Perceptual analysis of distance measures for color constancy algorithms

Arjan Gijsenij,* Theo Gevers, and Marcel P. Lucassen

Intelligent Systems Lab Amsterdam, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
**Corresponding author: a.gijsenij@uva.nl*

Received March 20, 2009; revised July 31, 2009; accepted August 20, 2009;
posted August 31, 2009 (Doc. ID 108990); published September 25, 2009

Color constancy algorithms are often evaluated by using a distance measure that is based on mathematical principles, such as the angular error. However, it is unknown whether these distance measures correlate to human vision. Therefore, the main goal of our paper is to analyze the correlation between several performance measures and the quality, obtained by using psychophysical experiments, of the output images generated by various color constancy algorithms. Subsequent issues that are addressed are the distribution of performance measures, suggesting additional and alternative information that can be provided to summarize the performance over a large set of images, and the perceptual significance of obtained improvements, i.e., the improvement that should be obtained before the difference becomes noticeable to a human observer. © 2009 Optical Society of America

OCIS codes: 150.0150, 330.1690, 330.5510.

1. INTRODUCTION

Color constancy is the ability of a visual system, either human or machine, to maintain stable object color appearances despite considerable changes in the color of the illuminant. Color constancy is a central topic in color and computer vision. The usual approach to solve the color constancy problem is by estimating the illuminant from the visual scene, after which reflectance may be recovered.

Many color constancy methods have been proposed, e.g., [1–4]. For benchmarking, the accuracy of color constancy algorithms is evaluated by computing a distance measure on the same data sets such as in [5,6]. In fact, these distance measures compute to what extent an original illuminant *vector* approximates the estimated one. Two commonly used distance measures are the Euclidean distance and the angular error, of which the latter is probably more widely used. However, as these distance measures themselves are based on mathematical principles and computed in normalized-*rgb* color space, it is unknown whether these distance measures correlate to human vision. Further, other distance measures could be defined based on the principles of human vision.

Therefore, in this paper, a taxonomy of different distance measures for color constancy algorithms is presented first, ranging from mathematics-based distances to perceptual and color constancy specific distances. Then, a perceptual comparison of these distance measures for color constancy is provided. To reveal the correlation between the distance measures and perception, color-corrected images are compared with the original images under reference illumination by visual inspection. In this way, distance measures are evaluated by psychophysical experiments involving paired comparisons of the color-corrected images. Further, following [7], a discussion of the distribution of performance measures is given, suggesting additional and alternative information that can be provided to give further insight into the performance of color constancy algorithms on a large set of images.

Finally, in addition to the psychophysical evaluation of performance measures, an analysis of the perceptual difference between color constancy algorithms is presented. This analysis is used to provide an indication of the perceptual significance of an obtained improvement in performance. In other words, the result of this analysis can be used to indicate whether an observer can actually see the difference between color-corrected images resulting from two color constancy algorithms.

The paper is organized as follows. In Section 2, color constancy and image transformation is discussed. Further, a set of color constancy methods is introduced. Then, the different distance measures are presented in Section 3. The first type concerns mathematical measures, including the angular error and Euclidean distance. The second type concerns measuring the distance in different color spaces, e.g., device-independent, perceptual, or intuitive color spaces. Third, two domain-specific distance measures are analyzed. In Section 4, the experimental setup of the psychophysical experiments is discussed, and the results of these experiments are presented in Section 5. In Section 6 various color constancy algorithms are compared to show the impact of several distance measures, and in Section 7 the perceptual significance of the difference between two algorithms is discussed. Finally, a discussion of the obtained results is presented in Section 8.

2. COLOR CONSTANCY

The image values $\mathbf{f}$ for a Lambertian surface depend on the color of the light source $e(\lambda)$, the surface reflectance $s(\mathbf{x},\lambda)$ and the camera sensitivity function $\mathbf{c}(\lambda)$, where λ is the wavelength of the light and $\mathbf{x}$ is the spatial coordinate:

$$\mathbf{f}(\mathbf{x}) = \int_{\omega} e(\lambda)\mathbf{c}(\lambda)s(\mathbf{x},\lambda)\mathrm{d}\lambda, \tag{1}$$

where ω is the visible spectrum. Assuming that the scene is illuminated by one light source and that the observed color of the light source $\mathbf{e}$ depends on the color of the light source $e(\lambda)$ as well as the camera sensitivity function $\mathbf{c}(\lambda)$, then color constancy is equivalent to the estimation of $\mathbf{e}$ by

$$\mathbf{e} = \int_{\omega} e(\lambda)\mathbf{c}(\lambda)\mathrm{d}\lambda, \tag{2}$$

given the image values of $\mathbf{f}$, since both $e(\lambda)$ and $\mathbf{c}(\lambda)$ are, in general, unknown. This is an underconstrained problem, and therefore it cannot be solved without further assumptions.

A. Color Constancy Algorithms

Several color constancy algorithms exist. Two well-established algorithms are based on the retinex theory proposed by [1]. The White-Patch algorithm is based on the white-patch assumption, i.e., the assumption that the maximum response in the RGB channels is caused by a white patch. The Gray-World algorithm by Buchsbaum [2] is based on the gray-world assumption, i.e., the assumption that average reflectance in a scene is achromatic. Finlayson and Trezzi [3] proved these two algorithms to be special instances of the more general Minkowski norm:

$$\mathcal{L}_p = \left(\frac{\int \mathbf{f}^p(\mathbf{x})\mathrm{d}\mathbf{x}}{\int \mathrm{d}\mathbf{x}} \right)^{1/p} = k\mathbf{e}. \tag{3}$$

When $p=1$ is substituted, Eq. (3) is equivalent to computing the average of $\mathbf{f}(\mathbf{x})$; i.e., $\mathcal{L}_1$ equals the Gray-World algorithm. When $p=\infty$, Eq. (3) results in computing the maximum of $\mathbf{f}(\mathbf{v})$; i.e., $\mathcal{L}_\infty$ equals the White-Patch algorithm. This algorithm is called the Shades-of-Gray algorithm.

Instead of using statistics of images for estimating the illuminant, more complex methods are developed that use information that is acquired in a learning phase. Possible light sources, distributions of possible reflectance colors, and prior probabilities on the combination of colors are learned and used for estimating the color of the light source. One of the first algorithms of this type is the gamut mapping algorithm by Forsyth [8]. This algorithm is based on the assumption that in real-world images, for a given illuminant, only a limited number of colors can be observed. Using this assumption, the illuminant can be estimated by comparing the distribution of colors in the current image to a prelearned distribution of colors (called the canonical gamut). Many algorithms have been derived from the original algorithm, including color by correlation [9] and the gamut-constrained illuminant estimation [10]. Other approaches that use a learning phase include probabilistic methods [11–13] and methods based on genetic algorithms [14].

Recently, promising results have been obtained with edge information used instead of pixel information. For instance, an extension of gamut mapping to incorporate any linear filter output has been shown to outperform the regular gamut mapping algorithm [15] by using a combination of pixel and edge information. Furthermore, an extension of the Gray-World algorithm is proposed by van de Weijer *et al.* [4], resulting in the Gray-Edge assumption, i.e., the assumption that the average reflectance difference in a scene is achromatic. They propose a general framework that incorporates algorithms based on zeroth-order statistics (i.e., pixel values) like the White-Patch, the Gray-World, and the Shades-of-Gray algorithms, as well as algorithms using higher-order (e.g., first- and second-order) statistics like the Gray-Edge and second-order Gray-Edge algorithm. The framework is given by

$$\left(\int \left| \frac{\partial^n \mathbf{f}_\sigma(\mathbf{x})}{\partial \mathbf{x}^n} \right|^p \mathrm{d}\mathbf{x} \right)^{1/p} = k\mathbf{e}_{n,p,\sigma}, \tag{4}$$

where n is the order of the derivative, p is the Minkowski norm and $\mathbf{f}^\sigma(\mathbf{x}) = \mathbf{f} \otimes \mathbf{G}_\sigma$ is the convolution of the image with a Gaussian filter with scale parameter σ. By use of this equation, many different color constancy algorithms can be generated. For the purpose of this paper, five instantiations are used, representing a wide variety of algorithms:

- White-Patch algorithm ($\mathbf{e}_{0,\infty,0}$)
- Gray-World algorithm ($\mathbf{e}_{0,1,0}$)
- General Gray-World algorithm ($\mathbf{e}_{0,13,2}$)
- First-order Gray-Edge algorithm ($\mathbf{e}_{1,1,6}$)
- Second-order Gray-Edge algorithm ($\mathbf{e}_{2,1,5}$).

