Color contributes to object-contour perception in natural scenes

Thorsten Hansen
Abteilung Allgemeine Psychologie,
Justus-Liebig-Universität Gießen, Gießen, Germany

Karl R. Gegenfurtner
Abteilung Allgemeine Psychologie,
Justus-Liebig-Universität Gießen, Gießen, Germany

The magnitudes of chromatic and achromatic edge contrast are statistically independent and thus provide independent information, which can be used for object-contour perception. However, it is unclear if and how much object-contour perception benefits from chromatic edge contrast. To address this question, we investigated how well human-marked object contours can be predicted from achromatic and chromatic edge contrast. We used four data sets of human-marked object contours with a total of 824 images. We converted the images to the Derrington–Krauskopf–Lennie color space to separate chromatic from achromatic information in a physiologically meaningful way. Edges were detected in the three dimensions of the color space (one achromatic and two chromatic) and compared to human-marked object contours using receiver operating-characteristic (ROC) analysis for a threshold-independent evaluation. Performance was quantified by the difference of the area under the ROC curves (ΔAUC). Results were consistent across different data sets and edge-detection methods. If chromatic edges were used in addition to achromatic edges, predictions were better for 83% of the images, with a prediction advantage of 3.5% ΔAUC, averaged across all data sets and edge detectors. For some images the prediction advantage was considerably higher, up to 52% ΔAUC. Interestingly, if achromatic edges were used in addition to chromatic edges, the average prediction advantage was smaller (2.4% ΔAUC). We interpret our results such that chromatic information is important for object-contour perception.

Introduction

The detection of edges and contours is one of the first major processing steps in artificial and natural vision systems (e.g., Marr, 1982). Edge and contour detection are traditionally regarded as achromatic processes. For example, in the classical neurophysio-logical study by Hubel and Wiesel (1968), neurons that responded to an oriented luminance contrast were readily classified as achromatic and orientation selective and not tested further for chromatic selectivity. Subsequent studies built on this notion and suggested that color and orientation are processed by specialized, distinct populations of neurons that project to different areas of the brain (Zeki, 1976). This simplistic view has been incorporated in models of attention that assumed an initial independent processing of color and shape by different modules (Treisman & Gelade, 1980). Further, most computational approaches to edge detection are based on grayscale images (e.g., Forsyth & Ponce, 2012).

However, there is now a growing consensus that color plays an important role in the segmentation, recognition, and memorization of objects (Gegenfurtner & Rieger, 2000; Geusebroek, van den Boomgaard, Smeulders, & Geerts, 2001; Tanaka, Weiskopf, & Williams, 2001; Wichmann, Sharpe, & Gegenfurtner, 2002; Wolfe, 1998). Gegenfurtner and Rieger (2000) used a delayed match-to-sample paradigm to study the role of color vision in natural scenes. They found that color vision plays "a general role in the processing of visual form, starting at the very earliest stages of analysis: color helps us to recognize things faster and to remember them better (p. 805)." Wichmann et al. (2002) found that images were recognized about 5%–10% better if presented in their natural color. The advantage disappeared for pseudocolored images, pointing toward the importance of learned knowledge about the colors in natural scenes for recognizing new scenes.

On the one hand, the role of color for human vision may be questioned based on the fact that color-vision deficiency often goes unnoticed in daily life unless diagnosed in a test like the Ishihara test. On the other hand, the disadvantage of dichromatic color vision
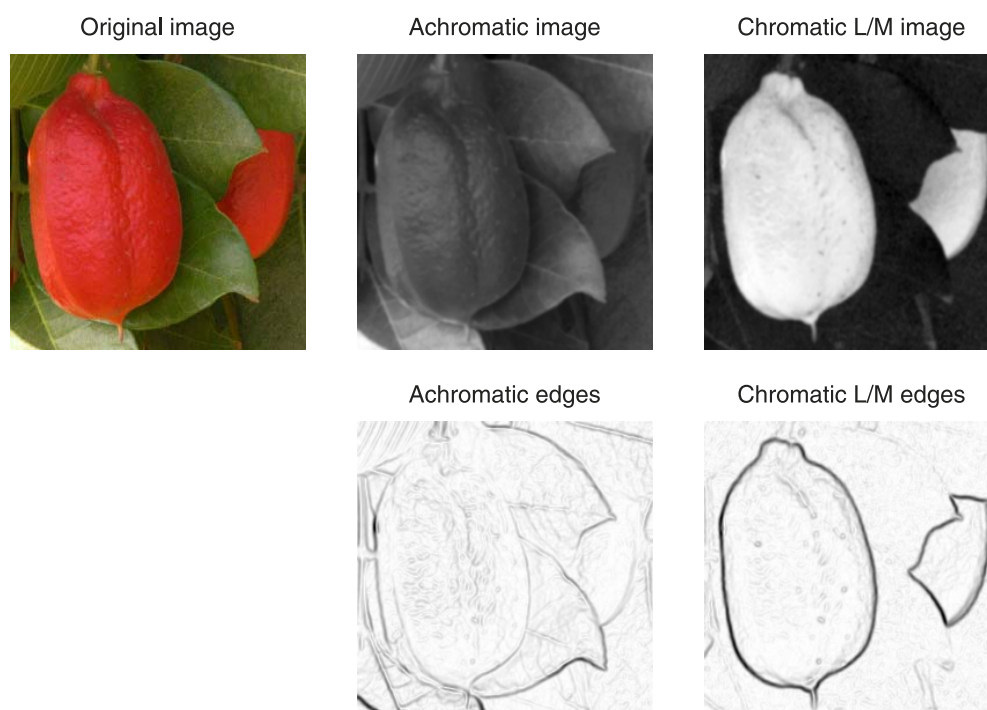
Figure 1. Image of an ackee fruit and edges detected based on achromatic and chromatic L/M information. The object contour is delineated faintly if at all by the achromatic edges but almost perfectly by the chromatic L/M edges. The fruit pops up in both the chromatic image and the chromatic edge map and can be easily separated from the background. It is the chromatic information in this image that allows us to detect the fruit fast and easily.

becomes obvious in special situations such as detecting fruits against dappled green leaves that vary in luminance (Mollon, 1989). Mollon concluded (p. 21), "The disabilities experienced by colour-blind people show us the biological advantages of colour vision in detecting targets, in segregating the visual field and in identifying particular objects or states." Along the same lines, Tanaka et al. (2001) have argued that color might be critical for object recognition, because objects that are represented by color and shape, such as a banana, can be recognized easier than objects represented by shape alone, in particular under occlusion conditions which frequently occur in everyday situations.

We have previously analyzed the the magnitude of co-occurring chromatic and achromatic edge contrast in natural scenes and found that the mutual information is minute (Hansen & Gegenfurtner, 2009). In other words, the magnitude of the achromatic edge contrast at a particular image location does not predict the magnitude of the chromatic edge contrast at this location. Nearly all edges combine luminance and color, and isoluminant edges in natural scenes are as likely as purely achromatic edges. Thus, information about object contours is sometimes represented only chromatically: Consider the image of a red fruit on green foliage (Figure 1). In the achromatic image, the edges of the fruit are hardly detectable, because the luminance of the fruit is almost the same as the luminance of the background foliage. Any natural or artificial vision system that tries to detect objects based on achromatic information alone would probably miss the fruit. Adding chromatic information changes the situation: In the chromatic $L - M$ dimension which codes reddish-turquoise signal variations, the object boundaries of the fruits are almost perfectly delineated. A vision system that can use this chromatic information will probably detect the fruit. Further, it has been long noticed that red–green chromatic edges, unlike achromatic edges, cannot result from shadows or shading, but indicate a change in surface reflectance that may signal an object contour (Párraga, Troscianko, & Tolhurst, 2002; Rubin & Richards, 1982). Algorithms have been proposed based on this notion to estimate the intrinsic reflectance and shading image from red–green–blue (RGB) color images (Geusebroek et al., 2001; Olmos & Kingdom, 2004; Tappen, Freeman, & Adelson, 2005).

In our previous work we showed that chromatic edges provide an independent source of information (Hansen & Gegenfurtner, 2009). We showed that the mutual information between achromatic edges and chromatic edges along the cardinal dimensions $L - M$ and $S - (L + M)$ was negligible. Here we investigate to what degree humans use this information and benefit from chromatic contrast information in the perception of high-level object contours in natural scenes.

Our main idea is to use human-marked object contours in natural scenes as ground truth and use the response of edge detectors to achromatic and chromatic dimensions of the image to investigate to what degree the prediction of the human-marked object contours based on achromatic edges can be improved if chromatic edges are also considered.

In general, image contrasts may be classified along a hierarchy ranging from localized circumscribed contrasts via edges—that is, straight, collinear contrast changes—to longer curved contours that signal object boundaries. The edge images we compared reside on different levels of this hierarchy: The edge detectors detect small edges based on local image statistics that arise from various sources, such as noise, textures, shadows, and object boundaries, while the human-marked contours are based on a fairly high-level scene segmentation. In fact, the human observers were instructed to mark "distinguished things" in the image or salient objects or the outline of an animal in the scene. In this work we were interested in the contribution of color to precisely these types of contours, namely the main and important contours in an image; it is therefore important to use ground-truth images that were labeled at a fairly high level, instead of images where observers were instructed to mark any minute contrast they perceive. Here and in the following, we use the term *contours* to refer to the high-levels contours at object boundaries that humans marked in the images.

Methods

We analyzed the contribution of chromatic information to predict human-marked object contours in natural scenes within a receiver operating-characteristic (ROC) framework. First, we converted the images to the Derrington–Krauskopf–Lennie (DKL) color space to separate chromatic and achromatic information in a physiological meaningful way. Second, edges were detected using an edge detector (e.g., the Sobel operator) in each dimension of the DKL color space—that is, in the achromatic and the two chromatic dimensions. Third, we used an ROC analysis to compare how well human-marked object contours could be predicted based on achromatic or chromatic edge information or a combination of both. To combine the edge information, we added the outputs of the operator for the different layers.

