V1-based modeling of discrimination between natural scenes within the luminance and isoluminant color planes

Michelle P. S. To

Department of Psychology, Lancaster University, Lancaster, UK

David J. Tolhurst

Emmanuel College, University of Cambridge, Cambridge, UK

We have been developing a computational visual difference predictor model that can predict how human observers rate the perceived magnitude of suprathreshold differences between pairs of full-color naturalistic scenes (To, Lovell, Troscianko, & Tolhurst, 2010). The model is based closely on V1 neurophysiology and has recently been updated to more realistically implement sequential application of nonlinear inhibitions (contrast normalization followed by surround suppression; To, Chimimuuta, & Tolhurst, 2017). The model is based originally on a reliable luminance model (Watson & Solomon, 1997) which we have extended to the red/green and blue/yellow opponent planes, assuming that the three planes (luminance, red/green, and blue/yellow) can be modeled similarly to each other with narrow-band oriented filters. This paper examines whether this may be a false assumption, by decomposing our original full-color stimulus images into monochromatic and isoluminant variants, which observers rate separately and which we model separately. The ratings for the original full-color scenes correlate better with the new ratings for the monochromatic variants than for the isoluminant ones, suggesting that luminance cues carry more weight in observers' ratings to full-color images. The ratings for the original full-color stimuli can be predicted from the new monochromatic and isoluminant rating data by combining them by Minkowski summation with power $m = 2.71$, consistent with other studies involving feature summation. The model performed well at predicting ratings for monochromatic stimuli, but was weaker for isoluminant stimuli, indicating that mirroring the monochromatic models is not sufficient to model the color planes. We discuss several alternative strategies to improve the color modeling.

Introduction

One strand of vision research has been to ask whether psychophysical studies of human thresholds or discrimination can be interpreted quantitatively with the response properties of single neurons in model experimental animals; such comparisons have a long history (e.g., De Valois, 1965; Ratliff, 1965). We have been investigating the perception of spatiochromatic differences in naturalistic images and movies, and we have asked whether a neurophysiologically based computational model (after Watson, 1987) can explain the perceived magnitudes of such changes (To, Gilchrist, & Tolhurst, 2015; To, Gilchrist, Troscianko, & Tolhurst, 2011; To, Lovell, Troscianko, & Tolhurst, 2010). It is our aim to ask whether such a model will better explain human performance if it simulates neuronal response behavior with greater fidelity.

In our experiments, human observers provide magnitude estimation ratings of the suprathreshold differences they perceive between pairs of natural images (To et al., 2010). Some of the image pairs show truly natural differences: they comprise two photographs of the same scene taken at different times. Other image differences are imposed by computational postprocessing. Thus, images could change in whole or in part in terms of color (hue and/or saturation), spatial frequency distribution (blur or sharpening), content (objects moving, changing aspect, appearing or disappearing), texture, and shadows. To et al. (2010) used full-color (normal) scenes but also inverted pixel-reversed variants of these scenes, whose purpose was to remove any higher-level features and content that may influence observers’ ratings.

Our visual difference predictor model is originally based on other physiological models of visual discrimination (e.g., Watson, 1987; Watson & Ahumada, 2005; Watson & Solomon, 1997), and incorporates neurophysiological findings from quantitative studies of V1 neuron receptive fields (DeAngelis, Ohzawa & Freeman, 1993; De Valois, Albrect, & Thorell, 1982; Field & Tolhurst, 1986; Jones & Palmer, 1987; Movshon, Thompson, & Tolhurst, 1978a, 1978b, 1978c; Ringach,
in addition to linear summation mechanisms, the model includes two key nonlinearities: nonspecific contrast normalization (Bonds, 1989; Carandini, Heeger, & Movshon, 1997; DeAngelis, Robson, Ohzawa, & Freeman, 1992; Foley, 1994; Heeger, 1992; Tolhurst & Heeger, 1997), and orientation-specific surround suppression (Blakemore & Tobin, 1972; Cavanaugh, Bair, & Movshon, 2002; DeAngelis et al., 1994; Henry, Joshi, Xing, Shapley, & Hawken, 2013; Meese, 2004; Sceniak, Ringach, Hawken, & Shapley, 1999). Models of this kind have been good at explaining patterns of detection thresholds for monochromatic gratings and Gabor patches (To et al., 2017; Watson & Ahumada, 2005; Watson & Solomon, 1997) and detection experiments with monochromatic natural images (Párraga, Troscianko, & Tolhurst, 2005; Rohaly, Ahumada, & Watson, 1997; Tolhurst et al., 2010). In particular, the two nonlinearities are necessary to model contrast discriminations even with quite simple sinusoidal grating stimuli (Foley, 1994; Meese, 2004; To et al., 2017) and it would seem logical to begin any model of natural image discriminations by including them (Rohaly et al., 1997). One aim of the present paper is to investigate whether model performance is affected by the order in which the two nonlinearities (contrast normalization and surround suppression) are imposed. In their study of sinusoidal grating dipper functions, To et al. (2017), found that the best model predictions were obtained when the two inhibitory mechanisms were implemented sequentially rather than in parallel. This change to our 2010 model is more consistent with neurophysiological findings (e.g., Henry et al., 2013).

While such models are a credible description of the early foveal coding of monochromatic information, our interest is in studying the perception of differences in natural images shown in color. We have always assumed that we should transform the RGB images, and then model three planes—a luminance plane, a red/green opponent plane and a blue/yellow opponent plane (De Valois, 1965; Hurvich & Jameson, 1957). We chose to recode the stimulus images according to MacLeod and Boynton (1979) weightings to give a luminance plane and two isoluminant cone-opponent planes: L/M and S/(L + M). This follows Párraga, Brelstaff, Troscianko, and Moorehead (1998) and a large body of psychophysical evidence for parallel, near-independent processing of luminance and isoluminant cone-opponent gratings (e.g., Losada & Mullen, 1994; Mullen, 1985). In our modeling, the isoluminant cone-opponent planes are processed by receptive fields with the same orientation and frequency tuning as the luminance channels (Beaudot & Mullen, 2005). Our luminance plane model (an extension of Watson & Solomon, 1997) is based on numerous detailed quantitative studies of receptive field shape and bandwidth, and clear models of normalization and surround suppression (see above). By contrast, very little is agreed about the neurophysiology of color coding in V1 (Shapley & Hawken, 2011). Thus, the modeling of the isoluminant opponent planes in our model is subject to many assumptions, which we shall consider in the Discussion.

To et al. (2010) found that their best model, with all its assumptions, was moderately successful at predicting the suprathreshold ratings for full-color naturalistic stimuli: \( r = 0.59 \) for the normal images and \( r = 0.72 \) for the inverted pixel-reversed ones. We argued that the better correlation for pixel-reversed images results from observers making decisions solely on the basis of low-level visual differences (which our model attempts to explain) rather than any semantic content. On one hand, even a correlation of 0.59 is impressive for a simplistic model where the behavior of several million neurons is defined by just seven free parameters and a few fixed features that could have been free parameters, such as the specific orientation tuning bandwidth, receptive-field aspect ratio (Tolhurst & Thompson, 1981), or the spacing in octaves between successive frequency bands. However, we do wish to understand why the correlations are not better!