Many other algorithms can be generated by varying the Minkowski norm for different orders of deviations on different scales.

The main purpose of this paper is not to propose a new color constancy algorithm, nor to compare the performance of the different algorithms. The goal in this paper is to psychophysically analyze the several performance measures that are used for comparing color constancy algorithms. To this end, the framework proposed by van de Weijer [4] is used to construct several result images. The main advantages of this framework are its simplicity (i.e., all algorithms are derived from a similar assumption), repeatability (i.e., the methods are easy to implement, e.g., see [16] for source code, and no learning step is required), and variability (i.e., many different methods can be systematically created, including pixel-based methods, edge-based methods and higher-order methods, with varying performance). Since the experiments involve human subjects, the number of observations that can be made by the subjects are limited. Therefore, the methods that are used are restricted to the five instantiations of this framework mentioned earlier.

B. Image Transformation

Once the color of the light source is estimated, this estimate can be used to transform the input image to be taken under a reference (often white) light source. This transformation can be modeled by a diagonal mapping or von Kries model [17]. This model is an approximation and

might not be able to model photometric changes accurately because of disturbing effects like highlights and interreflections. However, it is widely accepted as a color correction model [18–20], and it underpins many color constancy algorithms (e.g., gamut mapping [8] and the framework of methods used [4]). The model is given by

$$\mathbf{f}^c = \mathcal{D}^{u,c}\mathbf{f}^u \Rightarrow \begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix}, \quad (5)$$

where $\mathbf{f}^u$ is the image taken under an unknown light source, $\mathbf{f}^c$ is the same image transformed, so that it appears as if it were taken under the reference light, and $\mathcal{D}^{u,c}$ is a diagonal matrix that maps colors that are taken under an unknown light source u to their corresponding colors under the canonical illuminant c. The diagonal mapping is used throughout this paper to create output images after correction by a color constancy algorithm.

3. DISTANCE MEASURES

Performance measures evaluate the performance of an illuminant estimation algorithm by comparing the estimated illuminant to a ground truth, which is known *a priori*. Since color constancy algorithms can recover the color of the light source only up to a multiplicative constant (i.e., the intensity of the light source is not estimated), distance measures compute the degree of resemblance in normalized *rgb*:

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B}. \quad (6)$$

In color constancy research, two frequently used performance measures are the Euclidean distance and the angular error, of which the latter is probably more widely used. The Euclidean distance d_{euc} between the estimated light source $\mathbf{e}_e$ and the ground truth light source $\mathbf{e}_u$ is given by

$$d_{\text{euc}}(\mathbf{e}_e,\mathbf{e}_u) = \sqrt{(r_e - r_u)^2 + (g_e - g_u)^2 + (b_e - b_u)^2}. \quad (7)$$

The angular error measures the angular distance between the estimated illuminant $\mathbf{e}_e$ and the ground truth $\mathbf{e}_u$ and is defined as

$$d_{\text{angle}}(\mathbf{e}_e,\mathbf{e}_u) = \cos^{-1}\left(\frac{\mathbf{e}_e \cdot \mathbf{e}_u}{\|\mathbf{e}_e\| \cdot \|\mathbf{e}_u\|}\right), \quad (8)$$

where $\mathbf{e}_e \cdot \mathbf{e}_u$ is the dot product of the two illuminants and $\|\cdot\|$ is the Euclidean norm of a vector.

Although the value of these two distance measures indicates how closely an original illuminant vector is approximated by the estimated one (after intensity normalization), it remains unclear how these errors correspond to the perceived difference between the output of a color constancy algorithm and the ground truth. Further, other distances can be derived. To this end, in this section, a taxonomy of different distance measures for color constancy algorithms is presented. The different distance measures are defined ranging from mathematics-based distance measures (Subsection 3.A) to perceptual measures (Subsection 3.B) and color constancy specific measures (Subsection 3.C).

A. Mathematical Distance

The two distance measures that have been discussed so far (i.e., the angular error and the Euclidean distance) can be considered to be mathematical measures. In this Subsection, other mathematical measures are introduced by considering the more general Minkowski family of distances, denoted d_{Mink}, of which the Euclidean distance is a member:

$$d_{\text{Mink}}(\mathbf{e}_e,\mathbf{e}_u) = (|r_e - r_u|^p + |g_e - g_u|^p + |b_e - b_u|^p)^{1/p}, \quad (9)$$

where p is the corresponding Minkowski norm. In this paper, three special cases of this distance measure are evaluated. These three measures are the Manhattan distance (d_{man}) for $p=1$, the Euclidean distance (d_{euc}) for $p=2$, and the Chebychev distance (d_{sup}) for $p=\infty$.

B. Perceptual Distances

The goal of the color constancy algorithms is to obtain an output image that is identical to a reference image, i.e., an image of the same scene taken under a canonical, often white, light source. Therefore, perceptual distance measures as well as mathematical distance measures are included in the analysis. For this purpose, the estimated color of the light source and the ground truth are first transformed to different (human vision) color spaces, after which they are compared. Therefore, in this section, the distance is measured in the (approximately) perceptually uniform color spaces CIELAB and CIELUV [21], as well as in the more intuitive color channels chroma C and hue h. Further, in addition to the Euclidean distance between CIELAB colors, the CIEDE2000 [22] is computed, since the metric is shown to be more uniform and is considered to be state of the art in industrial applications.

Most color constancy algorithms are restricted to estimating the chromaticity values of the illuminant. To evaluate the performance of the light source estimations in different color spaces, both the (intensity normalized) estimate and the ground truth light source are applied to a perfect reflecting diffuser. Hence, two sets of (R,G,B) values are obtained, representing the nominally white object-color stimulus under the estimated light source and under the true light source. These (R,G,B) values can consequently be converted to different color spaces. Conversion from RGB to XYZ is device dependent, e.g., depending on the camera settings. Many different RGB working spaces can be defined, but since the monitor that is used in the experiments closely approximates the sRGB standard monitor profile (see Subsection 4.B), the conversion matrix is based on the sRGB color model [23]:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.4125 & 0.3576 & 0.1804 \\ 0.2127 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9502 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (10)$$

Differences in the conversion matrix can occur if the RGB working space of an image is known to be different from sRGB, e.g. Adobe RGB, NTSC RGB or CIE RGB. In Section 5, the effects of using other conversion matrices,

based on several RGB working spaces, are discussed.

After conversion to XYZ, the two (approximately) perceptual color models $L^*a^*b^*$ and $L^*u^*v^*$ are defined using $(X_w, Y_w, Z_w) = (0.9505, 1.0000, 1.0888)$ as reference white, which is the appropriate reference white for the sRGB color model [24]. From these perceptual color spaces, different color channels can be computed, such as chroma C and hue h. The transformation from $L^*a^*b^*$ to C^* and h is given by

$$C^*_{ab} = \sqrt{(a^*)^2 + (b^*)^2}, \quad h_{ab} = \tan^{-1}\left(\frac{b^*}{a^*}\right), \tag{11}$$

and analogously for $L^*u^*v^*$.

Finally, it is known that the spectral sensitivity of the human eye is nonuniform. This important property of the human visual system is used, for instance, in the conversion of RGB images to luminance images [25]. A deviation in one color channel might have a stronger effect on the perceived difference between two images than a deviation in another channel. This leads us to the introduction of the weighted Euclidean distance, or perceptual Euclidean distance (PED). The weights for the different color channels are described as sensitivity measures as follows:

$$\mathrm{PED}(\mathbf{e}_e, \mathbf{e}_u) = \sqrt{w_R(r_e - r_u)^2 + w_G(g_e - g_u)^2 + w_B(b_e - b_u)^2}, \tag{12}$$

where $w_R + w_G + w_B = 1$. Note that CIELAB and CIELUV also have weighting terms modifying different dimensions. However, these color spaces are just two instantiations, while the weighted Euclidean distance covers a large range of instantiations.

