the uniform photometric transformations.

To illustrate how different is our sensitivity to affine geometric transformations compared to added white noise, consider Figure 5. The top image has been transformed into the two images shown beneath, the one on the left by

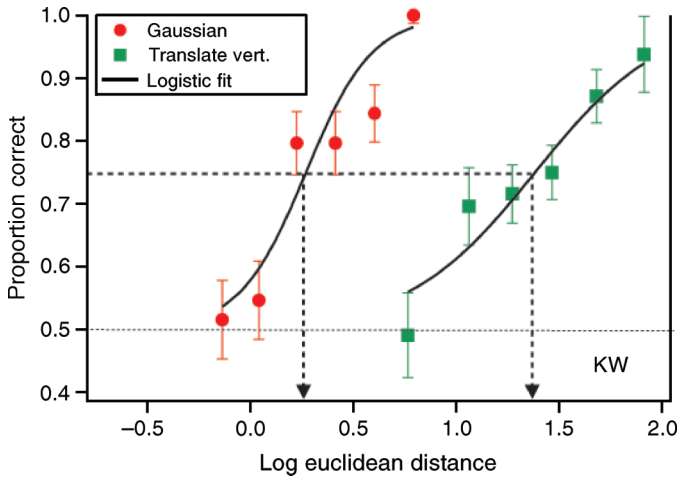

Figure 3. Example psychometric functions for KWs Gaussian noise (red symbols) and vertical translation (green squares) conditions from Experiment 1. The proportion of correctly detected transformations is plotted against the log Euclidean distance between the transformed and untransformed image. Error bars are binomial standard deviations. Continuous lines are best fitting logistic functions. The horizontal black line shows the 75% correct level, and the vertical dashed lines show the threshold log Euclidean distance for each condition.

adding white noise, the one on the right by stretching the image horizontally. The amount of transformation however is identical in terms of Euclidean distance. While it is easy to see the changes in the left image, the changes in the right image can only be seen with careful scrutiny.

Why are the geometric transformations so much more difficult to detect compared to added noise? One possibility is that the answer lies in the shape of the histogram of pixel differences between the original and transformed images. The pixel-difference histogram captures the first-order (point-wise) statistical differences between two images. Figure 6 shows the pixel-difference histograms for an image transformed by the same Euclidean distance in one of three ways: translation, brightening, and addition of Gaussian noise. The pixel-difference histogram is by definition a Gaussian for the added Gaussian noise condition. For translation it is more kurtotic, and for brightening it is a single-point function (all pixels are incremented by the same value) making it highly kurtotic. The marked difference in the shape of these pixel-difference histograms for the various transformations raises the possibility that pixel histogram shape is a factor determining thresholds. At face value, however, it would seem unlikely that kurtosis is the critical statistic since the relative magnitudes of thresholds are geometric > photometric > noise, whereas for the image