Data sets

We used four data sets of human-marked object contours: the All Natural Image Database (ANID), the Berkeley Segmentation Dataset (BSD), the McGill Color Calibrated Contour Dataset (MGCCCD) and the Salient Objects Dataset (SOD). The images in these data sets were taken by professional photographers (BSD, SOD) or vision scientists (ANID and MGCCCD).

ANID

The ANID is a collection of 294 images of animals (Drewes, Trommershäuser, & Gegenfurtner, 2011). The images in the data set show both the animal and the background in focus, in contrast to professional images where the animal is usually in sharp focus against a blurred background. A student assistant marked the contour of the animal and, if visible, the animal's head with the aid of a commercial image-processing program (Adobe Photoshop, Adobe Systems, San Jose, CA). For the present analysis we first cropped the images to a square containing the animal, to speed up computations, and then converted the segmented body of the animal to an outline edge with a width of 1 pixel. We ignored the segmented animal's head, because in most cases the head was not separated by a difference in luminance or color from the rest of the body. Sample images and human-marked contours of the animal in the scene are shown in Figure 2. The ANID is freely available from http://www.allpsych.uni-giessen.de/ANID.

BSD

The BSD contains hand-labeled segmentations of various images from a commercial digital image library (Corel Stock Photo Libraries, Corel Corporation, Ottawa, Ontario, Canada) to provide an empirical basis for research on image segmentation and contour detection (Martin, Fowlkes, Tal, & Malik, 2001). A portion of the data set is freely available for noncommercial research. The number of these publicly available images increased over the years from 100 to 300 to 500. We use the notation BSD to refer the most recent data set of 500 images, and BSD100 and BSD300 to refer to the other data sets; BSD500 will be used when necessary for contrast with former versions. We used the current version of the data set (as of January 2016), which contains 500 images with 12,000 hand-labeled segmentations from 30 human observers. Sample images of the BSD are depicted in Figure 2.

To obtain the segmentations, observers were asked to divide the image into "distinguished things." More precisely, observers were given the following intentionally vague instruction to break up the scene in a natural manner: "Divide each image into pieces, where each piece represents a distinguished thing in the image. It is important that all of the pieces have approximately equal importance. The number of things in each image

Figure 2. Sample images and human-marked object contours of the data sets used. The Salient Object Dataset is not shown because it is based on the same images as the 300-image Berkeley Segmentation Dataset, which is a subset of the full 500-image set. The original human-marked object contours have a width of 1 pixel and have been broadened here to increase visibility. Note that the vertical line in the image of the bears is an artifact in the original image which has been marked by one observer.

is up to you. Something between 2 and 20 should be reasonable for any of our images."

MGCCCD

The MGCCCD is a subset of 30 images from two categories (fruits and landscape) of the McGill Calibrated Colour Image Database. A single observer marked the contours. We scaled the images to a quarter of their original size (from 1920×2560 pixels to 480×640 pixels) to match the size of the images in the other data sets and to speed up computations. Aaron Johnson collected the MGCCCD. Sample images and human-marked contours are shown in Figure 2.

SOD

The SOD is a collection of salient-object contours for the 300 images of the BSD300 (Movahedi & Elder, 2010), which is a subset of the BSD500. Seven observers viewed an image together with an overlay of human-marked contours from the BSD300 and marked the objects they perceived as most salient by clicking on the corresponding segment or segments. The SOD is freely available from http://elderlab.yorku.ca/~vida/SOD.

Conversion to DKL color space

We transformed the images from RGB to DKL color space to separate chromatic and achromatic information in a physiologically meaningful way.

The DKL color space has three so-called cardinal directions: an achromatic direction L/M/S and two chromatic directions—one labeled L/M, where only L and M cones change at a constant sum, and the other labeled S, where only the S cones change. The corresponding cardinal mechanisms are defined as being orthogonal to a pair of cardinal directions. The cardinal mechanisms are conventionally labeled $L + M$, $L - M$, and $S - (L + M)$, although they are generally not computed as such from the cone responses *L*, *M*, and *S*. In the DKL color space, the cardinal directions and mechanism axes are aligned, which is not the case in general (Stockman & Brainard, 2009). The preferences of the chromatic cardinal mechanisms model the chromatic preferences of retinal ganglion cells and cells in the lateral geniculate nucleus (LGN).

We used two different conversion methods to convert from the RGB input image to DKL: monitor based and camera based. The monitor-based conversion is based on a particular calibrated monitor and transforms the image such that the properties of the DKL axes hold if the *image* would be viewed on this monitor. The camera-based calibration transforms the image such that the properties of the DKL axes hold if the *natural scene* depicted by the image would be viewed. For a camera-based conversion the images have to be taken by a calibrated camera; this is true only for the MGCCCD.

Monitor-based conversion

For the monitor-based conversion we use calibration routines for a standard CRT monitor (Sony GDM 20se

II). We used a conversion method where the axes of the DKL space are scaled to unity at the limit of the monitor gamut. Details of this conversion can be found in Hansen and Gegenfurtner (2013). The conversion can be expressed by a matrix multiplication:

$$\begin{pmatrix} L+M \\ L-M \\ S-(L+M) \end{pmatrix} = \begin{pmatrix} 0.2182 & 0.6889 & 0.0928 \\ 0.8217 & -0.5718 & -0.2499 \\ 0.2318 & 0.6795 & -0.9113 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}.$$

Camera-based conversion

For the camera-based conversion we first simulated how the three cone mechanisms S, M, and L of a human observer would respond to the image using a camera-specific conversion routine that is supplied with the McGill data set. Next, the response of the three cone mechanisms S, M, and L were transformed into the response of the three cardinal mechanisms of DKL color space. Following Johnson, Kingdom, & Baker, 2005 and Párraga et al., 2002, we computed the achromatic mechanism $L + M$ as the sum of the L and M cones, and the two chromatic mechanisms based on the Michelson contrast between different cone types. We use Roman to denote the cardinal mechanisms L + M, L − M and S − (L + M) and italics to denote the cone responses L, M, and S.

$$\begin{pmatrix} L+M \\ L-M \\ S-(L+M) \end{pmatrix} = \begin{pmatrix} L+M \\ c(L,M) \\ c(S,L+M) \end{pmatrix}$$
$$= \begin{pmatrix} L+M \\ \frac{L+M}{L-M} \\ \frac{S-(L+M)}{S+L+M} \end{pmatrix},$$

with the Michelson contrast $c(x, y)$ defined as

$$c(x,y) = (x-y)/(x+y).$$

In the numerical computation of the ratios, we added a tiny constant to the divisor to avoid division by zero. The value of this constant was $\varepsilon = 22^{-52}$, that is the distance from 1.0 to the next larger double precision number, as returned by MATLAB's eps command (The MathWorks, Natick, MA).

Finally we normalized the mechanism responses to the interval [0, 1]. We normalized the chromatic responses globally to ensure that high chromatic edge responses are not artifacts of a local normalization. A global normalization of the chromatic mechanism was feasible because the range of the chromatic contrast responses was limited due to the divisive normalization in the equation. Because the equation of the achromatic mechanism does not involve a divisive normalization, the range of the achromatic responses varied over several orders of magnitude for some data sets. We thus

normalized the achromatic responses locally for each image. The two different normalization methods result in an overall higher achromatic contrast. We have used the same normalization scheme in previous work (Hansen & Gegenfurtner, 2009).

Edge detection

We used different edge-detection algorithms to compute edges for each dimension of DKL color space and compared them to human-marked object contours.

Edge-detection algorithms

To detect edges, we used the Sobel operators and and three biologically motivated algorithms which model simple cell responses in primary visual cortex (V1): Gabor filters (Jones & Palmer, 1987; Pollen & Ronner, 1983), a simple cell model with dominating opponent inhibition (SimpleCell; Hansen & Neumann, 2004a), and a simple cell model that relies on LGN input (CORF; Azzopardi & Petkow, 2012). We ran the SimpleCell and CORF operators with their default parameters. We used different edge-detection algorithms to investigate to what degree the results depend on the algorithm.

Performance analysis

Our main question is if and to what degree the predictions of an edge detector based on achromatic information can be improved by adding chromatic information. Note that we were not interested in the absolute performance of a particular edge detector, but rather in the relative contribution of chromatic information for human contour perception.

We used two different framework, receiver-operating characteristic (ROC) analysis and Precision-Recall (PR) analysis. Both methods are based on ground-truth verification—that is, the comparison of a detection result with ground truth—, and both methods provide a threshold-independent analysis to capture the trade-off between hits and false-alarms.

Ground-truth images

Since the both ROC and PR analysis are based on ground-truth verification, the first step in the analysis is to specify the ground truth. We used the human-marked object contours as ground truth against which the responses of the edge detectors can be compared. In general, we set a pixel in the ground truth image to 1, indicating an object contour, if the pixel was marked as an object contour by any observer. If more than one

Number of image pixels $\|A \cup \bar{A}\| = \|B \cup \bar{B}\|$		Ground-truth image	
		Edge A	Background $\bar{A}$
Operator	Positive response: edge detected (B)	Hits $t_p = \|A \cap B\|$	False alarms $f_p = \|\bar{A} \cap B\|$
	Negative response: no edge detected ($\bar{B}$)	Misses $f_n = \|A \cap \bar{B}\|$	Correct rejections $t_n = \|\bar{A} \cap \bar{B}\|$

Table 1. The so-called contingency table is a formal way to represent the four possible outcomes when an edge operator detects a signal. There are two correct responses of the operator, namely a "hit" when a ground-truth edge is detected, and a "correct rejection" when no edge is detected at a background location; and there are two incorrect responses of the operator, namely a "false alarm" when an edge is signaled at a background location and a "miss" when no edge is signaled at a ground-truth edge. More formally, the ground-truth image divides the set of pixels into two subsets of edge pixels A and background pixels $\bar{A}$. Likewise, the operator divides the set of pixels into two subsets of edge pixels B and all other pixels $\bar{B}$. The number of elements in both the unions of A and $\bar{A}$ and of B and $\bar{A}$ are the number of pixels in the image.

observer labeled the image, other criteria are possible—for example, a ground-truth object-contour pixel has to be marked by all observers. These two criteria constitute the extremes, and any intermediate criteria may be chosen—for example, a ground-truth object-contour pixel have to have been marked by at least 50% of all observers. We investigated the effect of observer consensus on the results in a separate section, where we computed different ground-truth images. More precisely, if $N > 1$ observers labeled the image, we first summed the binary object-contour images of all observers, and then obtained ground-truth images by thresholding the summed image at 0, 1, ..., $N − 1$ (where N is the maximum number of observers that marked a pixel as an object contour).