C. Color Constancy Distances

In this Subsection, two color constancy specific distances are discussed. The first is the color constancy index CCI [26], also called the Brunswik ratio [27], and is generally used to measure perceptual color constancy [28,29]. It is defined as the ratio of the amount of adaptation that is obtained by a human observer versus no adaptation at all:

$$\mathrm{CCI} = b/a, \tag{13}$$

where b is defined as the distance from the estimated light source to the true light source and a is defined as the distance from the true light source to a white reference light. During evaluation, several different color spaces are used to compute the values a and b.

The second is a new measure, called the *gamut intersection*, which makes use of the gamuts of the colors that can occur under a given light source. It is based on the assumption underlying the gamut mapping algorithm; i.e., *under a given light source, onlya limited number of colors are observed*. The difference between the full gamuts of two light sources is an indication of the difference between these two light sources. For instance, if two light sources are identical, then the gamuts of colors that can occur under these two light sources will coincide, while the similarity of the gamuts will be smaller if the difference between the two light sources is larger. The gamut intersection is measured as the fraction of colors that occur under the estimated light source, with respect to the colors that occur under the true, ground truth, light source:

$$d_{\mathrm{gamut}}(\mathbf{e}_e, \mathbf{e}_u) = \frac{\mathrm{vol}(\mathcal{G}_e \cap \mathcal{G}_u)}{\mathrm{vol}(\mathcal{G}_u)}, \tag{14}$$

where $\mathcal{G}_i$ is the gamut of all possible colors under illuminant i and $\mathrm{vol}(\mathcal{G}_i)$ is the volume of this gamut. The gamut $\mathcal{G}_i$ is computed by applying the diagonal mapping, corresponding to light source i, to a canonical gamut.

4. EXPERIMENTAL SETUP

In this section, the experimental setup of the psychophysical experiments is discussed. The experiments are performed on two data sets, one containing hyperspectral recordings of natural and rural scenes, and the other containing a range of indoor and outdoor scenes, measured in RGB. The images are shown on a calibrated color monitor, and observers are shown images in a pairwise comparison paradigm. For each pair of color-corrected images, the observers have to specify which of the two images is closer to the ideal result (which is also shown). In this way, comparison of the distance measures (objective performance) and visual judgment (subjective performance) is carried out by computing the correlation between the two performance measures.

A. Data

Two data sets are used for the psychophysical experiments. The first data set consists of hyperspectral images and is used to perform a thorough, i.e., colorimetrically correct, analysis. The second data set consists of RGB images and is used to analyze the results of the experiments with the first data set.

Hyperspectral data. The first data set, originating from [29], consists of eight hyperspectral images, of which four are shown in Figs. 1(a)–1(d). These images are chosen in order to be able to study realistic, i.e., colorimetrically correct, and naturally occurring changes in daylight illumination.

Similar to the work of Delahunt and Brainard [28], one neutral illuminant (CIE $D65$) and four chromatic illuminants (red, green, yellow, blue) are selected to render images under different light sources. The spectral power distributions of the selected illuminants are shown in Fig. 2(a) and are created with the use of the CIE daylight basis function, as described in [24]. In Fig. 2(b), images of scene 3 rendered under these four illuminants are shown.

RGB images. The second data set consists of 50 RGB images, both indoor and outdoor. These images are taken from [5], which is a large data set (originally containing over 11,000 images) that is well known in color constancy research. For all images, the ground truth of the color of the light source is known from a gray sphere that was mounted on top of the camera. This gray sphere is cropped during the experiments. Some example images are shown in Figs. 1(e)–1(h). Images from this data set are not as well calibrated as the hyperspectral set and are therefore used mostly to confirm the results on the hyperspectral data.

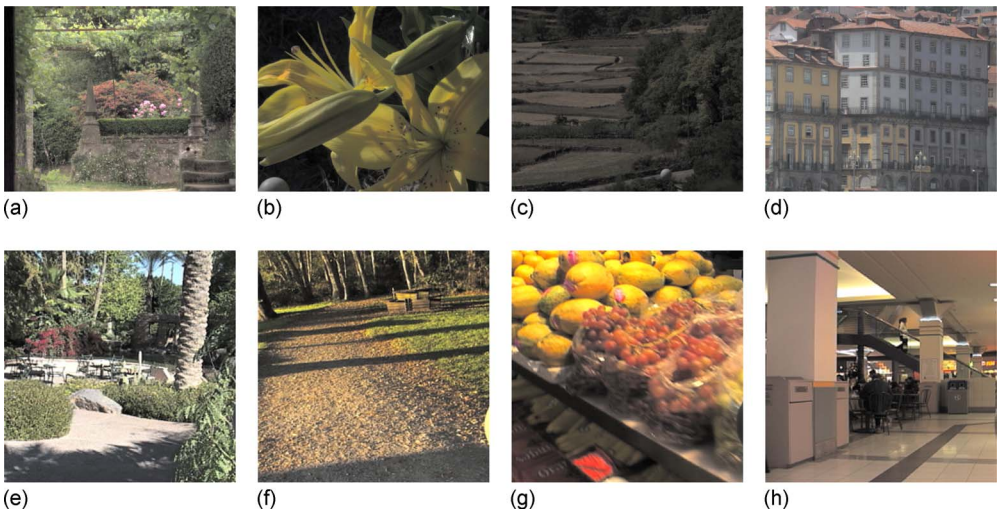

Fig. 1. (Color online) Four examples of the hyperspectral scenes used in this study are shown in (a)–(d), rendered under the neutral *D*65 illuminant. In (e)–(h), four examples of the RGB images are shown.

B. Monitor

Images are viewed on a high-resolution (1600 × 1200 pixels, 0.27 mm dot pitch) calibrated LCD monitor, an Eizo ColorEdge CG211. The monitor is driven by a computer system having a 24 bit (RGB) color graphics card operating at a 60 Hz refresh rate. Colorimetric calibration of the LCD is performed before each experimental session by using a GretagMacbeth Eye-one spectrophotometer. Combined with ColorNavigator software from Eizo, this setup allows self-calibration of the monitor at specified target settings for the white point, black level, and gamma values. The monitor is calibrated to a *D*65 white point of 80 cd/m^2, with gamma 2.2 for each of the three color primaries. CIE 1931 x,y chromaticities coordinates of the primaries were $(x,y)=(0.638,0.322)$ for red, (0.299,0.611) for green, and (0.145,0.058) for blue, respectively. These settings closely approximate the sRGB standard monitor profile [23], which is used for rendering the spectral scenes under our illuminants. Spatial uniformity of the display, measured relative to the center of the monitor, is $\Delta E_{ab}<1.5$ according to the manufacturer's calibration certificates.

C. Observers

All observers that participated in the experiments have normal color vision and normal or corrected-to-normal visual acuity. Subjects are screened for color vision deficiencies with the HRR pseudo-isochromatic plates (fourth edition), allowing color vision testing along both the red–green and yellow–blue axes of color space [30]. After taking the color vision test, our subjects first adapted for about 5 min to the light level in a dim room that only received some daylight from a window that is covered with sunscreens (both inside and outside). In the meantime they were made familiar with the experimental procedure.