One possibility is that we have made too many false assumptions in trying to extend the good monochromatic models (To et al., 2017; Tolhurst et al., 2010; Watson & Solomon, 1997) to the full-color case. There are questions about how the three cones contribute to red/green and blue/yellow opponency (De Valois & De Valois, 1993; Mollon & Cavonius, 1987; Schmidt, Neitz, & Neitz, 2014; Schmidt, Touch, Neitz, & Neitz, 2016; Stockman & Brainard, 2010) and there is little consensus on the receptive-field organization of V1 neurons responsible for color coding (Conway, 2001; Shapley & Hawken, 2011). Therefore, in this study, we have decomposed our original full-color images (both the normal and the pixel-reversed) into monochromatic and isoluminant variants and have asked observers to rate the perceived differences between pairs of monochromatic scenes and isoluminant scenes separately. First, we ask whether the different planes contribute equally to observers’ ratings. Second, we examine whether Minkowski summation can model the integration of the luminance and cone-opponent planes into a single rating for full-color stimuli. We have previously reported that Minkowski summation with power \( m = 2.5–3.0 \) can be used to model how differences along different feature dimension are combined (To, Baddeley, Troscianko, & Tolhurst, 2011; To, Lovell, Troscianko, & Tolhurst, 2008). Finally, we will model the new monochromatic and isoluminant ratings separately to determine whether the luminance-only model with all its neurophysiological and psychophysical backing is good enough for
monochromatic natural images, and whether the poor overall performance of our 2010 model is, indeed, due to weakness in the modeling of the two isoluminant color opponent planes.

**Methods**

**Observers**

Seven observers participated in all four separate experiments, and they remained naive to the purpose of each. The observers were students or researchers at Lancaster University, UK. To ensure that they had normal or corrected-to-normal vision, we assessed their spatial acuity with the Snellen acuity chart and color vision with the Ishihara color test (13th Ed.) prior to all testing. Informed consent was obtained from all observers.

**Display equipment and stimulus construction**

The stimuli were presented on a NEC MultiSync FP2141SB CRT 22 in. display driven at 800 × 600 pixels and a frame rate of 100 Hz by a ViSaGe system (Cambridge Research Systems: Rochester, UK).

The stimuli in the present four experiments were monochrome and isoluminant variants of the full-color 900 normal and 900 pixel-reversed image pairs previously used in To et al. (2010). The 900 original normal images contained animals, landscapes, objects, people, plants, and/or garden or still-life scenes (e.g., Figures 1 and 2). The differences between the images in a pair could include changes in content (with objects appearing or moving location), the spatial frequency distribution (images sharpened or blurred), color (saturation and/or hue), shape, texture, and shadows; see To et al. (2010) for examples. Of these, 325 of the pairs consisted of two photographs of the same scene taken at different times, and we call these truly natural or ecologically valid (e.g., Figure 1). The interval between taking the photographs could be a few seconds to 10s of minutes. The image differences could arise, for example, from changing shadows, melting snow, or the effects of wind; they could involve the movement of animals or vehicles; or they could involve the photographer rearranging objects within the scene. Thus, many of the pairs would have involved some affine transform of an object or objects in the scene; unfortunately, we did not construct pairs in which the whole scene changed, as if the observer had changed their viewpoint.

The remaining image pairs involved some kind of post processing, usually involving MATLAB (MathWorks, Natick, MA) programming, and some of these pairs contained combinations of two types of imposed change (To et al., 2008). There were 273 processed pairs that involved a change only in the hue and/or saturation of part or all of one image (color-only change; e.g., Figure 2); these color changes were not guaranteed to be isoluminant.

The pixel-reversed images were modified versions of the originals, in which the content was inverted and pixel-level values were reversed so the brightest pixels in the original were swapped in location with the dimmest
Colour-only changes

![Image of image pairs showing colour-only changes]

The purpose of modifying the normal images was to reduce the higher-level semantic content of scenes, while maintaining the lower-level visually discriminable elements intact. They were similar in appearance to inverted negatives of the originals, except that the pixel-reversal algorithm (To et al., 2010) retained the same overall luminance.

For Experiments 1 and 2 in the present study, monochromatic stimuli were generated by averaging the R, G, and B planes in the original full-color To et al. (2010) stimuli (e.g., Figures 1 and 2, middle rows in panels). For Experiments 3 and 4, isoluminant stimuli were produced by first measuring the CIE XYZ coordinates of the three phosphors on the NEC display, then transforming the To et al. (2010) stimuli to XYZ and subsequently to L*a*b* space. “L” in all the pixels was set to the same average value, before the L*a*b* images were transformed back to XYZ and then to RGB space (see Figures 1 and 2, lower row in panels). The “L” value was that given by a mid-gray on the display ([128,128,128]). Note that, since the monochromatic and isoluminant stimuli are derived from the original To et al. (2010) images, they also contained the same content and much of the same feature differences as the original normal images. The magnitude of changes along the color dimensions were typically affected differently by transforming the full-color photographs into monochromatic and isoluminant versions. Examples A and B from Figure 2 show that, in most cases, the isoluminant pairs preserve some of the changes from the full–color pairs, but the two images in a monochromatic pair appear very similar.

The images were 256 × 256 pixels square (covering an area of 3.2 degrees of visual angle), but the 30 pixels at the edges of the stimuli were blended with the gray surround by compressing the pixel values towards 128 with a Gaussian falloff with a standard deviation of 12 pixels. The surrounding gray of the display had a luminance of 88 cd/m².

Standard pairs

In the four experiments, observers were presented with pairs of images test pairs (TPs) and were asked to rate how different the images in a pair appeared to them based on a standard pair (SP), whose difference was set to 20:

- Image differences that were similar to the difference in the SP were rated 20.
- Image differences that were less than the difference in the SP were rated between 1 and 19.
- Image differences that were greater than the difference in the SP were rated over 20, with no imposed upper limit.
- Seemingly identical images were given a 0 (zero) rating.

Observers were told that all difference ratings should be proportional to the SP scale so that if, for example, the TP was half or twice as different as the standard, they should enter 10 (¼ 20/2) or 40 (¼ 20 × 2), respectively.

In the original To et al. (2010) study, the SP was a pair of lily photographs that differed in color saturation (see Figure 3A). In the current study, the SP for Experiments 1 and 2 were monochromatic variants of the original (see Figure 3B). However, when isoluminant versions of the original SP were produced, they appeared too similar, so the difference between the isoluminant SP was magnified (see Figure 3C).
The experimental protocol has been described in detail in To et al. (2010). Observers were expected to try to fixate the center of each image; a fixation spot was present between stimuli, but it was extinguished during the 833 ms when a stimulus image was actually present. After a number of practice trials, each experiment began with the sequential presentation of the two images in the SP with an interval between them: fixation point on otherwise mid-gray display (83 ms), Standard Image 1 (833 ms), fixation point (83 ms), and Standard Image 2 (833 ms). The SP was then shown after every subsequent 10 trials to remind observers of their reference point. Following this was the presentation of the TPs.

The presentation order of the TPs was randomized differently for the seven observers. In addition, the two images within each TP were also presented in random order in three 833 ms intervals: fixation point, first image from TP, fixation point, second image from TP, fixation point, and first image from TP again. The rationale behind this three-interval presentation was to allow observers to see change directions from first image to second image, and from second image to first image. Following presentation of a TP, a response screen displayed a random number between 10 and 30, which the observers were asked to modify into their judged rating of the perceived difference between the images in that pair.

