Effects of chromatic image statistics on illumination induced color differences

Marcel P. Lucassen,[1,*] Theo Gevers,[1] Arjan Gijsenij,[2] and Niels Dekker[2]

[1]*Intelligent System Laboratory Amsterdam, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

[2]*Akzo Nobel Coatings, Rijksstraatweg 31, 2171 AJ Sassenheim, The Netherlands*

**Corresponding author: marcel@lucr.nl*

Received February 20, 2013; revised July 17, 2013; accepted August 2, 2013; posted August 2, 2013 (Doc. ID 185571); published August 28, 2013

We measure the color fidelity of visual scenes that are rendered under different (simulated) illuminants and shown on a calibrated LCD display. Observers make triad illuminant comparisons involving the renderings from two chromatic test illuminants and one achromatic reference illuminant shown simultaneously. Four chromatic test illuminants are used: two along the daylight locus (yellow and blue), and two perpendicular to it (red and green). The observers select the rendering having the best color fidelity, thereby indirectly judging which of the two test illuminants induces the smallest color differences compared to the reference. Both multicolor test scenes and natural scenes are studied. The multicolor scenes are synthesized and represent ellipsoidal distributions in CIELAB chromaticity space having the same mean chromaticity but different chromatic orientations. We show that, for those distributions, color fidelity is best when the vector of the illuminant change (pointing from neutral to chromatic) is parallel to the major axis of the scene's chromatic distribution. For our selection of natural scenes, which generally have much broader chromatic distributions, we measure a higher color fidelity for the yellow and blue illuminants than for red and green. Scrambled versions of the natural images are also studied to exclude possible semantic effects. We quantitatively predict the average observer response (i.e., the illuminant probability) with four types of models, differing in the extent to which they incorporate information processing by the visual system. Results show different levels of performance for the models, and different levels for the multicolor scenes and the natural scenes. Overall, models based on the scene averaged color difference have the best performance. We discuss how color constancy algorithms may be improved by exploiting knowledge of the chromatic distribution of the visual scene.

OCIS codes: (330.1715) Color, rendering and metamerism; (330.1720) Color vision; (330.5510) Psychophysics.

http://dx.doi.org/10.1364/JOSAA.30.001871

1. INTRODUCTION

The background of this paper lies in our desire to improve our understanding of the relationship between performance measures of the human and computational color constancy approaches. As explained hereafter, these performance measures relate to very different aspects of color constancy. Here we study a new psychophysical method that yields observer judgments of the color fidelity of scenes under changing illumination. We expect these measurements to be more easily related to the outcome of color constancy algorithms and that they may help to tune such algorithms to accommodate the results from perceptual studies.

Color constancy is the property of a visual system (either human or machine) to maintain stable object color appearances despite considerable changes in the spectral composition of the illuminant. For quite some time it has been recognized as one of the central themes in color research. The key issue is how to disentangle the wavelength-by-wavelength product of the illuminant spectral power distribution (SPD) and the object reflectance function that is sampled by the visual system. Yet, the methodological approaches and performance measures in perception (human vision) studies and computational (computer vision) studies are very different. In computer vision, the main approach to solving the color constancy problem is by estimating the illuminant from the visual scene after which reflectance may be recovered [1–5] or the color balance of images may be corrected for display or to support object recognition [6]. The performance of such color constancy algorithms is usually quantified by the angular error [7], a measure for the chromatic mismatch between the estimated illuminant and the true illuminant, which is assumed to be known. So, the performance of computational color constancy algorithms that rely on illuminant estimation is quantified by a number relating to a global illuminant.

The degree of constancy exhibited by human observers is often quantified by a color constancy index [8]. A common finding in the many psychophysical studies on color constancy is that human color constancy is not perfect. It depends on the experimental method employed and the observer's state of adaptation, among other things. The main techniques are color matching, color naming, nulling to maintain neutral appearance, discriminating between a change in illumination and a change in surface reflectance (operational approach), and identification of surfaces across illuminants [9,10]. What these techniques have in common is that each measurement (i.e., each observer response) relates to the appearance of a single object or patch in the scene. This poses a problem for our desired comparison of color constancy performance

measures: the psychophysical measurement relates to a single (local) object, whereas the computational measure relates to a single (global) illuminant.

To solve this, we study a method for assessing human color constancy, featuring two new methodological elements. First, the observers are asked to judge the color fidelity of the whole scene, instead of a judgment on the appearance of a single color or object. Second, in each experimental trial the observers deal with a scene rendered under three illuminants (one reference and two test illuminants) instead of the usual two (one reference and one test illuminant). We denote this by the term "triad illuminant comparison". The advantage of this method is that an observer measurement is obtained relating to the scene as a whole and relating the color rendering properties of one illuminant to another. The downside of this is that, with the simultaneous display of the illuminant conditions, the state of adaptation is also affected. The extent that our results are influenced by this, and differ from fully natural viewing conditions, is unknown. We refer to Foster [10] for an overview and discussion of the many studies that employ different variants of simultaneous presentation. All we can say is that our measurements most probably arise from the fast components underlying the time dependency of chromatic adaptation [11].

Using our method, we measure the color fidelity on three image data sets. The first data set is composed of multicolor images, collections of colored rectangles with elliptical chromatic distributions having the same average (neutral) chromaticity, but with different orientations in color space. We have previously shown that these different chromatic orientations lead to different perceptual estimates of the color fidelity [12]. To investigate whether this effect is also found in natural images, we compose a second data set from natural images that, similar to the first data set, has specific ratios and orientations of the first and second principal components in chromaticity space. The natural images introduce an artifact that cannot be ignored: image semantics might bias observers toward responses that have no direct relation with the low-level contents of the image. Therefore, we use a third data set that is composed of pixelated and scrambled versions of the natural images in the second data set. In this way, the third data set has the same chromatic distribution as the second set, and the same granularity as the first data set. This allows us to investigate whether image semantics will have a major effect on the assessments of the global image fidelity.

We then proceed by quantitative modeling to account for the average observer data. Since there is no model for human color constancy available that handles images of visual scenes as a whole, we present four different models that take the spatio chromatic scene content into account. These four models differ in the extent to which they incorporate processing of the human visual system. Our first model predicts on the basis of the reflected light signal, a purely physical signal that we introduce in this paper. The second model predicts on the basis of the calculated overlap in rendered color gamut for the test and reference illumination. The third model predicts on the basis of the scene averaged ΔE color difference metric [13–15] between corresponding image parts under the test and reference illumination. The fourth model takes into account spatial information of the scene and of the visual system, either in the form of the S-CIELAB model [16] or as an image quality metric [17]. Finally, we discuss how color constancy algorithms may be improved by exploiting knowledge of the chromatic distribution of the visual scene.

2. METHODS

A. Triad Illuminant Comparison

The triad illuminant comparison method involves a test scene rendered under three illuminants (Fig. 1). One of these illuminants is the reference (R) while the other two are the test illuminants (T1 and T2). An observer has to select the test image (test illuminant) having the best color fidelity in comparison to the reference image shown above each test image. They do this by visually comparing the colors of the test scene rendered under the test illuminants with the colors rendered under the reference illuminant. In essence, the observer's task is to judge the differences between two image pairs, the first pair of images (T1, R) being the scene rendered under test illuminant T1 and under the reference illuminant R, the second pair (T2, R) being the scene under test illuminant T2 and R.

Although three images would, in principle, suffice here we deliberately use four images to maintain symmetry in eye movement patterns (see the Instructions section), anticipating a potential follow-up study employing eye tracking. Once an image pair is selected by the subject, the associated test illuminant is ranked higher by means of a scoring mechanism, hence is likely to give better color constancy. Here, we describe experiments in which we use four test illuminants to form six unique illuminant pairs (1 versus 2, 1 versus 3, 1 versus 4, 2 versus 3, 2 versus 4, 3 versus 4). The measurements from these six pairs allow a relative ranking of the four illuminants. This paper reports on the measurement from three image data sets, described below.

B. Data Set 1: Multicolor Test Scenes with Specified Chromatic Distributions

We synthesized multicolored images composed of 900 square color patches, arranged as a 30×30 matrix of adjacent patches (see Fig. 2), whose distribution in CIELAB color space was under control. In our reference condition, the 900 patches follow a Gaussian distribution, with standard

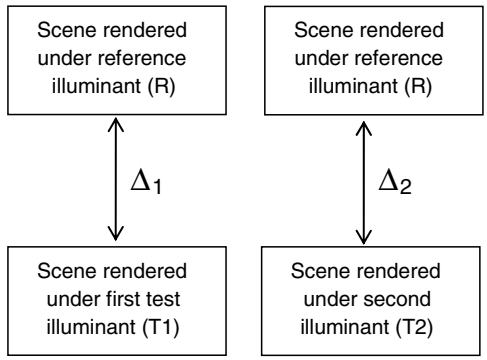

Fig. 1. Triad illuminant comparison method involves a test scene rendered under a reference illuminant (R) and two different test illuminants (T1 and T2). The visual differences between the scene rendered under reference and the two test illuminants are denoted by Δ_1 and Δ_2. Global color fidelity of the test scenes under T1 and T2 is measured by observers indicating which of the two differences (Δ_1 or Δ_2) appears smaller.