ROC analysis

ROC analysis has its origin in signal-detection theory (Green & Swets, 1966) and is now applied in a number of diverse fields, in particular the evaluation of medical diagnostic performance (Pepe, 2003). ROC analysis captures the trade-off between sensitivity and 1 − specificity—that is, between hit rate and false-alarm rate.

The hit rate and correct-rejection rate are threshold dependent: A sufficiently low threshold results in 100% hits, at the expense of a high false-alarm rate, because many pixels are considered as contour pixels that have not been marked by a human observer. For a sufficiently high threshold, the situation reverses, resulting in a low false-alarm rate but a also low hit rate. An ROC curve is a graphical plot that illustrates the performance of a binary classifier as its discrimination threshold is varied. The curve is the hit rate as a function of the false-alarm rate at various discrimination thresholds (Bowyer, Kranenburg, & Doughtery, 2001). The hit rate is also known as true positive rate (TPR), sensitivity, d', or recall (R); the false-alarm rate is also known as false positive rate (FPR), 1 − specificity, or fallout (F).

The ideal operator has to be sensitive to all signals (high hit rate) while at the same time responding specifically only to the signals and not the noise (low false alarm rate). In the context of the present work, a hit is a pixel in the human ground-truth image that is detected by the edge operator—that is, a pixel that has a value above threshold in the operator edge image. Similarly, a false alarm occurs if the detector signals an edge at a location that has not been marked as an object contour in the human ground-truth image. The four cases that occur when an operator responds to a signal are traditionally represented in a so-called contingency table (Table 1).

The hit rate or recall R is the number of correctly detected object-contour pixels (hits) divided by the number of ground-truth object-contour pixels:

$$R = \frac{|A \cap B|}{|A|}.$$

The false-alarm rate or fall-out F is the number of false detected object-contour pixels divided by the number of non-object-contour pixels in the ground-truth image:

$$F = \frac{|\bar{A} \cap B|}{|\bar{A}|}.$$

We determined the hit rate and the false-alarm rate for 101 discrete threshold values 0, 0.01, 0.02, ..., 1. For each threshold value we first binarized the operator edge image by setting all values above the threshold to 1 and all values below to 0. Second, we determined the number of hits (i.e., the number of pixels that are 1 both in the thresholded edge image and the human ground-truth image) and the number of false alarms (the number of pixels that are 1 in the thresholded edge image but zero in the ground-truth image). After we determined hits and false alarms for each threshold, we computed the *hit rate* by dividing the hits by the number of object-contour pixels in the ground-truth image, and the *false-alarm rate* by dividing the false alarms by the number of background pixels in the ground-truth image. The ROC curve can then be plotted as hit rate against false-alarm rate.

The ROC curves always start at (0, 0) and end at (1, 1). Curves that are higher represent better performance. An ROC curve that is equal to the main diagonal (TPR = FPR, or sensitivity = 1 − specificity) corresponds to a random operator where a positive test corresponds to flipping a coin with probability of a hit equal to a false alarm. An ROC curve that is below the main diagonal is worse than guessing, and hence can be improved upon by taking the opposite decision.

An ROC curve is a two-dimensional depiction of the performance of a classifier. It is often desirable and handy, for example, to compare different classifiers, to characterize the performance of the classifier by a single scalar value (Fawcett, 2006). Such a single measure is the area under the ROC curve (AUC, also AuROC or A').

The better the performance of the operator, the higher the AUC. The value of the AUC varies between $0.5 = 50\%$ (chance level) for the worst-informative operator and $1 = 100\%$ for the ideal operator. AUC values between 0 and 0.5 are possible but correspond to nondetectors—that is, detectors that respond better to the background (the nonsignal) than to the edge (the signal). Such a nondetector can of course be easily converted to a detector by negating its response. In the words of Fawcett (2006, p. 863), such a classifier "may be said to have useful information, but it is applying the information incorrectly (Flach & Wu, 2003)."

The AUC is an empirical estimate of the probability that the detector correctly responds to the signal plus noise but not to the noise (Green & Swets, 1966, pp. 45–49; cited in Hanley & McNeil, 1982). In the present context, the signal plus noise is the edge pixels in the ground-truth image, and the noise is the background pixels. The edge-detector response to the image can be considered as a signal-detection experiment using a rating paradigm, where the detector rates each pixel on a continuous scale as being an edge or not. The finding of Green and Swets (1966) tells us that the AUC obtained from such a rating experiment is equivalent to the AUC obtained from a signal-detection experiment using a two-alternative forced-choice paradigm. The AUC when computed by the trapezoidal rule is equivalent to the Wilcoxon–Mann–Whitney statistic (Bamber, 1975; Hanley & McNeil, 1982); the AUC is therefore equivalent to the frequency that in all pairs of pixels randomly picked from the edge map, the pixel with the higher value is a ground-truth contour pixel. This interpretation of the AUC holds independently of the distribution of the underlying data. The AUC is also a linear transformation of the Gini coefficient (Fawcett, 2006), which itself is equivalent to half of the relative mean absolute difference (Sen, 1977).

The AUC can be used to compare different ROC curves. If the ROC curve for one operator is higher than that for another operator, then the first operator's AUC will also be larger. The converse is not true: A larger AUC does not imply a uniformly better operator (Pepe, 2003).

Here we used the AUC as a single measure to characterize how well the human-marked object contours can be predicted by the achromatic or chromatic edge images or a combination of both. The AUC was computed from the empirical ROC curves by the trapezoidal rule; no curve fitting was used. The shapes of the empirical ROC curves could be approximated in most cases well by ideal ROC curves assuming normal distributions of equal variance. To grade the performance measured by an AUC we used the traditional academic point system: $0.90–1.00 =$ excellent (A), $0.80–0.89 =$ good (B), $0.70–0.79 =$ fair (C), $0.60–0.69 =$ poor (D), and $0–0.59 =$ fail (F). To compare the performance for different inputs (such as achromatic vs. chromatic and achromatic) we used the difference between the corresponding AUCs, denoted by ΔAUC.

Precision–recall analysis

An alternative approach to ROC analysis is to evaluate the performance in terms of precision and recall. Recall is a synonym for hit rate; precision or positive predictive value is the number of correctly detected edge pixels divided by number of detected edge pixels:

$$P = \frac{|A \cap \mathrm{B}|}{|B|}.$$

We computed precision–recall (PR) curves for each image analogously to ROC curves. For each image, we normalized the operator-detected edge image to the range [0, 1] and computed precision and recall for each of the 101 threshold values used to binarize the edge map. We computed a single curve by averaging precision and recall curves using the F-measure (van Rijsbergen, 1979), defined as

$$F_\alpha(P, R) = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1 - \alpha)\frac{1}{R}}.$$

The higher the curve of the F-measure, the better the performance of the operator. For an equal weighting of precision and recall with $\alpha = 1/2$, the F-measure is the harmonic mean of precision P and recall R:

$$F = F_{1/2}(P, R) = \frac{2PR}{P + R}.$$

Following Martin, Fowlkes, and Malik (2004), we used the maximum along the $F_{1/2}$ curve to analogous to the AUC in the ROC analysis to derive a single scalar value that characterizes the performance of the operator.
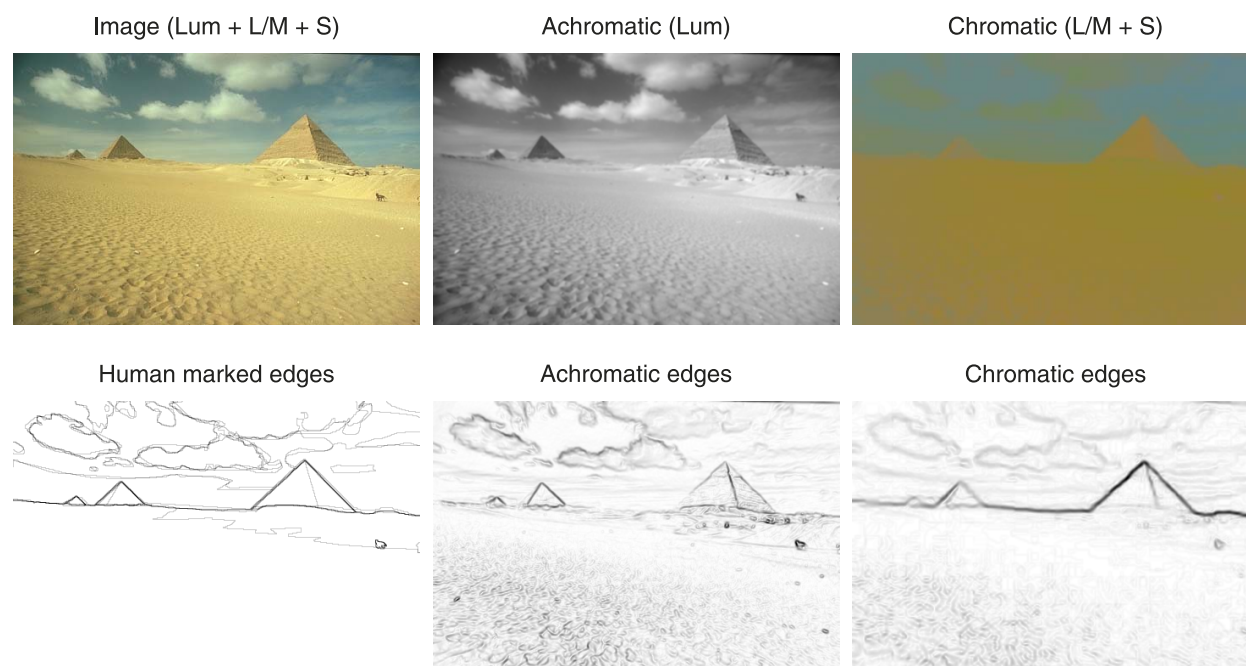
Figure 3. A sample image from the 100-image Berkeley Segmentation Dataset, where main contours are much better delineated in the chromatic than in the achromatic image.