Each experiment was divided into four sessions in which the observer had to rate the difference between 225 of the TPs. These experimental sessions could be completed on the same or different days.

Data collation and statistical analysis

In each of the four experiments, seven observers rated each TP once. The ratings of each observer in an experiment were normalized against that observer’s median rating within the experiment. The normalized ratings for each TP were then averaged across the seven observers, and these averaged ratings were then multiplied by the grand average of all ratings (for all TPs from all observers’ ratings in that experiment) so that the data roughly centered on the standard value of 20. The data in the graphs in the Results section therefore only show the mean ratings given to each TP, averaged across observers. The standard error of the mean rating averaged about 3.0, but tended to be higher for the higher averaged ratings and lower for the very low average ratings. We previously suggested that observers sometimes differed quite markedly when giving ratings for big perceptual changes, even when they agreed more consistently for small and moderate differences (To et al., 2010); however, this applies to no more than about 90 of the 5,400 ratings, those whose average was above 40 (twice the standard). The few over-exuberant ratings were outliers in each observer’s responses, so that standardizing their data to z scores, say, might not “correct” the problem (which affects few of the data points).

V1-based visual difference predictor modeling (VDP)

We have been developing a computational model of the perceived magnitude ratings in experiments with full-color naturalistic images, trying to model the responses of millions of V1 simple or complex cells in response to the two images in a pair (To, Gilchrist et al., 2011; To et al., 2010; To et al., 2015). As mentioned in the Introduction, this model derives from the seminal work of Rohaly et al., (1997), Watson (1987), and Watson and Solomon (1997). We have also elaborated the model in studies of contrast discrimination in monochromatic naturalistic images and sinusoidal gratings (Tolhurst et al., 2010; To et al., 2017). The details of the modeling and the physiological and psychophysical justification of the various steps are given in our previous papers.

The first step of the model is of particular interest to the present study. It is widely accepted that colored lights are encoded in three planes: luminance, red/green
opponent, and blue/yellow opponent (De Valois, 1965; Hurvich & Jameson, 1957; Losada & Mullen, 1994). Thus, the full-color images (normal and pixel-reversed) are recoded with a MacLeod and Boynton (1979) transform into a luminance plane, and two cone-opponent isoluminant planes: L/M opponent and S/(L + M) opponent. The complex cell model is then run in parallel on these three planes with identical receptive field code and identical parameters (see below). A plane is first convolved with odd- and even-symmetric Gabor functions of five optimal spatial frequencies (one octave interval) and six optimal orientations (60 receptive field shapes in all, 256 × 256 locations in each set). Division by local mean luminance gives contrast rather than luminance responses. Complex cell responses are calculated as the RMS of the responses of the odd- and even-symmetric fields to give 30 sets of complex-cell responses. In this study, all the Gabor functions are self-similar with bandwidth of about one octave, but the field length or aspect ratio (and, therefore, the orientation specificity) is a free parameter in the fitting procedure.

The quasi-linear responses of the many complex cells are then subject to two nonlinearity terms derived from physiological and psychophysical studies with gratings: within-field, nonspecific contrast normalization or gain control (Carandini et al., 1997; Foley, 1994; Heeger, 1992; Watson & Solomon, 1997) and orientation-specific surround suppression (Blakemore & Tobin, 1972; Cavanaugh et al., 2002; Meese, 2004; Sceniak et al. 1999; To et al., 2017). At each location \((x,y)\) in the stimulus, we calculate a nonspecific contrast normalization signal \(N_{x,y}\) by summing the quasi-linear contrast responses \((C)\) of the 30 complex cell fields exactly centered at that point (across frequency \(f\) and orientation \(o\)), each raised to a power \(q\):

\[
N_{x,y} = \sum_{f=1}^{5} \sum_{o=1}^{6} |C_{x,y,f,o}|^q
\]  

(1)

This one nonspecific signal will suppress the responses of all 30 fields at the location equally, and \(q\) is a free parameter in the fitting procedure.

We model surround suppression as coming from an elongated area (aspect ratio 1.6) centered on the receptive field, elongated along the complex cell’s optimal orientation. The spread of this elongated Gaussian blob is proportional to the period of a cell’s optimal spatial frequency, and is a free parameter in the fitting procedure (“surround spread”, expressed as a proportion of the period of the neuron’s best spatial frequency). A different surround signal \(S_{x,y,f,o}\) is calculated (see To et al., 2010) at each point and for each of the five spatial frequencies and six orientations. The calculation involves raising responses to a free parameter \(r\).

In To et al. (2010), the two nonlinearities were applied in parallel at the same point in the model. However, following evidence that surround suppression probably follows contrast normalization (Baker, Meese, & Summers, 2007; Durand, Freeman, & Carandini, 2007; Henry et al., 2013; Li, Thompson, Duong, Peterson, & Freeman, 2006; Petrov, Carandini, & McKee, 2005), we found that our model was more effective at explaining grating contrast discrimination if the application of the two nonlinearities was sequential rather than parallel (To et al., 2017).

**Parallel models**

In our original model (To et al., 2010), the responses of each of the millions of “neurons” in the model were raised to power \(p_1\), and were finally subjected to the two nonlinear suppressive effects by division at the same time, using a modified version of the Naka-Rushton equation, an elaboration of Heeger’s (1992) formulation for contrast normalization. In the case of the parallel model, the final response of the field at location \((x,y),\) frequency \(f\), orientation \(o\), and symmetry \(s\) is:

\[
response_{x,y,f,o} = \frac{\text{sign}(c_{x,y,f,o}) \cdot |c_{x,y,f,o}|^{p_1}}{1 + W_N \cdot N_{x,y} + W_S \cdot S_{x,y,f,o}}
\]  

(2)

where \(\text{sign}\) extracts the sign (+ or −) of \(c_{x,y,f,o}\), \(W_N\), and \(W_S\) are weights, and the calculations of \(N_{x,y}\) and \(S_{x,y,f,o}\) involve raising response values to powers \(q\) and \(r\), respectively, as described above. The surround suppressive signal \(S\) is calculated from the same quasi-linear contrast responses as the normalizing signal \(N\).

**Sequential models**

Here, an intermediate normalized response \((i\_response)\) is calculated based on the normalizing signal only, and then the surround suppressive signal is calculated from these normalized responses \((i\_response)\). There are two successive Naka-Rushton equations:

\[
i\_response_{x,y,f,o} = \frac{\text{sign}(c_{x,y,f,o}) \cdot |c_{x,y,f,o}|^{p_1}}{1 + W_N \cdot N_{x,y}}
\]  

(3)

\[
response_{x,y,f,o} = \frac{\text{sign}(i\_response_{x,y,f,o}) \cdot |i\_response_{x,y,f,o}|^{p_2}}{1 + W_S \cdot S_{x,y,f,o}}
\]  

(4)

It will be noted that there is an extra parameter here \((p_2)\).
Final pooling of all the difference cues

We finally have a model of the responses or outputs of all the neurons to one plane of one image in a pair. The process is repeated for the comparison image, and we subtract the model outputs for the two stimuli neuron-by-neuron. The many visibility cues across x, y, frequency, and orientation are combined into a single value by Minkowski summation with power m (Watson & Solomon, 1997). The n (1.97 million) individual visibility cues are raised to the power m, are summed and the mth root taken.

\[
\text{overall difference} = \left( \frac{\sum^n_{i=1} (\text{difference cue}_i)^m}{n} \right)^{1/m}
\]  

(5)

This generates a single number, which is predicted to be directly proportional to the magnitude rating of the perceived difference for that plane. For the ratings for monochromatic stimuli, we model only the luminance plane and this number should be proportional to the observers’ final rating. For the isoluminant stimuli, we model the L/M and S/(L + M) planes, so that the final rating prediction is gained by a Minkowski summation of the two plane cues with the same exponent m, with the cue in the S/(L + M) plane weighted against the L/M plane with a parameter \(W_B\). Finally, for the full-color stimuli, all three planes are modeled and the final rating obtained by a Minkowski summation of three cues, with the isoluminant planes weighted against the luminance plane with weights \(W_R\) and \(W_B\).

