Perception of suprathreshold naturalistic changes in colored natural images

Michelle P. S. To
School of Psychology, University of St. Andrews, St. Andrews, UK
Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, UK

P. George Lovell
Department of Experimental Psychology, University of Bristol, Bristol, UK

Tom Troscianko
Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, UK

David J. Tolhurst

Simple everyday tasks, such as visual search, require a visual system that is sensitive to differences. Here we report how observers perceive changes in natural image stimuli, and what happens if objects change color, position, or identity—i.e., when the external scene changes in a naturalistic manner. We investigated whether a V1-based difference-prediction model can predict the magnitude ratings given by observers to suprathreshold differences in numerous pairs of natural images. The model incorporated contrast normalization and surround suppression, and elongated receptive-fields. Observers’ ratings were better predicted when the model included phase invariance, and even more so when the stimuli were inverted and negated to lessen their semantic impact. Some feature changes were better predicted than others: the model systematically underpredicted observers’ perception of the magnitude of blur, but over-predicted their ability to report changes in textures.

Keywords: natural scenes, V1 models, suprathreshold discrimination, difference ratings, Minkowski summation


Introduction

The response properties of V1 neurons influence what we are able to perceive in the visual world. When presented with two different stimuli, it is presumably the difference in firing pattern elicited by each stimulus that allows the observer to differentiate them. If there is no difference in activity, we can be sure that an observer will not perceive a difference. But, is the converse necessarily true? Will an observer always perceive differences when the V1 activity changes? The behavior of V1 neurons and their presumed psychophysical homologues (“channels”) has been extensively studied for 50 years (see direct comparison by Boynton, Demb, Glover, & Heeger, 1999), and several researchers have developed models to describe the population activity of V1 neurons in response to various visual stimuli; e.g. models by Carlson and Cohen (1980), Daly (1993), Itti, Koch, and Braun (2000), Lubin (1995), Otazu, Vanrell, and Párraga (2008), Watson (1987), and Watson and Solomon (1997). Although these models of millions of neurons are based on rather few free parameters, their predictions are remarkably accurate in grating detection threshold experiments (Watson & Ahumada, 2005). We are interested in how the vast knowledge of visual processing derived from, say, the study of sinusoidal gratings both neurophysiologically and psychophysically can be applied to predicting the visibility or saliency of naturalistic changes in colored photographs of natural scenes. Can V1-based models (derived from grating work) explain performance in everyday visual tasks, such as noting changes in the locations, colors or shadowing of items in scenes? The experiments in the present paper examine how observers perceive changes in images which correspond to what would happen if objects changed color, or position, or identity—in other words, when the external scene changes in a naturalistic manner. We ask whether low-level computational models which have not yet been challenged with such naturalistic image changes can predict how observers perceive these kinds of changes.

While V1-based models have been applied most successfully to the visibility of grating stimuli, they have also been used to model the detectability of objects or
changes in natural scenes (Lovell, Párraga, Ripamonti, Troscianko, & Tolhurst, 2006; Párraga, Troscianko, & Tolhurst, 2005; Peters, Iyer, Itti, & Koch, 2005; Rohaly, Ahumada, & Watson, 1997; Tolhurst, Párraga, Lovell, Ripamonti, & Troscianko, 2005). Most particularly, V1-based models have been used to devise quality metrics for images or video clips of natural scenes that have been distorted by deliberate compression algorithms, or by noise or bit-loss in transmission (Barten, 1990; Chandler & Hemami, 2007; Daly, 1993; Feng & Daly, 2003; Ferwerda, Pattanaik, Shirley, & Greenberg, 1997; Lubin, 1995; Teo & Heeger, 1994; Wang, Bovik, Sheikh, & Simoncelli, 2004). Often, in such studies, the number and variety of images used has been small, or the types of changes presented have been limited to a single dimension, e.g. JPEG or blur degradation. Although our study bears some resemblance to earlier research on metric differences, we want to emphasize that the present work asks whether a model based on V1 properties can predict the magnitude of perceived changes in a variety of plausible suprathreshold manipulations to the contents of natural scenes. We do not aim to provide another image quality metric, but to test the validity of V1 models when operating on representations of natural changes within images such as might happen in time-lapse photography of a busy scene, and so this paper will examine how observers perceive differences in a very large number of naturalistic stimuli containing differences that span several feature dimensions, separately or in combination (e.g. blur, hue, saturation, or object shape, size or number). Very many of our stimuli exhibit totally natural changes in scenes, caused by changes in the weather, or the movement of objects, animals and people within the scenes. Unlike the change-blindness literature (Simons & Rensink, 2005), our scene changes are designed not to be challenging for attentional or memory systems.

We have used magnitude estimation ratings (Ahumada & Lovell, 1971; Gescheider, 1997) to estimate discriminability between pairs of naturalistic images. This extends the common use of ratings to assess image quality after, say, JPEG compression or blur (Barten, 1990; Chandler & Hemami, 2007; Feng & Daly, 2003; Lubin, 1995; Teo & Heeger, 1994; Wang, Bovik, Sheikh, & Simoncelli, 2004). Ratings allow examination of the impact of suprathreshold as well as threshold changes, and they allow the possibility that the perceived importance or salience of a change may not match its visibility. The use of this technique enables us to report on measurements of hundreds of image pairs, encompassing a great variety of naturalistic feature changes, so that the large and disparate data set can pose real challenges for modeling. From this data set, we can begin to identify which visual tasks are amenable to low-level modeling and which will require modeling of the functions of the many visual areas beyond V1.

We have previously presented a model of visual difference discriminations which considered each V1-like ‘neuron’ as an independent filter (Lovell et al., 2006; Párraga et al., 2005; Tolhurst et al., 2005). Now, like some studies (e.g. Lubin, 1995; Teo & Heeger, 1994; Watson & Ahumada, 2005; Watson & Solomon, 1997), we model the likely non-linear interactions between filters, such as non-specific contrast normalization (Bonds, 1989; Carandini, Heeger, & Movshon, 1997; DeAngelis, Robson, Ohzawa, & Freeman, 1992; Foley, 1994; Heeger, 1992; Li, Peterson, Thompson, Duong, & Freeman, 2005; Tolhurst & Heeger, 1997) and surround suppression (Blakemore & Tobin, 1972; Cavanaugh, Bair, & Movshon, 2002; Maffei & Fiorentini, 1976; Meese, 2004). However, even though we have striven for physiological realism, we are led to the conclusion that the saliency of some specific kinds of image difference will not be amenable to modeling only of the relatively “low-level” visual processes considered in conventional Visual Difference Predictor models (see Chandler & Hemami, 2007; Wang et al., 2004); the observer may ignore some potential cues while being particularly alerted by others—despite V1 models predicting that both cues should be equally salient.

**Experimental methods**

Extra details on methodology are given in the Auxiliary File.

**Display equipment and stimulus presentation protocols**

Stimuli were presented on a 19” SONY CRT display driven at 800 by 600 pixels and a frame rate of 120 Hz by a ViSaGe system (Cambridge Research Systems: Rochester, UK). The display was viewed in a darkened room from 2.28 m, so that the visible area subtended 10 by 7.5 degrees; each square pixel subtended 0.75 minutes. The screen was linearized using the OptiCal system. A pixel value of 128 in all 3 color planes gave 88 cd.m⁻², CIE [x, y] 0.30, 0.31.