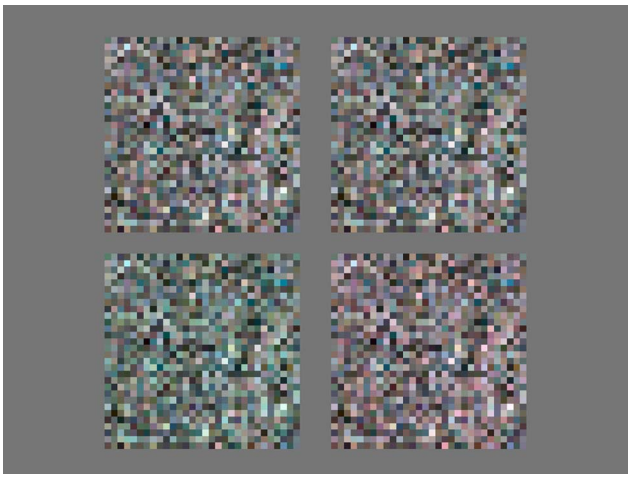

Fig. 2. Screenshot of an experimental trial for Data Set 1. The multicolor test scene on the top row represents a chromatic distribution specified in CIELAB color space. For each color element we derive a spectral reflectance function that is used to simulate the effect of illuminant changes. A greenish and a reddish test illuminant is used in this trial to render the bottom left and right images, respectively. Observers indicate which of these two renderings has the best color fidelity compared to the rendering under neutral reference illumination (in the top row). The size of the background is a $39.6° \times 30.2°$ visual angle. Images are $16.6° \times 16.6°$ each in Data Set 1, and $6.2° \times 6.2°$ in Data Sets 2 and 3. Horizontal and vertical separation is 2.0° and 0.9°, respectively.

deviations along the L* axis twice the standard deviations along the chromatic a* and b* axes. This is done to match the distribution of RGB derivatives of the Corel image database, which contains 40,000 images representative of the "real world" [18]. So, the chromatic distribution of our reference condition approximately conforms to the statistics of natural images. In addition to the reference distribution, we study four chromatic distributions having a 5:1 ratio in the standard deviations along the major and minor axes of the distribution in the a*, b* plane. These four distributions differ in the orientation of the major axis, denoted by θ (the angle between the distributions' major axes and the positive a* axis) in Fig. 3. Table 1 shows the parameters of the five chromatic distributions discussed above. It shows that the mean L* value lies around 49 and that the standard deviation in L* is larger than that of the chromatic components a* and b*. Also, the skewness and kurtosis values of the L*, a*, b* vary around 0 and 3, respectively. Skewness is an indicator for the asymmetry in the distribution, with a value of 0 for a normal distribution. The kurtosis is a measure for the peakedness of a distribution, with a value of 3 belonging to a normal distribution. Distributions 2–4 have a 5:1 ratio of the standard deviations along the major and minor axis in the a*, b* chromaticity plane. Angle θ denotes the orientation of the major axis. See Fig. 3 also.

C. Data Set 2: Natural Images with Selected Chromatic Distributions

The second image data set consists of 25 images that we selected from the gray ball image data set (11,346 images) described in [19]. The idea behind these selected images is that their chromatic distributions are oriented similarly to those of the synthesized images shown in Fig. 3, except that the scene averaged chromaticity is not restricted to coincide with the neutral point. Images from the data set were first transformed to CIELAB space using D65 as the white point, after which a principal component analysis was performed in the a*b* chromaticity plane. We selected images based on two criteria. First, we search for images having an approximate 5:1 ratio of the eigenvalues from the first and second principal components. This is inspired by the 5:1 ratio in the standard deviations along the major and minor axes of the chromatic distributions of Data Set 1. Second, we searched for images having target angles $\Phi = 0$, 45, 90, and 135° where Φ is the angle (in CIELAB a*b* chromaticity space) between the axis of the first principal component and the horizontal line passing through the average chromaticity (a line parallel to the a* axis). For each of the four target angles Φ, we selected 5 images. Another 5 images are selected that have a 1:1 ratio of the eigenvalues from the first and

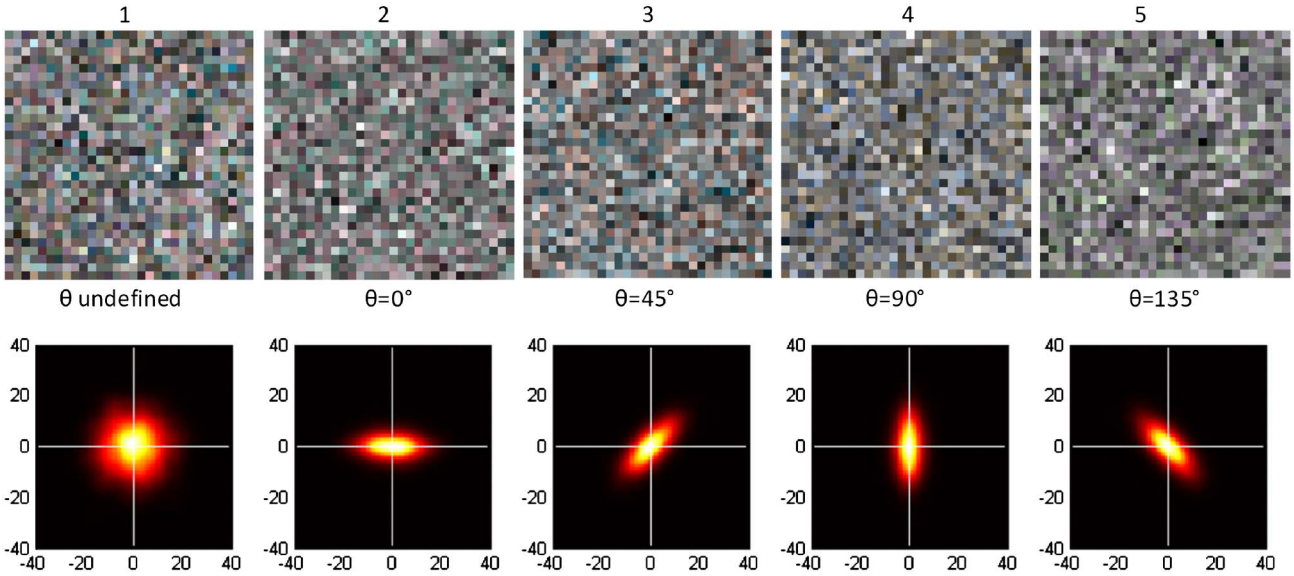

Fig. 3. Multicolor stimuli of Data Set 1 under D65 reference illumination (top row) and density plots of their chromatic distributions plotted in the CIELAB a*, b* chromaticity plane (bottom row, a* on the horizontal axis). The numbers above the top images label the distributions described in Table 1. Angle θ denotes the angle between the positive a* axis and the major axis of the distribution in the a*b* plane of CIELAB color space.

Table 1. Specification of the Five Chromatic Distributions (900 Samples) under D65 Reference Illumination

		Mean			Standard Deviation			Skewness			Kurtosis		
Distribution	θ	L*	a*	b*	L*	a*	b*	L*	a*	b*	L*	a*	b*
1	—	48.71	−0.01	0.09	16.18	7.93	7.80	−0.08	0.02	−0.02	2.69	2.82	3.24
2	0°	49.13	−0.14	0.13	15.70	7.79	1.67	0.08	0.00	0.01	2.93	2.92	3.06
3	45°	49.04	−0.08	0.15	16.15	5.65	5.68	−0.11	0.13	0.06	2.78	2.96	2.92
4	90°	48.76	−0.15	0.05	15.76	1.66	7.98	−0.09	0.05	0.02	2.75	3.06	3.02
5	135°	48.90	−0.06	−0.01	15.47	5.58	5.56	−0.03	−0.05	0.01	3.00	2.84	3.10

second principal components. So, for each of the five chromatic distributions defined in Data Set 1, we selected 5 natural images with similar (but not exactly equal) distributions. These images are 240 pixels × 240 pixels in size and shown in Fig. 4. Density plots of their distributions in CIELAB a*b* chromaticity space are shown in Fig. 5 and statistical parameters of the distributions are given in Table 2. To interpret the values of this Table, a positive skewness value indicates that the distributions' tail on the right side is longer than the tail on the left side, and vice versa for a negative skewness value. Kurtosis values larger than 3 indicate sharper peaked distributions, while values lower than 3 indicate a more flattened distribution. By comparing the graphical and statistical representations of the chromatic distributions (Figs. 3 and 5, and Tables 1 and 2) it is clear that the chromatic distributions of the natural images are different from the well-defined distributions of the multicolor images in Data Set 1, most notably perhaps in the tail of the distributions and the non-neutral average chromaticity. This is not a problem per se, since our model predictions make use of the actual image statistics. In Table 2, Φ is the angle (in CIELAB a*b* chromaticity space) between the axis of the first principal component and the horizontal line passing through the average chromaticity.

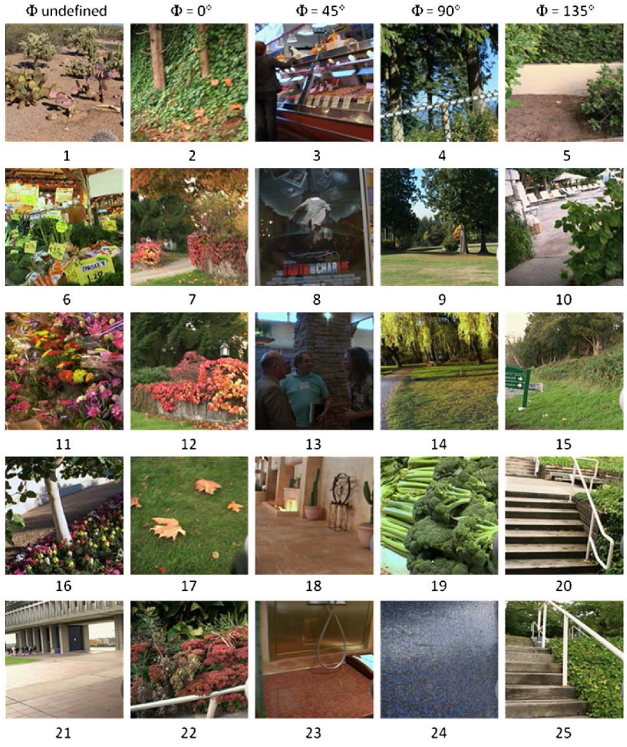

Fig. 4. Image Data Set 2, containing 5 natural images per chromatic distribution. Angle Φ denotes the angle in CIELAB a*b* chromaticity space between the axis of the first principal component and the horizontal line parallel to the a* axis. Numbers below the images are labels and correspond to the image numbers in Table 2.

D. Data Set 3: Scrambled Natural Images

The third data set is composed of pixelated and scrambled versions of the images of Data Set 2. Pixelation causes each 8 × 8 pixel block to get the average color of the original 64 pixels in the block. This results in lower resolution images of 30 × 30 blocks, just like the synthesized images of Data Set 1. Scrambling of the images causes the blocks to be spatially reorganized (randomly). In this way, the chromatic distributions of the corresponding images in Data Sets 3 and 2 are the same, but the image semantics are destroyed. In addition, the images of the third data set have the same granularity as the synthesized images in Data Set 1. Fig. 6 shows the scrambled images, and Table 3 presents the statistical parameters of their chromatic distributions. In Table 3, Φ is the angle (in CIELAB a*b* chromaticity space) between the axis of the first principal component and the horizontal line passing through the average chromaticity.