Finding model parameters

Depending upon which experiments are fitted and whether the model is parallel or sequential, there are 8–11 free parameters. These are found by iteratively searching for the combination of parameters that maximizes the correlation coefficient between the model output and the observers’ ratings (using \textit{fminsearch}() in MATLAB). For each image pair, the ratings of the participating observers were standardized and averaged (see above). We report single fits for the 900 normal and 900 pixel-reversed ratings together (n = 1,800). However, we previously suggested that the ratings for some kinds of image change will never be satisfactorily fit by the kind of V1 model that we implement. Our model neurons very literally compare the images point-by-point and can detect small changes in object location or texture; the observers barely notice these. As well as fitting all 1,800 data for each model, we have separately fit a subset of 1,324 data, after discarding the (“unfittable”) images pairs with small spatial changes (see To et al., 2010).

To compare the performances of different versions of models to a given data set, we calculate the corrected \textit{Akaike coefficient} for small sample-size from the residual sum of squares of the regression of actual rating plotted against model prediction, while taking into account the number of parameters (Motulsky & Christopoulos, 2004). Although the actual Akaike information criterion (AIC) number for any one fit is not very informative, the delta AIC (\(\Delta\text{AIC}\)) between two models weighs up a difference in residual sums of squares against any difference in the number of parameters, and can give some indication of the relative success of different models.

\[
\text{AIC} = n \ln \left( \frac{\text{ssq}}{n} \right) + 2k \quad (6a)
\]

\[
\text{Corrected AIC} = \text{AIC} + \frac{2k.(k + 1)}{(n - k - 1)}
\]

\[
= n \ln \left( \frac{\text{ssq}}{n} \right) + 2k + \frac{2k.(k + 1)}{(n - k - 1)} \quad (6b)
\]

where \(n\) is the number of data points to fit, \(k\) is one more than the number of model parameters, and \(\text{ssq}\) is the residual sum of squares deviation between model and data.

Results

Experimental observations

To et al. (2010) measured the perceived differences between 900 pairs of naturalistic images, and between 900 pairs of inverted pixel-reversed versions of those pairs. In this study, we have re-evaluated those 1,800 ratings measured in To et al. (2010) by recruiting and testing seven new observers who each provided a total of 3,600 magnitude estimation ratings for image pairs presented in four suprathreshold discrimination experiments: 900 monochrome variants of the original full-color normal pairs, 900 monochrome variants of the original pixel-reversed pairs, 900 isoluminant variants of the normal pairs, and 900 isoluminant variants of the pixel-reversed pairs.

Interobserver correlations and standard errors

Comparing each observer’s 900 ratings in an experiment with each of the other observers, the interobservers’ correlations ranged between 0.31 and 0.81. The correlations in Experiments 1 and 2 with
monochromatic variants were higher compared to those in Experiments 3 and 4 with isoluminant variants, and the latter were slightly lower than in the original experiments in To et al. (2010). Table 1 presents interobserver correlations for each experiment. The correlations for monochromatic stimuli are noticeably higher compared to those for To et al.’s (2010) original full-color and the isoluminant variants.

Comparing ratings for full-color, monochromatic, and isoluminant stimuli

We examined the correspondence between ratings for the full-color image pairs (normal and pixel reversed) from To et al. (2010) with the new ratings given for their monochromatic and isoluminant variants. For each of the experiments, the normalized ratings of the observers were averaged together to generate a single numerical rating for each image pair. In general, we would expect the ratings for the monochromatic and isoluminant pairs to be the same as or (more likely) lower than the original ratings for the full-color images, since they now contain only partial cues to differences. However, this was not always the case, and we will consider this in the Discussion.

When comparing the ratings for monochromatic normal pairs with the original full color normal pairs, we found a good correspondence ($r = 0.69, n = 900$) between the two (see Figure 4A). The ratings in Experiment 1 seemed to be confined to scores under 50 but this was not the case for the original 2010 experiment where ratings went up to 60. We identified the 273 original full-color image pairs containing color-only changes (red symbols). The color changes were not guaranteed to be isoluminant and so some of these pairs may have included luminance changes, but largely these changes would have been difficult to discern in the monochromatic versions (see Methods, Figure 2). Unsurprisingly, therefore, for these color-only change stimuli, the monochromatic ratings were much lower than for the original full-color pairs. Furthermore, these low ratings for the monochromatic versions were poorly correlated with the full-color ratings ($r = 0.42, n = 273$). The remaining 627 stimuli (gray symbols) gave ratings lying closer to the identity line.

There was a stronger correlation between the ratings for monochromatic and full-color inverted pixel-reversed image pairs ($r = 0.79; Figure 4B$): The gray data points are more closely clustered around the identity line. However, for the pairs with color-only changes in the originals, the ratings for the monochromatic versions again tend to be very low, as should be expected since the monochromatic versions show little of our applied color changes.

![Table 1](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAIcAAAAsCAYAAADz7mz8AAAABGdBTUEAALGPC/xhBQgAAAgAElEQVR42mXyX...)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>To et al. (2010) Full-color normal</td>
<td>0.59</td>
<td>0.44</td>
<td>0.72</td>
</tr>
<tr>
<td>To et al. (2010) Full-color pixel-reversed</td>
<td>0.57</td>
<td>0.30</td>
<td>0.70</td>
</tr>
<tr>
<td>1. Monochromatic variants of normal</td>
<td>0.70</td>
<td>0.60</td>
<td>0.81</td>
</tr>
<tr>
<td>2. Monochromatic variants of pixel-reversed</td>
<td>0.69</td>
<td>0.59</td>
<td>0.80</td>
</tr>
<tr>
<td>3. Isoluminant variants of normal</td>
<td>0.53</td>
<td>0.42</td>
<td>0.68</td>
</tr>
<tr>
<td>4. Isoluminant variants of pixel-reversed</td>
<td>0.49</td>
<td>0.31</td>
<td>0.64</td>
</tr>
</tbody>
</table>

The correspondence between the ratings for isoluminant and full-color ratings was noticeably weaker compared to the previous two comparisons. In the case of the normal pairs, the correlation between the two sets was $r = 0.63 (n = 900)$ and the data points are more widely spread (see Figure 4C). For color-only changes (red symbols), the ratings for isoluminant variants were higher than those for the full-color images, and now they are reasonably correlated with the full-color ratings ($r = 0.70, n = 273$). The isoluminant ratings for the remaining stimuli (gray symbols) are also correlated to some extent with the full-color ratings ($r = 0.66, n = 627$); this follows since many of these stimuli will have involved changes in the geometry, location, or presence of objects that had some color difference from the rest of the image. These are different trends from those shown for the monochromatic stimuli (Figure 4A and 4B).