Results

We determined ROC curves for the edge detector response to achromatic edges (Lum) and compared them to ROC curves for the edge detector response to achromatic and chromatic edges (Lum & Col). Object contours marked by human observers in the color images were used as ground truth. The ROC framework ensures that the increase in performance is independent of the threshold used to separate edges from the background.

A sample image where chromatic edges detected with the Sobel operator resulted in a considerable improvement is shown in Figure 3. Here the horizon and the sides of the pyramids are almost invisible in the achromatic edge map but well delineated in the chromatic L/M + S edge image.

The corresponding ROC curves are shown in Figure 4. Interestingly, the best prediction for the human-marked object contours in this image is obtained based on chromatic L/M edges alone; adding achromatic information worsens the prediction. More precisely, achromatic edges with an AUC of 57% failed to predict the human-marked object contours, while chromatic edges (Col ≡ L/M + S) gave a fair result (76%). The combination of achromatic and chromatic edges was poor (67%)—that is, worse than the prediction of chromatic edges alone. The same effect occurred when achromatic edges were added to the chromatic edges

detected in a single chromatic dimension (L/M or S). The reason for the bad prediction based on achromatic edges in this image is twofold: First, important contours in the image are almost isoluminant (such as the horizon or the pyramid contour against the sky) and thus missing in the achromatic edges. Second, unimportant contours such as the stones below the pyramid or the sand texture have a higher achromatic than chromatic contrast, leading to false positive responses. Human contour perception in this image clearly benefits from chromatic information.

Next, we evaluated the full BSD500 using the Sobel operator. In all cases, the average prediction was fair, with an average AUC of 70.8% for the achromatic edges and 75.2% for the chromatic edges. Two results obtained for the sample image in Figure 4 hold for the full data set: Performance was worst for the achromatic edges and best for the chromatic edges. Adding achromatic information to the chromatic edges had only a negligible effect on the performance. The average improvement was quantified by ΔAUC(Lum & Col, Lum) = AUC(Lum & Col) − AUC(Lum)—that is, the difference between the AUC based on the achromatic edges and the two chromatic edges (Lum & Col) and the AUC based on the achromatic edges (Lum). Its value was 4.3%.

We also analyzed the performance if only a single chromatic dimension was added to the achromatic dimension—that is, for simulated dichromats. The difference between the AUCs dropped to 1.8% for
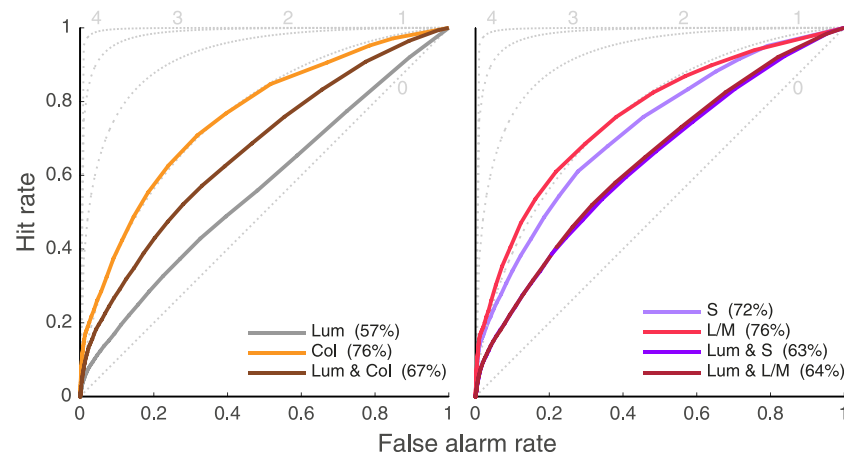
Figure 4. Receiver operating-characteristic (ROC) curves reflect the qualitative finding that for the pyramid image (Figure 3) chromatic edges can better predict human-marked object contours. ROC curves for all seven possible combinations of the three postreceptoral channels Lum, S, and L/M are shown. The ROC curve based on achromatic and chromatic edges (Lum & Col, brown) is well above the curve based on achromatic edges alone (Lum, gray). In fact, achromatic edges fail to predict human-marked object contours, and combined achromatic and chromatic edges yield a poor prediction. For this image, chromatic edges alone (Col) give the best prediction; adding achromatic information worsens the prediction (Lum & Col). Each chromatic channel alone results in fair performance (L/M and S); combining achromatic information with a single chromatic channel decreases the performance to poor (Lum & S, Lum & L/M). The area under the ROC curve is a single qualitative measure to characterize prediction advantage. Values of area under the curve for each curve are given in parentheses. Ideal ROC curves for $d' = 0$, 1, 2, 3, and 4 are shown for reference (dotted light gray).

protanope/deuteranope dichromats, lacking either the L or M cone type—that is, without the L/M dimension—and to 1.6% for tritanopes, lacking an S cone type—that is, without the S dimension.

Images that had the greatest color advantages and disadvantages are shown in Figure 5. Large advantages from using chromatic information occur if the observers marked largely isoluminant contours, such as the outline of the sports car, the red suit of the woman in front of the lawn, the bird's back against the lawn and its plumage coloration, or the back of the bears against the water. Small disadvantages occur for animals that use color to camouflage (like the shark, the moray eel, the cormorant, and the snake), for artificial objects with a chromatic texture that is not marked by the human observers (like the pattern on the Hawaiian shirt), and, trivially, for essentially achromatic images, like the fisherman in backlight.

So far, we have shown that chromatic information improves the prediction of human-marked object contours. We obtained the results for a particular choice of data set, edge detector, spatial scale, observer consensus used to define ground-truth images from the human-marked edges, and evaluation framework. In the following we shall investigate how robust our findings are against variation of these choices. We shall also run the analysis for color-calibrated images.

Performance for other data sets and edge detectors

To quantify the prediction advantage of chromatic edges, we computed the difference in the AUC between achromatic and full-color edges, ΔAUC(Lum & Col, Lum), and compared it to the prediction advantage for achromatic edges, ΔAUC(Lum & Col, Col). Histograms of both measures for the BSD and the Sobel operator and results averaged across all combinations of data sets and edge detectors are shown in Figure 6. The histogram for the color advantage has very few negative values, which would indicate a color disadvantage. On the contrary, the histogram for the luminance advantage has considerably more negative values, indicating a luminance disadvantage: For these images, adding achromatic edges to the chromatic edges worsens the prediction. For the BSD and the Sobel operator, color is advantageous for 99.6% of all images, while luminance is advantageous for only 47.4% of all images. The corresponding values for all data sets and edge detectors are 85.5% and 67.6%.

These differences are also reflected in the average prediction advantage. For the BSD and the Sobel operator the average color-prediction advantage is 4.3%, while the average luminance prediction advantage is even slightly negative (−0.06%), indicating a luminance disadvantage. For all data sets and edge detectors, the average color-prediction advantage is
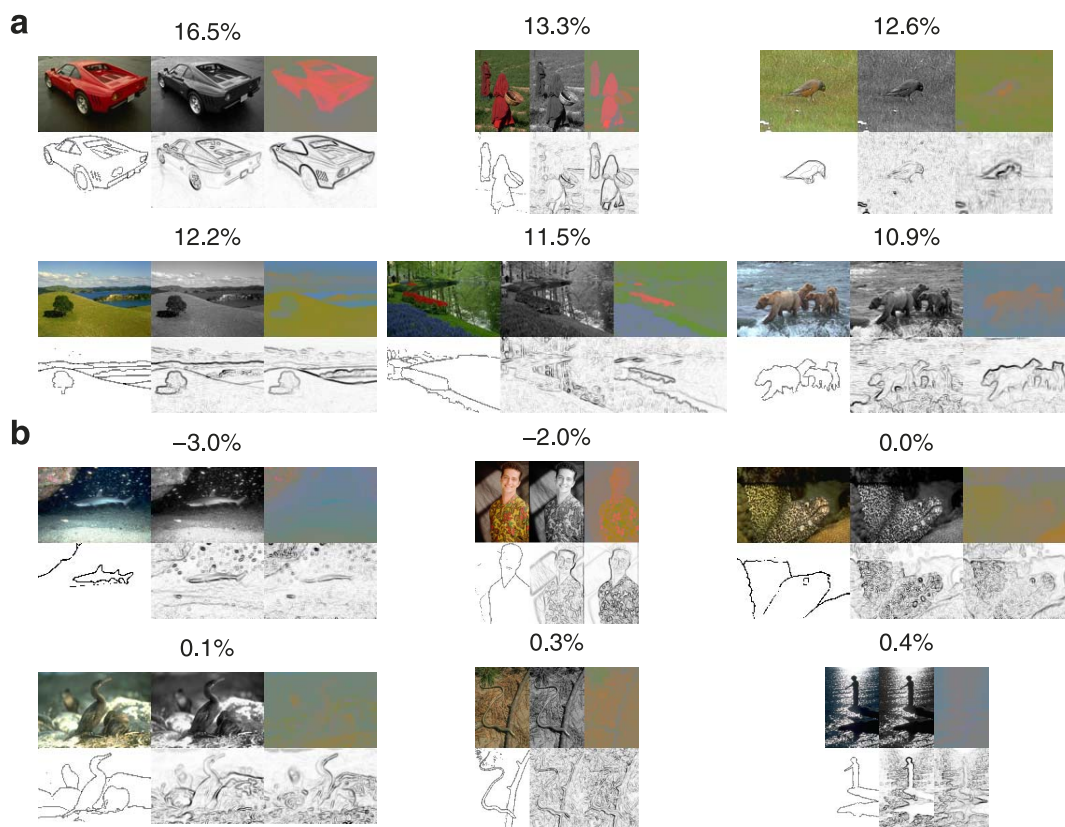
Figure 5. Images from the Berkeley Segmentation Dataset, where adding chromatic edges resulted in (a) the highest advantages and (b) the lowest advantages in predicting human-marked object contours. For each image, the original full-color image and the corresponding achromatic and chromatic image are shown together with the human-marked object contours and the edges detected in the achromatic dimension and the two chromatic dimensions.