From a comparison of Tables 2 and 3, it becomes clear that scrambling does have an effect on the chromatic distribution, albeit small.

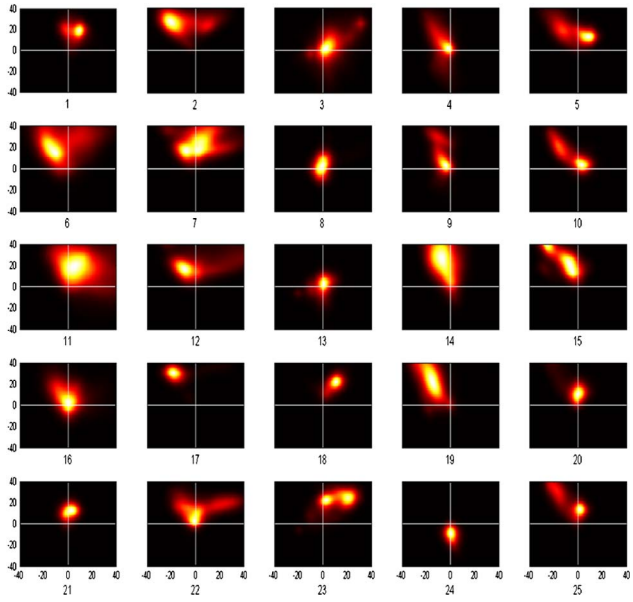

Fig. 5. Density plots of the chromatic distributions in CIELAB a*b* chromaticity space of the natural images shown in Fig. 4. Labels below the images correspond to those in Fig. 4 and the image numbers in Table 2.

Table 2. Parameters of the Chromatic Distributions of the 25 Images Shown in Fig. 4 (Data Set 2), under D65 Reference Illumination

		Mean			Standard Deviation			Skewness			Kurtosis		
Image	Φ	L*	a*	b*	L*	a*	b*	L*	a*	b*	L*	a*	b*
1	—	52.41	5.62	17.76	18.85	5.71	6.88	−0.95	0.00	−0.56	3.25	3.07	5.07
6	—	49.74	−5.68	27.45	26.65	15.45	16.93	0.09	0.71	0.44	1.75	3.46	3.24
11	—	35.49	12.25	19.67	19.70	16.31	15.43	0.59	0.75	0.45	2.99	3.11	4.42
16	—	34.28	0.70	8.41	28.56	10.66	10.99	0.86	1.84	1.03	2.61	9.01	6.32
21	—	62.85	1.11	11.56	23.52	4.46	5.95	−0.60	−2.11	1.67	2.64	15.48	22.58
2	0°	47.33	−9.95	26.03	19.79	13.61	7.39	−0.17	0.78	−0.23	2.42	2.72	3.94
7	0°	42.56	3.98	23.90	18.88	13.33	9.18	0.52	0.54	0.29	2.85	3.57	2.43
12	0°	39.32	4.33	21.94	20.14	20.16	9.94	0.37	0.85	0.82	2.33	2.76	3.24
17	0°	41.87	−14.31	29.51	11.84	9.93	5.29	1.30	2.51	−1.87	6.92	9.68	10.93
22	0°	29.94	3.56	13.72	21.56	13.85	8.12	0.78	0.76	0.03	3.45	2.87	2.80
3	45°	34.05	7.44	4.34	26.38	10.88	13.51	0.72	0.56	−0.09	2.61	3.46	4.41
8	45°	24.83	0.10	2.85	16.78	5.54	7.63	1.00	3.68	0.42	3.27	22.12	3.90
13	45°	20.99	−0.65	0.36	20.20	6.11	7.50	1.84	−1.13	−1.03	6.56	5.59	6.38
18	45°	47.56	8.78	19.81	12.85	5.32	5.72	0.07	−1.28	−0.42	4.30	6.46	5.34
23	45°	33.01	11.04	21.54	11.97	10.28	7.84	1.30	−0.36	−1.35	8.83	2.53	5.69
4	90°	32.36	−7.00	5.68	27.81	6.23	15.06	0.55	−0.37	−0.20	2.12	3.04	3.23
9	90°	33.12	−6.93	12.19	28.14	4.97	13.60	0.40	−0.12	0.11	1.71	3.48	2.43
14	90°	39.85	−6.12	27.19	21.08	7.60	16.31	0.17	−0.05	0.08	2.51	2.86	2.82
19	90°	47.81	−16.64	25.72	24.56	7.67	14.83	0.05	0.00	0.05	2.16	3.08	2.43
24	90°	43.35	0.92	−10.58	13.88	2.12	5.58	0.51	0.37	−0.82	2.42	3.80	4.14
5	135°	47.77	−3.92	17.62	23.86	11.58	8.33	0.30	−0.27	1.09	2.34	1.92	4.39
10	135°	44.47	−5.69	12.24	27.00	10.05	10.51	0.48	−0.13	0.54	2.02	1.75	2.77
15	135°	46.76	−12.57	25.47	19.57	9.78	10.85	0.33	−0.22	0.13	3.17	1.87	2.06
20	135°	46.84	−3.41	15.33	31.34	8.43	10.88	0.03	−1.33	1.57	1.56	4.52	5.05
25	135°	43.77	−8.34	21.46	21.92	9.67	11.49	0.56	−0.19	0.48	3.08	1.63	2.24

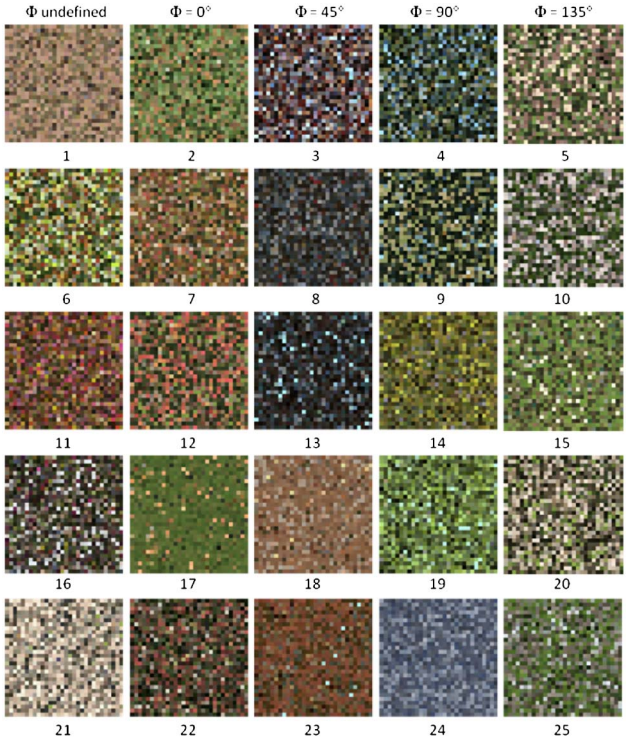

Fig. 6. Image Data Set 3. These images were obtained by first pixelating the images of Data Set 2 (8×8 pixel blocks receiving the average color) and then scrambling (spatial relocation). This leaves the global chromatic distribution intact but destroys image semantics.

E. Simulation of Illuminant Changes

The usual way to simulate object colors under different illuminants is to calculate XYZ tristimulus values according to

$$X = k\int_{\lambda} E(\lambda)\rho(\lambda)\bar{x}(\lambda)\mathrm{d}\lambda,$$
$$Y = k\int_{\lambda} E(\lambda)\rho(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda,$$
$$Z = k\int_{\lambda} E(\lambda)\rho(\lambda)\bar{z}(\lambda)\mathrm{d}\lambda, \quad (1)$$

in which $E(\lambda)$ represents the SPD of the illuminant, $\rho(\lambda)$ is the spectral reflectance function of the object and $\bar{x}$, $\bar{y}$, $\bar{z}$ represent the 1931 color matching functions for the 2° standard observer. The factor k is defined as

$$k = \frac{100}{\int_{\lambda} E(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda}, \quad (2)$$

and serves to normalize Y at 100 for a perfect white reflectance. However, our object colors are defined in terms of CIELAB values rather than spectral reflectance functions. This poses a problem since an infinite number of reflectance functions can result in identical XYZ tristimulus values (and hence identical CIELAB values) under one illuminant. Therefore, a selection criterion for picking one reflectance function is needed. We apply Van Trigt's method [20] to estimate the smoothest reflectance function from a set of tristimulus values, where the smoothness measure is defined as the square of the derivative of the reflectance function with respect to

Table 3. Parameters of the Chromatic Distributions of the 25 Images Shown in Fig. 6 (Data Set 3), under D65 Reference Illumination