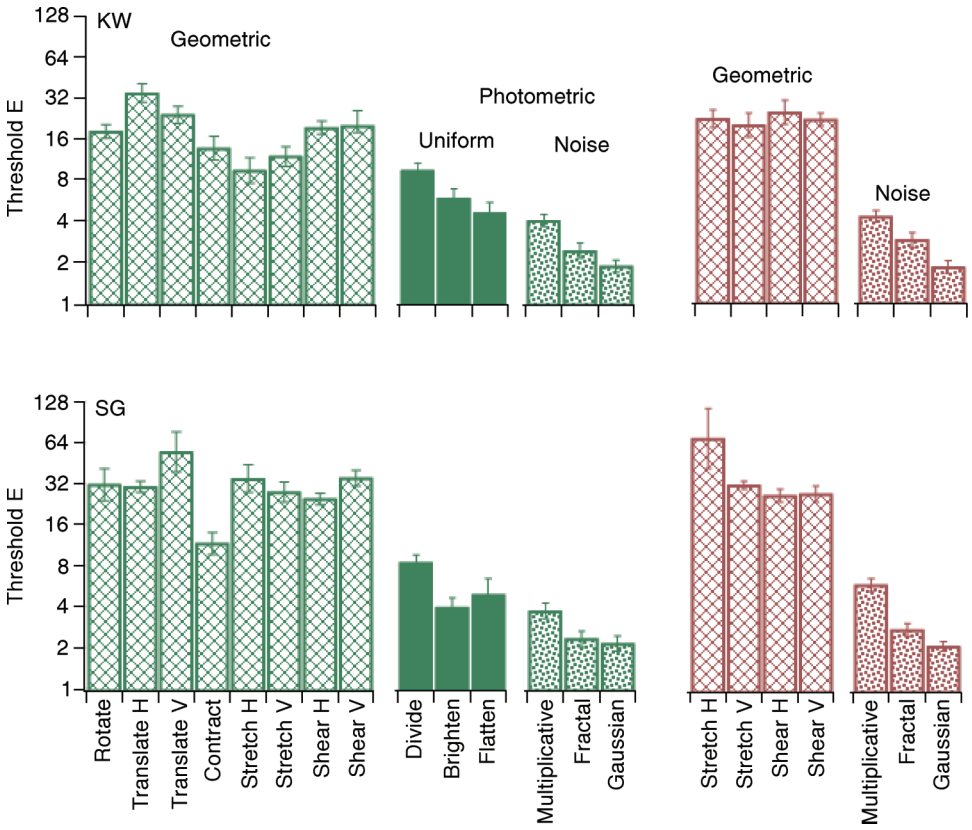

Figure 4. Euclidean distance *E* thresholds for all types of transformation from both experiments and for both subjects. Results from Experiment 1 are shown in green, Experiment 2 in red. H = horizontal, V = vertical. Subject KW top, SG bottom. Viewing distance was 1 m.

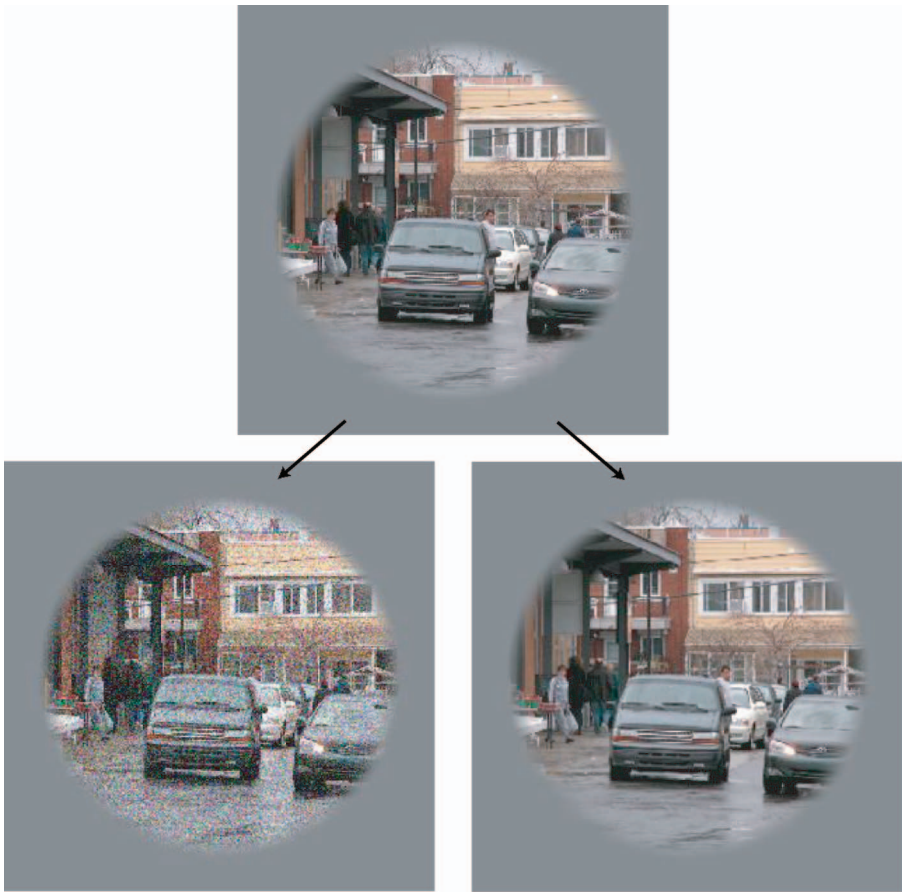

Figure 5. The top image has been transformed into the bottom two images by (left) the addition of white noise and (right) by stretching the image horizontally. The Euclidean distance between the top and each of the two transformed images is identical.

in Figure 6, which is typical, the magnitudes of kurtosis are photometric > geometric > noise.