In the case of the inverted pixel-reversed stimuli (Figure 4D), the correspondence between two sets of ratings was weaker ($r = 0.60$) and the isoluminant ratings were generally lower than the full-color ratings. However, the isoluminant color-only change data (red symbols) are still well correlated with the full-color versions ($r = 0.76, n = 273$).

Integration of monochromatic and isoluminant cues

Given that the original images can be decomposed into the monochromatic and isoluminant images, we questioned whether the full-color ratings from To et al. (2010) could be predicted by combining the present monochromatic and isoluminant ratings. We have previously shown that a Minkowski summation with $m = 2.5–3.0$ was able to model how different features such as object movement, blur, and color change are integrated (To et al., 2008, To, Baddeley et al., 2011). We attempted to fit the 1,800 full-color ratings (normal together with pixel-reversed) by a Minkowski summation of their corresponding monochromatic and isoluminant variants. We minimized the summed squared error between actual and predicted ratings with three parameters: in addition to the Minkowski exponent $m$...
as a free parameter, we required non-unity weights for the monochromatic and isoluminant ratings (see Discussion). Similar to other studies of feature combination (To et al., 2008; To, Baddeley et al., 2011), we found that the best-fit Minkowski exponent was 2.71:

\[
\text{fullcolor} = \sqrt[2.71]{(0.78 \cdot \text{monochromatic})^{2.71} + (0.76 \cdot \text{isoluminant})^{2.71}}.
\]  

The correlation between the actual and modeled ratings was 0.85 \((n = 1,800;\) see Figure 5), though the correlation was slightly higher for the pixel-reversed images. There is noticeable curvature for the higher ratings, as if the Minkowski sum of the components is not great enough to explain the full-color ratings.

It is of interest to ask the relative contribution of the monochromatic and isoluminant cues to the overall perception of image differences. Unfortunately, the weights 0.78 and 0.76 in Equation 7 are arbitrary (see Discussion). Furthermore, 675 of the 900 parent image pairs involved some kind of image post-processing such as painting out of features, imposing blur, or color changes that are potentially unnatural. Therefore, we have examined the Minkowski model performance for just those normal image pairs made from two unprocessed photographs of the same scene, taken at different times.

In the original experiment with normal images, there were 325 pairs that were real differences that did not include any artificial changes (color, bandwidth, objects appearing/disappearing). We do not include the pixel-reversed variants in the following analysis, since those images are clearly unnatural. For this subset of 325 pairs, there is a strong correspondence between the monochromatic ratings and full-color ratings (see Figure 6A; \(r = 0.81, n = 325\). This is much higher than the correspondence for the remaining post-processed pairs (\(r = 0.68, n = 675\); not shown). For the isoluminant ratings, the data were less well correlated with the original full-color ratings (see Figure 6B; \(r = 0.68\), but this correspondence is still superior compared to ratings for the post-processed pairs (\(r = 0.61, n = 325\)). The results suggest that ratings for full-color ecologically valid pairs are better correlated with monochromatic ratings than with isoluminant ratings. This also demonstrates that, in general, the monochromatic and isoluminant ratings are better correlated with ratings for pairs with real differences rather than processed ones.

We fitted the 325 full-color “real” pair ratings by Minkowski summation of the appropriate monochromatic and isoluminant ratings (Figure 6C). The best fit was given with a Minkowski exponent of 1.93 and a correlation coefficient of 0.849 \((n = 325)\):

\[
\text{fullcolor} = \sqrt[1.93]{(0.71 \cdot \text{monochromatic})^{1.93} + (0.68 \cdot \text{isoluminant})^{1.93}}.
\]

The weights 0.71 and 0.68 are arbitrary. That the correlation between the monochromatic and full color ratings \((r = 0.81)\) is almost as high as that between the Minkowski sum and the full-color ecologically valid ratings \((r = 0.85)\) suggests that luminance-based cues contribute more to the perception of differences in natural images, in general, than do pure color ones.
V1-based modeling of perceptual ratings

Full-color stimuli: In To et al. (2010), we fitted our first attempts at a V1-based discriminator model to the normal and the pixel-reversed images separately, with best correlations between model predictions and actual ratings of $r = 0.59$ ($n = 900$) and $r = 0.73$ ($n = 900$), respectively. Here, we have recoded some details of the model such as reverting to the more usual self-similar receptive-field shapes and allowing receptive-field aspect ratio to be a new free parameter. Table 2 (columns 1 and 2) shows the parameter values resulting from iteratively fitting our present coding to all 1,800 full-color image pairs at once. The table shows the fits for two variant models: (a) where two key non-linearities are applied in parallel (Equation 2) and (b)

<table>
<thead>
<tr>
<th></th>
<th>Full-color</th>
<th>Monochromatic</th>
<th>Isoluminant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parallel</td>
<td>Sequential</td>
<td>Parallel</td>
</tr>
<tr>
<td>$N$ parameters</td>
<td>10</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>$W_N$</td>
<td>1.66</td>
<td>1.12</td>
<td>2.55</td>
</tr>
<tr>
<td>$W_S$</td>
<td>0.085</td>
<td>0.43</td>
<td>0.073</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0.21</td>
<td>3.40</td>
<td>0.17</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.00037</td>
<td>0.015</td>
<td>0.00055</td>
</tr>
<tr>
<td>$p_2$</td>
<td>—</td>
<td>0.078</td>
<td>—</td>
</tr>
<tr>
<td>$R$</td>
<td>2.89</td>
<td>0.73</td>
<td>2.41</td>
</tr>
<tr>
<td>Surround spread</td>
<td>1.6672</td>
<td>0.9938</td>
<td>0.2965</td>
</tr>
<tr>
<td>Field aspect</td>
<td>1.54</td>
<td>1.68</td>
<td>1.22</td>
</tr>
<tr>
<td>Minkowski</td>
<td>4.03</td>
<td>4.29</td>
<td>4.93</td>
</tr>
<tr>
<td>$W_B$</td>
<td>1.27</td>
<td>1.15</td>
<td>—</td>
</tr>
<tr>
<td>$W_R$</td>
<td>1.34</td>
<td>1.08</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 2. The best fitting values of the various parameters (defined in Methods) of the main VDP models discussed here. Parallel and sequential versions of the model were fit to the full-color and monochromatic rating data, but the isoluminant rating data were fit only with a sequential model. The number of parameters depends on model type and on the experimental data set (see Methods). These fits are for $n = 1,800$, with all the normal and all the pixel-reversed data together.
where they are applied sequentially (Equations 3 and 4).

Table 3A (columns 1 and 2) shows the statistics of those best fits, and Figure 7A plots the 1,800 actual full-color ratings against the model predictions (in arbitrary units) for the sequential model variant. The data for the pixel-reversed pairs (purple) seem to be closer to the regression line than the normal image data. Pearson’s $r$ is higher for the sequential model than the parallel model; the difference (0.66 vs. 0.71) is significant at $p = 0.002$. Furthermore, the difference in Akaike criterion ($-251$) is very large, implying that the sequential model is very much “better” than the parallel model, even given that the sequential model has an extra free parameter. It is also the case that the full-color ratings are correlated with the Euclidean distance (Kingdom, Field, & Olmos, 2007) or root-mean-square difference between pixel values; however, Pearson’s $r$ was only 0.345.