		Mean			Standard Deviation			Skewness			Kurtosis		
Image	Φ	L*	a*	b*	L*	a*	b*	L*	a*	b*	L*	a*	b*
1	—	52.92	5.51	17.97	12.47	5.18	5.57	−0.85	−0.16	−0.40	3.39	2.40	5.30
6	—	50.17	−6.15	28.36	22.00	13.75	14.36	0.19	0.63	0.56	1.95	2.99	3.45
11	—	35.67	12.61	20.23	14.78	14.28	13.79	0.38	0.71	0.46	2.84	3.10	4.10
16	—	34.75	0.58	9.42	23.23	9.73	9.58	0.98	1.54	0.65	3.18	6.76	4.67
21	—	63.07	1.11	11.54	21.12	3.74	5.02	−0.46	−1.83	1.33	2.54	13.42	19.15
2	0°	47.68	−10.06	26.25	14.88	12.06	5.90	−0.61	0.77	−0.38	3.03	2.52	3.78
7	0°	42.82	4.01	23.97	16.00	12.33	8.30	0.41	0.51	0.28	2.86	3.47	2.23
12	0°	39.54	4.45	22.01	16.65	19.18	8.88	−0.01	0.76	0.93	1.88	2.56	3.31
17	0°	41.94	−14.22	29.48	9.88	9.04	4.92	1.44	2.46	−2.33	8.27	9.36	13.11
22	0°	30.20	3.32	14.22	17.35	12.93	6.77	0.64	0.75	0.13	3.73	2.73	2.74
3	45°	34.32	7.34	4.15	23.31	9.50	11.91	0.53	0.72	−0.02	2.61	3.49	3.67
8	45°	25.01	0.01	2.79	14.67	5.11	6.67	0.75	3.82	0.50	2.72	22.80	3.75
13	45°	21.06	−0.79	0.53	19.40	5.66	6.48	1.79	−1.09	−0.95	6.32	5.19	3.67
18	45°	47.74	8.68	19.78	10.84	4.79	5.11	0.45	−0.98	−0.47	4.05	4.70	4.20
23	45°	33.03	10.94	21.56	10.53	9.85	6.48	1.15	−0.36	−1.48	9.58	2.34	5.78
4	90°	32.84	−7.92	6.64	22.08	4.87	13.96	0.40	−0.41	−0.28	2.36	2.77	3.08
9	90°	33.43	−7.43	12.70	25.52	3.99	12.82	0.34	0.15	0.03	1.67	3.55	2.27
14	90°	40.31	−6.77	29.09	15.68	5.95	13.74	0.13	−0.11	0.10	3.13	2.47	3.56
19	90°	48.34	−17.53	27.19	19.98	6.24	12.04	−0.05	0.06	−0.04	2.74	3.17	2.56
24	90°	43.46	0.81	−10.63	12.16	1.15	3.24	0.51	0.36	−0.71	2.15	2.59	4.03
5	135°	47.89	−4.05	17.80	21.85	10.94	7.31	0.46	−0.09	1.05	2.53	1.72	3.86
10	135°	44.90	−5.81	12.63	23.63	9.29	9.78	0.38	−0.16	0.49	1.97	1.59	2.29
15	135°	47.05	−12.60	25.59	16.89	9.24	10.35	0.51	−0.28	0.19	3.99	1.80	1.96
20	135°	47.85	−3.58	15.90	25.17	7.73	10.04	−0.07	−1.56	1.70	1.94	4.34	5.20
25	135°	44.24	−8.61	22.34	15.49	9.22	10.86	0.48	−0.11	0.48	3.83	1.50	1.96

wavelength, integrated over the visual range. We thus assume that the CIELAB values of our object colors in the reference condition result from illumination of the estimated reflectance functions by D65, our reference illuminant. Conversion of the resulting XYZ values to RGB drive values for displaying the colors on the color monitor is done using the sRGB profile (our monitor is calibrated to sRGB). Since the simulated illuminant changes may shift object colors out of the monitor's color gamut, we used Data Set 2 (the natural images) to calculate the percentage of pixels that are out of gamut under the chromatic illuminants. Averaged over the 25 images, these percentages are 3.2, 4.0, 1.0, and 2.0 for the red, green, yellow, and blue illuminants. For these pixels, the average color difference between the target color (what it should be) and actual color (what was displayed on the monitor) is $\Delta E_{00} = 0.092$, 0.040, 0.003, and 0.064, respectively. These color differences (reproduction errors) are very small and below visual threshold so that we may safely assume that the restrictions of the color gamut have no effect on the outcome of the experiments presented here.

F. Illuminants that Induce Equal Changes in the Reflected Light Signal Distribution

In line with previous studies [21,22] we select a neutral reference illuminant (D65) and four chromatic illuminants, all composed with the CIE basis functions for spectral variations in natural daylight. In [21] it is investigated whether color constancy would be better for illuminant changes along the daylight locus than for illuminant changes perpendicular to it, but no experimental evidence was found that unambiguously supports this hypothesis. Here, we use the same paradigm with the yellow and blue illuminants along the daylight locus, and another two (red and green) perpendicular to it, but without the constraint that they are perceptually equidistant from the neutral reference point. Perceptual equidistance of object colors rendered under these illuminants would only be guaranteed for spectrally nonselective (achromatic) samples seen in isolation, but it does not necessarily imply perceptual equidistance for individual chromatic samples or a distribution of chromatic samples. Therefore, we modify the purity of the chromatic illuminants such that when changing the illumination from neutral to one of the four chromatic illuminants, equal distributions of physical shifts in the reflected light signal are obtained, as explained below. In Fig. 7 the positions of the illuminants are plotted in the CIE 1931 x, y chromaticity diagram and the CIE 1976 a*b* space together with the daylight locus, and Fig. 8 shows the SPDs of the illuminants.

We define the reflected light signal L as the wavelength-by-wavelength product of the illuminant SPD E and the reflectance function ρ:

$$L = \int_{\lambda} E(\lambda)\rho(\lambda)\mathrm{d}\lambda. \tag{3}$$

This signal does not include the sensitivity of the visual system and is therefore regarded as a pure physical signal. When changing the illuminant from E_1 (neutral) to E_2 (chromatic), the associated change in the reflected light signal is given by

$$\Delta L = L_2 - L_1 = \int_{\lambda} [E_2(\lambda) - E_1(\lambda)]\rho(\lambda)\mathrm{d}\lambda. \tag{4}$$

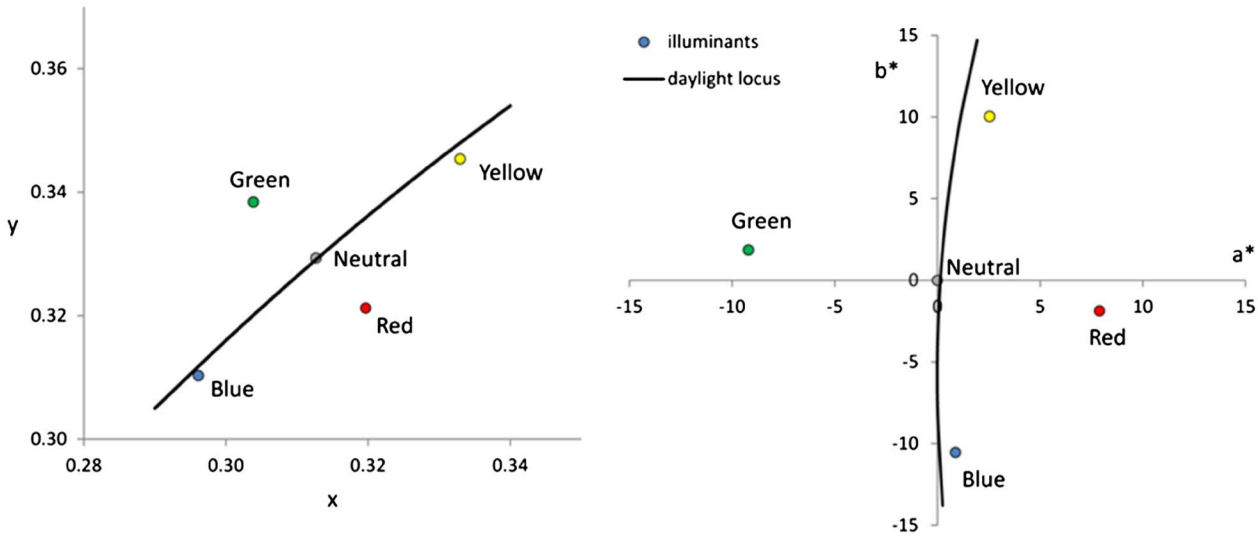

Fig. 7. Daylight locus and positions of the neutral illuminant and the four chromatic illuminants in CIE 1931 x, y chromaticity space (left) and CIE 1976 a*b* space (right).

We tune the purity of the four chromatic illuminants to arrive at almost identical cumulative distributions of $(\Delta L)^2$ by adjusting the mixture of the SPD of the original four chromatic illuminants with the SPD of the neutral illuminant D65. This process is described by

$$E'(\lambda) = \frac{E(\lambda) + xE_{\mathrm{D65}}(\lambda)}{1+x}, \tag{5}$$

where $E'(\lambda)$ represents the spectral power of the adjusted illuminant at wavelength λ, $E(\lambda)$ is the spectral power of the original illuminant, $E_{\mathrm{D65}}(\lambda)$ is the spectral power of illuminant D65, and x is the mixing factor that regulates the mixing of $E(\lambda)$ and $E_{\mathrm{D65}}(\lambda)$. We derived mixing factors of 3.55, 2.54, 2.0, and 2.71 for the red, green, yellow, and blue illuminants, respectively. The similarity in the cumulative distributions of the reflected light signal for the chromatic illuminants, obtained for the first chromatic distribution defined in Table 1, is shown in Fig. 9. For the other chromatic distributions of Data Set 1, similar cumulative distributions are obtained.

To summarize, we have defined four chromatic illuminants that induce equal physical shifts for the test scenes of Data Set 1. The advantage of using these illuminants is that they enable us to separate perceptual effects from physical effects when changing the illuminant from neutral to one of the chromatic illuminants. For a vision system without any spectral preference, i.e., a flat spectral sensitivity profile, these illuminant changes would be indistinguishable.

G. Color Calibration of the Monitor

The images are presented on a calibrated LCD monitor (Eizo, ColorEdge CG211) operating at 1600 pixels × 1200 pixels (0.27 mm dot pitch) and with a 24-bit color resolution. Using a spectrophotometer (GretagMacbeth, Eye-one) the monitor is calibrated to a D65 white point of 80 cd/m^2, with a gamma of 2.2 for each of the three color primaries. CIE 1931 x, y chromaticity coordinates of the primaries were $(x, y) = (0.638, 0.322)$ for red, (0.299, 0.611) for green and (0.145,

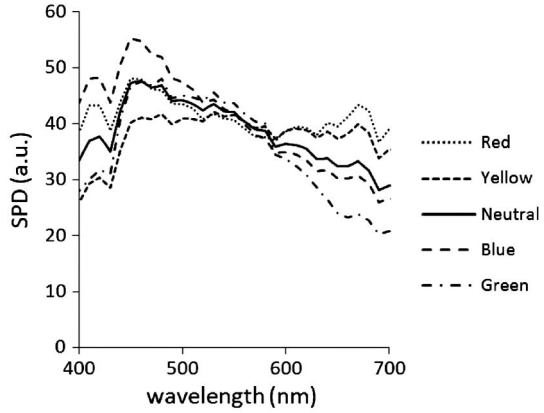

Fig. 8. Relative SPD (in arbitrary units, a.u.) of the neutral reference illuminant and the four chromatic illuminants used in the experiments. The illuminants were created with the CIE basis functions for spectral variations in natural daylight, and were modified in purity such that they elicit equal distributions of changes in the reflected light signal (see Fig. 7 also and text for explanation).