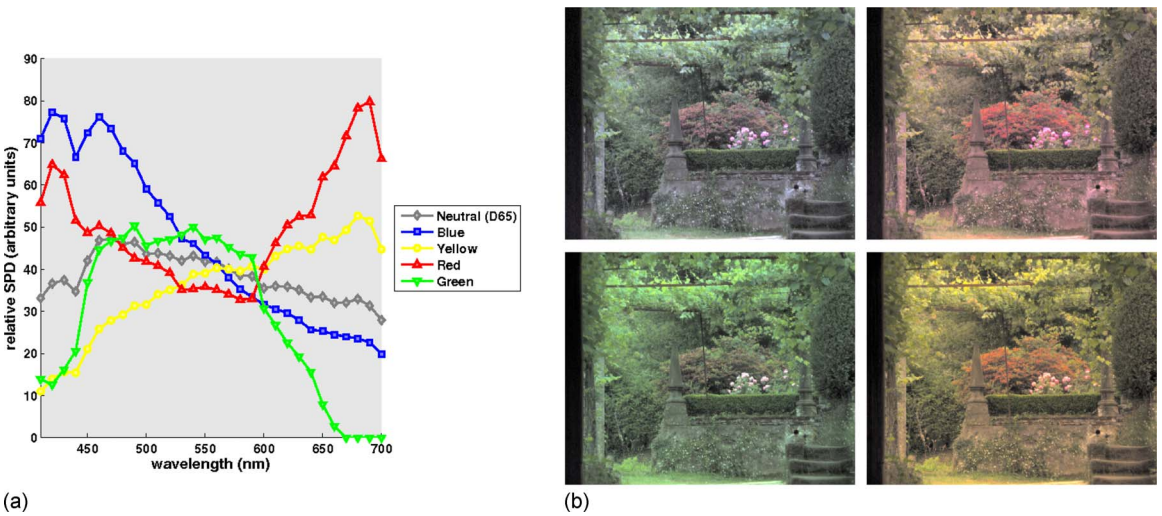

Fig. 2. (Color online) Relative spectral power distribution of the illuminants used in the experiments. Left, illuminant spectra; right, illuminants applied to scene 3. The illuminants are created with the CIE basis functions for spectral variations in natural daylight and were scaled such that a perfectly white reflector would have a luminance of 40 cd/m^2. The four chromatic illuminants red, green, yellow, and blue are perceptually at an equal distance (28 ΔE_{ab}) from the neutral (D65) illuminant.

D. Experimental Procedure

The experimental procedure consists of a sequence of image comparisons. The subjects are shown four images at once, arranged in a square layout, on a gray background having $L^*=50$ and $a^*=b^*=0$, see Fig. 3. The upper two images are (identical) reference images, representing the test scene. The lower two images correspond to the resulting output of two different color constancy algorithms, applied to the original test scene (i.e., the scene under a certain light source). Subjects are instructed to compare the color reproduction of each of the lower images with the upper references. Both the global color impression of the scene and the colors of local image details are to be addressed. Subjects then indicate (by pressing a key on the computer's keyboard) which of the two lower images has the best color reproduction. If the color reproduction of the two test images is identical (as good or as bad), the subjects have the possibility of indicating this. Subjects are told that response time would be measured, but that they are not under time pressure; they can use as much time as they need to come to a decision.

In each trial of our paired-comparison experiment, two color constancy algorithms are competing, the result of which can be interpreted in terms of a win, a loss, or a tie. Each of the five color constancy algorithms is competing with every other algorithm once, for every image and illuminant, in tournament language known as a single round-robin schedule [31]. We apply a scoring mechanism in which the color constancy algorithm underlying a win is awarded with 1 point and the algorithm underlying a loss with no points. In case of a tie, the competing algorithms both receive 0.5 point. Ranking of the algorithms can then be performed by simply comparing the total number of points. The above scoring mechanism is straightforward and makes no distributional assumptions.

5. RESULTS

Experimental results are processed on an average-observer basis. The interobserver variability is analyzed first, after which the results of the observers are averaged to come to robust subjective scores. Next, correlation between these subjective scores and the several objective measures is determined by using linear regression. Since the objective measures are absolute error values and the subjective measure depicts a relative relation between the algorithms, the objective measures are converted to relative values. This is done by using the same round-robin schedule as used with the human observers, this time using the error values as the criterion to decide which result is better.

A. Hyperspectral Data

The experiments on the hyperspectral data are run in two sessions, with four scenes per session. Per session, a total of 160 comparisons are made (4 scenes $\times$ 4 illuminants $\times$ 10 algorithm combinations). Half of the subjects started with the first set, the other half with the second set. The two images that are to be compared in a trial always belong to the same chromatic illuminant. The sequence of the trials is randomized, and the two test images are randomly assigned to left and right positions.

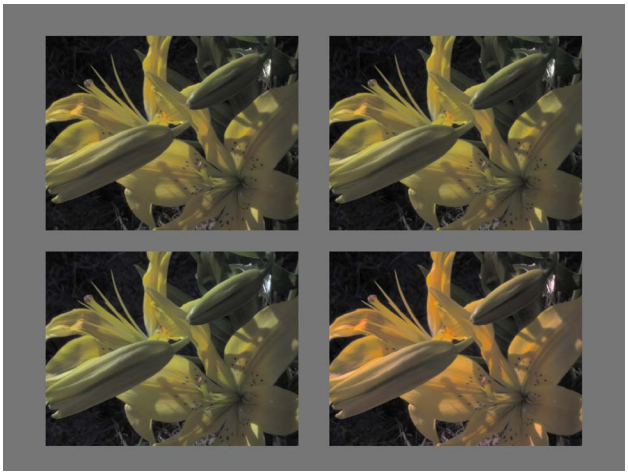

Fig. 3. (Color online) Screen capture of an experimental trial. Subjects indicate which of the two bottom images (resulting from two different color constancy algorithms) is the best match to the upper reference image. Background dimensions are 39.6° $\times$ 30.2° visual angle. Horizontal and vertical separation between the images was 2.1° and 1.4°, respectively. The hyperspectral images are 16.6° $\times$ 12.7°, the RGB images are 6.2° $\times$ 6.2°.

Eight observers participated in this experiment, four men and four women, with ages ranging from 24 to 43 years (an average of 34.6). At a viewing distance of about 60 cm, each of the four images subtended a visual angle of 16.6° $\times$ 12.7°. Horizontal and vertical separation between images was 2.1° and 1.4°, respectively.

Interobserver variability. As a measure of the interobserver variability, the individual differences from the mean observer scores are computed, a procedure that is often used in studies involving visual judgements, e.g., [32,33]. For each observer, the correlation coefficient of his or her average algorithm scores (averaged over scenes and illuminants) with the algorithm scores of the average observer is computed. The correlation coefficients so obtained vary from 0.974 to 0.999, with an average of 0.990. Correlation coefficients between scores of the individual observers range from 0.937 to 0.997. The significance of this result becomes clear when these high values are compared with the values that can be obtained from random data. Based on random generated responses for each trial, with 45%, 45%, 10% chances for a win, loss, or tie, respectively, the correlation coefficients of the simulated individual observers range from 0.074 to 0.948, with an average of 0.396. Correlation coefficients between actual individual observers in this case range from -0.693 to 0.945. Since the agreement between observers is considered good, in the remainder we will discuss the results only for the average observer.

Mathematical measures versus subjective scores. First, the angular error d_{angle} is analyzed, since this measure is probably the most widely used performance measure in color constancy research. Overall, the correlation between the angular error and the perception of the human observer is reasonably high, with an average correlation coefficient of 0.895; see Table 1, where the correlation coefficients on the spectral data set for all distance measures are summarized. Also shown in this table are the results

Table 1. Correlation Coefficients ρ for Several Distance Measures and Color Spaces with Respect to the Subjective Measure[a]

Measure	Hyperspectral Data		Images	
	ρ	t test	ρ	t test
d_{angle}	0.895	3	0.926	3
d_{man}	0.893	3	0.930	3
d_{euc}	0.890	3	0.928	3
d_{sup}	0.817	3	0.906	3
$d_{\text{euc}}-L^*a^*b^*$	0.894	4	0.921	3
$\Delta E^*_{00}-L^*a^*b^*$	0.896	4	0.916	3
$d_{\text{euc}}-L^*u^*v^*$	0.864	3	0.925	3
$d_{\text{euc}}-C+h$	0.646	0	0.593	1
$d_{\text{euc}}-C$	0.619	0	0.562	1
$d_{\text{euc}}-h$	0.541	0	0.348	0
$\text{PED}_{\text{hyperspectral}}$	0.963	14	—	—
PED_{RGB}	—	—	0.961	15
$\text{PED}_{\text{proposed}}$	**0.960**	14	**0.957**	15
$\text{CCI}(d_{\text{angle}})$	0.895	3	0.931	3
$\text{CCI}(d_{\text{euc},RGB})$	0.893	3	0.929	3
$\text{CCI}(d_{\text{euc},L^*a^*b^*})$	0.905	4	0.921	3
$\text{CCI}(d_{\text{euc},L^*u^*v^*})$	0.880	3	0.927	3
d_{gamut}	0.965	14	0.908	3

[a]The subjective measure is derived from human observers. Significance is shown using a Student's t test (at the 95% confidence level). The score in the column t test represents the number of times the null hypothesis (i.e., that two distance measures have a similar mean correlation coefficient) is rejected.

of a paired comparison between the different measures. A Student's t test (at 95% confidence level) is used to test the null hypothesis that the mean correlation coefficients of two distance measures are equal, against the alternative hypothesis that measure A correlates higher with the human observer than measure B. When every distance measure is compared with all others, a score is generated representing the number of times the null hypothesis is rejected, i.e., the number of times that the correlation coefficient of the given distance measure is significantly better than the other measures.