The stimuli were square (256 by 256 pixel) colored images constructed from digitized photographs of natural scenes, occupying 3.2 degrees square in the center of the display. Each pixel in the stimuli was represented with 8 bits each of red, green and blue, and the pixel values were fed through linearizing look-up tables to be displayed through 14 bit DACs so that each color plane was presented with 256 equally spaced precisely defined luminance steps (Pelli & Zhang, 1991). When and where the display was not occupied by a stimulus, it was held at a mid-brightness gray, except for a small dark fixation dot in the center of the screen. The 30 pixels around the 4 stimulus edges were blended into the gray surround, by squeezing the pixel values towards 128; the Gaussian fall-off had a standard deviation of 12 pixels.
On each trial a randomly selected image from the current comparison-pair was presented for 833 ms; then the fixation point was presented in the center of the gray screen for 83 ms; the other image from the pair was presented for 833 ms; the fixation point was presented again for 83 ms, and finally the first image of the pair was presented again (833 ms). To ensure that the fixation dot did not act as a marker to any small object shifts near the center of the stimulus, it was extinguished when stimuli were present. The fixation dot was present during each of the 83 ms gray screen intervals—and prior to the presentation of the first image. At the end of the trial the observer made a numerical magnitude estimation rating of the perceived difference between the two images. A random number between 10 and 30 appeared at the center of the screen, and the observer modified this number using a CRS CB6 response box until their choice of difference rating was reached.

Observers viewed the stimuli binocularly; they were asked to gaze at the central fixation point between image presentations and to maintain their gaze on the center of each square image during presentations. We could have presented the two images under comparison simultaneously, side by side (Charrrier, Maloney, Cherifi, & Knoblauch, 2007; Maloney & Yang, 2003; Wang & Simoncelli, 2008) and the observer could have looked back and forth until they were happy with their chosen rating magnitude; and, indeed, this is presumably how the reader would look at the examples of our stimuli. However, particularly for the modeling, we wished to constrain the time available to view each image and to constrain the observer’s gaze to one known point, i.e. the center of the image.

The 83 ms interval between presentations was long enough that observers could not gain any cue about potential image differences from apparent motion of objects in changed positions. The featureless display during the interval was not intended to be a distractor, and the image changes were generally clear and unsurprising. Although the presence of the blank interval made some image pairs harder to detect, we were not trying to imitate some kinds of “change blindness” paradigm (Simons & Rensink, 2005) where large changes are disguised by the nature of the image transition or interval. Indeed, the observers over-rated image pairs where objects appeared or disappeared (Figures 4 and 5). Of course, our observers had to hold a memory of one scene and compare it to a subsequent modified scene; change blindness experiments reveal that the function of this memory is rather limited.

It might have been adequate to present each of the stimuli under comparison only once each. However, in our experiments, the subject matter of the images and the nature of the image change varied widely. During the presentation of the first member of each image pair, the observer would generally have had little idea as to what the likely change would be, so that the state of the second image might have been surprising; we decided, therefore, to re-present the first image again after the second image, to allow the observer a fairer opportunity to make a decision about the change. Furthermore, the 3 presentations might have allayed any order effects. Because the order in which images from each pair (say A and B) are presented is randomized, a two-presentation paradigm meant that observers were presented with A → B → A or B → A → B, i.e. both transitions.

Observer training and instructions, ratings and experimental design

Before an experiment, each observer underwent a training session, when they were asked to rate 40–60 image pairs of the various types of image differences that could be presented to them later. During both the training and the testing phases, observers were frequently presented with one particular standard image pair (Figure 1 bottom) whose magnitude difference was defined as ‘20’. Observers were instructed that their ratings of the subjective difference between any other image pair must be based on this standard pair using a ratio scale, even when the test pairs differed along different stimulus dimensions from the standard pair (details of the instructions and training are in To, Lovell, Troscianko, & Tolhurst, 2008 and in the Auxiliary File). All other images (apart from the standard image pair) used in demonstration or training phases were different from those in the testing phase proper.

The testing phase was divided into blocks of 100–150 image pairs, lasting 30–45 minutes. Each block started with the presentation of the standard image pair, which was subsequently presented after every 10 further trials to remind the observers of the standard difference of ‘20’. The image presentation sequence was randomized differently for each observer.

Observers were recruited from the student or post-doctoral researcher populations at the University of Cambridge, UK. They all gave informed consent. They had normal vision after prescription correction, as verified using Landolt C acuity chart and the Ishihara color test (10th Edition). While some observers participated in more than one of our experiments, they remained naïve to the purpose of each. Table 1 in the Auxiliary File lists some details of the various experiments, including the number of observers.

Construction of stimuli

Photographs of natural scenes were taken with various calibrated Nikon still digital cameras and a JVC camcorder (see Auxiliary File), and 256 by 256 pixel segments of
these photos were used to make stimuli. Figures 1 and 3 give thumbnail examples of stimuli (larger images and further examples are in the Auxiliary File). The stimulus pixel values were corrected for the luminance nonlinearities of the camera used to take the parent photograph.

Normal image pairs. Stimulus images were made from photographs of 6 broad and partly overlapping thematic categories, to have a well-balanced variety of image types: animals, landscapes, objects, people, plants and garden/still-life scenes (see Figure 1 left, the Auxiliary File, and further examples in To et al., 2008). Each category contained 30 parent images, each matched with 5 variants (all different from the parent), giving 900 different image pairs in total. For 325 pairs, the variant was from a second photograph of the same scene taken when, say, an item had moved or when natural changes in the illumination had changed the shadowing; these image changes were entirely natural. Other variants were made from originals using PaintShopPro v.8 (JASC Software) or code written in Matlab (The Mathworks). We have retrospectively grouped the different changes into the following partially overlapping categories (Table 2 gives numbers in each category):

Appear: Image pairs where the only change was in prominent objects that appeared, particularly at fixation.
Making negatives. We also wanted to make stimuli with the same spatial and color complexity as the natural ones, but with the semantic content difficult to discern. Inverting or negating images of faces or some objects makes them difficult to recognize (see Discussion). It is easy enough to present images upside down and we did, in fact, run a variant of the main experiment where 180 of the image pairs (one pair per parent image) were presented both upright and inverted, randomly interleaved. However, we could not make negatives by simply subtracting each pixel value in the linearized stimuli of the main experiment from 255, since this would make most images look overly bright and desaturated. Instead, we made pixel-reversed images (To et al., 2008) by reversing the rank order of the pixel values in the R, G and B planes of the original image (e.g. Figure 1B). These images had the same first-order statistics as the originals but there were some changes in the spatial organization. In images with a wide color gamut, the process would have resulted in saturated negative-like stimuli, but the changes in some image pairs were largely in luminance rather than hue (see Auxiliary File). We also attempted to resolve the saturation problem by transforming (see Auxiliary File) the pixel values of the desaturated true negatives to give pseudo-negative stimuli that looked almost as saturated as the original images. In experiments with pseudo-negatives or pixel-reversed images, the standard image pair (defined as having a difference of ‘20’) was the same, normally colored “Lily” pair shown in Figure 1 bottom; its color values were not distorted.