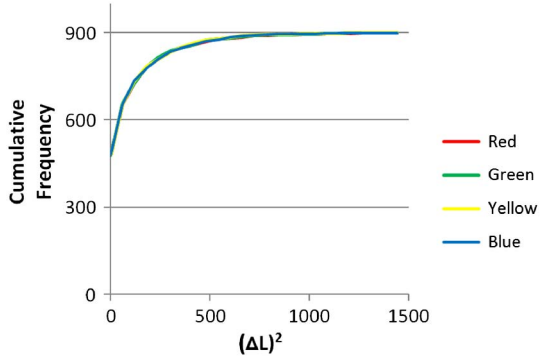

Fig. 9. Cumulative distributions of the squared changes in the reflected light signal (ΔL), due to a change from neutral illumination to one of the four chromatic illuminants (indicated in the legend). The change in the reflected light signal is defined in Eq. (4). The figure shows that the distributions are very similar, which was achieved by adjusting the purity and distance of the chromatic illuminants to the neutral point, as described in the text.

0.058) for blue, respectively, closely approximating those of the sRGB standard monitor profile (IEC, 1999). Spatial uniformity of the display, measured relative to the center of the monitor, was $\Delta E^*_{ab} < 1.5$, according to the manufacturer's calibration certificates. This type of display was previously shown to provide an average color reproduction error of about $\Delta E_{00} = 0.6$, in the order of one just noticeable difference [23], accurate enough for the type of experiment described in this paper.

H. Subjects

The subjects that participated in the experiments all have normal color vision and normal or corrected-to-normal visual acuity. Screening on color vision deficiencies was done by testing on the HRR (Hardy Rand and Rittler) pseudoisochromatic plates (4th edition), which were shown under the prescribed illumination. Eight subjects (including the first three authors) participated in the experiment with images from Data Set 1, and six subjects for Data Sets 2 and 3. All subjects were male, ranging in age from 27 to 45 years (average 32). They worked in our laboratory, and did not receive a separate financial reward for their participation.

I. Instructions and Procedure

In each trial four images are shown, as illustrated in Fig. 2. The subjects had to indicate which vertical image pair (left or right) had the highest color fidelity. The subjects were told that in each trial the two upper images were identical references. They had to visually compare the colors of the lower images with the reference on top by making vertical eye movements. Both global and local aspects of color comparison had to be taken into account. They did this for one vertical image pair (test-illuminant) and subsequently for the other image pair, and they were instructed not to compare the two test images (test illuminants) horizontally. So, they switched from making vertical comparisons on the left side of the screen to making vertical comparisons on the right side of the screen, until they came to a decision. In our implementation, the observers pressed "Q" on the computer keyboard to indicate their choice for the image pair on the left, "P" for the image pair on the right, and spacebar when they could not choose left or right. In the latter case it meant that the observer judged the color fidelity of the left image pair to be as good or as bad as the right image pair. The subjects were encouraged to select an image pair, and only indicate "no selection" when they really could not decide. In between trials, the neutral background was shown for 5 s. Apart from the authors, the subjects did not know that the lower test images were obtained by simulated illuminant changes.

3. RESULTS

We first provide a more qualitative account of the results. In the section on data modeling a quantitative analysis of the data in terms of the predicted illuminant probability is presented.

A. Visual Scores

We recall that in each experimental trial our observers selected one of two competing illuminant renderings. The illuminant associated with the rendering that was indicated as having higher color fidelity received 1 point, the other received no points in that trial. In cases where the observer could not choose one or the other, meaning that the two illuminant renderings were considered equally good or bad, both test illuminants received 0.5 points. Given that each test illuminant appeared three times in competition against the other illuminants, the maximum visual score for an illuminant to obtain was 3. For Data Set 1, trials were replicated, leading to a maximum visual score of 6 per illuminant. On average, in 73% of the trials the repetition resulted in the same response. For Data Sets 2 and 3 we used 5 images per chromatic distribution, leading to a maximum visual score of 15 per illuminant.

The results obtained on the three image data sets are presented in Fig. 9. Shown are the visual scores per illuminant, labeled in the corresponding illuminant color, for the five chromatic distributions. Visual scores are averaged across observers, and error bars represent the standard error of the mean (SEM). As a rule of thumb, nonoverlapping standard errors indicate a statistically significant difference. Plots in the left column show the visual scores averaged across observers, and the plots in the right column show these scores normalized to the scores of the first chromatic distribution (the Gaussian distribution without preference of the chromatic orientation). As a result, all scores equal 100% for the first chromatic distribution.

B. Data Set 1: Multicolor Scenes

Results for the first data set (the synthesized chromatic distributions) are shown in the top row of Fig. 10. Two aspects are salient: the unequal scores for the first chromatic distribution, and the change in the scores for the distributions with varying chromatic orientation (chromatic distributions 2–4). The data for the first chromatic distribution clearly shows a large difference between the visual score for the red illuminant and the other illuminants. Apparently, our multicolor test scene under the red illuminant results in a higher color fidelity than the other illuminants. Second best is the blue illuminant, and green and yellow have the lowest fidelity scores. Why do the red and the blue illuminants lead to higher color fidelity than yellow and green? Recall that we deliberately adjusted the purity of the chromatic illuminants to get identical distributions of the changes in the reflected light signal, when the illumination changed from neutral to chromatic. The mixing factors for red (3.55) and blue (2.54) mentioned in the previous section are larger than for yellow (2.0) and green (2.71), resulting in lower purity of the first two illuminants. When comparing, for example, the scene rendered under red and green illumination to the scene rendered under the reference illumination, as illustrated in Fig. 2, one may notice the slightly lower purity of the red illuminant, resulting in higher color fidelity. So, although the chromatic illuminants induce similar distributions of physical light changes, they lead to perceptually different estimates of color fidelity. As will be shown in the section on data modeling, these different estimates are only partly explained by perceptual color difference metrics.

The results for the other four chromatic distributions are best explained using the right hand side of Fig. 10, showing the visual scores normalized to those of the first distribution. The fact that the visual scores change compared to those of the first distribution indicates that color fidelity depends on the shape of the chromatic distribution. In addition, the visual scores change with the chromatic orientation. For

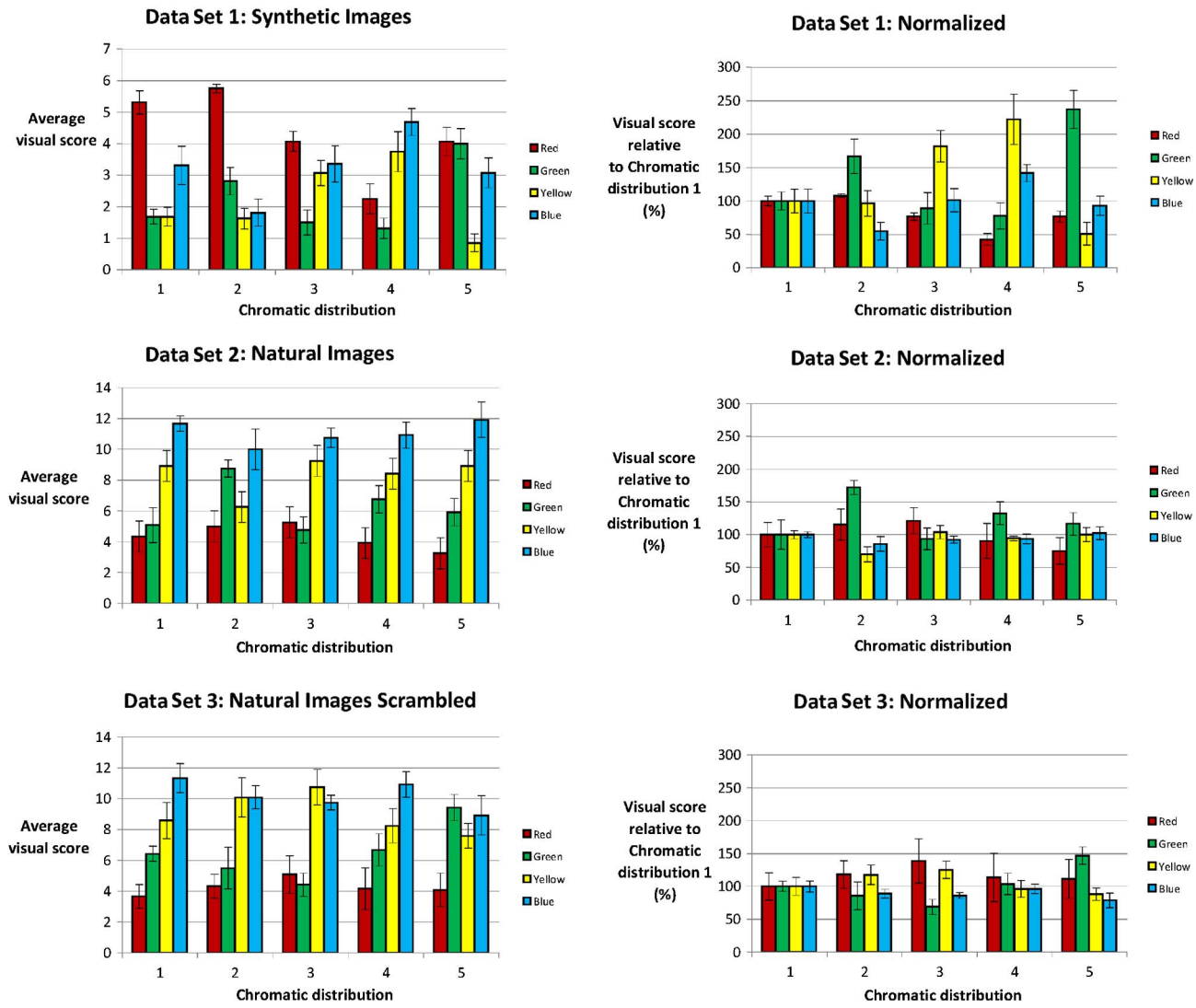

Fig. 10. Visual scores per chromatic illuminant (red, green, yellow, blue), averaged across subjects, obtained for the five different chromatic distributions. The plots in the left column show the visual scores for the three image data sets, and the plots in the right column show the same data normalized to the visual scores for the first chromatic distribution. The higher the visual score, the more often the illuminant was (indirectly) judged as having higher color fidelity. Error bars denote ±1 SEM.

the second chromatic distribution, having its major axis along the a* axis of CIELAB color space (roughly the red–green axis), the color fidelity for the red and green illuminants increase, and decrease for yellow and blue. In opposition, for chromatic distribution number 4, having its major axis along the b* axis of CIELAB color space (roughly the yellow–blue axis), the color fidelity for the yellow and blue illuminants increases while it decreases for red and green. For chromatic distributions 3 and 5, having their major axes in between the a* and b* axes, color fidelity for red and blue takes on values in between those for distributions 2 and 4. However, the relative scores for yellow and green illumination do not follow this systematic pattern. These findings are summarized as follows:

1. The shape of the chromatic distribution affects the color fidelity.
2. The orientation of the chromatic distribution affects the color fidelity.
3. When the vector of the illuminant change (pointing from neutral to chromatic) is parallel to the orientation of the chromatic distribution, color fidelity of the rendered scene is judged better than when the direction of the illuminant change is orthogonal to it.