3.5%, while the average luminance-prediction advantage is significantly smaller, namely only 2.4 (paired-samples *t* test), $t(4,615) = -8.25$, $p < 2.0002^{-16}$. This reflects the importance of color for human contour perception.

Next we show separately for each combination of data sets and edge detectors the percentage of images with a color advantage and the average color advantage $\Delta AUC(Lum \& Col, Lum)$ (Figure 7). We found that the main result was independent of the edge detector: For all data sets and edge detectors, adding chromatic information increased the performance compared to achromatic information alone. Color information is advantageous for more than 50% of the images in each data set, with an average value of 83%. The difference between the average AUCs for achromatic and full-color images were always positive, with values ranging from 0.7% (CORF for the ANID) to 8.4% (SimpleCell for the BSD). This difference was significant in 15 out of 20 cases (nonsignificant with CORF for the ANID, MGCCCD for the SOD, and Sobel or Gabor for the MGCCCD).

Dependence on spatial scale

We also investigated the dependence of the results on the spatial scale. The images were blurred by 2-D Gaussians of different standard deviations from $\sigma = 1$ to 16 pixels. We found that the color advantage as measured by $\Delta AUC(Lum \& Col, Lum)$ increased with spatial scale from 3.5% without blurring to a maximum of 4.9% for $\sigma = 12$ and decreased for higher values of σ.

Dependence of performance on observer consensus in defining ground truth

In this section we analyze how variations in defining the ground-truth image affect the results. In particular, we investigate the influence of observer consensus and edge thickness.

For the ROC analysis, binary ground-truth images are needed where a pixel is either an edge or not. In the results presented so far, a pixel has been considered a ground-truth edge pixel if it was marked by any observer (defined as 0% consensus). Alternatively,
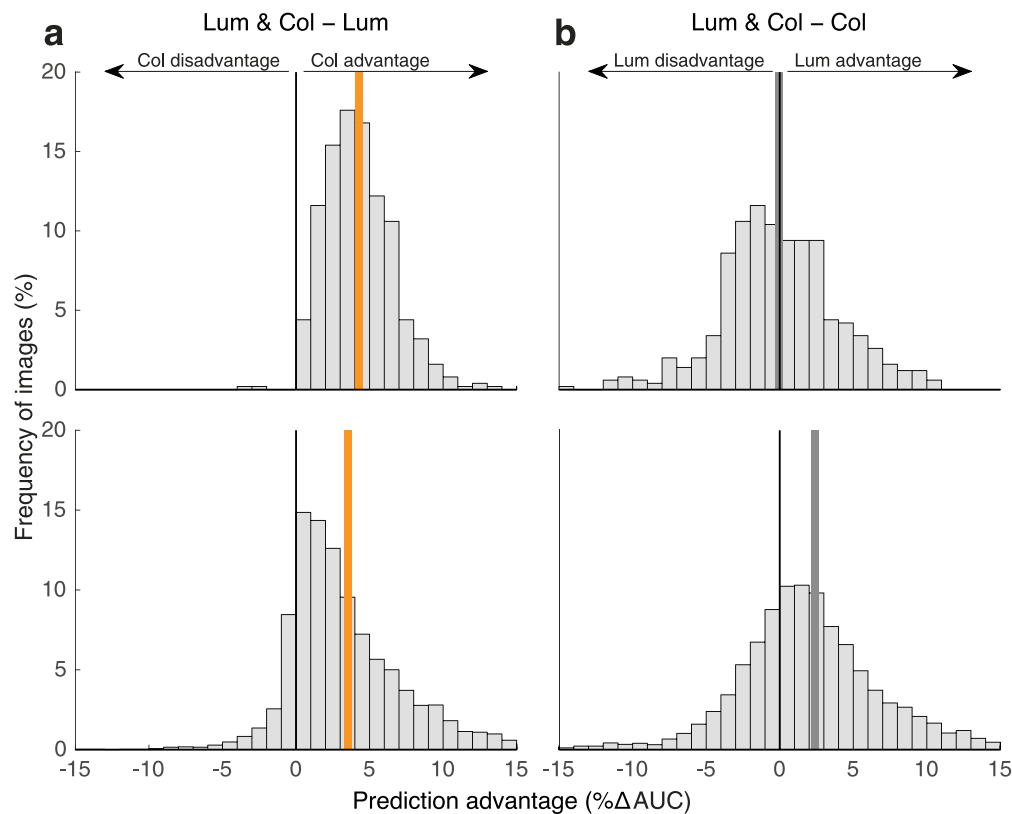
Figure 6. Histogram of differences in area under the receiver operating-characteristic curve. These are computed to assess (a) the color advantage ΔAUC(Lum & Col, Lum) and (b) the luminance advantage ΔAUC(Lum & Col, Col); data are shown for the Berkeley Segmentation Dataset and the Sobel operator (top row) and averaged across all data sets and operators (bottom row). Histograms are normalized to show the probability; data outside the interval [−15, 15] are not shown. Bold vertical lines mark the average.

ground-truth images could have been generated with a more severe criterion—that is, if they were marked by at least a fraction of all N observers, with the fraction ranging from $1/N$ to N. Does changing this criterion affect performance? If so, how? One hypothesis is that the performance remains the same, because increasing the criterion increases both hits and misses. An alternative hypothesis is that increasing the criterion reduces the fuzziness in the ground truth such that the operator performance would be better because the number of misses decreases. To affect the comparison of chromatic and achromatic edges, one must further assume that lowering the criterion boosts performance selectively for either chromatic or achromatic edges but not both.

To address this question we analyzed the data sets where more than a single observer has labeled the images—that is, the BSD and the SOD. We ran the ROC analysis separately for each of the n possible ground-truth images, where n is the maximum number of observers who labeled a pixel as an edge. Results are shown in Figure 8. We found that the number of images with a color advantage drops as the consensus increases. This could be interpreted as meaning that

observers agree more on achromatic edges compared to chromatic edges. The average improvement depending on consensus differed between data sets: For the BSD the advantage was largely robust against consensus changes, while it dropped for the SOD. These differences could be due to the different nature of contours in the two data sets: While the observers were asked to label "distinguished things" in the BSD, for the SOD they were asked to select just the salient contours.

The human-marked object contours have a width of only 1 pixel. This implies that observers have a very high level of spatial precision when defining object contours, which may or may not reflect their percept. To test whether our results are affected by the line width in the ground-truth data, we varied the criterion for a pixel to be classified as an edge by dilating the ground-truth maps with a disk of radius 3 or 6, corresponding to counting a pixel as hit if it is within a distance of 3 or 6 pixels from a human-marked object-contour pixel. We found no substantial difference in the overall color advantage. Values were 3.7% for a radius of 3 pixels and 4.5% for a radius of 6 pixels, compared to 3.5% for no dilation.
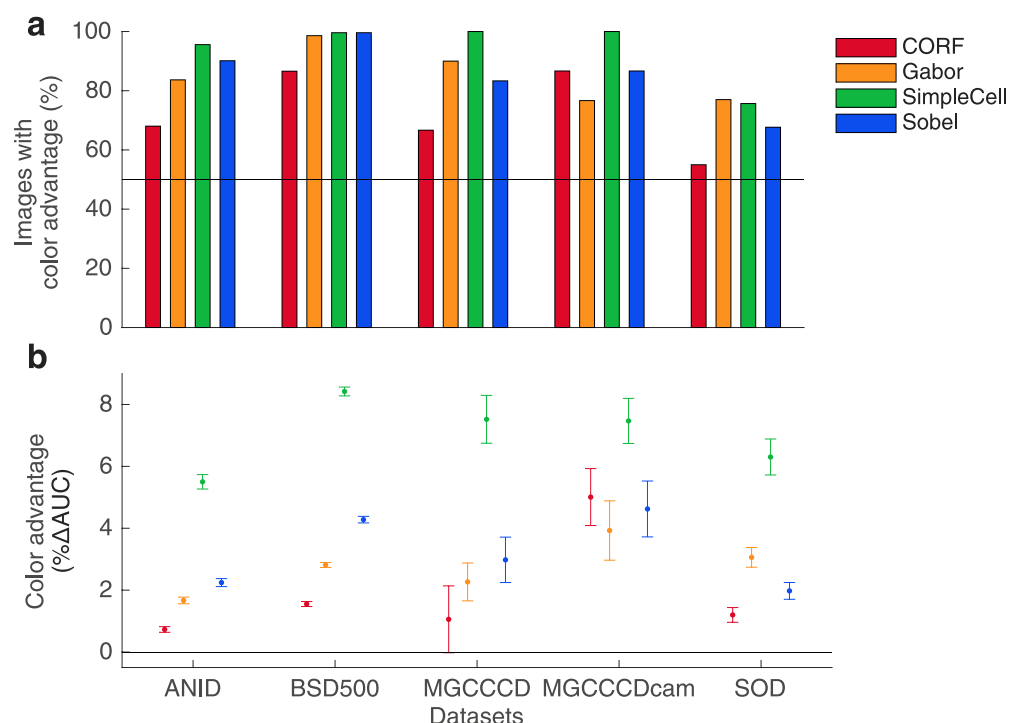
Figure 7. Color advantage for the different data sets and edge detectors. Color advantage is quantified for each combination of data set and edge detector by (a) the percentage of images where chromatic information was advantageous and (b) the difference between the area under the receiver operating-characteristic curve for achromatic and chromatic edges (Lum & Col) and achromatic edges (Lum). Error bars denote the standard error of the mean.