**Monochromatic stimuli:** Tables 2 and 3A (columns 3 and 4) show the best fit parameters and fitting statistics of parallel and sequential models to fit the 1,800 monochromatic ratings collected for this paper. Since the monochromatic stimuli occupy only one of the three luminance/color-opponent planes of the full-color stimuli, these models have two fewer parameters than the fits for the full-color stimuli (see Methods). Figure 7B plots the experimental ratings for monochromatic stimuli against the predictions of the sequential model. It is very clear from Figure 7B (confirmed by Table 3A) that the monochromatic ratings are fitted much better than the full-color ratings (Figure 7A). Again, the sequential model is much “better” than the parallel one.

![Image](image_url)

**Table 3.** (A) Summary statistics of the five model fits shown in Table 2 (i.e., for all 1,800 normal and pixel-reversed data). The table shows the correlation between ratings and model predictions, and the Akaike criterion (Equation 6) calculated from the residual sum of squares after fitting a regression to the experiment/model plot. Delta AIC is shown for the full-color and monochromatic models; it summarizes the difference in the fits of the parallel and sequential models. The correlation between rating and Euclidean distance is also shown. (B) The same, but for fits to a subset of the ratings data ($n = 1,324$ out of 1,800), after discarding the ratings given to image pairs that differed by a small object movement or a texture change (To et al., 2010).

<table>
<thead>
<tr>
<th>Panel</th>
<th>Full-color</th>
<th>Monochromatic</th>
<th>Isoluminant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parallel</td>
<td>Sequential</td>
<td>Sequential</td>
</tr>
<tr>
<td>A.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of data</td>
<td>1,800</td>
<td>1,800</td>
<td>1,800</td>
</tr>
<tr>
<td>Euclidean distance</td>
<td>0.345</td>
<td>0.620</td>
<td>0.594</td>
</tr>
<tr>
<td>$N$ parameters</td>
<td>10</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Pearson’s $r$</td>
<td>0.657</td>
<td>0.712</td>
<td>0.830</td>
</tr>
<tr>
<td>Akaike criterion</td>
<td>7,227</td>
<td>6,976</td>
<td>6,664</td>
</tr>
<tr>
<td>Delta AIC</td>
<td>$-251$</td>
<td></td>
<td>$-153$</td>
</tr>
<tr>
<td>B.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of data</td>
<td>1,324</td>
<td>1,324</td>
<td>1,324</td>
</tr>
<tr>
<td>$N$ parameters</td>
<td>10</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Pearson’s $r$</td>
<td>0.738</td>
<td>0.764</td>
<td>0.885</td>
</tr>
<tr>
<td>Akaike criterion</td>
<td>5,104</td>
<td>4,989</td>
<td>4,617</td>
</tr>
<tr>
<td>Delta AIC</td>
<td>$-115$</td>
<td></td>
<td>$-95$</td>
</tr>
</tbody>
</table>

Figure 7. Experimental data plotted against the sequential model predictions. (A) Ratings from the original experiment with full-color images (To et al. 2010); (B) for Experiments 1 and 2 with monochromatic images; and (C) for Experiments 3 and 4 with isoluminant images, respectively. The regression lines of best fit are shown. Data corresponding to the normal images (original or variant) are shown in blue, and the data corresponding to the pixel-reversed images (original or variant) are shown in purple.
The purpose of this study was to investigate how human observers perceive and rate changes in the monochromatic and isoluminant components in naturalistic scenes, and to determine the extent to which a V1-based model can be used to predict these ratings. In particular, we were interested in whether the isoluminant data would be as well-modeled as the monochromatic data. We took the full-color (900 normal and 900 pixel-reversed) natural scenes from our original study (To et al., 2010), decomposed them into monochromatic and isoluminant scenes, and repeated the experiments with the monochromatic and isoluminant versions separately.

**Discussion**

The stimulus pairs differed from each other in a variety of ways (see To et al., 2010), but an interesting subset consisted of pairs where a color-only change was applied by computer processing of original photographs (e.g., Figure 2). Although these changes were not guaranteed to be isoluminant, the difference between the pairs was primarily chromatic. Some pairs might also contain small luminance changes, but these were generally difficult to detect. The ratings for the monochromatic variants of these color-only stimuli, as would be expected, were considerably lower compared to those for the original full-color and isoluminant pairs (see Figure 4). Furthermore, these low ratings for the monochromatic versions were also more poorly correlated with the original color-only ratings ($r = 0.43$ and $0.55$ for normal and pixel-reversed, respectively) compared to ratings for pairs containing other changes, such as content and spatial frequency distribution ($r = 0.78$ and $0.82$ for normal and pixel-reversed, respectively). The opposite trend is seen for isoluminant color-only pairs: these ratings were better correlated with the original color-only ratings ($r = 0.70$ and $0.76$ for normal and pixel-reversed, respectively) compared to ratings for pairs containing other changes ($r = 0.66$ for both normal and pixel-reversed).

In addition to normally colored naturalistic scenes, To et al. (2010) also studied inverted, pixel-reversed versions (akin to inverted negatives) to disguise the semantic content of the scenes. Here, we also studied monochromatic and isoluminant versions of those.

($>$AIC $= -95$) even though the correlation coefficients (0.83 and 0.85) are not significantly different. These correlations coefficients are highly significantly better than those describing the fits to the full-color stimuli. The monochromatic ratings had a correlation of 0.62 against Euclidean distance; while this is higher than the equivalent correlation for full-color ratings, it is substantially less than the correlation with a biologically driven model.

**Isoluminant stimuli:** Tables 2 and 3A (column 5) show the best fit parameters and fitting statistics for a sequential model only to fit the 180 isoluminant ratings collected for this paper. Since the isoluminant stimuli occupy only two of the three luminance/color-opponent planes of the full-color stimuli, this model has one fewer parameter than the fit for the full-color stimuli (see Methods). Figure 7C plots the experimental ratings for isoluminant stimuli against the predictions of the sequential model. The correlation between ratings and sequential model ($r = 0.702$) is the lowest of the three experiments shown in Figure 7. The isoluminant ratings had a correlation of 0.59 against Euclidean distance.

**Discarding stimuli with only small spatial changes:** Of the 900 basal full-color image pairs, we suggested that some 238 would never be fit well by models based on point-by-point comparison of neuronal responses (To et al., 2010). These stimuli have small spatial changes that are well detected by the models, but not by the observers. We have fitted parallel and sequential models to the remaining 662 normal and 662 pixel-reversed pairs. The parameters and graphs are not shown, but the fitting statistics are shown in Table 3B. For the full-color and monochromatic stimuli, discarding the “unfittable” stimuli does indeed lead to highly significant increases in the correlation coefficients (compare Tables 3B and 3A). Interestingly the fit to the isoluminant stimuli is not improved. The sequential model fitted to the 1324 monochromatic stimuli ($r = 0.89$) is particularly good. As for the full set of 1,800 stimuli, the Akaikke coefficient shows that the sequential models are much “better” for the full-color and monochromatic stimuli than the parallel models.
Without the distraction of semantic content, observers’ ratings are presumably dependent just on simple shape and color cues. The correlation between full-color and monochromatic ratings is closer for the pixel reversed than for the normal images retaining semantic content. Minkowski prediction of full-color ratings from combination of monochromatic and isoluminant ratings is also closer for the pixel-reversed versions.