Nevertheless, it would be prudent to test whether the statistics of the difference image, whose first-order properties are reflected in the pixel-difference histogram, is the reason for the relatively high geometric transformation thresholds. An anonymous reviewer suggested a way of doing this.

Take a baseline image I_B and transform it, say by rotation, to image I_T. Call the difference between these two images $I_D = I_T - I_B$. Any difference between two images (even a difference caused by an affine transformation) can be described in terms of this difference image. In the third control experiment, we compare the thresholds for detecting the increment versus the decrement of this difference image. That is we compare the thresholds for

$$I_T \text{ versus } I_B \quad \text{and } I_C \text{ versus } I_B, \tag{5}$$

where I_T is the incremental and I_C the control, decremental image, defined as

$$\begin{aligned} I_T &= I_B + I_D \\ I_C &= I_B - I_D. \end{aligned} \tag{6}$$

Given that

$$I_D = I_T - I_B, \tag{7}$$

the control image can also be written as

$$\begin{aligned} I_C &= I_B - (I_T - I_B) \\ &= 2I_B - I_T. \end{aligned} \tag{8}$$

Figure 7 provides an example of the two images that are created by adding and subtracting the difference image. The difference image in the two cases is identical. We can therefore ask whether an increment (which corresponds to

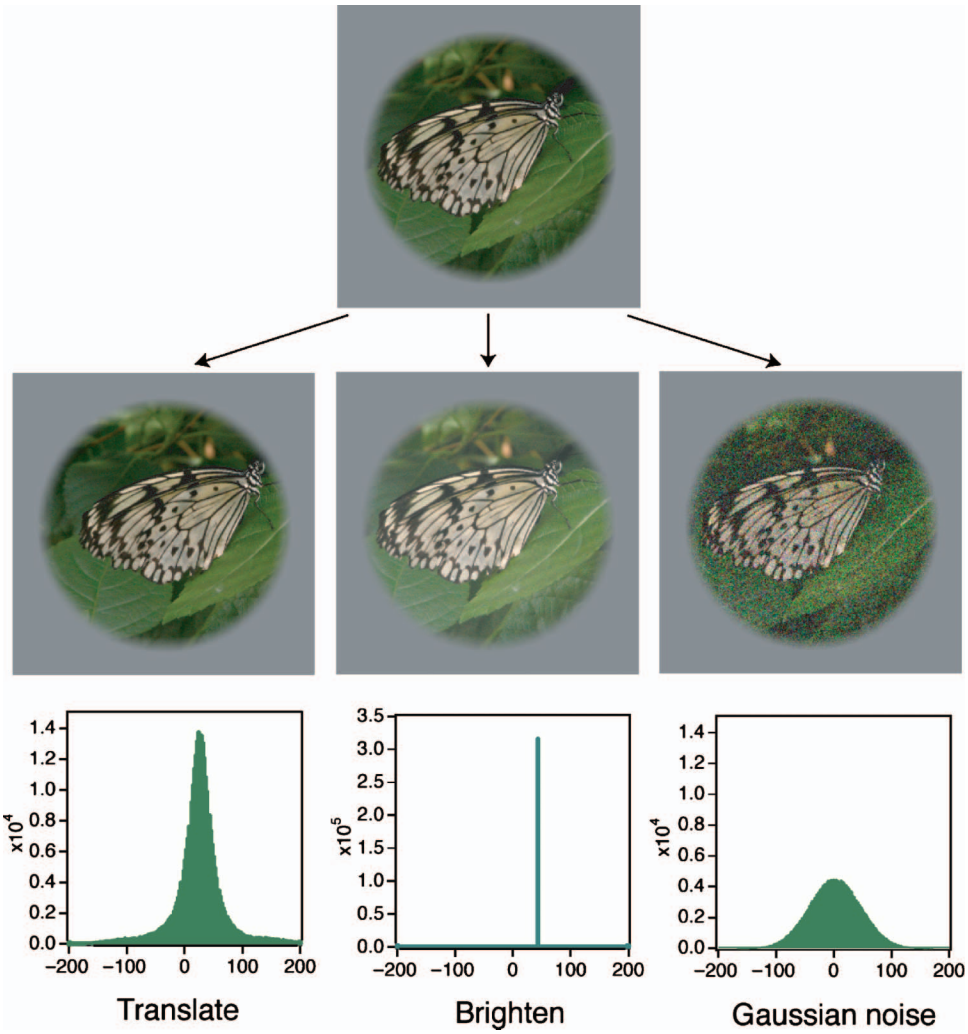

Figure 6. Image (top) transformed by the same Euclidean distance (E = 40) by translation (left), brighten (middle), and addition of Gaussian noise (right). The resulting pixel-difference histograms are shown below. Note that the range on the y-axis for the brighten histogram is greater than for the other two histograms. Kurtoses for the three histograms are (left to right) 7.6, 12512, and 3.0.

an affine transform) is equally distinguishable from a decrement of equal magnitude. The image I_T shows a 1 deg rotation, whereas its counterpart image I_C appears to be edge-sharpened and is clearly much easier to detect. If indeed easier to detect, this suggests that the structure of the difference image is not in itself the main factor producing the relatively high thresholds for the geometric transformations.

In order to create the I_C images, it was necessary to reduce the contrasts of the images by a factor of 3 to prevent pixel values going outside the 0–255 range (underflow/overflow) (see Figure 7). Pilot studies confirmed the impression obtained in Figure 7 that we are much more sensitive to the I_C transformation, and in order to obtain meaningful psychometric functions, we had to make the task more difficult by increasing the viewing distance to 2.8 m. We tested three geometric transformations—rotation, translation, and shear. Two of the authors (AO and FK) served as subjects.

The results are shown in Figure 8. Thresholds for the I_C transformations are about 4 times lower than their conventional geometric transformation counterparts (note again the logarithmic spacing on the y-axis). This is conclusive evidence that the relatively high thresholds for the geometric transformations are not caused by the statistics of the difference between the baseline and transformed images.

As far as we are aware, these results also demonstrate the largest difference ever obtained between a threshold for detecting an increment and a threshold for detecting an equivalent decrement.

Discussion

These results support the notion that the human visual system is relatively insensitive to the types of image