PR analysis

We also used PR curves and the *F*-measure to evaluate the contribution of chromatic information. Analogously to the ROC framework, we quantified the advantage of chromatic information by the average difference ΔF(Lum & Col, Lum) in the *F*-measure between achromatic and full-color images. The average ΔF-value was small (1.1%), but for some images it was considerably higher, up to 40%. The ΔF-value differed across data sets. For the ANID and the SOD there was almost no effect (0.1%), while for the BSD, MGCCCD (monitor-based conversion), and MGCCCD (camera-based conversion), respectively, the ΔF-value was 2.2%, 0.8%, and 2.6%. An analysis at different spatial scales revealed that the effect remained absent for the ANID (0.1%) but increased for the SOD (2.1%) and the other data sets (2.5%, 2.7%, and 4.6%, respectively). Values are given for blurring with a 2-D Gaussian of standard deviation 4; similar values were obtained for the other
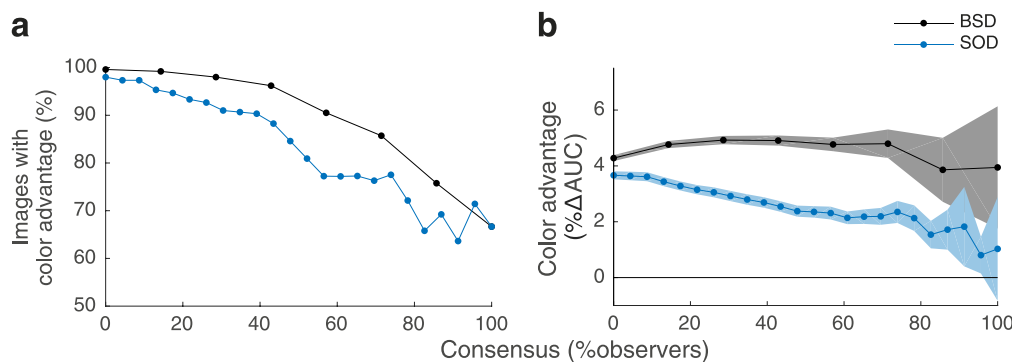

Figure 8. Dependence of color advantage on observer consensus for the Berkeley Segmentation Dataset and Salient Object Dataset. (a) The percentage of images with a color advantage drops as the consensus increases. (b) The average color advantage as quantified by the difference between the area under the receiver operating-characteristic curve for achromatic and chromatic edges (Lum & Col) and achromatic edges (Lum) is largely independent of observer consensus for the Berkeley Segmentation Dataset but drops for the Salient Object Dataset. Shaded areas denote the standard error of the mean.

tested standard deviations of 1 and 2. Overall, we found a color advantage across different scales and data sets in 15 out of 20 cases.

The mixed results show that the advantage of chromatic information is not completely independent of the evaluation framework. The absence of a color advantage for the ANID data set in the PR framework may be due to the fact that only a subset of all contours was marked in this data set, namely the animals' contours.

Discussion

We investigated the contribution of color and luminance information to predict human-marked object contours. We used different operators to detect edges in three dimensions of the DKL color space and compared them to human-marked object contours from four different data sets. We used an ROC framework for a threshold-independent comparison of edge-detector responses to the ground truth given by the human-marked object contours. Adding chromatic information was advantageous for 83% of the images. The improvement as quantified by the difference between the AUC for achromatic and chromatic edges versus achromatic edges, ΔAUC(Lum & Col, Lum), was small on average (3.5%) but considerably higher for some images, up to 52%. Interestingly, the luminance advantage ΔAUC(Lum & Col, Col) was smaller than the color advantage (2.4%). The reason for this might be that strong chromatic edges likely signal an object boundary, while strong achromatic edges can also result from shadow.

Effect of image format

The images of the BSD and the SOD—which builds upon the BSD—are provided in JPG format. The JPG image format defines a stronger compression of the chromatic channels compared to the achromatic channel, to reflect different sensitivities of the human visual system. This compression might reduce chromatic noise, and this might influence the responses of small operators like the Sobel operator. This may explain the particular result for the Sobel operator applied to the BSD, where chromatic information alone was as good as achromatic information combined with chromatic information (Figure 7).

Biases of data sets

In our analysis we used the images in the data sets, so any bias in these data sets may influence our results.

The data sets are not an unbiased random sample of the world. We used several data sets to minimize the influence of any one bias. For example, most images in the data sets show some objects. Because we focus on the contributions of color for human contour perception, the presence of objects in the image—rather than images of, for example, soil, sand, or lawn—is not problematic. Further, objects that attract attention are primarily focused and are therefore in general of high importance for human vision.

Image statistics versus scene statistics

In our investigation we were interested in the contribution of chromatic information in contour perception when natural scenes, not photographs on a monitor, are viewed. This requires a data set with object contours marked on calibrated images. A calibrated camera was used to obtain the images of the MGCCCD data set; the other data sets are based on uncalibrated images. For these uncalibrated images we could use only a monitor-based conversion from RGB to DKL. To investigate the degree to which the use of uncalibrated images may affect the results, we used the MGCCCD to compare results obtained with a camera- versus a monitor-based conversion. We found that results were largely independent of the conversion method: For both methods, color was advantageous for 90% of the images, and the ΔAUC values for achromatic vs. achromatic and chromatic information were similar (3.9% for the monitor-based conversion and 5.5% for the camera-based conversion, using the Sobel operator). We interpret these findings such that the results are largely independent of the conversion routine.

Small effect size of the chromatic advantage

The area under the ROC curve varies between 50% and 100%. This limits the theoretically maximal advantage to 50%. In practice, most operators do not fail (AUC < 60%), so even a poor detector has an AUC above 60%. The value that humans reach in predicting human-marked object contours (obtained from a leave-one-out ROC analysis for the BSD) is 85% for the BSD. Therefore, the AUC values in the domain that we studied here varies between 60% and 85%—that is, in a range of only 25%. The average advantage values below 5% that we found in the present study have to be judged based on this practically reachable maximal advantage.

A similar argument can be made for the *F*-measure. The small improvement that we found has to be related to the generally small value in improvement in the *F*-

measure in this domain. For example, 30 years of contour-detection research resulted in an improvement of the *F*-measure of 11%, from 60% for the Canny operator (Canny, 1986) to 71% for a biologically motivated model based on surround modulation (Akbarinia & Párraga, 2016).

Natural-scene-statistics findings on the independence of achromatic and chromatic edges

The highly correlated cone responses to natural images are decorrelated by the second-stage mechanisms $L + M$, $L - M$, and $S - (L + M)$, which show some resemblance to the response properties of retinal ganglion cells. The ganglion cells project along the optic tract to the LGN; a decorrelated representation allows for an optimal transmission of chromatic information from the retina to the LGN and further to primary visual cortex V1. The first study that showed this decorrelation was based on information theory (Buchsbaum & Gottschalk, 1983). Subsequent work by Atick, Li, and Redlich (1992) extended the analysis to the spatiotemporal domain. Ruderman, Cronin, and Chiao (1998) investigated the statistics of cone responses to natural images and confirmed the findings of Buchsbaum and Gottschalk.

Investigation of the spatiochromatic structure using principal-components analysis or independent-component analysis revealed that the achromatic and chromatic dimensions of the second-stage channels were entirely decorrelated—that is, based on activity in one channel, one cannot predict the activity in another channel (Heidemann, 2006; Ruderman et al., 1998; van Hateren, 1993; Wachtler, Lee, & Sejnowski, 2001; Webster & Mollon, 1997). The basis functions found in these studies are in most cases not uniform, but spatially structured. This provides evidence that the physical structure of the world is coded in color and luminance by the human visual system.

In previous work we investigated the joint distribution of chromatic and achromatic edges in natural scenes (Hansen & Gegenfurtner, 2009). We found that most edges combine luminance and color contrast. This is reflected in the observation that if one views a single dimension of a typical image, one can observe much if not all of the same spatial structure in each dimension. However, the magnitude and the sign of the luminance and color contrast at the edges are independent. Therefore, chromatic edge contrast is an independent source of information that can thus be linearly combined with other cues for subsequent processing, such as object segmentation. The present study provides evidence that human observers use chromatic information to delineate object contours.

Zhou and Mel (2008) collected joint responses of red–green and blue–yellow edge detectors both for ON- and OFF-edges using the BSD300 as ground truth. They investigated the rules for combining edge cues. Because the conditions for a linear combination of cues were not fulfilled for the color edge data they collected (statistical independence and exponential ratio), the combination rule was complex and nonlinear. Here we used a linear, additive combination of cues.

Note that these findings do not contradict findings in a different domain, where the dependence of the filter responses to achromatic natural scenes has been studied (Karklin & Lewicki, 2006; Schwarz & Simoncelli, 2001). These studies have found that the strength of the dependence varies depending on the difference in orientation and displacement between the filters and the complexity of the natural scene: The more similar the filter and the more regular the scene, the higher the dependency.