Data collation and statistical analysis

In each experiment, the ratings given by the several observers were averaged together for further analysis. The results for each observer were first divided by that observer’s median value for the experiment (typically, the median was about 20, close to that demanded for the standard pair); we normalized against the median instead of the mean because we could not assume that observers’ ratings were normally distributed. Then the normalized ratings of the several observers were averaged together stimulus-by-stimulus; for convenience, these averages were finally multiplied by the grand average of all the ratings of all the observers in that experiment to bring the values into the range of rating values actually given by observers during the experiments. Standard errors are shown in Figure 5, being greater for higher averaged ratings, as if different observers used different scales when rating very large perceived differences. We did not feel it useful to compress or stretch each observer’s ratings to fit within a specified range.

The models cannot predict the actual rating magnitudes, but only numbers that we hypothesize will be directly proportional to the ratings; our results (see especially Figure 7) are compatible with this. To summarize model fits, we use Pearson’s correlation coefficient (r) between the experimentally measured ratings and the numerical output of models. To examine whether the ratings for some kinds of image difference are better fit than other kinds by a model, we use a z-score difference (Figure 5). The averaged ratings of the observers were reduced to z-scores (subtract the mean of the 900 rating values, and divide by their standard deviation). The numerical outputs of the model were also reduced to z-scores, and these were subtracted from the z-scores of the ratings for the
appropriate image pairs. Then, for a given image change (e.g., “blur” or “color” changes), we average the z-score differences for relevant image pairs. An average z-score difference of zero implies that the model was not, in general, over- or underestimating the ratings for that image category of change.

### An optimized V1-based model

The model involves two major steps in the processing of the images: linear convolutions, followed by non-linear effects. A linear first stage seems justifiable physiologically (Jones & Palmer, 1987; Movshon, Thompson, & Tolhurst, 1978a), even with natural images (Ringach, Hawken, & Shapley, 2002; Smyth, Willmore, Thompson, Baker, & Tolhurst, 2003).

The R, G and B planes of each colored image in a pair were transformed into “cone space”—L, M and S planes (Smith & Pokorny, 1975) from knowledge of the emission spectra of the 3 phosphors on our Sony display CRT. It is widely accepted that human vision works on a “luminance channel” and two “color opponent channels”, R/G and B/Y (Hurvich & Jameson, 1957) although it difficult to see such a clear distinction between “luminance” and “R/G” in the lower levels of the monkey visual system (DeMonasterio, Gouras, & Tolhurst, 1975; Derrington, Krauskopf, & Lennie, 1984; Johnson, Hawken, & Shapley, 2001; Lennie, Krauskopf, & Sclar, 1990); we presume that the color-opponent channels are those studied with isoluminant gratings (Mullen, 1985). We therefore transformed the L, M and S planes into a luminance (L + M) plane and two MacLeod and Boynton (1979) color-opponent planes (L/(L + M) and S/(L + M)). Since there is little sub-threshold summation between isoluminant and luminance gratings (García, Nieves, Valero, & Romero, 2000; Gegenfurtner & Kiper, 1992; Giulianini & Eskew, 1998; Mullen & Losada, 1994), we subsequently processed the three planes entirely separately until the final Minkowski summation step (Equation 4; Ferwerda et al., 1997).

We chose the same orientation and spatial-frequency bandwidths for the channels in the luminance as for the chromatic planes (Beaudot & Mullen, 2005; Bradley, Switkes, & De Valois, 1988; Losada & Mullen, 1994, 1995; Werner, 2003).

### Receptive field models

In processing each of the chromatic planes, the first stage was to convolve the plane with 60 Gabor-function receptive-fields. At each of 6 orientations (30 deg steps), there were 5 spatial frequencies ranging from 4 to 64 cycles across the 256 pixel stimuli (1.25, 2.5, 5, 10, 20 cycles per degree), with even- and odd-symmetric fields for each of the 30 bands. The Gabor function is a reasonable model of real simple-cell receptive-fields (Daugman, 1985; Field & Tolhurst, 1986; Jones & Palmer, 1987; Marcelja, 1980; Ringach, 2002), but there is little neurophysiological support for the common idea that fields might fall into discrete odd- and even-symmetric classes (but see Ringach, 2002) or other quadrature pairings (Pollen & Ronner, 1981). The linear outputs of the convolutions were divided by measures of the local mean value in the image plane (after convolution with 2D Gaussian blobs with the same spread as the Gabor receptive fields) to ensure that the “response” of each element in the model was related to stimulus contrast (Peli, 1990; Tadmor & Tolhurst, 1994).

We did not implement pyramid models with fewer fields at low-frequencies than at high (for which there is scant neurophysiological evidence; De Valois, Albrecht, & Thorell, 1982; Movshon, Thompson, & Tolhurst, 1978c; Tolhurst & Thompson, 1981). A pyramid would not alter a model’s outcome, but would require differential weighting of the information provided by the different frequency bands. We weighted the contrast output of each frequency or chromatic band according to a human observer’s foveal contrast sensitivity for luminance or isoluminant sinusoidal gratings of the field’s center spatial frequency (from Mullen, 1985; Mullen & Kingdom, 2002). The output was also weighted according to how sensitivity to sinusoidal gratings falls off with eccentricity from the fovea (Foley, Varadharajan, Koh, & Farias, 2007; Pointer & Hess, 1989; Robson & Graham, 1981); we modeled a frequency-dependent, radially symmetric exponential fall-off in sensitivity with a factor of 10 for every 40 cycles of the field’s center frequency.

**Elongated receptive fields.** It seems conventional to model Gabor fields with a circularly symmetric Gaussian envelope, whose standard deviation is scaled in proportion to the period of the carrier sinusoid (see Figure 2A). However, neurophysiological studies of tuning bandwidths in V1 of cat, monkey and ferret suggest that fields are neither circular overall nor self-similar (Baker, Thompson, Krug, Smyth, & Tolhurst, 1998; De Valois et al., 1982; Movshon, 1979; Tolhurst & Thompson, 1981). First, neurons are typically more sharply tuned for orientation than for spatial frequency, implying that their fields are about 1.5 times longer along the axis of their preferred orientation than across the orthogonal axis. Second, neurons with low optimal spatial frequencies have broader bandwidths than neurons with high optimal (see also Blakemore & Campbell, 1969; Daugman, 1984; Wilson, McFarlane, & Phillips, 1983), as if there is a tendency for all neurons to have the same overall field size whatever their optimal frequency. Thus, we have modeled receptive fields with more realistic elongated shapes (Figure 2B). The Gaussian envelope was 1.5 times longer along the axis of preferred orientation; Foley et al. (2007)
modeled psychophysical thresholds with length ratios of about 1.2. The bandwidths were graded with optimal frequency: for fields with optima of 1.25, 2.5, 5, 10, 20 cycles per degree, the frequency bandwidths were 2.12, 1.43, 0.93, 0.64, 0.45 octaves and the orientation bandwidths were 43.4, 34.5, 22.8, 17.7, 11.6 degrees. These values lie in the ranges found for single neurons and deduced for psychophysical channels.