C. Data Set 2: Natural Images

Compared to the visual scores of Data Set 1, the data for the natural images (middle row in Fig. 10) are different in three ways. First, the average visual scores for the first chromatic distribution show a much lower value for the red illuminant, and a higher value for the yellow illuminant. Second, the change in the pattern of these values for the other chromatic distributions is much smaller. Third, illuminant blue has the highest visual scores for all chromatic distributions, followed by the yellow (with the exception of the green illuminant in the second chromatic distribution). It is remarkable that the yellow and blue illuminants lead to a higher color fidelity structure for the natural images. This might indicate a preference for illuminant changes along the daylight locus. We comment on this topic in more detail in the Discussion section.

D. Data Set 3: Scrambled Natural Images

The data for the scrambled natural images (bottom row in Fig. 10) are quite similar to those of the second data set, except for the green illuminant in chromatic distributions 2 and 5. Here, also, the yellow and blue illuminants are dominating the visual score. Taking out image semantics by scrambling the natural images does not have a strong effect on the color fidelity of the renderings of the visual scenes.

E. Modeling the Illuminant Probability

In Fig. 10 the visual scores were shown as having been obtained by summing the number of times that each illuminant rendering was selected by the observers as having the best color fidelity. Here, we model the data in terms of illuminant probability, i.e., the observed probability that an illuminant is selected in a given combination of test illuminants. To illustrate, Table 4 shows these observed probabilities for Data Set 1 across observers. According to the choice modeling theory, we model these probabilities using binary logistic regression, in which the predicted illuminant probability P is given by

$$P = \frac{e^{\eta}}{1 + e^{\eta}}, \tag{6}$$

$$\eta = c(\mathrm{var}_i - \mathrm{var}_j), \tag{7}$$

Table 4. Observed Probability of Test Illuminants R, G, Y, B as Being Selected by the Observers, for Each of the Five Chromatic Distributions and Six Illuminant Combinations of Data Set 1

Chromatic Distribution	Illuminant Combination	P [R]	P [G]	P [Y]	P [B]
1	R–G	0.9375	0.0625	0	0
	R–Y	0.875	0	0.125	0
	R–B	0.625	0	0	0.375
	G–Y	0	0.375	0.625	0
	G–B	0	0.625	0	0.375
	Y–B	0	0	0.1875	0.8125
2	R–G	0.9375	0.0625	0	0
	R–Y	0.9375	0	0.0625	0
	R–B	1	0	0	0
	G–Y	0	0.6875	0.3125	0
	G–B	0	0.6875	0	0.3125
	Y–B	0	0	0.625	0.375
3	R–G	0.8125	0.1875	0	0
	R–Y	0.6875	0	0.3125	0
	R–B	0.5625	0	0	0.4375
	G–Y	0	0.25	0.75	0
	G–B	0	0.375	0	0.625
	Y–B	0	0	0.3125	0.6875
4	R–G	0.625	0.375	0	0
	R–Y	0.3125	0	0.6875	0
	R–B	0.1875	0	0	0.8125
	G–Y	0	0.3125	0.6875	0
	G–B	0	0.125	0	0.875
	Y–B	0	0	0.25	0.75
5	R–G	0.375	0.625	0	0
	R–Y	0.9375	0	0.0625	0
	R–B	0.875	0	0	0.125
	G–Y	0	0.75	0.25	0
	G–B	0	0.5	0	0.5
	Y–B	0	0	0.375	0.625

where c is a coefficient to be estimated and var is an independent variable characterizing the experimental trial according to some sort of "model". Subscripts i and j on var label the two test illuminants in the trial. Actually, we use four different types of models to feed var in Eq. (7), as discussed below.

Our first model predicts based on the reflected light signal that we introduced in this paper. The second model predicts based on the calculated overlap in the rendered color gamut for the test and reference illumination. The third model predicts based on the scene averaged color difference between corresponding image parts (pixels) under the test and reference illumination. The fourth model predicts based on two image quality metrics. These four models differ in the extent to which they incorporate information processing by the human visual system.

We first describe each of the models before presenting the results.

1. Model 1: Reflected Light Signal

The reflected light signal L is described by Eq. (3) and captures the product of the SPD of the illuminant and the spectral reflectance of an object. The change in L, described by ΔL in Eq. (4), is a purely physical measure that we use to describe the change in the light reflected from an object when the SPD of the illuminant is changed. For a scene in which each image pixel is considered as an individual object having an individual spectral reflectance, a cumulative distribution of $(\Delta L)^2$ can be calculated as shown in Fig. 9. In general, the four distribution curves (referring to the four illuminants) will not be identical, as illustrated in Fig. 11 for scene 1 from Data Set 2.

These distributions can be fitted by a model using two parameters, c and d,

$$\mathrm{cdf} = 1 - ce^{(-d\,\mathrm{bin})}, \tag{8}$$

in which cdf is the cumulative distribution function and bin is the bin number in Fig. 11 encoding the square of the change in the reflected light signal, $(\Delta L)^2$. The ratio d/c is indicative of the speed with which the cumulative distribution function saturates, with the highest d/c value belonging to the upper curve. Higher values of d/c indicate higher frequencies in

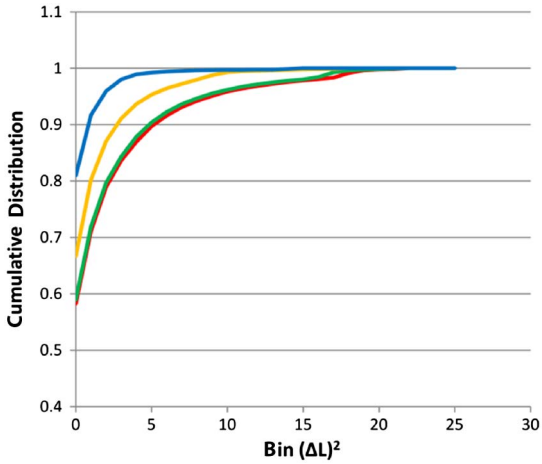

Fig. 11. Cumulative distribution of $(\Delta L)^2$ for image 1 of Data Set 2 (see Fig. 4 also). The line colors code the illuminant color (blue, yellow, green, red from top to bottom).

the lower bins, which belong to smaller differences. Illuminants that are characterized with higher d/c values are thus expected to have a higher probability of being selected in our experimental trials.

2. Model 2: Chromatic Overlap

The Chromatic Overlap model calculates the overlap in a test scene's chromatic distribution (in the a*, b* chromaticity plane of CIELAB color space) when rendered under the test and reference illumination. The idea is that, when a change in illumination is small, the associated change in rendered color gamut is also small. Likewise, a large change in illumination will result in a larger change in the color gamut. Hence, the overlap between the two gamuts may act as a predictor for the color fidelity. The assumption is that the larger the overlap, the higher the color fidelity of the scene under the test illuminant. We compute the chromatic overlap as the normalized histogram intersection, a measure for the similarity of two histograms [24]. When there is no overlap between two histograms, the normalized histogram index NHI = 0. When the two histograms are identical, NHI = 1. The index is calculated as follows:

$$\mathrm{NHI} = \frac{\sum_{j=1}^{n} \min(R_j, T_j)}{\sum_{j=1}^{n} T_j}, \tag{9}$$

in which R and T represent the histograms of n bins each of the scene under the reference and test illumination, respectively. The numerator in Eq. (9) is the intersection between R and T, and the denominator is the number of pixels in the image (scene). In our case the histogram bins represent a quantization of the a*, b* chromaticity plane in CIELAB color space, which has been shown to be the preferred color space for evaluating image similarities using the histogram intersection (Lee *et al.* 2005). For each a*, b* cell (having $\Delta a^* = 1$ and $\Delta b^* = 1$) within the a*, b*-plane we count the number of colors present in that cell and construct a histogram of these counts. We compute the normalized histogram intersections with Eq. (9) for the image pair formed by the scene under the reference illumination and the test illuminant.

3. Model 3: Scene Averaged Color Difference

The scene averaged color difference, ΔE_{avg}, is calculated as the average of N (indicating the number of pixels) color differences, each color difference ΔE_j obtained from two corresponding scene pixels under reference and test illumination:

$$\Delta E_{\mathrm{avg}} = \frac{1}{N} \sum_{j=1}^{N} \Delta E_j. \tag{10}$$

Three color difference models are tested: CIELAB [13], CIE94 [14], and CIEDE2000 [15]. The first, CIELAB, is based on the Euclidian distance between two color points in L*a*b* space. The second and third (CIE94 and CIEDE2000) use differences in L*, C*, and h (C* stands for chroma and h for hue angle) which are weighted by scaling factors that depend on the location in CIELAB space and by parametric scaling factors that account for deviations from the preferred (reference) viewing conditions. The equations required for the computation of ΔE_j are presented in the given references and in many textbooks on color. An online overview of these equations is given in [25]. For the CIE94 and CIEDE2000 models the standard values for the parametric scaling factors $k_L = k_C = k_H = 1$ were used.