Instead of looking at the average correlation over all images, as for the hyperspectral data in Table 1, we now analyze the correlation of all images individually. For most images, the correlation is relatively high (correlation coefficient $\rho > 0.95$), while for some images the correlation is somewhat lower, but still acceptable ($\rho > 0.8$). In a few cases, however, the correlation is rather low ($\rho < 0.7$). When the results of the images with such a low correlation are observed, the weakness of the angular error becomes apparent. For these images, results of some images are judged worse than indicated by the angular error, meaning that human observers do not agree with the angular error. The angular errors for the corresponding images are similar, but visual inspection of the results show that the estimated illuminants (and hence the resulting images) are far from similar. In conclusion, from a perceptual point of view, the direction in which the estimated color of the light source deviates from the ground truth is important. Yet, the angular error, by nature, ignores this direction completely.

The correlation between the Euclidean distance and the human observer is similar to the correlation of the angular error, i.e., $\rho = 0.890$. The other two instantiations of the Minkowski distance, i.e., the Manhattan distance (d_{man}) and the Chebychev distance (d_{sup}), have a correlation coefficient of $\rho = 0.893$ and $\rho = 0.817$, respectively. The correlation coefficients of other Minkowski-type distance measures are not shown here, but vary between $\rho = 0.89$ and $\rho = 0.82$. In conclusion, none of these mathematical distance measures is significantly different from the others.

Perceptual measures versus subjective scores. First, the estimated illuminant and the ground truth are converted from normalized-*rgb* to RGB values. This is done by computing the two corresponding diagonal mappings to a perfect white reflectance, in order to obtain the RGB values of a perfect reflectance under the two light sources. These RGB values are then converted to *XYZ* and the other color spaces, after which they are compared by using any of the mathematical measures. For simplicity, the Euclidean distance is used.

For comparison, recall that the correlation between the human observers and the Euclidean distance of the normalized-*rgb* values is 0.895. When the correlation of the human observers with the Euclidean distance in different color spaces is computed, the lightness channel L^* is omitted, since the intensity of all estimates is artificially imposed and similar for all light sources. Correlations of human observers and distance measured in the perceptual spaces $L^*a^*b^*$ ($\rho = 0.902$) and $L^*u^*v^*$ ($\rho = 0.872$) are similar to the correlation of the human observers with the Euclidean distance in normalized-*rgb* space. When computing the Euclidean distance in color spaces such as hue and chroma, the correlation is remarkably low; considering both chroma and hue, the correlation is 0.646, which is significantly lower than the correlation of other color spaces. Considering chroma or hue alone, the correlation drops even further to $\rho = 0.619$ and $\rho = 0.541$, respectively. In conclusion, using perceptual uniform spaces provides similar or lower correlation than *rgb*.

As is derived from the analysis of the results of the angular error, it can be beneficial to take the direction of a change in color into consideration. In this paper, this property is computed by the perceptual Euclidean distance (PED), by assigning higher weights to different color channels. The question remains, however, what values to use for the weights. For this purpose, an exhaustive search was performed to find the optimal weighting scheme, denoted $\text{PED}_{\text{hyperspectral}}$ in Table 1. The weight combination $(w_R, w_G, w_B) = (0.20, 0.79, 0.01)$ results in the highest correlation ($\rho = 0.963$); see Fig. 4(a).

Color constancy measures versus subjective scores. The color constancy index makes use of a distance measure as defined by Eq. (13), where b is defined as the distance from the estimated light source to the true light source and a is defined as the distance from the true light source to a white reference light. To compute the distance, the angular error in normalized-*rgb* and the Euclidean distance in RGB, $L^*a^*b^*$, and $L^*u^*v^*$ are used. From Table 1, it is derived that the highest correlation with the human observers is obtained when the color constancy index is measured with $L^*a^*b^*$ ($\rho = 0.905$). However, differences

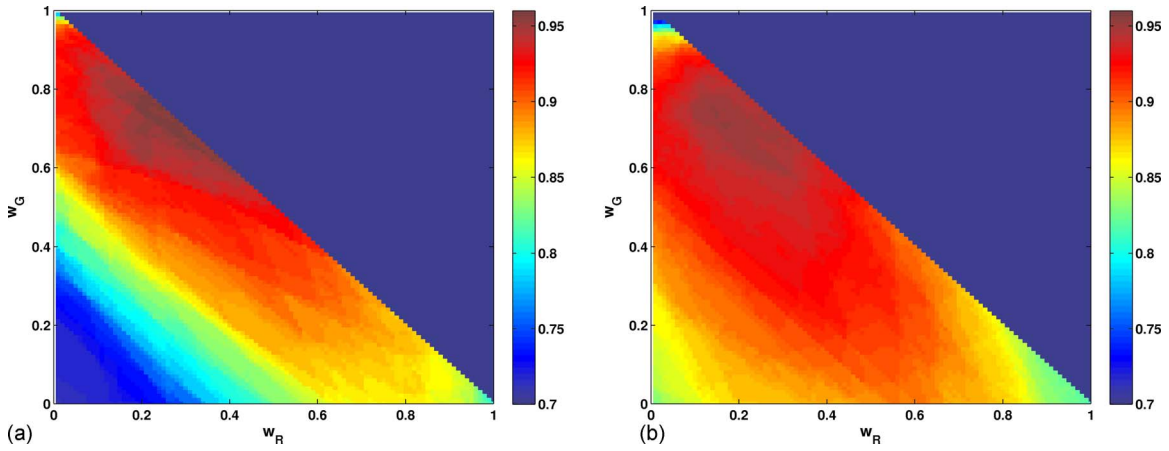

Fig. 4. (Color online) Plot of the correlation coefficients of the weighted Euclidean distance with respect to the human observer (psychophysical data). Only the dependency on weight coefficients w_R and w_G are shown here; w_B follows from $w_B=1-w_R-w_G$. Left, the results of experiments using the hyperspectral data are demonstrated; right, results of the experiments with the RGB images.

between other distance measures are small. In conclusion, the color constancy index does not correlate better with human observers than the mathematical measures.

The gamut intersection distance measures the distance of the gamuts under the estimated light source and the ground truth. These gamuts are created by applying the corresponding diagonal mappings to a canonical gamut. This canonical gamut is defined as the gamut of all colors under a known, often white, light source and is constructed by using a widely used set of 1995 surface spectra [6] combined with a perfect white illuminant. The correlation of this measure is surprisingly high (see Table 1): $\rho = 0.965$, which is even slightly higher than the correlation of the perceptual Euclidean distance (PED).

Discussion. From Table 1, (hyperspectral data), it is derived that the correlation of the angular error with the judgment of the human observers is reasonable and similar to the other mathematical measures; i.e., there is no significant difference at the 95% confidence level. Measuring the distance in perceptual color spaces such as $L^*a^*b^*$ and $L^*u^*v^*$ does not increase the correlation with human observers. Using chroma C and hue h significantly decreases the correlation with the human observers. The gamut intersection distance and the perceptual Euclidean distance have the highest correlation with the human observers. In fact, they have significantly higher (at the 95% confidence level) correlation than all other distance measures. Hence, the gamut and perceptual Euclidean distances are significantly better than all other distance measures on the spectral data set.

B. RGB Images

The experiments on the RGB images are run in three sessions, with the number of images equally divided into three parts. The sequence of the sets is randomized for every observer. In this experiment, seven observers participated (four men and three women), with ages ranging from 24 to 43 years. The difference between the observers is analyzed similarly to the experiments on the hyperspectral data, and again the agreement of the individual observers is found to be sufficiently high: the correlation coefficients vary from 0.894 to 0.977, with an average of 0.953. Correlation coefficients between scores of the individual observers range from 0.638 to 0.980. For this experiment, the correlation coefficients based on random generated responses vary from −0.634 to 0.772, with an average of 0.280. Correlation between random individual observers ranges from −0.923 to 0.889. The agreement is considered good, and consequently, only the results for the average observer are discussed.