“Complex cells”. In an attempt to explain the “texture or small movement” outlier class in the experiments (Figures 4, 5, and 6), we modeled phase-invariance by taking the r.m.s. of the outputs of the paired even- and odd-symmetric fields in each orientation/frequency band (Adelson & Bergen, 1985; Heeger, 1992), perhaps reflecting the behavior of complex cells (Hubel & Wiesel, 1962; Movshon, Thompson, & Tolhurst, 1978b).

Non-linear interactions

Non-specific suppression, or contrast normalization. There are many neurophysiological observations which imply (Heeger, 1992) that a given V1 neuron is inhibited by all other neurons whose receptive fields lie within the same part of visual field (e.g. Bonds, 1989; Carandini et al., 1997; DeAngelis et al., 1992; Li et al., 2005). Inhibition that is relatively unspecific for orientation

Figure 2. Gray-level representations of the receptive-field shapes used. A (self-similar fields) and B (bandwidth graded with optimal frequency) show even-symmetric, vertically oriented fields at the 5 different optimal spatial-frequencies. The 3 leftmost thumbnails show the full 256 by 256 pixel representations of the fields. The rightmost 2 are for the highest spatial frequencies and, to show the tiny fields, the thumbnails have double the magnification and show only 128 by 128 pixels. C. The geometry of the annular “surrounds” (with rad of 0.77 periods) used to calculate the orientation-specific and spatial-frequency specific surround suppression.
or spatial frequency has been incorporated into models of psychophysical performance, particularly to model “crossed-orientation masking” (Foley, 1994; Peters et al., 2005; Rohaly et al., 1997; Teo & Heeger, 1994; Watson & Ahumada, 2005; Watson & Solomon, 1997). At each location (x, y) in the image, we calculated a suppressing signal by summing the contrast responses (e_{f,o,s}) of the 60 fields centered at that point (across frequency f, orientation o, and symmetry s), each raised to a power q:

\[ N_{x,y} = \sum_{f=1}^{f=5} \sum_{o=1}^{o=6} \sum_{s=1}^{s=2} |c_{x,y,f,o,s}|^q. \]

This one non-specific signal will suppress the responses of all 60 fields at the location equally (Equation 3).

**Orientation-specific surround suppression.** While the suppression within the receptive field may be modeled to a first approximation as being orientation non-specific, it is clear that the suppression that arises from stimuli in the surrounding visual field (i.e. outside the receptive field) is specific for orientation (Blakemore & Tobin, 1972; Cavanaugh et al., 2002). It may be that there are several overlapping classes of suppression, which have not yet been clearly distinguished or parameterized. Peters et al. (2005) have a saliency model that includes surround suppression, and several psychophysical masking paradigms are explained by surround suppression (e.g. Meese, 2004; Polat & Sagi, 1993). We modeled surround suppression as if it is strictly specific for orientation and spatial frequency, and arises with equal strength from all directions surrounding the field although this symmetry may not fit with neurophysiology (Cavanaugh et al., 2002). For a receptive field centered at a given (x, y) location, the surround strength was modeled as a radially symmetric annulus centered on that point:

\[ \text{surround strength}_f = d \cdot e^{-d^2/(2rad_f^2)}, \]

where \( d \) is distance from the center of the field, and \( rad_f \) is the radius of the annulus which was set to be directly proportional to the period of the center spatial frequency of the receptive field (Figure 2C). The annulus was normalized to have a volume of unity and was the same shape for the luminance as for the color-opponent planes. It will be noted from Figure 2C that \( rad_f \) was small (less than 1 period), so that “surround” suppression is probably not the best terminology for the process (see Cavanaugh et al., 2002). For each of the 30 orientation/spectral frequencies, we first took the r.m.s. of the odd- and even-symmetric field responses. Then, these r.m.s. values were raised to a power r, before being convolved with the annulus appropriate to that spatial frequency. This gives 30 different maps of surround-suppression signals, \( S_{x,y,f,o} \),

**Sigmoidal transducer function, image comparison, and the contrast “dipper”**

The contrast responses of the 60 receptive field types were finally subjected to these two non-linear suppressive effects, using a modified version of the Naka-Rushton equation (Foley, 1994; Heeger, 1992; Watson & Solomon, 1997). The response of the field at location (x, y), frequency f, orientation o, and symmetry s was:

\[ \text{response}_{x,y,f,o} = \frac{\text{sign}(c_{x,y,f,o}) \cdot |c_{x,y,f,o}|^p}{1 + W_N \cdot N_{x,y} + W_S \cdot S_{x,y,f,o}}, \]

where \( W_N \) and \( W_S \) are weights, and the calculation of \( N_{x,y} \) and \( S_{x,y,f,o} \) involved raising response values to powers q and r respectively, as described above. The powers p, q and r ensure a quasi-sigmoidal “response transducer function” (Legge & Foley, 1980). The parameters were the same for all orientations and spatial frequencies and for all 3 chromatic planes. We suppose that this is not the relationship between response and contrast for a single neuron, but is a pooling (Chirimuuta & Tolhurst, 2005; Heeger, Huk, Geisler, & Albrecht, 2000; Watson & Solomon, 1997) of the transducers of a population of neurons each with a limited dynamic range and with response variance that increases with increasing response level (Albrecht & Hamilton, 1982; Geisler & Albrecht, 1997; Tolhurst, Movshon, & Dean, 1983; Tolhurst, Movshon, & Thompson, 1981; Tolhurst, Smyth, & Thompson, 2009).

**Final pooling of all the difference cues**

We finally have a model of the responses or outputs of all the neurons to a given image. The process was repeated for the comparison image, and we subtracted the model outputs for the two images point-by-point through the 3 chromatic planes, 5 spatial frequencies, 6 orientations and 2 symmetries (see comments in Auxiliary File). Subtraction of responses that have been transduced through a sigmoid (Equation 3) leads naturally to the familiar contrast discrimination “dipper”.

The many visibility cues across x, y, frequency, orientation, symmetry, and color plane were combined into a single value by Minkowski summation (Graham, 1989; Quick, 1974; Robson & Graham, 1981; Rohaly et al., 1997; Wang et al., 2004; Watson & Solomon, 1997) with power m.

\[ \text{overall difference} = \left( \sum_i (\text{difference cue}_i)^m \right)^{1/m}. \]

The n (11.8 million) individual visibility cues were raised to the power m, summed and the nth root taken (compare Equation 5). Minkowski summation, at least at threshold,
Fixing the parameters of the models

The Minkowski power, power exponents, weights and surround radii for the 2 different receptive field models are listed in Table 1. Within a model, the same parameter values applied to all spatial frequencies, all orientations and all chromatic planes. The Minkowski power and the Naka-Rushton power exponents $p$, $q$ and $r$ are similar to values reported in other models (Foley et al., 2007; Legge & Foley, 1980; Watson & Ahumada, 2005; Watson & Solomon, 1997). The parameter values were obtained by least-squares fitting of the models to contrast discrimination “dipper” experiments conducted by changing the contrast of patches of monochrome images of natural scenes (Chirimuuta, Jiwa, & Tolhurst, 2007; Chirimuuta & Tolhurst, 2004). The values were then refined (though this made little difference to the final fits) by fitting the models to a subset of the ratings data shown in Figure 4.