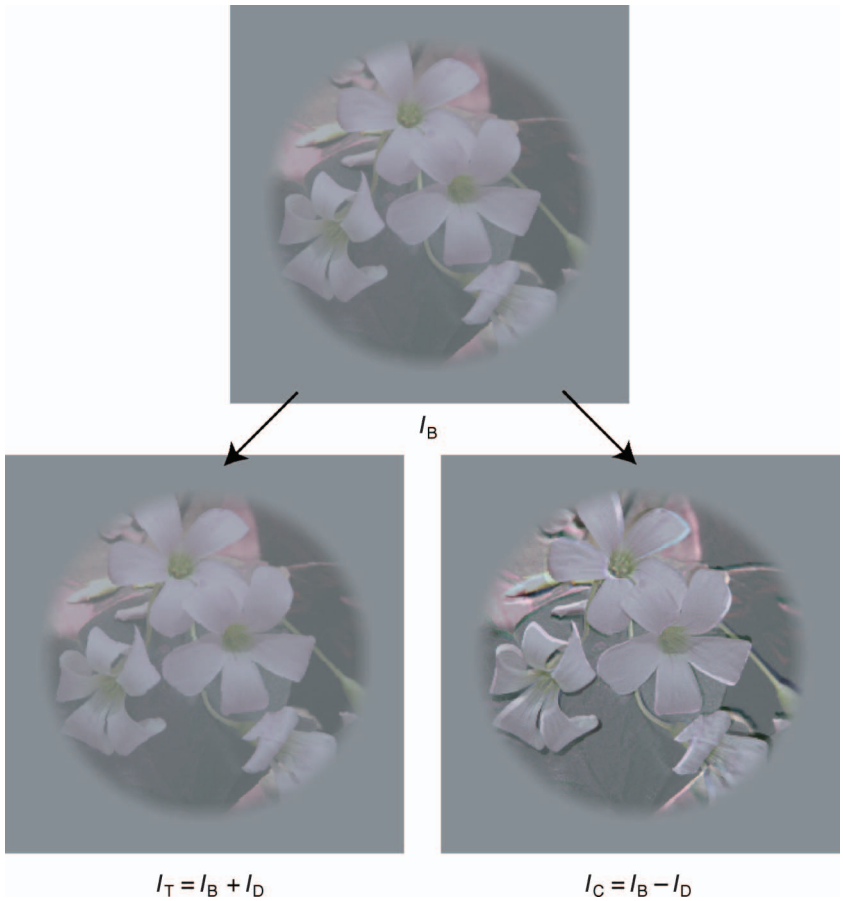

Figure 7. The baseline image I_B (top) has been rotated by 1 deg to produce the transformed image I_T (bottom left). As with any image transformation, this rotation can be treated as a sum of the original image and the difference image I_D. The control image I_C (bottom right) is obtained by *subtracting* the difference image from the standard. Both transformed images have identical Euclidean distances from the baseline (E = 9.1) and identical pixel-difference histograms. Note that the contrasts of the images have been reduced to prevent pixel overflow/underflow in the I_C image.

transformation that are common to our visual experience. The work implies that invariance is partly achieved through a loss of information. Observers were at least 10 times less sensitive to the types of geometric transformation that would likely be involved in perceptual invariance than they were to added white noise.

As argued in the Introduction, Euclidean distance E is not an accurate perceptual measure of image difference. However, as a physical measure, it can be used to compare sensitivities across all types of transformation. It is precisely the fact that threshold Es are so different for the geometric and noise transformations that we can appreciate that E is an inadequate predictor of perceived image difference.

It is also true that different images produce different values of E for a given unit of transformation (e.g., for a 5 deg rotation), though this fact in itself does not preclude the possibility that E could predict performance better than the unit of transformation itself. Thus, it is conceivable that a different set of images would produce different threshold Es. However, given that we sampled a large number of images with a variety of different types of natural scene, we are confident that had we used a completely different image set, the pattern of results would nevertheless be the same.