Objective versus subjective scores. In general, the same trends in this data set as in the hyperspectral data are observed; see Table 1, RGB images. The correlation coefficients are slightly higher than the spectral data set, but the ordering between the different measures remains the same. For the mathematical measures, the angular distance ($\rho = 0.926$) the Manhattan distance ($\rho = 0.930$), and the Euclidean distance ($\rho = 0.928$) are similar, while the Chebychev distance has a lower correlation with human observers ($\rho = 0.906$). Results of the perceptual measures also show a similar trend. Correlation coefficients of the perceptual color spaces are similar to the mathematical measures, while the intuitive color spaces are significantly lower. Again, the perceptual Euclidean distance has the highest correlation ($\rho = 0.961$). This correlation is obtained with the weights $(w_R,w_G,w_B)=(0.21,0.71,0.08)$, denoted PED_{RGB} in Table 1; see also Fig. 4(b). The results for the color constancy specific distances are slightly different from the results obtained from the hyperspectral data. The results of the color constancy index are similar, but the correlation of the gamut intersection distance with the human observers is considerably lower for this data set.

Device dependency. As explained in Subsection 3.B, the transformation from RGB to $L^*a^*b^*$ and $L^*u^*v^*$ is dependent on the conversion from RGB to *XYZ*. If the RGB working space is known, as in the case of the hyperspectral data in Subsection 5.A, then the conversion from RGB to *XYZ* can be performed accurately. However, the correct conversion from RGB to *XYZ* for the images that are currently used, i.e., the RGB images, is unknown. In order to analyze the effect of the *XYZ* transformation, we used 16 frequently used RGB working spaces (of which sRGB is the most widely used) to compute the transformation from RGB to *XYZ*, adapted from [34]. As a result, we obtained 16 different values for the correlation coeffi-

cients of the distance measures based on the $L^*a^*b^*$ and $L^*u^*v^*$ color spaces. The results that are reported in Table 1, RGB images, are obtained by using the conversion from sRGB to *XYZ*, but differences with the other RGB working spaces are small. For instance, the average correlation coefficient over 16 working spaces for the Euclidean distance of the $L^*u^*v^*$ values is 0.920 (with a standard deviation of 0.006), while the correlation coefficient when assuming the sRGB color space is 0.916. From these results it is concluded that the conversion from RGB to *XYZ* has only a marginal effect on the correlation coefficients.

Discussion. The results of the experiments on the RGB images, Table 1, correspond to the results of the experiments on the hyperspectral data. Note, though, that the images in this data set are gamma corrected (with an unknown value for gamma) before the color constancy algorithms are used to color correct the images. Applying gamma correction previously to the color constancy algorithms affects the performance of the algorithms, but this effect was not investigated in this paper.

The most noticeable difference between the results for this data set and the results for the previous data set is the correlation of the gamut intersection distance. This distance has the highest correlation with the human observers for the hyperspectral data. However, for the RGB images, the correlation is considerably lower, though not significantly lower, than for the other measures. The correlation of the perceptual Euclidean distance for the RGB images is still significantly higher than the correlation of all other distance measures. To obtain a robust, stable combination of weights, the results of the exhaustive search on the hyperspectral data and the RGB images are averaged. The optimal correlation is found for the weight combination $(w_R, w_G, w_B) = (0.26, 0.7, 0.04)$. With these weights, the correlation of the perceptual Euclidean distance with human observers for the hyperspectral data is 0.960, and for the RGB images is 0.957, denoted $\text{PED}_{\text{proposed}}$ in Table 1. Both are still significantly higher (at the 95% confidence level) than all other distance measures.

6. COMPARING ALGORITHM PERFORMANCE

The different error measures that are discussed in this paper allow a comparison of different color constancy algorithms used on an image data set. However, as was shown by Hordley and Finlayson [7], different summarizing statistics can lead to different conclusions. For instance, if the distribution of errors of a specific data set is severely skewed, then the mean error is not an accurate summary of the underlying distribution, and consequently comparing the mean error of two color constancy algorithms might result in wrong conclusions about the performance of those algorithms. This section provides an analysis of the proposed perceptual Euclidean distance, to identify which summarizing statistic is most suited. Further, some characteristics are presented and compared with the characteristics of the angular error.

A. Distribution of Errors

When evaluating the performance of color constancy algorithms on a whole data set instead of on a single image, the performances on all individual images need to be summarized into a single statistic. This is often done by taking the mean, root mean square, or median of, for instance, the angular errors of all images in the data set. If the error measures are normally distributed, then the mean is the most commonly used measure for describing the distribution, and the root mean square provides an estimate of the standard deviation. However, if the metric is not normally distributed, for instance, if the distribution is heavily skewed or contains many outliers, then the median is more appropriate for summarizing the underlying distribution [35].

From previous work, it is known that the angular error is not normally distributed [7]. To test whether the perceptual Euclidean distance is normally distributed, a similar experiment as in [7] is conducted. In Fig. 5, the errors for the White-Patch algorithm on the 11,000 images from the RGB images data set [5] are plotted, from which it is clear that both the angular error and the perceptual Euclidean distance are not normally distributed. The distributions of both metrics have a high peak at lower error rates, and a fairly long tail. For such distributions, it is known that the mean is a poor summary statistic, and hence, previously, it was proposed to use the median to describe the central tendency [7]. Alternatively, to provide more insight into the complete distribution of errors, one can calculate box plots or compute the trimean instead of the median. Box plots are used to visualize the underlying distributions of the error metric of a given color constancy method, as an addition to a summarizing statistic. This summarizing statistic can be the median, as proposed by Hordley and Finlayson [7], or it can be the trimean, a statistic that is robust to outliers (the main advantage of the median over a statistic like the root mean square), but still has attention to the extreme values in the distribution [36,37]. The trimean (TM) can be calculated as the weighted average of the first, second, and third quantile Q_1, Q_2, and Q_3, respectively:

$$\text{TM} = 0.25Q_1 + 0.5Q_2 + 0.25Q_3. \tag{15}$$

The second quantile Q_2 is the median of the distribution, and the first and third quantiles Q_1 and Q_3 are called hinges. In other words, the trimean can be described as the average of the median and the midhinge.

B. Analysis of Results

In this section, two comparisons of color constancy algorithms are presented, to analyze the effects of the proposed perceptual Euclidean distance and the different summarizing statistics. The first comparison is based on methods from the color constancy framework proposed by van de Weijer *et al.* [4]. The second comparison uses pixel-based and edge-based gamut mapping algorithms proposed by Gijsenij *et al.* [15]. The data set that is used to compare the methods is the full RGB images data set with over 11,000 images [5]. This set is chosen because it is well known and widely used in color constancy research [4,13,15,38–42]. Note that the purpose of this section is *not* to provide a large-scale comparison, but to gain insight into the behavior of the perceptual Euclidean distance with respect to the angular error.

Low-level color constancy. Eight methods are created

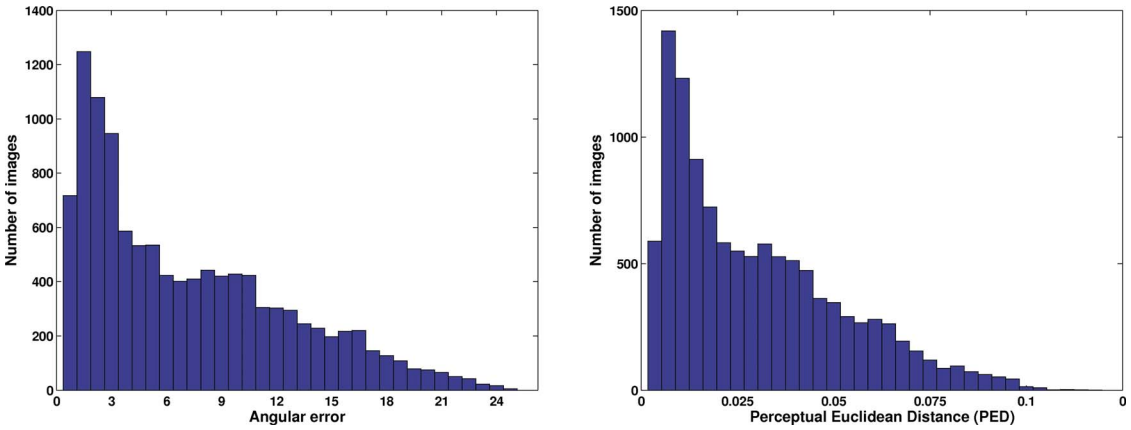

Fig. 5. (Color online) Distribution of estimated illuminant errors for the White-Patch algorithm, obtained for a set of over 11,000 images.