Experimental results

Reliability of the ratings measurements

If ratings are subjective, do they provide reliable and consistent measurements, which are similar between observers? The Auxiliary File discusses measures of within-observer and between-observer consistency. In summary, when individual observers repeated experiments or performed experiments on two very similar sets of images (e.g. upright versus inverted images), the correlation between the 2 sets of ratings was typically about 0.72. The correlation between the ratings given by any one observer and those given by every other observer in an experiment was generally lower, typically about 0.55.

Most of our experiments have used 8 or more observers, and we have scaled and averaged their ratings to each image pair (see Methods). While this is intended to average out within-observer variability, it will also obscure any between-observer differences in strategy, if such exist. Averaging together the results of observers produces useful data sets for modeling: these correlation coefficients are all substantially higher than the correlations between the experimental results and the predictions of our models.

The robustness of the averaged ratings across observers can be demonstrated by how clearly they allow us to test the rules by which observers combine difference cues when image pairs differ in more than one way (discussed in detail by To et al., 2008; To, Troscianko, & Tolhurst, 2009). Can the rating to an image pair with two differences be predicted well by knowing the ratings to the two image pairs where the changes are made singly? From the main and the pixel-reversed experiments, we extracted 272 image-combination groups where there was one change from the reference in the first pair, a second change in the second pair, and both changes in the third pair (e.g. Figure 3A and Auxiliary File). The nature of the two changes differed considerably across the data sets. Using the grand averages of the ratings given by the several observers to each of the 3 image pairs in a combination set, we investigated a number of possible summation rules (e.g. city-block summation, Euclidean summation and maximum; Shepard, 1987), but were most impressed by Minkowski summation whose exponent we found by least squares fitting (To et al., 2008):

$$\text{predicted combo rating} = \sqrt[2.78]{\text{rating}_{1}^{2.78} + \text{rating}_{2}^{2.78}}.$$  

The exponent of 2.78, is of course, very similar to that used in V1-based models that predict the detection thresholds for combinations of gratings (see especially Foley et al., 2007; Watson & Ahumada, 2005) and to the values implemented in our own models (Table 1). The correlation between the predicted and measured ratings for the combination sets was very high ($r = 0.93$; Figure 3B) implying very robust data.

Ratings of perceived differences between normal image pairs

Difference ratings for 900 image pairs were obtained from 11 naïve observers. The average of the normalized ratings for each image pair is plotted against the simple-cell model predictions in Figure 4. Similar patterns were observed for the different thematic categories (animals, landscapes, people, plants, objects and scenes; see Methods). Overall, the correlation between the average
normalized ratings and model predictions was 0.51; although this is low, the high degrees of freedom (898) make this statistically very significant.

We examined model performance for image pairs that changed along only one dimension (see Methods), such as "appear", "blur", "color", "shape" and "texture" (Table 2). Model predictions for "color" (Figure 4A cyan) and "texture" (Figure 4B red) were satisfactory, with correlation coefficients against the model of 0.548 (n = 273) and 0.565 (n = 134) respectively, higher than for the overall trend. The fit is particularly good for the class "appear" (Figure 4C magenta) with r of 0.698 (n = 82), but the fits to "blur" and "shape" are disappointing. In fact, the correlation with "blur" is negative (r = -0.234; n = 36) perhaps because the ratings and the model outputs for the blurred image pairs covered too small a range to show a straight-line relation. The results for the different image change classes do not all lie along the same trend line. The observers’ ratings for image pairs containing "texture" changes (and small object movements) were clearly overestimated by the model (Figure 4B), while the ratings for pairs with "appear" and "blur" changes were generally underestimated (Figure 4C).

The systematic mismatches between the observers’ averaged ratings and the predictions of the model are summarized in Figure 5A. For various categories of image pair, the gray histogram blocks (±1 standard deviation) show the z-score difference (see Methods) averaged across the number of pairs in that category. The measure essentially shows whether data for a particular image-change tended to be above or below a single overall trend-line relating rating to model prediction. "Color" and "shape" changes are moderately well predicted by the model on average, but the ratings for "blur" and "appear" are substantially underestimated by the model (positive z-score differences) while "texture and small movement" changes are substantially overestimated (negative value). The category "other" (see Methods) includes 31 image pairs where the major change was in shadowing caused by changes in natural illumination (e.g. Figure 1F). There were also 15 image pairs based on "faces" (see Auxiliary File for an example). The average z-score differences for "shadows" and "faces" are also shown in Figure 5. Observers gave lower ratings for changes in shadowing that predicted by the model, while they gave larger ratings for changes in faces.

We repeated this experiment with 100 ms image presentations instead of 833 ms to ensure that observers did make judgements without gazing around the stimuli. The 10 new naïve observers’ ratings for 100 ms presentations were highly correlated with those obtained in the main experiment (833 ms): r = 0.90. The only difference was that some image changes seemed not to be detectable at the shorter viewing time.

**Attempts to remove “high-level” semantic content from the stimuli**

The remainder of this paper asks why certain categories of image change are systematically wrongly predicted. One possibility is that the “high-level” semantic content of...
these naturalistic images influences the rating, while we have attempted to model only low-level visual processes. Since image inversion interferes with recognition and interpretation of faces and some objects or scenes (see references in Discussion), we ran a reduced experiment in which 180 of the normal image pairs were presented inverted as well as upright, randomly interleaved. The normalized averaged ratings of the 7 observers for the upright and inverted versions of image pairs were highly correlated \((r = 0.91)\). A paired Student’s \(t\)-test comparing the averaged ratings for inverted and upright images showed that image inversion by itself had little effect \((t = 1.42; n = 180; P = 0.08)\).

**Inverted negative images.** Negating images is another method of removing higher-level information. We therefore tried 2 different ways to make inverted negatives of the image pairs (see Methods). Figure 6A plots the ratings for inverted pixel-reversed images against the predictions of the simple-cell model. The overall correlation between ratings and model rose from 0.51 in Figure 4 to 0.64 \((z = 4.14, P < 0.001; \text{Howell, 1992})\), but the several image-change types are still systematically misplaced. The pink histogram blocks in Figure 5A show the \(z\)-score differences for the ratings and modeling of these color-distorted images, while Table 2 lists the correlations between rating and model for the different image-change categories. The averaged \(z\)-score differences with the pixel-reversed images are mostly closer to zero than for the normal pairs, but the systematic divergences remain. The results show that our attempt to remove high-level content has led to improvement of the model’s performance in all measures. Interestingly, interference with the semantic content of the images has caused “shadow” changes to be better fit by the model; a paired \(t\)-test of the \(z\)-score difference for the 31 shadow pairs in the normal and pixel-reversed experiments was very significant \((t = 6.34; n = 31)\). Furthermore, there has been a substantial change in the pattern for the “faces” category (paired \(t = 4.85; n = 15\)) as we would expect (Bruce & Langton, 1994; Haxby et al., 1999; Perrett et al., 1988; Thompson, 1980; Yin, 1969) from inverting and negating face images.