4. Model 4: Color Image Quality

Many image difference metrics have been proposed in the last decade. Recently, in [26] 29 metrics were evaluated using 6 image databases. It was shown that metric performance is still dependent on the image data set, and therefore selecting one metric is not straightforward. We selected two of them, described below as models 4a and 4b, both incorporating spatial processing of the test scene.

5. Model 4a: Spatial CIELAB (S-CIELAB)

The first is the well-known S-CIELAB metric [16], which features a spatial extension of the standardized ΔE_{ab} color difference metric. Prior to calculating color differences between corresponding parts in two images, spatial filtering on the (achromatic) luminance channel and the two chrominance channels is performed, with filter specifications that mimic the spatial sensitivities of the human visual system. For large uniform patches this spatial preprocessing has no effect and S-CIELAB results in identical calculations as the CIELAB color difference ΔE_{ab}. We customized the publicly available S-CIELAB MATLAB routines with the spatial and spectral characterizations of our color monitor. On output, S-CIELAB has a computed image difference map from which we calculate the average ΔE_{ab} value.

6. Model 4b: Toet–Lucassen Color Image Fidelity

The second metric we selected is the color image fidelity metric developed by Toet and Lucassen [17]. It is the color analog of a gray scale image fidelity metric, introduced by Wang and Bovik [27], capable of quantifying a wide range of local distortions in image pairs. The gray scale metric of [27] is indicated by symbol Q and is calculated as

$$Q = \left(\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right)\left(\frac{2\bar{x}\,\bar{y}}{\bar{x}^2 + \bar{y}^2}\right)\left(\frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right), \tag{11}$$

in which the three terms between parentheses represent loss of correlation, luminance distortion, and contrast distortion, respectively. The symbols $\bar{x}$ and $\bar{y}$ represent averages in image x and y, σ_x and σ_y represent the variance in image x and y, and σ_{xy} is the covariance. The values of Q lie between −1 and 1. The optimum value of 1 is obtained when the original image and test image are identical, and the worst value of −1 is obtained when the test image is twice the mean of the original image subtracted by the original image. The Q value for an image is obtained from averaging the Q-values for a window of 8 pixels × 8 pixels, sliding over the image. The color fidelity metric by Toet and Lucassen [17] first computes three such Q values, one for each of the separate L, α, and β channels of a decorrelated color space [28]. This leads to the three quality metrics Q_L, Q_α, and Q_β. Then, to obtain a single overall quality measure, the three metrics are combined into a weighted vector mean,

$$Q_{\mathrm{color}} = \sqrt{w_L Q_L^2 + w_\alpha Q_\alpha^2 + w_\beta Q_\beta^2}, \tag{12}$$

where w_L, w_α and w_β are the weighing coefficients for the three channels. The fidelity metric Q_{color} is computed for

Table 5. Overview of Models and Output Variables Used for Predicting the Illuminant Probability with Eqs. (6) and (7)

Model	Var in Eq. (7)
1. Reflected light signal	d/c
2. Gamut overlap	NHI
3. Scene averaged delta E	dE_{ab}, dE_{94}, dE_{00}
4a. S-CIELAB	Average dE of image difference map
4b. TL-image fidelity	Q_{color}

the image pair formed by the test illuminant and the reference illuminant.

To summarize, the models and their output used as the variable var in Eq. (7) are listed in Table 5.

F. Comparison of Model Performance

Using the models described above, we estimated the coefficients of Eq. (7) for each of the three image data sets separately. Shown in Table 6 are the results in terms of % OK, which is the percentage of experimental trials for which the predicted probability with Eq. (6) is on the same side of 0.5 as the observer data, i.e., higher or lower than 0.5.

For the TL-Image Fidelity model, the parameters w_L, w_α, and w_β in Eq. (12) were optimized to obtain the maximum %OK. The optimum values are $w_L = 1$, $w_\alpha = 0.01$, and $w_\beta = 2$ for all three data sets, indicating that for our experiments the importance of the quality metric is amplified in the β-channel and suppressed in the α-channel.

From a comparison of the values in Table 6 it can be seen that, for Data Set 1, the Gamut Overlap model leads to the highest percentage of correct trial predictions (82%), followed by the TL-Image Fidelity model (73%) and S-CIELAB (68%). The Reflected Light Signal model is clearly the worst (55%), and the scene averaged ΔE models (60%–65%) are in between. This is different for Data Sets 2 and 3, where the Reflected Light Signal model and the Gamut Overlap model have comparable performance levels and outperform S-CIELAB. Also, the scene averaged ΔE models have higher performance levels, with ΔE_{00} being the best model for these two data sets. Averaged over the three data sets, the ΔE_{00} model performs best (%OK = 77), followed by the ΔE_{94} and the TL-Image Fidelity model (%OK = 73). Surprising, perhaps, is the lowest performance of the S-CIELAB model, which may be related to the fact that the images being compared (under reference and test illumination) are different in chromatic content only and are not spatially disturbed.

Finally, we note that the performance increase of the three color difference models (in the order from ΔE_{ab} to ΔE_{94} to ΔE_{00}) is in line with the historical progress in the development of the color difference formulas.

4. DISCUSSION

We have shown that when multicolor scenes are rendered under different (simulated) illuminants, the perceived color fidelity of the rendered test scenes depends on the shape and orientation of the chromatic distribution. Distributions having the same mean chromaticity but different chromatic orientations are judged differently in terms of their color fidelity. How does this finding relate to other studies? It is clear that a model based on the gray world assumption (e.g., [1]) would fail to predict the effect of chromatic orientation, since it only relies on the mean chromaticity in the scene. There are a number of studies, however, in which the chromatic distribution of the visual scene is probed to derive information about the chromaticity of the illuminant (e.g., [2,29–32]). But even if we suppose that this would result in the correct estimation of the exact illuminant color, it would still not predict that one illuminant leads to a higher color fidelity than another. From our attempts to model the data of this study, it is not completely clear, either, what mechanism is responsible for the effects reported here. The best performing three models differ completely in the way they are related to information processing in the visual system. The Chromatic Overlap model is a computational measure that has nothing to do with the visual system. The ΔE_{00} is an advanced model for the perception of color differences, but spatially averaged in a straightforward manner, and the TL-image Fidelity model is more specialized, incorporating a decorrelated color space and spatial sampling. The fact that these three different models give such comparable performance levels leaves us with the question of what mechanism is involved in the image quality comparison. Perhaps more data from more test images and/or illuminants is needed here to make that distinction. The fact is that all models leave a considerable amount of data variance unexplained. We should not forget, however, that the models try to describe the average observer response, which itself has some uncertainty. To quantify this, we calculated for each observer the correlation with the group average, while first removing the subject in question from the group. For the eight observers we obtained correlations of 0.82, 0.36, 0.85, 0.71, 0.79, 0.77, 0.87, and 0.46, with an average of 0.70. This serves to illustrate that we may expect an upper limit to the correlation coefficient that is set by the interobserver variability.

A. Preference for Natural Illuminant Changes?

The idea that the human visual system may be better equipped for compensating for illuminant changes along the locus of natural daylight variations has been studied several times, but so far without convincing evidence in support of that idea. Lucassen and Walraven [33], using successive haploscopic color matching on a computer display, studied color constancy for six colored illuminants equidistant from the neutral

Table 6. Model Performances (in %OK) for the Three Data Sets Separately and as Weighted Average

			Scene Averaged ΔE				
Data Set	RLS (Reflected Light Signal)	CO (Chromatic Overlap)	ΔE_{ab}	ΔE_{94}	ΔE_{00}	S-CIELAB	TL-Image Fidelity
1	55.1	81.7	60.0	65.0	63.3	68.3	73.3
2	65.3	68.0	68.7	73.3	78.7	63.3	74.7
3	74.7	72.7	76.0	78.0	80.7	63.3	71.3
avg	67.5	72.2	70.3	73.9	77.0	64.1	73.1

point, including a red, green, yellow and blue illuminant. From a reanalysis of the data from their Experiment 1, it follows that constancy indices were found (averaged across samples), descending in the order of blue (0.74), green (0.66), red (0.61), and yellow (0.58) illumination. In Lucassen and Walraven (1996), again higher constancy indices for blue (4000 K) illuminant changes were found (average constancy index = 0.74) than for yellow (25,000 K) illuminant changes (average index = 0.64). Using more natural viewing conditions and objects illuminated by real light sources, Brainard [34] investigated achromatic appearance settings of a test patch under varying illumination, and reported no dependency on the chromaticity of the illuminant change. Delahunt and Brainard [21], using yellow and blue illuminants along the daylight locus and red and green perpendicular to it, reported color constancy indices descending in the order of blue (0.82), green (0.76), yellow (0.69), and red (0.67) illumination. Brainard *et al.* [35] investigated a Bayesian model of human color constancy, capable of testing whether a prior daylight for the illumination would optimally predict experimental data. Better results, however, were obtained for a prior broad illumination. Hansen *et al.* [36] used color naming of color patches under simulated illuminant changes. They obtained color constancy indices that were similar across illumination conditions, so without any clear preference for a certain illumination change. In studies with (rapid) temporal changes in which observers have to discriminate between a change in illumination and a change in surface material, it is found that color constancy is best for blue and green, and less for yellow and red [37], although this order may vary for individual observers.

The pattern that seems to emerge from the pooled results of these experimental methods is that highest constancy indices are found for illuminants blue and green, and lower indices for yellow and red. But how do the results of our study relate to this? From Fig. 10, it is easily seen that for Data Sets 2 and 3, the visual scores for illuminants yellow and blue are higher than for red and green. For Data Set 1, however, this is not the case. To quantify, the pooled visual scores for red and green, and for yellow and blue, expressed as a percentage of the sum of the visual scores are 55% (R + G) and 45% (Y + B) for Data Set 1, and 36% (R + G) and 64% (Y + B) for both Data Sets 2 and 3. Apparently, these figures are strongly dependent on the data set. Also, from a small experiment using the scenes in Fig. 5 of [21], presented as images in our experimental paradigm, the results for 3 observers indicate a strong preference for yellow and blue (83% of the visual score) over red and green (17%). These findings shed some new light on the issue and seem worthwhile to further investigate. Again, it shows that scenes rendered under perceptual equidistant illuminants may result in completely different judgments about color fidelity. Another interesting detail to mention here is that the illuminants used in this study have different values for the light source quality measure known as the color rendering index (CRI) [38], being the maximum value of 100 for D65, 89 for red and green, 99 for yellow, and 96 for blue. This would predict better color fidelity for yellow and blue as compared to red and green, as generally measured for our Data Sets 2 and 3, but cannot explain the strong visual scores for the red illuminant in the chromatic distributions in Data Set 1. It should be noted, however, that the CRI is determined from a very limited number of test samples (8 to be precise). In contrast, in the ΔE models all pixels are considered as test samples that constitute a much stronger quality indicator for the color rendering properties.