using the framework of [4], all with different properties. Four methods use pixel values, two methods use edges, and another two methods use higher-order statistics for estimating the illuminant, all constructed by applying different parameters to Eq. (4). The different parameter settings obviously result in different performances. However, as Table 2 shows, the ranking of the methods is quite dependent on the summarizing statistic and evaluation metric that are used. When the angular error is compared with the perceptual Euclidean distance, no large differences in ranking are observed. The White-Patch algorithm (i.e., $e_{0,\infty,0}$) ranks higher when the median perceptual Euclidean distance is considered as compared with the median angular error, and the ordering of some algorithms is reversed when the trimean is considered. However, comparing the median and the trimean as measures for central tendency reveals some changes, even though both statistics are insensitive to outliers. The trimean, with a higher focus on extreme values than the median, ranks the second-order Gray-Edge lower than the first-order Gray-Edge, while the median inverts this ranking. This difference is caused by a larger spread in performance of the second-order Gray-Edge; see Fig. 6. Even though the first-order Gray-Edge method has outliers with higher errors, the spread from the first to the third quantile is larger for the second-order Gray-Edge. This indicates that the errors of the first-order Grey-Edge are more condensed around the median, with the exception of a few outliers.

Table 2. Ranking of Methods Created by Using Color Constancy Framework of [4]

	Angular Error		Perceptual Euclidean Distance	
Method	Median	Trimean	Median	Trimean
$\mathbf{e}_{0,\infty,0}$	7	7	5	7
$\mathbf{e}_{0,\infty,1}$	5	6	6	5
$\mathbf{e}_{0,1,0}$	8	8	8	8
$\mathbf{e}_{0,9,0}$	6	5	7	6
$\mathbf{e}_{1,1,1}$	1	1	1	1
$\mathbf{e}_{1,1,2}$	3	2	3	2
$\mathbf{e}_{2,1,1}$	2	4	2	3
$\mathbf{e}_{2,1,2}$	4	3	4	4

Gamut mapping. Five gamut mapping methods are compared, two using pixel values [$\mathcal{G}^{\sigma=3}(\mathbf{f})$ and $\mathcal{G}^{\sigma=5}(\mathbf{f})$, differing only in the size of the filter that is used to smooth the image], and three using edges [$\mathcal{G}^{\sigma=1}(\nabla\mathbf{f})$, $\mathcal{G}^{\sigma=2}(\nabla\mathbf{f})$, and $\mathcal{G}^{\sigma=3}(\nabla\mathbf{f})$, again differing only in the size of the filter that is used to compute the edges]. Again, completely different ranking results are obtained when different summarizing statistics are used; see Table 3. For the median, the best-performing method is the edge-based gamut mapping with a filter size of $\sigma = 1$. However, when considering the trimean, it can be derived that perhaps it is better to use a filter size of $\sigma = 2$. An explanation for this shift can be found in Fig. 7, which shows that using a filter size of $\sigma = 2$ results in a distribution that is more densely sampled around the median, so this filter size is more appropriate for a larger set of images.

When comparing the angular error with the perceptual Euclidean distance, it is noticed that the differences are small but that the rankings are shifted in favor of pixel-based gamut mapping. For the perceptual Euclidean distance, the difference between the median and the trimean is minor, which is also reflected in the minor differences between the box plots shown in Fig. 7.

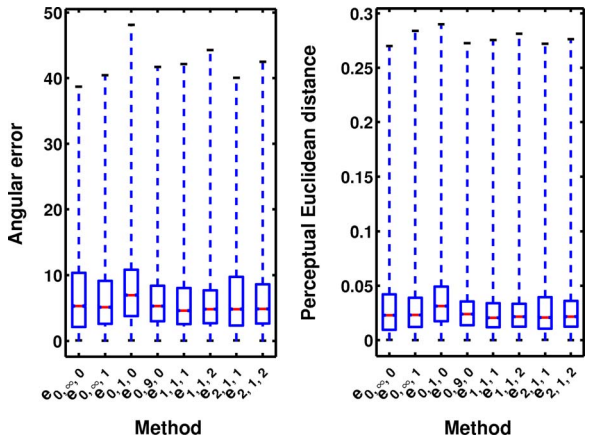

Fig. 6. (Color online) Box plots of the angular error and the perceptual Euclidean distance for several color constancy methods of the framework from [4].

Table 3. Ranking of Several Gamut Mapping Methods, from [15]

	Angular Error		Perceptual Euclidean Distance	
Method	Median	Trimean	Median	Trimean
$\mathcal{G}^{\sigma=3}(\mathbf{f})$	2	4	1	1
$\mathcal{G}^{\sigma=5}(\mathbf{f})$	5	5	3	2
$\mathcal{G}^{\sigma=1}(\nabla\mathbf{f})$	1	3	2	3
$\mathcal{G}^{\sigma=2}(\nabla\mathbf{f})$	3	1	4	4
$\mathcal{G}^{\sigma=3}(\nabla\mathbf{f})$	4	2	5	5

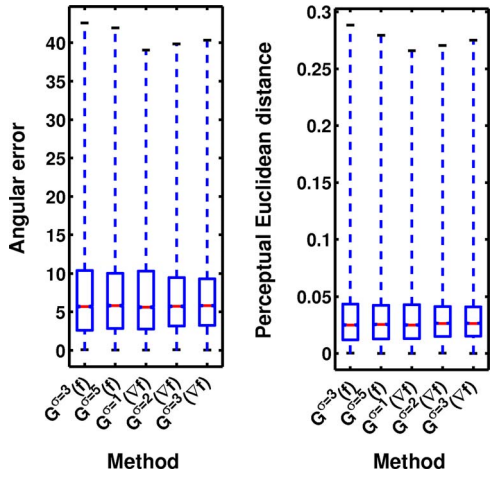

Fig. 7. (Color online) Box plots of the angular error and the perceptual Euclidean distance for several gamut-mapping methods taken from [15].

In conclusion, the tail of the distribution of estimated illuminant errors can play an important role in evaluating color constancy performance.

7. PERCEPTUAL SIGNIFICANCE

This section is devoted to the notion of the perceptual significance of the performance difference between two algorithms. The fact that the difference between two algorithms is statistically significant might not always justify the conclusion that one algorithm is better than the other. For instance, using the Wilcoxon sign test (or some other hypothesis test) to show that algorithm A performs significantly better than algorithm B merely shows that the error of algorithm A is often lower than the error of algorithm B. It does not show *how much* lower, nor does it tell if this difference is noticeable to a human observer.

Color constancy performance evaluation is often done with respect to a ground truth, i.e., computing the error for a number of methods on a large set of images. The methods are then compared by analyzing the summarizing statistic of the distribution of errors, sometimes accompanied by significance testing. However, significance testing is limited to hypothesis testing, with which the distributions of the errors are compared. Consequently, in the literature the differences between two methods have not been analyzed psychophysically yet. The question remains whether an observer would even notice the difference between the results of two color constancy methods.

A few attempts have been made to quantify the term "acceptable color reproduction." For instance, Funt *et al.* [43] stated that the root mean squared Euclidean error of the estimated chromaticity value should be 0.04 at most, for accurate color-based object recognition. In terms of angular error, a deviation of 1° with respect to the ground truth was found to be not noticeable, while an angular error of 3° was found noticeable but acceptable [44,45]. From an analysis, Hordley [46] derives that an angular error of 2° represents good enough color constancy for complex images. However, these values are all with respect to the ground truth; the perceptual difference between two algorithms is not discussed.