The pixel-reversed images were not simple negatives and there were spatial differences as well as color or luminance differences. Furthermore, different observers participated in the experiments with normal images (Figure 4) and pixel-reversed images (Figure 6A). We were therefore unable to compare results for the two kinds of image directly. Thus, we conducted another experi-
ment, where the stimulus set comprised 450 normal image pairs (normal N pairs) from the first experiment and their 450 inverted pseudo-negative variants (IN pairs), created in a way that did not distort the spatial structure of the images (see Methods); the N and IN image pairs were randomly interleaved. The ratings for IN pairs are plotted against the ratings for normal pairs in Figure 6B along with a line of equality. The averaged ratings for normal

<table>
<thead>
<tr>
<th>Change</th>
<th>n</th>
<th>Simple cell</th>
<th>Complex cell</th>
<th>Simple cell</th>
<th>Complex cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>900</td>
<td>0.514</td>
<td>0.589</td>
<td>0.638</td>
<td>0.725</td>
</tr>
<tr>
<td>Appear</td>
<td>82</td>
<td>0.698</td>
<td>0.761</td>
<td>0.620</td>
<td>0.723</td>
</tr>
<tr>
<td>Blur</td>
<td>36</td>
<td>−0.234</td>
<td>−0.184</td>
<td>0.043</td>
<td>0.186</td>
</tr>
<tr>
<td>Color</td>
<td>273</td>
<td>0.548</td>
<td>0.582</td>
<td>0.619</td>
<td>0.749</td>
</tr>
<tr>
<td>Shape</td>
<td>114</td>
<td>0.344</td>
<td>0.426</td>
<td>0.522</td>
<td>0.608</td>
</tr>
<tr>
<td>Texture</td>
<td>134</td>
<td>0.565</td>
<td>0.592</td>
<td>0.719</td>
<td>0.754</td>
</tr>
<tr>
<td>Other</td>
<td>261</td>
<td>0.561</td>
<td>0.582</td>
<td>0.667</td>
<td>0.714</td>
</tr>
<tr>
<td>Shadows</td>
<td>31</td>
<td>0.647</td>
<td>0.704</td>
<td>0.674</td>
<td>0.632</td>
</tr>
<tr>
<td>Faces</td>
<td>15</td>
<td>0.711</td>
<td>0.722</td>
<td>0.696</td>
<td>0.804</td>
</tr>
</tbody>
</table>

Table 2. The correlation coefficients between the ratings for particular categories of image change and the predictions of the simple cell model and the complex cell model, for the 900 normal image pairs and for the 900 inverted and pixel-reversed images. The image pairs categorized as “shadows” or “faces” are subsets of other categories.

Figure 5. A, For several categories of image change, the z-scores of the observers’ averaged rating are compared with the z-score of the simple-cell model’s predicted ranking of the rating. A positive z-score difference indicates that the model underestimated the observers’ rating. The bars show the averaged z-score difference for the category, ±1 standard deviation. Gray blocks are for normal image pairs, while pink blocks are for inverted pixel-reversed images. The numbers in brackets show the number of image pairs in each category.

B: 900 normal images

C: 900 pixel-reversed images
pairs were highly correlated with those for IN pairs \((r = 0.84)\), but a paired Student’s t-test showed that inverting and pseudo-negating the image pairs significantly lowered the ratings \((t = 15.73; n = 450; P < 0.01)\), perhaps because negatives tend to have lower saturation and contrast, despite our attempt to equate the pixel distributions. The correlation between the model and average observers’ ratings was 0.54 for the N pairs (much as in the main experiment) and rose for the IN pairs but only to 0.56. This increase in correlation coefficient is in the same direction as for the pixel-reversed images but is not statistically significant \((z = 0.27, P = 0.54)\). The z-score differences for different image categories (including “faces”) were affected by inversion and pseudo-negation in the same way as for the pixel-reversed images of Figure 5A.

One reason that negation and inversion give a small improvement in correlation with the model is that the observers seemed to be more consistent in their ratings. The standard errors of the ratings were slightly smaller in general for the inverted, pixel-reversed images (Figure 5C) than for the normal pairs (Figure 5B): 10%, 50%, 90% percentiles of 1.24, 2.30, 3.64 “rating units” compared to 1.38, 2.49, 5.02 (Kolmogorov–Smirnov 2-sample, \(P \approx 0\)). The difference in the 90th percentile reflects the presence of a “tail” of large standard errors for some normal image pairs, mostly ones for which the average rating was especially high. Presumably, individual observers have maintained reasonably consistent rating scales and strategies but, when the images have recognizable content, different observers tend to have differently exuberant scales for some of the image pairs with large differences. The standard errors for the ratings of the pseudo-negatives (Figure 6B) also had less of tail at high values than did the errors for normal images.

A phase-invariant model

Figure 7A shows how the ratings for the 900 normal image pairs were predicted by the phase-invariant (“complex cell”) version of the model, while Figure 7B shows how the ratings for the inverted pixel-reversed images were predicted by the complex cell model. The phase-invariance has significantly improved the model’s predictions compared to the elongated simple-cell model (Table 2). For the 900 normal image pairs, the correlation has improved from 0.52 to 0.59 \((P < 0.01)\) while, for the pixel-reversed images, the improvement is from 0.64 to 0.73 \((P \approx 0)\). The correlations between the ratings and the complex-cell model for different image changes are listed in Table 2. The correlation for all changes was improved compared to the simple cell model, and is generally high enough that we can see from the graphs of Figure 7 that the relation between ratings and model predictions is roughly linear. This justifies our hypothesis that the numerical output of the computer models will be directly proportional to the rating given by the observer.

![Graph A](image1.png)

**Figure 6. A.** The ratings for the 900 inverted pixel-reversed images are plotted against the predictions of the simple-cell model. Ratings for image pairs with “appear” (magenta) and “blur” (blue) changes were still generally underestimated by the model, while pairs with “texture” changes (red) were still overestimated. **B.** The normalized averages of 11 observers’ ratings for 450 inverted pseudo-negative images are plotted against the same observers’ ratings for the normal versions of the images. Ratings for pseudo-negative images were generally lower than those for normal natural scenes. The line of equality is shown.
We implemented the phase-invariant model to test whether it would deal better with “texture” or “small movement” changes in images. For the 900 normal image pairs, Figure 7C plots the z-score difference for the complex cell model against that for the simple cell model. The lower left quadrant shows image pairs where both models overestimated observers’ ratings (the red symbols show “texture” change pairs), while the upper right quadrant shows pairs where both models underestimated the ratings (“blur” in blue, “appear” in magenta). If the complex-cell model had given a consistent improvement, the data would be rotated clockwise about the [0,0] point with respect to the line of identity. There seems to be little overall systematic rotation, but there has been a small but consistent improvement (a clockwise rotation) in the fits for the “texture” class (red symbols). The fits for 40 “others improved” image pairs (shown in green) changed similarly; 6 of these were combination pairs where “texture” was one of the elements, while 4 were subtle changes in dappled “shadows” which might be considered as “texture” changes (see example in Auxiliary File). However, 25 out of the remaining 30 improved pairs were in the “shape” class (either alone or in combination), but many of these changes were subtle, although some seemed rather obvious (see the Auxiliary File for some examples of improved image pairs). Thus, phase-invariance in the model did improve the model’s performance with small spatial changes, but the improvement was small. We did not find any image pairs where phase-invariance was detrimental because, say, bright features were confused with dark.