In general, we would expect the ratings for the monochromatic and isoluminant pairs to be the same as or (more likely) lower than the original ratings for the full-color images, since they now contain only partial cues to differences. However, this was not always the case for two main reasons. First, we used different standard pairs to anchor observers’ ratings in the different experiments (see Figure 3) so that the rating scales for monochromatic, isoluminant, and full-color scenes are not directly comparable. Second, even though we attempted to fix the scales by reference to the standards, there is a tendency of observers to self-normalize (Gescheider, 1997). Bearing this in mind, the ratings can still be compared if they are given weights to compensate for the different standard pairs and different self-normalization.

We decomposed the original full-color images from To et al. (2010) into their monochromatic and isoluminant components in the current study. These components might be equivalent to the achromatic and chromatic planes which underlie the independent coding of simple colors (e.g., Hurvich & Jameson, 1957). We considered whether ratings for differences along achromatic and chromatic dimensions could be combined to predict ratings for full-color stimuli in the same way that independent channels have often been modeled (e.g., To et al., 2009; To, Gilchrist et al., 2011; Watson, 1987; Watson & Solomon, 1997). Here we attempted to fit the 1,800 original full-color ratings (normal together with pixel-reversed) by Minkowski summation of their corresponding monochromatic (achromatic) and isoluminant (chromatic) components. The best predictions were obtained with a Minkowski summation model with exponent \( m = 2.71 \) (refer to Equation 7). The correlation between the actual and modeled ratings was 0.85, with the correlation slightly higher for the pixel-reversed images. The optimal model weighed the ratings for the monochromatic and isoluminant components similarly (0.78 and 0.76, respectively). These weights parameter were included as the ratings were based on different standards and were normalized within each experiment (see above). While not a definitive proof, this is consistent with a model where “luminance based shape” and “color” are processed separately.

The original normal full-color scenes included 325 pairs that contained only real differences; that is, no artificial manipulation of color, bandwidth, and content. For these original real difference pairs, there was a strong correlation between their ratings and the ratings for the monochromatic \((r = 0.81)\), much less so for the isoluminant variants \((r = 0.68)\). This is much higher than the correspondence for the remaining artificial post-processed pairs \((r = 0.68 \text{ and } 0.61 \text{ for monochromatic and isoluminant, respectively})\). The ratings for this subset of normal full-color images can be predicted by Minkowski summation of the monochromatic and isoluminant pairs. The best exponent was lower than for the whole data set \((1.93, \text{Equation 8})\). The correlation between actual ratings and Minkowski sum \((r = 0.85)\) was only slightly higher than that between the monochromatic ratings alone and the full-color ratings \((r = 0.81)\). The color cues have added little in general to the perception of differences in everyday natural scenes. The monochromatic component preserves most of the spatial information (Eskew & Boynton, 1987; Tansley & Boynton, 1976; see also Stockman & Brainard, 2010). In the isoluminant component, spatial details are indistinct and difficult to identify (see the loss of shadow information in the isoluminant examples of Figure 1). If luminance plays a more central important role in the identification of, and therefore changes in, content in a scene, perhaps the visual system has evolved to be better and more accurate at processing achromatic information. This increased reliance on the luminance channels could explain why correlation between ratings for full-color scenes and monochromatic scenes is higher compared to correlations between full-color scenes and isoluminant scenes. This complements the findings of Yoonessi and Kingdom (2008) who demonstrated that luminance contributes more to the perception of changes in complex images, compared to the red-green channel. We do, however, recognize that there are instances where color cues are vitally important for those few but specific scenes involving fruits, edible leaves, and sexual display (Donomy & Lucas, 2001; Parraga, Troscianko, & Tolhurst, 2002; Sumner & Mollon, 2000).

**V1-based modeling of ratings**

We have been developing a multineuronal visual difference predictor model to explain the perceived magnitudes of spatial, chromatic, and temporal differences in full-color natural images in terms of the response properties of single V1 neurons (To, Gilchrist et al., 2011; To et al., 2010; To et al., 2015). The model is based on Watson (1987) and Daly (1993), and it has proven to be very successful at explaining detection and contrast-discrimination thresholds in sinusoidal gratings and Gabor patches of various configurations (To et al., 2017; Watson & Ahumada, 2005; Watson & Solomon, 1997). It was early extended to detection of
objects in monochromatic natural scenes (Rohaly et al., 1997), and we also made early attempts at modeling the detection of changes in monochromatic natural images (Párraga, Troscianko, & Tolhurst, 2000, 2005; Tadmor & Tolhurst, 1994) but these were quite crude compared to the work of Watson and colleagues. Our present interest is to extend the modeling of thresholds in monochromatic gratings to explain the perception of supra-threshold differences in full-color natural images.

Here, we have applied two versions of our model to the full-color, monochromatic and isoluminant rating data separately (summarized in Tables 2 and 3). The models have just 8–11 explicit numerical parameters covering the behavior of millions of model neurons, although there are at least an equal number of programming decisions that we have fixed rather than allowing them to float. Given that there are so few parameters and so many neurons, a correlation coefficient between model and full-color ratings of 0.71 is a cause for optimism. There are two key nonlinear inhibitory processes involved, and our better model for the full-color (and monochromatic) ratings applies surround suppression (Equation 4) after contrast normalization (Equation 3), rather than together at the same stage (a single Equation 2). This is more consistent with recent neurophysiological studies (e.g., Henry et al., 2013) and confirms our findings when modeling contrast discrimination in gratings and Gabors (To et al., 2017). That the order of application matters shows that we need to include such nonlinear behaviors for full fidelity.

These V1-inspired models were substantially better at explaining the observers’ ratings for an 1,800-set of our stimuli than was the simple physical Euclidean distance between images in the pairs. This is likely to be because we have so many image pairs with differences of so many different unrelated kinds and magnitudes. A physical metric may well rank the order of change in a set of highly related stimuli that differ stepwise in just one way (as, for instance, in the progression of stimuli in a psychometric function) but it does not explain the relative visually perceived differences between different kinds of stimulus. This is consistent with the experience of Kingdom et al. (2007) who showed that Euclidean distance was poor at predicting the difference in thresholds between natural affine transforms of natural images and added noise. We have performed unpublished rating experiments where the image pairs were based on only 15 photographs, but each was subject to 15 different levels of jpeg compression. Not surprisingly, the average rating of the perceived difference between an image and its jpeg variant was highly correlated with the amount of compression; the ratings also had a high Pearson’s $r$ of 0.88 ($n = 225$) against Euclidean distance, leaving little space for a V1-based model to show its superiority ($r = 0.93$). As Kingdom et al. (2007) found, a good challenge to quantitative modeling of perception or detection in natural images must involve comparison among a variety of image types and transforms.

Table 2 lists the values of the several parameters in the best fitting models. We have previously noted that it is difficult to interpret the specific values of some of these, such as the powers and the weights (To et al., 2017). However, the combination of values can lead to a model of a single neuron that displays many of the properties of real single neurons in V1. The surround spread, at first sight, seems to be rather small, as if the surround suppression arises very close to the receptive field centre. To et al. (2017) discuss similar fits (to grating detection data) and show that the small surround radius in these fits is still closely compatible with real neuronal data (Cavanaugh et al., 2002; Sceniak et al., 1999). The receptive-field aspect ratio of the best models (1.6–1.8) is greater than that we reported for gratings (To et al., 2017) but is, perhaps, more compatible with real neuronal data (Tolhurst & Thompson, 1981). The Minkowski parameter has no neurophysiological equivalent (but see To, Baddeley et al., 2011). The values of 3.72–4.29 are close to those long used in studies that model the combination of independent detectability cues (e.g., Robson & Graham, 1981).