Psychophysical findings of chromatic contrast processing, contour detection, and scene recognition

The spatial resolution for the processing of achromatic stimuli is generally higher than that of chromatic stimuli. Contrast sensitivity for gratings with a spatial frequency above 0.5 c/° is higher for achromatic than for chromatic gratings (Kelly, 1983; Mullen, 1985; Sekiguchi, Williams, & Brainard, 1993). The spatial contrast sensitivity function is high-pass for achromatic contrast and low-pass for chromatic contrast. Besides these differences in the contrast sensitivity function, the processing of achromatic and chromatic form is highly similar (for a review, see Shevell & Kingdom, 2008). It has been proposed that achromatic and chromatic contrast are processed in multiple channels sensitive to different spatial frequencies that are similar for achromatic and chromatic processing (Bradley, Switkes, & De Valois, 1988; Losada & Mullen, 1995; Reisbeck & Gegenfurtner, 1998; Switkes, Bradley, & De Valois, 1988; Webster, De Valois, & Switkes, 1990). Chromatic and achromatic processing are also highly similar at a higher processing stage, where local orientations are grouped into coherent contours (McIlhagga & Mullen, 1996; Mullen, Beaudot, & McIlhagga, 2000) and object contours are extracted (Gheorghiu & Kingdom, 2007; Mullen & Beaudot, 2002).

While isolated chromatic and achromatic processing have been studied extensively, comparatively few studies have investigated the combined effect of chromatic and achromatic processing. The overall view is that chromatic and achromatic information are combined nonlinearly. Kingdom (2003) studied the

effect of combining achromatic and chromatic gratings and found that a 2-D color grating appeared as a 3-D object if an achromatic grating was added in phase but appeared flat if the n grating was added out of phase. Sharman, McGraw, and Peirce (2013b) studied the detection of achromatic and chromatic blur. They found that in natural images, chromatic blur was consistently harder to detect if combined with an unblurred achromatic image. The effect persisted if the chromatic and achromatic images were equated for contrast, showing that the masking was not due to the fact that in natural scenes, achromatic contrast is higher than chromatic contrast (Rivest & Cavanagh, 1996). The effect also persisted in pseudocolored versions of the images where the information in the chromatic channels was taken from the achromatic channel and vice versa, ruling out the possibility that the effect could be attributed to differences in the achromatic and chromatic image statistics.

A dominant effect of achromatic information in boundary processing has been found in the Boynton illusion, where straight chromatic edges are perceived to be aligned with irregular achromatic edges, and in the water-color effect (Pinna, Brelstaff, & Spillmann, 2001), where color is perceived to spread between achromatic contours. These studies suggest that color is primarily a surface property and plays only a secondary role in edge detection and localization (Mollon, 1989). However, a recent study by Sharman et al. (2013a) challenged this view. Those researchers asked observers to mark the edge location in superimposed bipartite achromatic and chromatic fields that were offset by a small amount (3 arcmin) but appeared to be fused. When the achromatic and chromatic contrasts were equated to be equally reliable for localizing the edge, chromatic information dominated the edge localization. The study found that the localization in the combined stimuli could be predicted from isolated conditions by a maximum-likelihood model that was adjusted by giving the chromatic component a higher weight.

The role of color in edge detection and form extraction is further substantiated by the fact that many visual shape illusions persist in purely chromatic versions. In the tilt aftereffect, observers adapt to a grating of stripes and perceive the orientation of a subsequently shown grating to be illusorily tilted in the direction away from the adapting orientation. The tilt aftereffect is believed to depend on the interaction of selectively adaptable orientation channels and has been found not only in achromatic but also in purely chromatic—that is, isoluminant—stimuli (Clifford, Spehar, Solomon, Martin, & Zaidi, 2003). Further, numerous geometric-optical illusions such as the Horizontal–Vertical, Poggendorff, Ponzo, and Zöllner illusions are as strong in the isoluminant version as in

the original achromatic version (Hamburger, Hansen, & Gegenfurtner, 2007).

Color also contributes to higher-level tasks such as image recognition. Gegenfurtner and Rieger (2000) used a delayed-match-to-sample paradigm to investigate the sensory and cognitive contribution of color to the recognition of natural scenes. Targets were presented at four different intervals (between 16 and 66 ms) to tap into the two stages of encoding and retrieval in the recognition process. It was found that color contributes to both stages. For short presentation times (16 and 33 ms), color images were encoded better: The percentage of correctly matched targets was higher if the color images were initially presented compared to luminance images. For longer presentation times (49 and 66 ms), color images were retrieved better: The recognition performance increased if the images were both presented and tested in color. The advantage of color in the encoding stage suggests a fast processing of chromatic information that may have its neural correlate in cell assemblies in primary and secondary visual cortex devoted to the processing of chromatic contours (Shapley & Hawken, 2011).

Design and evaluation of contour detectors based on human-marked object contours

Other studies have used human-marked object-contour maps to evaluate the performance of various edge detectors and design an optimized contour detector (Martin et al., 2004), or to investigate the features used for occlusion detection in natural scenes and design an occlusion detector (DiMattina, Fox, & Lewicki, 2012).

Martin et al. (2004) trained a classifier based on human-labeled images to obtain a detector that combines brightness, color, and texture cues to signal the posterior probability of a contour in an image. They found that a linear combination of cues resulted in the best detector. They quantified the performance using an *F*-measure based on precision and recall. Within the range of *F*-values marked by 0.58 for a simple detector based on Gaussian derivatives and 0.8 for the median human observer, their best detector based on brightness and texture yielded an *F*-measure of 0.65 that could be improved to 0.67 if color cues were added. The small benefit of color—0.02—closely resembles the value we find in the present work.

None of the edge detectors we investigated could perfectly predict the human-marked object contours. Finding an optimal contour detector is an active area of research, and the ROC framework we present here can be used to evaluate new models of contour detection.

ROC analysis versus PR curves

There are two established evaluation methods to characterize the trade-off between hits and false alarms in a detection task. One method is based on ROC curves that plot hit rate versus false-alarm rate (or recall vs. fallout, in image-retrieval terminology); the other method uses PR curves that plot precision versus recall (or sensitivity vs. hit rate). In the context of edge detection, precision is the number of correctly detected edge pixels divided by the number of *detected* edge pixels, and recall (or hit rate) is the number of correctly detected edge pixels divided by the number of *ground-truth* edge pixels.

Both ROC and PR curves show qualitatively the same trade-off between misses and false positives. ROC curves have been used to evaluate edge and boundary detectors (Abdou & Pratt, 1979; Bowyer et al., 2001) and junction detectors (Hansen & Neumann, 2004b). PR curves have been used to evaluate edge and boundary detectors (Martin et al., 2004).

Martin et al. (2004) argued that "ROC curves are not appropriate for quantifying boundary detection" (p. 536). The axes for an ROC curve are false-alarm rate (or fallout) and hit rate (or recall). Fallout is the probability that a true negative was labeled a false positive. In the context of edge detection, it is the number of falsely detected edge pixels divided by the number of ground-truth *background* pixels. Martin et al. continue: "Fallout is not a meaningful quantity for a boundary detector since it depends on the size of the pixels [i.e., the image resolution]. . . . Since boundaries are 1D . . . the number of true negatives will grow as n^2 while the number of true positives will grow as slow as n [the scaling of the pixel radius]. Thus, the fallout will decline as much as $1/n$" (2004, p. 536).

The argument of Martin et al. may be valid if the *absolute performances* of detectors are compared *across images of different sizes*. However, this reasoning is unclear because all operators are evaluated for the same images, namely the BSD, and are not compared across different data sets with images of different sizes. Unlike Martin et al., we were not interested in the absolute performance of detectors rather than in the benefit of using chromatic information to predict human contour perception. Any postulated effects of a stronger decline of the false-positive rate compared to the hit rate would affect the ROC curves based on achromatic information and on combined achromatic and chromatic information in the same way. Further, all comparisons were made on the same image, never on images of different sizes.

Moreover, the ROC is a threshold-free comparison, and the AUC is a well-established measure to characterize the ROC curve *with a single number*. For the PR curves, a different measure is used to characterize the curve with a single number, namely the maximum of the *F*-measure. The maximum of the *F*-measure differs in two ways from the AUC: First, the maximum along the F_α curve is just a single position, and not an integral measure as the AUC. Second, the *F*-measure involves an additional parameter α. This additional parameter is usually set to 0.5, corresponding to an equal weighting of precision and recall. Of course, other values would also be possible, resulting in different values extracted from the F_α curve. Different α-values may crucially affect the analysis.. Moreover, due to the need to define the parameter α, the PR framework does not have a main advantage of the ROC framework, namely being parameter free.

To sum up, we think that the argument of Martin et al. (2004) does not apply to the present work, and that the PR framework in general may have no advantage over the ROC framework.

Conclusions

In previous work (Hansen & Gegenfurtner, 2009) we have found that chromatic edge contrast is an independent source of information. Here we investigated the contribution of color to the detection of object contours in natural scenes. We found that color is advantageous for 83% of the images, and that this advantage can be very high in some cases. We found the result to be robust against variations in data set, edge detector, spatial scale, evaluation framework, and observer consensus. We conclude that chromatic information is helpful in contour detection.

Keywords: color, luminance, object-contour perception, natural scenes

Acknowledgments

Preliminary results have been reported in abstract form (Hansen & Gegenfurtner, 2013). Supplementary data and software are available at http://dx.doi.org/10.5281/zenodo.159566.

References

Abdou, I., & Pratt, W. (1979). Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proceedings of the IEEE*, *67*(5), 753–763.