**Discussion**

We have measured the perceived difference for a great variety of naturalistic changes in very many disparate images of natural scenes, so that we have a large and challenging data set to evaluate neurophysiologically inspired models of visual difference perception. However,

Figure 7. A, The ratings for the 900 normal image pairs are plotted against the predictions of a phase-invariant (“complex cell”) V1 model. “blur” blue symbols, “appear” magenta symbols, “texture” red symbols. Compare with Figure 4. B, The ratings for the 900 inverted pixel-reversed image pairs are plotted against the predictions of the phase-invariant (“complex cell”) V1 model. “blur” blue symbols, “appear” magenta symbols, “texture” red symbols. Compare with Figure 6A. C, The z-score differences for normal image pairs for the “complex cell” model are plotted against those for the elongated simple-cell model. “blur” blue symbols, “appear” magenta symbols, “texture” red symbols. The line of equality is drawn. The “others improved” green symbols are for image pairs with negative z-score differences that were improved by the “complex cell” model but were not in the “texture” class.
our model predictions were only moderately correlated with observers’ difference ratings, with correlation coefficients below those reported in other modeling studies (e.g. Barten, 1990; Feng & Daly, 2003) which usually have reported results for a limited variety of changes in a limited number of parent images. We too have performed experiments with a limited variety of naturalistic change types in families of stimuli based on few parent images (Experiment 3 of To et al., 2008), and the correlation between the model predictions and observers’ ratings was 0.85 (unpublished observation) showing that our methodology and modeling is capable of producing good correlations. It may therefore be that frequent changes in the high-level composition of the images, as occurs in the present experiments, may hinder the observer to maintain a constant internal reference against which to estimate difference ratings. The models, although they may not generalize well, can thus be usefully applied to quantifying visual performance in practical tasks where the stimuli or the natural changes tend to be of one or a few kinds.

The correlation in our experiments is not low simply because the greater variety in the task has led the observers to become more erratic. Rather, the model performed consistently differently for different kinds of image change: underestimating observers’ ratings for image pairs containing “appear” and “blur” changes and overestimating “texture and small movement” changes. The consistency and reliability of our observers’ rating scales is shown most impressively when they were used to distinguish between subtly different models of how observers combine cues when they give ratings for images that change in two ways at once (see Figure 3, and detailed discussions in To et al., 2008, 2009). The same cue combination rules applied even for those image change types such as “blur” which were the most obvious outliers when comparing ratings with the model.

The lower correlations between ratings and model found with normal image pairs may also result from the “semantic” content in the stimuli. Our model relies solely on low-level information to generate its predictions. It would be difficult to find naturalistic images devoid of high-level content, and it might be possible to synthesize surrogate images with the correct statistics from, say, Gabor patches (see Arsenault, Yoonessi, & Baker, 2009; Baker, Yoonessi, & Arsenault, 2008). We attempted to reduce observers’ being influenced by high-level cues by inverting and negating the image set because these procedures affect, e.g., face and shadow processing (e.g. Bruce & Langton, 1994; Haxby et al., 1999; Lovell, Gilchrist, Tolhurst, & Troscianko, 2009; Perrett et al., 1988; Rensink & Cavanagh, 2004; Thompson, 1980; Yin, 1969; but see Nederhouser, Yue, Mangini, & Biederman, 2007). Inversion alone had no significant effect on observers’ ratings but, when colors were also distorted, the ratings were slightly better predicted by the model. The improvement seemed to result mostly because the observers tended to agree more about the ratings to give to large image differences when the semantic content was harder to discern.

Observers’ ratings for “face” and “shadows” changes were noticeably affected by our attempts to remove high-level cues (Figure 5), but the manipulations made only a modest improvement to the overall model fit. Some observers did mention recognizing objects in the inverted and/or negated images, suggesting that these manipulations alone were unable to remove all higher-level cues. While the expression on an inverted face may be difficult to interpret (Thompson, 1980), it is still easy enough to tell that it is a face rather than, say, a garden scene. Perhaps, the effects of inversion are strongest for objects that are generally seen at one consistent orientation especially after much practice (Martelli, Majaj, & Pelli, 2005); our disparate stimuli probably contained many objects that appear naturally in many orientations.

Details of the V1-based models

In An optimized V1-based model section, we have listed the assumptions of our modeling and we have detailed the primary experimental evidence on which they are based. We have modeled elongated receptive fields with bandwidths dependent on their center spatial frequency, since this has greater physiological realism (e.g. Tolhurst & Thompson, 1981). However, we also modeled the more familiar circular and self-similar fields, producing much the same correlation with the experimental ratings: \( r = 0.47 \) for the normal image pairs and 0.64 for the inverted pixel-reversed images (compared to the values 0.52 and 0.64 we report for Figures 4 and 6). The phase-invariant complex cell model is a distinct improvement over the simple-cell models for our stimuli. It is worth pointing out that, when we tried to predict the ratings from the root mean squared difference between the pixel values of the images in the pairs, the correlations were only 0.28 and 0.47 for the two kinds of stimuli.

Our simple-cell and complex-cell models each contain 7 numerical parameters whose values we sought by iterative search. However, there are other numbers that we decided on and fixed (based on neurophysiological evidence), such as the orientation and frequency bandwidths of the neurons and how they depend on spatial frequency. Such numbers could have been additional variable parameters (Watson & Solomon, 1997). Certainly, the models have more parameters than our 7 explicit ones, but the real elegance of such models (Watson & Ahumada, 2005; Watson & Solomon, 1997) is how astonishingly few parameters they need to make a reasonable model incorporating millions of neurons. In fact, real neurons in V1 vary very considerably in such features as their bandwidths, contrast thresholds and phase symmetries (Field & Tolhurst, 1986; Tolhurst et al., 1983; Tolhurst &
Thompson, 1981), and their responses to natural images are heterogeneous in magnitude and sparseness (Tolhurst et al., 2009; Yen, Baker, & Gray, 2007). Thus, each neuron in our models might have had its unique behavior specified by its own set of 10–20 parameters; we could add an unmanageably large number of parameters to the models to greatly enhance their fits to the present experiments, but the predictive value of the models would likely fall.

We have made simplifications and generalizations in the models which are not strictly compatible with published psychophysical data. For instance, we have assumed that the sigmoidal transducer function has the same shape (same parameter values) at all frequencies, orientations and locations, and in all 3 chromatic planes. Medina and Mullen (2009) and Meese and Holmes (2007) have shown that the strength of crossed-orientation masking varies with spatial frequency and chromatic plane. Since such masking is speculated to arise from the non-specific normalization term in the transducer equation (Equation 3), we must presume that the equation’s parameters will also vary with spatial frequency and chromatic plane. Furthermore, we have assumed that the non-specific masking term is totally non-specific in orientation and frequency but that it is highly localized spatially (but see Watson & Solomon, 1997). Conversely, we have supposed that the surround-suppression term is entirely specific in orientation and spatial-frequency but that it arises equally from all directions around the receptive field; the true geometry of such suppression may be much more complex than this (Cavanaugh et al., 2002). We added the surround-suppression term to the Watson and Solomon based model, following Meese (2004), to explain the different shapes of contrast-discrimination dippers measured with gratings or Gabor patches of different geometry (Chirimuuta & Tolhurst, 2005; Meese, 2004).