It has been argued that our ability to view 2D images and movies from a wide variety of viewing angles follows from our insensitivity to the transformations produced by off-angle projections (Busey et al., 1990; Cutting, 1987), producing the apparent invariance to viewing angle. Indeed Cutting (1987) found that with rotating wire frames, we are more sensitive to non-affine changes in comparison to affine. We are never likely to encounter a 3D object twice from the same exact viewing angle stimulating the same collection of neurons in V1. It is therefore critical that at some point in the visual system, the system becomes invariant to common forms of transformation.

The studies by Cutting and colleagues, as well as others dealing with perceptual invariance (see Introduction), have considered how objects are recognized under various transformations. In the present study, rather than measuring

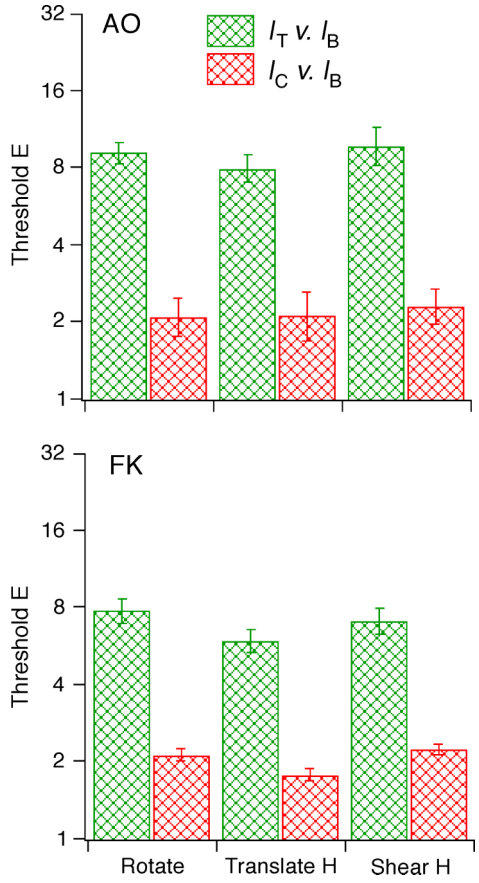

Figure 8. Results from control experiment. I_B = baseline image; I_T = difference-added transformed image; I_C = control, difference-subtracted transformed image. H = horizontal. Subject AO top, FK bottom. See text for details.

the recognition of objects undergoing various transformations, we have measured sensitivity to the transformations themselves. That this is the other side of the coin of perceptual invariance is evidenced by our finding that we are relatively insensitive to precisely those transformations that have little negative impact on object recognition, such as rotation, scaling, and translation.