Our VDP originates from antecedents successful at modeling monochromatic stimuli, and it is based largely on single neuron studies with monochromatic stimuli (see citations in Introduction). It was a major aim of this study to investigate the success of our extension of the model to deal with color stimuli. Thus, we constructed two new sets of stimuli from our original full-color ones: monochromatic variants and isoluminant color variants. As we might hope from its origins, the VDP was good at predicting the magnitude ratings for monochromatic stimuli ($r = 0.845$). The VDP was much less effective at modeling the isoluminant ratings ($r = 0.702$), and the overall moderate performance on the original full-color images ($r = 0.712$) can be blamed on a weaker model of color processing. To et al. (2010) noted that some of our image pairs would likely never be well fit by a V1-based model: the single-neuron receptive-fields are sensitive to very small changes in object location or texture, whereas in an 800 ms presentation time, human observers generally fail to perceive such differences (for some examples, see the Supplementary Material for To et al., 2010). If we discard these “unfit” stimuli from our data sets, the model fits to the full-color and monochromatic stimuli improve; the monochromatic fit has a gratifying correlation of 0.894. Interestingly, discarding these stimuli from the isoluminant set did not provide a better model fit, perhaps because colors tend to be more uniform over larger areas than
brightness so that small changes in the locations of similar objects (texture) would not much change the overall color organization. We argued previously (To et al., 2010) that the ratings given to some kinds of stimuli (e.g., faces, shadows) might be influenced by “higher” cognitive processes and not just the low-level visual differences, which we are capable of modeling. For that reason, we also studied inverted pixel-reversed image pairs to obscure such cognitive cues. Perhaps if we had studied other kinds of image difference, we might have had lower correlations than 0.845 or 0.894. None of our pairs, for example, consisted of affine changes in the whole scene, as if the observer had changed their viewpoint. However, given we did include such a variety of types and magnitudes of change, we would not expect our modeling of these to be less successful than our modeling of affine changes of objects within an otherwise-constant scene.

Modeling the color planes

Given the many detailed and systematic quantitative studies of linear and nonlinear summation in V1 neurons (see citations in Introduction), it is possible to write computer program to simulate the key neuronal processes. The VDP is successful for monochromatic stimuli, but less so for colored ones. Retinal processing of colored lights and the connectivity of cones has been well studied, but neurophysiology has not provided a quantitative (and computable) consensus of V1 processing (Shapley & Hawken, 2011). In the absence of the required detail, we chose uncontroversially to decompose our color stimuli into three planes and to run a VDP on each plane separately. The way in which we chose to do this has not been totally successful, and there are a number of steps that may need revisiting.

We decomposed the stimuli onto a luminance and two isoluminant cone-opponent planes (MacLeod-Boynton transform). Others have used CIE L*u*v or CIELAB spaces (Jin, Feng, & Newell, 1998; Lubin, 1995). Mollon and Cavonius (1987) and Schmidt et al. (2014) point out that the well-studied L/M cone opponency in retina and LGN is not equivalent to the perceptual red-green opponency of Hurvich and Jameson (1957), despite the two being often conflated. Neither is retinal S/(L+M) opponency equivalent to perceptual blue/yellow opponency. De Valois and De Valois (1993) proposed that the cone-opponent streams from the LGN must be repackaged in the cerebral cortex to give new opponent axes: (S + L) versus M for red/green, and (S + M) versus L for blue/yellow. There is as yet no evidence for such reorganization of cone inputs in V1. In fact, Schmidt et al. (2014, 2016) propose that these axes already appear in subsets of retinal ganglion cells. The (S + M)/L and (S + L)/M combinations were reported by DeMonasterio, Gouras, and Tolhurst (1975), and have been given much more credence by Tailby, Solomon, and Lennie (2008) and Field et al. (2010) particularly for the connectivity of OFF neurons. We await full descriptions and cortical fates of these ignored neurons, which Schmidt et al. (2014, 2016) propose are the basis of color perception while going so far as suggesting that the majority L/M P-cell pathway is not so involved. Psychophysical studies also imply an S-cone involvement in the perceptual red/green axis (Danilova & Mollon, 2012; Wuerger, Atkinson, & Cropper, 2005). A future modeling project would look at different ways of transforming stimuli into luminance and color planes. We would certainly expect better performance in modeling a perceptual task if we tried to use opponent axes that matched perceptual color opponency rather than cone opponency.

In our model, the two opponent planes (which may be questionable) are processed by narrowly tuned, orientation-specific neurons just like those processing luminance information. This was partially motivated by the large body of psychophysical work suggesting that luminance, L/M and S/(L + M) might be processed in parallel, more or less independently (e.g., Beaudot & Mullen, 2005; Losada & Mullen, 1994). However, there is very little evidence to support such independence in V1; most neurons respond to both luminance-modulated and L/M-modulated gratings with a great range of bias between the two (Hass & Horwitz, 2013; Johnson, Hawken, & Shapley, 2008). An alternative view, with little quantitative backing, is that a small subset of V1 neurons process “color” (Hubel & Livingstone, 1987), and these might be broadly tuned for orientation, and responsive primarily to low spatial frequencies (e.g., Conway, 2001). Johnson et al. (2008) did report that the V1 neurons most biased towards L/M gratings and away from luminance gratings did prefer lower spatial frequencies, and some were poorly tuned for orientation. This has also been reported psychophysically (Gheiratmand & Mullen, 2014). While most V1 neurons might respond differentially to color stimuli to some extent, it is still not clear whether all these neurons contribute to color perception, or whether this is the role of some (yet to be characterized) subset of neurons (Schmidt et al., 2014, 2016). The receptive field organization, tuning, and nonlinear interactions of these putative neurons await documentation, but a future VDP model might investigate whether broadly-tuned, non-oriented, double-opponent neurons would better describe the isoluminant rating data than the multiple banks of narrow-band, oriented simple and complex cells that we presently use. Our present study, by showing the success of the monochromatic part of the VDP, allows us now to focus on...
searching for better ways of modeling just the “color” planes in natural stimuli.

**Keywords:** natural scenes, suprathreshold discrimination, magnitude estimation ratings, luminance and isoluminant planes, feature integration

### Acknowledgments

This research was funded by an ESRC: Transforming Social Science Programme Grant and an Early Career Researcher Grant from the Faculty of Science and Technology at Lancaster University.

Commercial relationships: none. Corresponding author: Michelle P.S. To.
Email: m.to@lancaster.ac.uk.
Address: Department of Psychology, Lancaster University, Lancaster, UK.

### References


Eskew, R. T., Jr., & Boynton, R. M. (1987). Effects of field area and configuration on chromatic and...


Movshon, J. A., Thompson, I. D., & Tolhurst, D. J.

Párraga, C. A., Troscianko, T., & Tolhurst, D. J. (2000). The human visual system is optimised for processing the spatial information in natural visual images. *Current Biology, 10*(1), 35–38.


To, M. P. S., Gilchrist, I. D., Troscianko, T., & Tolhurst, D. J. (2011). Discrimination of natural