Akbarinia, A., & Párraga, C. A. (2016). Biologically-inspired edge detection through surround modulation. In Richard C. Wilson, Edwin R. Hancock and William A. P. Smith (Eds.), *Proceedings of the British Machine Vision Conference 2016* (pp. 1–13), Durham, UK: BMVA Press. Retrieved from http://bmvc2016.cs.york.ac.uk/

Atick, J. J., Li, Z., & Redlich, A. N. (1992). Understanding retinal color coding from first principles. *Neural Computation*, *4*, 559–572.

Azzopardi, G., & Petkow, N. (2012). A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model. *Biological Cybernetics*, *106*, 177–189.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Bowyer, K., Kranenburg, C., & Doughtery, S. (2001). Edge detector evaluation using empirical ROC curves. *Computer Vision and Image Understanding*, *84*(1), 77–103.

Bradley, A., Switkes, E., & De Valois, K. D. (1988). Orientation and spatial frequency selectivity of adaptation to color and luminance gratings. *Vision Research*, *28*, 841–856.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society B: Biological Sciences*, *220*(1218), 89–113.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 679–698.

Clifford, C. W., Spehar, B., Solomon, S. G., Martin, P. R., & Zaidi, Q. (2003). Interactions between color and luminance in the perception of orientation. *Journal of Vision*, *3*(2):1, 106–115, doi:10.1167/3.2.1. [PubMed] [Article]

DiMattina, C., Fox, S. A., & Lewicki, M. S. (2012). Detecting natural occlusion boundaries using local cues. *Journal of Vision*, *12*(13):15, 1–21, doi:10.1167/12.13.15. [PubMed] [Article]

Drewes, J., Trommershäuser, J., & Gegenfurtner, K. R. (2011). Parallel visual search and rapid animal detection in natural scenes. *Journal of Vision*, *11*(2):20, 1–21, doi:10.1167/11.2.20. [PubMed] [Article]

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Flach, P., & Wu, S. (2003). Repairing concavities in ROC curves. In J. Editor (Ed.), *Proceedings of the 2003 UK Workshop on Computational Intelligence* (pp. 38–44). Bristol, UK: University of Bristol.

Forsyth, D. A., & Ponce, J. (2012). *Computer vision—A modern approach*. London: Pearson.

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, *10*, 805–808.

Geusebroek, J.-M., van den Boomgaard, R., Smeulders, A. W. M., & Geerts, H. (2001). Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 1338–1350.

Gheorghiu, E., & Kingdom, F. A. (2007). Chromatic tuning of contour-shape mechanisms revealed through the shape-frequency and shape-amplitude after-effects. *Vision Research*, *47*, 1935–1949.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons.

Hamburger, K., Hansen, T., & Gegenfurtner, K. R. (2007). Geometric-optical illusions at isoluminance. *Vision Research*, *47*(26), 3276–3285.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Hansen, T., & Gegenfurtner, K. R. (2009). Independence of color and luminance edges in natural scenes. *Visual Neuroscience*, *26*, 35–49.

Hansen, T., & Gegenfurtner, K. R. (2013). Higher order color mechanisms: Evidence from noise-masking experiments in cone contrast space. *Journal of Vision*, *13*(1):26, 1–21, doi:10.1167/13.1.26. [PubMed] [Article]

Hansen, T., & Neumann, H. (2004 a). A simple cell model with dominating opponent inhibition for robust image processing. *Neural Networks*, *17*(5–6), 647–662.

Hansen, T., & Neumann, H. (2004 b). Neural mechanisms for the robust representation of junctions. *Neural Computation*, *16*(5), 1013–1037.

Heidemann, G. (2006). The principal components of natural images revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*, 822–826.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*, 215–234.

Johnson, A. P., Kingdom, F. A., & Baker, C. L., Jr. (2005). Spatiochromatic statistics of natural scenes: First and second-order information and their correlational structure. *Journal of the Optical Society of America A*, *22*(10), 2050–2059.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1233–1258.

Karklin, Y., & Lewicki, M. S. (2006). Is early vision optimized for extracting higher-order dependencies? In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 635–642). Cambridge, MA: MIT Press. Retrieved from http://papers.nips.cc/paper/2901-is-early-vision-optimized-for-extracting-higher-order-dependencies.pdf.

Kelly, D. H. (1983). Spatiotemporal variation of chromatic and achromatic contrast thresholds. *Journal of the Optical Society of America*, *73*, 742–750.

Kingdom, F. A. A. (2003). Color brings relief to human vision. *Nature Neuroscience*, *6*(6), 641–644.

Losada, M. A., & Mullen, K. T. (1995). Color and luminance spatial tuning estimated by noise masking in the absence of off-frequency looking. *Journal of the Optical Society of America A*, *12*, 250–260.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freemann & Co.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(1), 530–549.

Martin, D., Fowlkes, C., Tal. D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological *statistics*. In *Proceedings Eighth International Conference on Computer Vision ICCV* 2001 (Vol. 2, pp. 416–423). Washington, DC: IEEE Computer Society

McIlhagga, W. H., & Mullen, K. T. (1996). Contour integration with colour and luminance contrast. *Vision Research*, *36*, 1265–1279.

Mollon, J. D. (1989). "Tho' she kneel'd in that place where they grew...": The uses and origins of primate colour vision. *Journal of Experimental Psychology*, *146*, 21–38.

Movahedi, V., & Elder, J. H. (2010). Design and perceptual validation of performance measures for salient object segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops* (pp. 49–56). Washington, DC: IEEE Computer Society

Mullen, K. T. (1985). The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *Journal of Physiology*, *359*, 381–400.

Mullen, K. T., & Beaudot, W. H. (2002). Comparison of color and luminance vision on a global shape discrimination task. *Vision Research*, *42*, 565–575.

Mullen, K. T., Beaudot, W. H., & McIlhagga, W. H. (2000). Contour integration in color vision: A common process for the blue-yellow, red-green and luminance mechanisms? *Vision Research*, *40*, 639–655.

Olmos, A., & Kingdom, F. A. A. (2004). A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, *33*, 1463–1473.

Párraga, C. A., Troscianko, T., & Tolhurst, D. J. (2002). Spatiochromatic properties of natural images and human vision. *Current Biology*, *12*, 483–487.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press.

Pinna, B., Brelstaff, G., & Spillmann, L. (2001). Surface color from boundaries: A new "watercolor" illusion. *Vision Research*, *41*(20), 2669–2676.

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

Reisbeck, T. E., & Gegenfurtner, K. R. (1998). Effects of contrast and temporal frequency on orientation discrimination for luminance and isoluminant stimuli. *Vision Research*, *38*, 1105–1117.

Rivest, J., & Cavanagh, P. (1996). Localizing contours defined by more than one attribute. *Vision Research*, *36*(1), 53–66.

Rubin, J. M., & Richards, W. A. (1982). Color vision and image intensities: When are changes material. *Biological Cybernetics*, *45*, 215–226.

Ruderman, D. L., Cronin, T. W., & Chiao, C. C. (1998). Statistics of cone responses to natural

images: Implications for visual coding. *Journal of the Optical Society of America A*, *15*, 2036–2045.

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

Sekiguchi, N., Williams, D. R., & Brainard, D. H. (1993). Aberration-free measurements of the visibility of isoluminant gratings. *Journal of the Optical Society of America A*, *10*, 2105–2117.

Sen, A. (1977). *On economic inequality* (2nd ed.). Oxford, UK: Oxford University Press.

Shapley, R., & Hawken, M. J. (2011). Color in the cortex: Single- and double-opponent cells. *Vision Research*, *51*, 701–717.

Sharman, R. J., McGraw, P. V., & Peirce, J. W. (2013 a). Cue combination of conflicting color and luminance edges. *Journal of Vision*, *13*(9): 1257, doi: 10.1167/13.9.1257. [Abstract]

Sharman, R. J., McGraw, P. V., & Peirce, J. W. (2013 b). Luminance cues constrain chromatic blur discrimination in natural scene stimuli. *Journal of Vision*, *13*(4):14, 1–10, doi:10.1167/13.4.14. [PubMed] [Article]

Shevell, S. K., & Kingdom, F. A. (2008). Color in complex scenes. *Annual Reviews in Psychology*, *59*, 143–166.

Stockman, A., & Brainard, D. H. (2009). Color vision mechanisms. In M. Bass (ed.), *OSA handbook of optics: Vol. III. Vision and vision optics* (3rd ed., pp. 11.1–11.104). New York: McGraw-Hill.

Switkes, E., Bradley, A., & De Valois, K. K. (1988). Contrast dependence and mechanisms of masking interactions among chromatic and luminance gratings. *Journal of the Optical Society of America A*, *5*, 1149–1162.

Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of color in high-level vision. *Trends in Cognitive Sciences*, *5*(5), 211–215.

Tappen, M. F., Freeman, W. T., & Adelson, E. H. (2005). Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(9), 1459–1472.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

van Hateren, J. H. (1993). Spatial, temporal and spectral pre-processing for colour vision. *Proceedings of the Royal Society of London B: Biological Sciences*, *251*(1330), 61–68.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworth. Retrieved from http://www.dcs.gla.ac.uk/Keith/Preface.html

Wachtler, T., Lee, T. W., & Sejnowski, T. J. (2001). Chromatic structure of natural scenes. *Journal of the Optical Society of America A*, *18*, 65–77.

Webster, M. A., De Valois, K. K., & Switkes, E. (1990). Orientation and spatial-frequency discrimination for luminance and chromatic gratings. *Journal of the Optical Society of America A*, *7*, 1034–1049.

Webster, M. A., & Mollon, J. D. (1997). Adaptation and the color statistics of natural images. *Vision Research*, *37*, 3283–3298.

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 509–520.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, *9*(1), 33–39.

Zeki, S. M. (1976). The functional organization of projections from striate to prestriate visual cortex in the rhesus monkey. *Cold Spring Harbor Symposia on Quantitative Biology*, *15*, 591–600.

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