Wang and Simoncelli (2005) have proposed an image similarity metric that is simultaneously insensitive to luminance change, contrast change, and spatial translation. The key idea behind the metric is that it makes use of the fact that these image changes lead to *consistent* magnitude and/or phase changes in local wavelet coefficients. It would be interesting to see how well the similarity metric predicts the results of the present study. Wang and Simoncelli point out that small scaling and rotation of images can be locally approximated by translation. This may be the reason why our stretch and shear conditions, which unlike the other geometric transformations distort the images in a less common way, nevertheless produce comparable thresholds. In other words, the process involved in perceptual invariance may be computed over relatively small regions of the image.

The results from single unit recordings of higher level neurons in the mammalian visual system (e.g., inferotemporal cortex) have demonstrated high degrees of selectivity (e.g., to faces) while maintaining relatively high invariance to position, lighting, pose, etc. (Desimone, 1991; Ito et al., 1995; Logothetis et al., 1995). The insensitivity to geometric image transformations shown in this study may be a result of the processes used to build these higher level invariances. If this is true, we would predict that higher level neurons would show quite similar responses to the various transformations shown here, even though lower level neurons might show very different responses.

The 10-fold or more decrease in sensitivity to the geometric distortions might be described as a form of "change blindness" (Simons & Ambinder, 2005). However, the fact that similar sensitivities were found for the distortion transformations in the same-scene and different-scene comparisons (Figure 4) suggests different perceptual processes from those involved in change blindness. Moreover, the high sensitivity to noise is not predicted by change blindness theories. High sensitivity to noise also cannot be explained away on the grounds that the amplitude spectra of the added noise is so different from that of the test images that the test images fail as maskers. Sensitivity to added fractal noise, whose amplitude spectra was matched to those of our test images (see Methods), was also much higher.

The interpretation we prefer is one that considers transformations as movements along an image space manifold. In such a model, the images and objects in the world are points in high-dimensional state space (Field, 1994; Field & Wu, 2005). The geometric and uniform photometric transformations form a collection of curved manifolds (i.e., smooth surfaces) in this state space. The addition of random noise is equivalent to a movement in a random direction and will likely be a movement away from the manifold of common transformations. Our results suggest that we are very sensitive to movements away from (or towards) the probable manifold (e.g., addition of noise), but relatively insensitive to movements along the probable manifold (e.g., the common transformations) (we are presumably also insensitive to movements between points that are off the manifold, for example between identical images with independent samples of white noise). Therefore, under this interpretation, our results suggest that we are an order of magnitude more sensitive to movements away/towards the manifold relative to movements along the manifold.

There are two important caveats to this conclusion. The first is that we have not explored all types of transformation, and there may be ones for which our pattern of results does not apply. For example, there is the "fish-eye" transformation, in which the region around fixation is selectively expanded (Lau et al., 2004). The fish-eye transformation can be considered a combination of scaling and shear, so we might expect thresholds for the fish-eye to be similar to these transformations. A second caveat is

that our results are likely only applicable to transformations that are significantly separated over time or space. If the two images in the translation forced-choice pairs were presented consecutively without the 500-ms inter-stimulus interval, they would constitute a two-frame apparent motion sequence, and displacement thresholds for detecting the motion direction of such sequences are known to be less than a minute of arc (Baker & Braddick, 1984; Nakayama & Tyler, 1981). Therefore, our results are arguably relevant primarily to situations where the spatial properties of natural scenes have been coded into some form of short-term memory. We are currently investigating how our sensitivity to these distortions is affected by the temporal interval between them.

On the other hand, our 10-fold difference in sensitivity may be an underestimate of that obtained had we used natural scenes that filled the visual field. Our images were presented in a fixed-size circular window 11 deg in diameter on a neutral gray background. For the geometric transformations, the accretion and deletion at the edges of the stimulus window provided a visual cue that would be unavailable in a full-field scene because in the far visual periphery spatial resolution is so poor. A similar argument applies to the uniform photometric transformations. The luminance-contrast between the image and its background in the "brightness" and "divide" conditions and the contrast–contrast between the image and its background in the "flatten" condition are cues that would also be unavailable in a full-field scene, again because sensitivity to these dimensions in the far periphery is poor. Indeed Schubert and Gilchrist (1992) have shown that when viewing a Ganzfeld, or homogenous visual field, human test subjects are insensitive to the slow changes in light level exemplified by our brighten and divide transformations. None of these edge cues would have been available for the noise transformations, yet still their thresholds were the lowest.