B. Improvements in Methodology

Color fidelity was measured using the triad illuminant comparison method. In this study we employed 4 colored test illuminants from which 6 unique pairs can be created. Increasing the number of illuminants is an obvious way to further improve on the method, but at the cost of measurement time [n illuminants lead to $n!/(2!(n-2)!)$ unique pairs]. We have previously used the triad comparison method in studies for perceptual evaluation of color constancy algorithms [22,39]. In those studies, the observers had the same task as in our study, namely, to indicate which of the two images in the bottom row best resembled the image on the top row. However, those images were color corrected images from different color constancy algorithms, instead of renderings under different illuminants.

Another way for possible improvement is the presentation mode. Foster *et al.* [40] showed that with successive presentation the degree of color constancy obtained with surface color matching is higher and leads to less variance between observers. Whether this would also apply to the color fidelity judgments remains to be investigated.

The images containing the chromatic distributions in Data Set 1 were generated with software that contains a random component used for sampling the Gaussian distribution and for assigning the spatial location to a color patch. It is thus possible that one instance of a chromatic distribution is visually very different from another instance of that same distribution. We therefore ran a control experiment in which we checked ten instances of the same chromatic distribution, but did not find systematic differences in the visual scores.

C. Improvements to Color Constancy Algorithms

We started this paper by noting that performance measures of human color constancy and color constancy algorithms are quite different. Here, we discuss how the current findings of this paper may help to narrow that gap. Most color constancy algorithms in computer vision are directed at estimating the unknown illuminant from the scene. After estimating the illuminant, which is assumed to be spatially uniform across the visual scene, the images are color corrected to compensate for the effect of the color of the illuminant. Any mismatch in the illuminant estimation will thus show up as a global color cast in the corrected image, whose color fidelity can be measured with the triad comparison method used in this study. That would help to perceptually evaluate the different algorithms, as reported previously [22]. We have shown in this paper that, for ellipsoidal chromatic distributions, illuminant changes are perceptually less well noticed when the direction of the illuminant change is parallel to the major axis of the chromatic distribution. Since many color constancy algorithms are available, selecting one with a mismatch along the major axis in the chromatic distribution would be preferable.

Another application concerns the color correction of images of outdoor scenes. The success rate of the classification of a scene as indoor or as outdoor is above 90% [41]. Once a daylight scene is classified as “outdoor,” we know that the true illuminant lies somewhere along the daylight locus in

chromaticity space. The estimated illuminants from different color constancy algorithms will scatter around this locus. Here, also, there will be a preference for an algorithm, depending on the chromatic distribution of the scene, and the positions of the estimated illuminants in chromaticity space. Likewise, when a scene is classified as indoor, the range of possible illuminants is also restricted and this situation may also ask for a preferred algorithm.

ACKNOWLEDGMENTS

This research is supported by VICI grant 639.023.705 from the Dutch Organization for Scientific Research (NWO).

REFERENCES

1. G. Buchsbaum, "A spatial processor model for object color perception," J. Franklin Inst. **310**, 1–26 (1980).
2. D. A. Forsyth, "A novel algorithm for color constancy," Int. J. Comput. Vis. **5**(1), 5–35 (1990).
3. G. D. Finlayson and G. Schaefer, "Solving for colour constancy using a constrained dichromatic reflection model," Int. J. Comput. Vis. **42**, 127–144 (2001).
4. L. T. Maloney and B. A. Wandell, "Color constancy: a method for recovering surface spectral reflectance," J. Opt. Soc. Am. A **3**, 29–33 (1986).
5. D. H. Brainard and W. T. Freeman, "Bayesian color constancy," J. Opt. Soc. Am. A **14**, 1393–1411 (1997).
6. B. Funt, K. Barnard, and L. Martin, "Is colour constancy good enough?" *Proceedings of the 5th European Conference on Computer Vision* (Springer-Verlag, 1998), pp. 445–459.
7. S. D. Hordley and G. D. Finlayson, "Reevaluation of color constancy algorithm performance," J. Opt. Soc. Am. A **23**, 1008–1020 (2006).
8. L. Arend, A. Reeves, J. Schirillo, and R. Goldstein, "Simultaneous color constancy: patterns with diverse Munsell values," J. Opt. Soc. Am. A **8**, 661–672 (1991).
9. H. E. Smithson, "Sensory, computational and cognitive components of human colour constancy," Phil. Trans. R. Soc. B **360**, 1329–1346 (2005).
10. D. H. Foster, "Color constancy," Vis. Res. **51**, 674–700 (2011).
11. O. Rinner and K. R. Gegenfurtner, "Time course of chromatic adaptation for color appearance and discrimination," Vis. Res. **40**, 1813–1826 (2000).
12. M. P. Lucassen, T. Gevers, and A. Gijsenij, "Color fidelity of chromatic distributions by triad illuminant comparison," in *IEEE Image, Video, and Multidimensional Signal Processing (IVMSP) Workshop: Perception and Visual Signal Analysis* (2011), pp. 1–6.
13. "Colorimetry," CIE Technical Report 15.2-1986, 2nd ed. (Central Bureau of the CIE, 1986) .
14. "Industrial Colour-Difference Evaluation," CIE Technical Report 116-1995 (Central Bureau of the CIE, 1995).
15. "Improvement to Industrial Colour-Difference Evaluation," CIE Technical Report 142-2000 (Central Bureau of the CIE, 2000).
16. X. Zhang and B. A. Wandell, "A spatial extension of CIELAB for digital color image reproduction," in *Society for Information Display 96 Digest*, San Diego, California, 1996, pp. 731–734.
17. A. Toet and M. P. Lucassen, "A new universal colour image fidelity metric," Displays **24**, 197–207 (2003).
18. J. Van de Weijer, T. Gevers, and A. Gijsenij, "Edge based color constancy," IEEE Trans. Image Process. **16**, 2207–2214 (2007).
19. F. Ciurea and B. Funt, "A large image database for color constancy research," in *IS&T 11th Color Imaging Conference*, Scottsdale, Arizona, 2003, pp. 160–164.
20. C. Van Trigt, "Smoothest reflectance functions. I. Definition and main results," J. Opt. Soc. Am. A **7**, 1891–1904 (1990).
21. P. B. Delahunt and D. H. Brainard, "Does human color constancy incorporate the statistical regularity of natural daylight?" J. Vis. **4**(2):1, 57–81 (2004).
22. A. Gijsenij, T. Gevers, and M. P. Lucassen, "A perceptual analysis of distance measures for color constancy," J. Opt. Soc. Am. A **26**, 2243–2256 (2009).
23. M. P. Lucassen, P. Bijl, and J. Roelofsen, "The perception of static colored noise: detection and masking described by CIE94," Color Res. Appl. **33**, 178–191 (2008).
24. M. J. Swain and D. H. Ballard, "Color indexing," Int. J. Comput. Vis. **7**(1), 11–32 (1991).
25. Bruce Lindbloom, http://www.brucelindbloom.com.
26. S. A. Ajagamelle, M. Pedersen, and G. Simone, "Analysis of the difference of Gaussians model in image difference metrics," in *IS&T 5th European Conference on Colour in Graphics, Imaging, and Vision*, Joensuu, Finland, 2010, pp. 489–496.
27. Z. Wang and A. C. Bovik, "A universal image quality index," IEEE Signal Process. Lett. **9**, 81–84 (2002).
28. D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: implications for visual coding," J. Opt. Soc. Am. A **15**, 2036–2045 (1998).
29. J. Golz and D. I. A. MacLeod, "Influence of scene statistics on colour constancy," Nature **415**, 637–640 (2002).
30. R. Mausfeld and J. Andres, "Second-order statistics of colour codes modulate transformations that effectuate varying degrees of scene invariance and illumination invariance," Perception **31**, 209–224 (2002).
31. J. Golz, "The role of chromatic scene statistics in color constancy: spatial integration," J. Vis. **8**(13):6, 1–16 (2008).
32. J. J. M. Granzier, E. Brenner, and J. B. J. Smeets, "Can illumination estimates provide the basis for color constancy?" J. Vis. **9**(3):18, 1–11 (2009).
33. M. P. Lucassen and J. Walraven, "Quantifying color constancy: evidence for nonlinear processing of cone-specific contrast," Vis. Res. **33**, 739–757 (1993).
34. D. H. Brainard, "Color constancy in the nearly natural image. 2. Achromatic loci," J. Opt. Soc. Am. A **15**, 307–325 (1998).
35. D. H. Brainard, P. Longère, P. B. Delahunt, W. T. Freeman, J. M. Kraft, and B. Xiao, "Bayesian model of human color constancy," J. Vis. **6**(11):10, 1267–1281 (2006).
36. T. Hansen, S. Walter, and K. R. Gegenfurtner, "Effects of spatial and temporal context on color categories and color constancy," J. Vis. **7**(4):2, 1–15 (2007).
37. C. Arnold, "Surface color perception under different illuminants and surface collections," Unveröffentlichte Dissertation (Universität Regensburg, 2009).
38. "Colour Rendering (TC 1-33 closing remarks)," CIE 135/2 (CIE Central Bureau, Vienna, 1999).
39. M. P. Lucassen, A. Gijsenij, and T. Gevers, "Comparing objective and subjective error measures for color constancy," in *IS&T 4th European Conference on Colour in Graphics, Imaging, and Vision*, Terrassa, Spain, 2008.
40. D. H. Foster, K. Amano, and S. M. C. Nascimento, "Colour constancy from temporal cues: better matches with less variability under fast illuminant changes," Vis. Res. **41**, 285–293 (2001).
41. N. Serrano, A. Savakis, and J. Luo, "A computationally efficient approach to indoor/outdoor scene classification," in *16th International Conference on Pattern Recognition* (2002), Vol. **4**, pp. 146–149.