It seems obvious that a model which deals with colored images should run 3 more-or-less parallel sets of computations: on a luminance, a red-green opponent and a blue-yellow opponent transform of the images (Hurvich & Jameson, 1957). But it is not really clear how those planes should be specified, and there is little neurophysiological evidence to suggest that the luminance and RG planes are really distinct. We used the MacLeod and Boynton (1979) formulation, where the red-green opponent plane is isoluminant. We suppose that the lowpass CSF of the chromatic planes is the envelope of the individual sensitivities of bandpass channels with the same bandwidths as the luminance-contrast channels (e.g. Losada & Mullen, 1994, 1995), but we wonder whether some chromatic processing might be performed by a spatial-frequency lowpass mechanism (Johnson et al., 2001; Vimal, 1997). We processed the 3 chromatic planes entirely separately, but there is some evidence of suprathreshold masking between the luminance and opponent planes (Mullen & Losada, 1994). Perhaps, the non-specific suppression term in Equation 1 should be summed across all 3 chromatic planes before its application (Equation 3).

While some of the details of the models are over-simplified, we suggest that their rectification will actually cause little improvement in the predictions of ratings for natural image changes. Although we have striven to include much physiological realism in the models of this paper, we find that these models are only marginally better at predicting ratings than our previous independent-filters model (Lovell et al., 2006; Tolhurst et al., 2005) which did not model contrast normalization or surround suppression. The correlations between that previous model and the ratings were 0.45 and 0.61 for the two kinds of stimuli. We suggest that it is the underlying philosophy of modeling low-level visual features that needs attention rather than the model details. Others have noted the inadequacies of such modeling. For instance, Osberger, Bergmann, and Maeder (1998) proposed that the models should concentrate only on the parts of images where there are objects of interest to observers; Wang et al. (2004) concentrate on the “structural similarity” of the images rather than low level features such as luminance and contrast.

### Systematic problems with low-level models, and future directions

We have described some systematic problems with the way that the observers’ ratings were predicted by the V1 models. The observers’ ratings for image pairs containing “texture” changes were mostly overestimated by the models, while ratings for image pairs with “appear” and “blur” changes were generally underestimated. The discrepancy for “texture” changes was slightly rectified by moving to a phase-invariant model. These 3 discrepant image-change classes raise questions about the fundamental assumptions of low-level V1 modeling.

**Blur:** The failure to model “blur” might be because the models do not give sufficient importance to changes in the high spatial-frequency content of the images. Georgeson and Sullivan (1975) found that, unlike lower frequency gratings, when high frequency gratings were just above threshold, they were immediately perceived as having high contrast. This suggests that weighting frequencies in the models only by their contrast thresholds may not properly represent the visibility and appearance of high frequency components in a scene. Thus far, the models have been based on the contrast thresholds for gratings, appropriately since they have been used to model detection processes (Rohaly et al., 1997; Watson & Ahumada, 2005; Watson & Solomon, 1997); however, a model based on suprathreshold grating appearance (Cannon & Fullenkamp, 1991) would seem more appropriate for the perception of suprathreshold stimulus changes.
Topography within models. The high ratings for changes in the “appear” category point to real problems in the philosophy of the modeling. The predicted rating of a difference depends on how many neurons in the model change their activity when presented with the second image and by how much their activity changes. It does not matter which neurons change their activity. If an object disappears, one set of neurons in the model will change their responses. But, if the object moves within the scene, there will be an additional cue: a second set of neurons at the destination location will change their responses. The model will propose that a moderate object movement is twice as salient as complete disappearance!

Furthermore, the model would predict much the same rating for a moderate object movement as for a large one, since the same number of neurons change their activity in the two cases; analogously it would predict much the same rating for a moderate change in a feature’s orientation as for a large change. Almost certainly, a human observer would give a bigger rating to what is clearly a bigger step through the “topography” of the neurons’ properties. Thus, whatever the exact details of a V1 model, the predictions of ratings for small movements, moderate movements and large movements (or total disappearance) may not rank properly with a person’s judgment.

Texture and small movements. The models are also very literal in that they compare two images point-by-point as if a neuron’s receptive-field views exactly the same point in the images. The less low-level metrics of Osberger et al. (1998) and Wang et al. (2004) still operate on point-by-point comparison between two images. Given drift in fixational eye movements, this is unrealistic and, furthermore, it may be that literal information about minute differences is discarded when making judgments about the visual world. Often, it matters only that there are some pebbles on the ground rather than, say, tree bark (see Figure 1 and a demonstration in the Auxiliary File), but it hardly matters exactly where each pebble is. Thus, V1-based models will over emphasize small image movements and, perhaps analogously, small changes in object size. In everyday visual tasks, the content of the scene is often more important than the exact detail. Incorporating phase invariance (“complex cell”) behavior into the model improved the predictions for differences in texture and small movements as hoped, but only to a small degree (Figure 7). Incorporating larger receptive fields with less-specific positional requirements (such as those reported beyond V1; Maunsell & Newsome, 1987) might be a necessary step in building a generalizable suprathreshold discrimination visual model. Moreover, an intelligent computational model must be able to locate areas of texture patterns (Hill, Canagarajah, & Bull, 2003; Lu, Dorsey, & Rushmeier, 2009; Portilla & Simoncelli, 2000) and treat them as tokens, so that the model compares classes of textures rather than detailed instantiations of textures. The model must ultimately identify salient objects and determine whether the detail within them provides useful information or whether they too can be treated as tokens (depending on the task).

Acknowledgments

This project was supported by grants to TT and DJT from the EPSRC/Dstl on the Joint Grant Scheme (GR/S56399/01, GR/S56405/01, EP/E037097/1 and EP/E037372/1). MPST and PGL were employed on those grants. We are very grateful to Dr. C. Ripamonti for designing and programming the pilot experiments in this series. We are very grateful to Dr. M.A. Gilmore, Dr. I.R. Moorhead, Prof. K.T. Mullen, Dr. C.K. Jones and Dr. C.A. Párraga for their helpful suggestions. Some of these results have been reported briefly (To, Lovell, Troscianko, & Tolhurst, 2006, 2007a, 2007b). The visual stimuli and rating results are available for download.

Commercial relationships: none.
Corresponding author: Dr. Michelle P. S. To.
Email: to@cantab.net.
Address: Department of Physiology, Development & Neuroscience, Downing Street, Cambridge CB2 3EG, UK.

References


Field, D. J., & Tolhurst, D. J. (1986). The structure and symmetry of simple-cell receptive-field profiles in the


Lovell, P. G., Gilchrist, I. D., Tolhurst, D. J., & Troschianto, T. (2009). Search for gross illumination...