Notwithstanding the possibility that the thresholds for the uniform photometric transformations may have been higher had we used full-field stimuli, they were nevertheless lower than the geometric transformation thresholds. This may be because the geometric transformations are more ubiquitous—they occur whenever we move our bodies or eyes. The uniform photometric transformations would normally arise from physical changes in the scene itself and are thus less frequently experienced. Another possibility is that the visual system prefers not to discard information about uniform photometric changes because they provide important information about the illuminant, which some recent studies have suggested is encoded for the purpose of color constancy (Golz & MacLeod, 2002; Maloney, 2002; Smithson, 2005; Zaidi, 2001).

There is currently a large literature that attempts to model human visual sensitivity with regard to the distortions created by various compression algorithms (Chandler & Hemami, 2003; Teo & Heeger, 1994; Wang, Bovik, Sheikh, & Simoncelli, 2004). We believe our results support the notion that perceptual space must be considered in terms of a non-Euclidean transform of pixel space. However, we feel that our results make a further and rather unusual claim: that we are least sensitive to the most common types of transformation and most sensitive to the highly unlikely random changes. We believe that the 10-fold increase in sensitivity to additive noise reveals a very important process involved in building invariant representations in the visual system. It argues that at least for small affine changes, the visual system does sacrifice information in order to achieve this invariance. Furthermore, this loss of information does not result from an overall insensitivity to change but to the precise types of changes most likely to occur under natural viewing.

Is this insensitivity learned from exposure to natural transformations? Or would we find evidence for this insensitivity in newborns? The results here do not provide an answer. Certainly some degree of invariance is required at birth so that learning can occur across multiple instances. However, if it is learned, these results imply that experience is helping the visual system to become insensitive to the most common forms of transformation and distortion.

Conclusion

The human visual system appears to be relatively insensitive to the types of image transformation that are common to our visual experience. This implies that spatial invariances are partly achieved through a loss of information.

Acknowledgments

FK was supported by a Canadian Institute of Heath Research Grant #11554. DF was supported by NGA contract HM 1582-05-C-0007. Special thanks to the anonymous reviewer for suggesting the control experiment. Thanks also to Frans Verstratten and Aaron Johnson for useful comments.

Commercial relationships: none.
Corresponding author: Frederick Kingdom.
Email: fred.kingdom@mcgill.ca.
Address: McGill Vision Research, 687 Pine Av. W. Rm. H4-14, Montréal, Québec, H3A 1A1, Canada.

References

Baker, C. L., Jr., & Braddick, O. J. (1984). Eccentricity-dependent scaling of the limits for short-range

apparent motion perception. *Vision Research, 25,* 803–812. [PubMed]

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception, 1,* 371–394. [PubMed]

Barlow, H. B. (2001). Redundancy reduction revisited. *Network, 12,* 241–253. [PubMed]

Brady, N., & Field, D. J. (1990). What's constant in contrast constancy? The effects of scaling on the perceived contrast of bandpass patterns. *Vision Research, 35,* 739–756. [PubMed]

Brainard, D. H. (2004). Color constancy. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 948–961). Cambridge, MA: MIT Press.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colour coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London B: Biological Sciences, 220,* 89–113. [PubMed]

Busey, T. A., Brady, N. P., & Cutting, J. E. (1990). Compensation is unnecessary for the perception of faces in slanted pictures. *Perception & Psychophysics, 44,* 339–347. [PubMed]

Chandler, D. M., & Hemami, S. S. (2003). Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions. *Journal of the Optical Society of America A, 20,* 1164–1180. [PubMed]

Cutting, J. E. (1987). Rigidity in cinema scenes from the front row, side aisle. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 323–334. [PubMed]

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience, 3,* 1–8.

DeValois, R. L., & DeValois, K. K. (1991). *Spatial vision.* Oxford University Press.

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

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation, 6,* 559–601.

Field, D. J., & Wu, M., (2005). An attempt towards a unified account of non-linearities in visual neurons [Abstract]. *Journal of Vision, 4*(8):283, 283a, http://journalofvision.org/4/8/283/, doi:10.1167/4.8.283.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks, 1,* 119–130.

Golz, J., & MacLeod, D. I. A. (2002). Influence of scene statistics on colour constancy. *Nature, 415,* 637–640. [PubMed]

Horn, R. A., & Johnson, C. R. (1990). *Norms for vectors and matrices.* Cambridge: University Press.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology, 73,* 218–226. [PubMed]

Jacobsen, A., & Gilchrist, A. (1988). The ratio principle holds over a million-to-one range of illumination. *Perception & Psychophysics, 43,* 1–6. [PubMed]

Johnson, A. P., & Baker, C. L., Jr. (2004). First- and second-order information in natural images: A new view of what second-order sees. *Journal of the Optical Society of America A, 21,* 913–925. [PubMed]

Lau, K., Rensink, R. A., & Munzner, T. (2004). Perceptual invariance of nonlinear focus + context transformations. In *Proceedings of the First Symposium on Applied Perception in Graphics and Visualization (APGV 2004)* (pp. 65–72). Los Angeles, CA: SIGGRAPH.

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

Maloney, L. T. (2002). Illuminant estimation as cue combination. *Journal of Vision 2*(6):6, 493–504, http://journalofvision.org/2/6/6/, doi:10.1167/2.6.6. [PubMed] [Article]

Nakayama, K., & Tyler, C. W. (1981). Psychophysical isolation of movement sensitivity by removal of familiar position cues. *Vision Research, 21,* 427–433. [PubMed]

Olmos, A., & Kingdom, F. A. A. (2004). *McGill Calibrated Colour Image Database.* Retrieved from http://tabby.vision.mcgill.ca

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1995). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Computational Neuroscience, 2,* 45–62. [PubMed] [Article]

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology, 14,* 481–487. [PubMed]

Rensink, R. A. (2004). The invariance of visual search to geometric transformation [Abstract]. *Journal of Vision, 4*(8):178, 178a, http://journalofvision.org/4/8/178/, doi:10.1167/4.8.178.

Ruderman, D. L., Cronin, T. W., & Chiao, C.-C. (1998). Statistics of cone responses to natural images: Implications for visual coding. *Journal of the Optical Society of America A, 15,* 2036–2045.

Rutherford, M. D., & Brainard, D. H. (2002). Lightness constancy: A direct test of the illumination estimation hypothesis. *Psychological Science, 13,* 142–149. [PubMed]

Schubert, J., & Gilchrist, A. L. (1992). Relative luminance is not derived from absolute luminance. *Investigative Ophthalmology & Visual Science, 33,* 1258.

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

Simoncelli, E. P. (2003). *matlabPyrTools—Matlab source code for multi-scale image processing*.

Simons, D., & Ambinder, M. (2005). Change blindness. *Current Directions in Psychological Science, 14,* 44–48.

Smithson, H. E. (2005). Sensory, computational and cognitive components of human colour constancy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360,* 1329–1346. [PubMed] [Article]

Tang, S. M., Wolf, R., Xu, S. P., & Heisenberg, M. (2004). Visual pattern recognition in *Drosophila* is invariant for retinal position. *Science, 305,* 1020–1022. [PubMed]

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21,* 233–282. [PubMed]

Teo, P., & Heeger, D. J. (1994). Perceptual image distortion. *IEEE International Conference on Image Processing, 2,* 982–986.

Thomson, M. G. A. (1999). Higher-order structure in natural scenes. *Journal of the Optical Society of America A, 16,* 1549–1553.

Wallis, G., & Rolls, E. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51,* 167–194. [PubMed]

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error measurement to structural similarity. *IEEE Transactions on Image Processing, 13,* 1–14. [PubMed]

Wang, Z., & Simoncelli, E. P. (2005). Translation insensitive image similarity in complex wavelet domain. *IEEE International Conference on Acoustics, Speech & Signal Processing, Vol. II,* 573–576.

Watt, A. H. (2000). *Computer graphics*. Redwood City, Calif: Addison Wesley.

Wiskott, L. (2004). How does the visual system achieve shift and size invariance? In J. L. van Hemmen & T. J. Sejnowski (Eds.), *Problems in systems neuroscience*. New York: Oxford University Press.

Zaidi, Q. (2001). Color constancy in a rough world. *Color Research and Application, 26,* 192–200.