A. Just Noticeable Difference

In this section, the data that are obtained from the psychophysical experiments are used to obtain a measure for the notion of *just noticeable difference*. As was explained in Section 4, the observers had the possibility of indicating that the quality of two color constancy reproductions is the same (as good or as bad). These responses are used and analyzed here. When an observer indicates that the color reproductions are identical, this does not necessarily mean that the considered images are close enough to the ground truth. It means that the observer could not observe the difference between the result of algorithm A and the result of algorithm B. Hence, from these responses it can be extracted whether the difference between two algorithms is psychophysically significant. However, note that the observers were not explicitly instructed to indicate whether or not they could see the difference between two color reproductions.

Following Weber's law [47], it is to be expected that as the absolute error of two algorithms increases, the just noticeable difference between these two algorithms increases too. So, if two algorithms would have angular errors of 3° and 4°, then the difference between these two algorithms will most likely be apparent to most people. However, if these two algorithms would have errors of 15° and 16°, then it is likely that the difference between these two algorithms is less noticeable (if noticeable at all).

For the analysis of the hyperspectral data, the difference between two algorithms is defined as not noticeable if at least three of the eight observers agreed that the color reproductions are identical. From our data, this results in 36 comparisons corresponding to approximately 11% of all comparisons for one observer. Every comparison is characterized by the tuple $\langle \epsilon_{\max}, \epsilon_{\min}, \Delta\epsilon \rangle$, where $\epsilon_{\max}$ is the maximum error of the two methods, $\epsilon_{\min}$ is the minimum error, and $\Delta\epsilon = \epsilon_{\max} - \epsilon_{\min}$ is the difference between the two methods, called the relative error level. From the set of 36 comparisons, consider those comparisons with an absolute error level $\epsilon_{\max}$ between ϵ_i and ϵ_j. From these comparisons, the average of the relative error level $\overline{\Delta\epsilon}$ is computed. The results for the angular error and the perceptual Euclidean distance, together with linear regression lines, are shown in Fig. 8. As was expected from Weber's law, the just noticeable difference increases linearly with the absolute error level. For the angular error, the correlation coefficient is even as high as 0.9, with

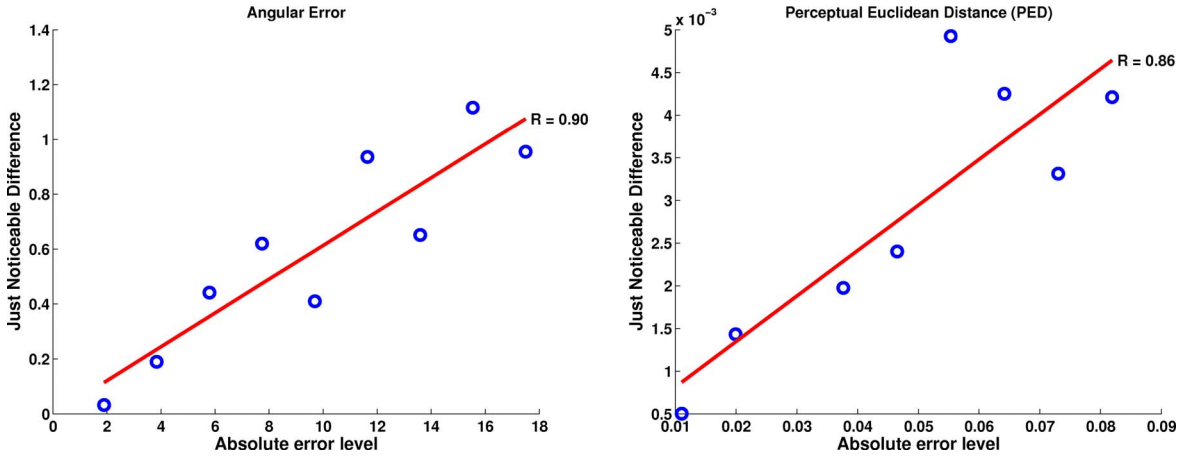

Fig. 8. (Color online) Indication of the just noticeable difference with respect to the absolute error level.

a p value of only 8.4×10^{-4}, which means that the correlation is considered to be highly significant.

Based on the analysis of the hyperspectral data, the difference in terms of angular error between two methods should be at least $0.06\times\epsilon_{\max}$ to be noticeable. For instance, if method A has an angular error of 10°, then an improvement of at least 0.6° is necessary; otherwise the improvement will be not visible to a human observer. In terms of perceptual Euclidean distance, the difference between two methods should be at least $0.05\times\epsilon_{\max}$ before it is noticeable.

B. Implications

The notion of the just noticeable difference can be used to indicate whether some proposed improvement is perceptually significant, i.e., whether or not a human observer is likely to see the difference between the original result and the result of the proposed improvement. To this end, some recently proposed methods are examined, based on the performance that is reported. The just noticeable difference can be computed by using the linear regression analysis that was discussed in Subsection 7.A. For the angular error, this results in the following formula to compute the just noticeable difference (JND) between methods A and B:

$$\mathrm{JND}_{\mathrm{angular}} = 0.06 \times \epsilon_{\max}, \tag{16}$$

where $\epsilon_{\max}=\max(\epsilon_A,\epsilon_B)$ is the maximum error of the two methods. Note that the fact that a proposed improvement is not perceptually significant does not justify the conclusion that the proposed method is without merit. Sometimes progress is made in little steps; so two or three small improvements eventually might result in the same increase in performance as one large improvement. The results are summarized in Table 4.

Low-level framework. The framework that is used in this paper to create the different output images for the psychophysical experiments is proposed by van de Weijer *et al.* [4]. In the original paper, several instantiations are evaluated on a subset of the RGB images data set [5] that is also used in this paper. From the experiments, it is concluded that the first-order Gray-Edge performs best with a median angular error of 4.1°. However, the performance of the second-order Gray-Edge is very similar (median angular error 4.3°). Consequently it can be concluded that the difference between the first-order and the second-order Gray-Edge is not perceptually significant, as the just noticeable difference is $0.06 \times 4.3° = 0.26°$.

Gamut-constrained illuminant estimation. The gamut mapping algorithm [8] is still one of the best-performing

Table 4. Relative Differences between the Best-Performing Algorithm and the Other Methods[a]

Method	Relative Difference
Low-level framework [4]	
Proposed 1st-order Gray-Edge	—
Proposed 2nd-order Gray-Edge	+4.7%
General Gray-World	+12.8%
Max-RGB	+38.8%
Gray-World	+43.8%
Gamut-constrained [10]	
Proposed GCIE	—
Gamut mapping	+11.0%
Max-RGB	+35.3%
Gray-World	+70.6%
High-level information [39]	
Indoor	
Proposed BU + TD	—
Proposed BU	+0%
Proposed TD	+5.3%
Best single algorithm	+13.1%
Worst single algorithm	+56.9%
Outdoor	
Proposed BU + TD	—
Proposed BU	+4.3%
Proposed TD	+4.3%
Best single algorithm	+8.2%
Worst single algorithm	+39.2%
Using indoor-outdoor classification	
Proposed CDA	—
Proposed CDP	+4.1%
1st-order Gray-Edge	+15.4%
2nd-order Gray-Edge	+18.7%
White-Patch	+31.0%
General Gray-World	+34.8%
Gray-World	+36.5%

[a]Performances are taken from the corresponding papers.

algorithms. An extension to this gamut mapping approach is proposed by Finlayson *et al.* [10] and effectively reduces the problem of illuminant estimation to illuminant classification. In its most general form, i.e., assuming as little *a priori* information as possible, the median angular error improves from 2.92° for the regular gamut mapping to 2.60°. Given the initial performance of 2.92°, the just noticeable difference is $0.06 \times 2.92° = 0.18°$, so it can be concluded that the obtained improvement is perceptually significant.

Using high-level visual information. The idea of illuminant classification is also present in the work of van de Weijer *et al.* [39], where semantic information is incorporated into the illuminant estimation process. Given a number of illuminant estimates, the most appropriate one is selected by using the visual information that is present in the input image. The initial set of estimates can be based on the result of various illuminant estimation algorithms. Alternatively, the visual information can be used to estimate a plausible illuminant by using a top-down approach. Experiments on both indoor and outdoor images show that the difference between the two alternative sets of illuminant hypotheses is small (i.e., perceptually not significant), but the combination of the two sets of illuminant hypotheses results in a perceptually significant improvement over the best-performing single algorithm. For indoor images, the angular error is reduced from 6.1° to 5.3°, while for outdoor images the error can be reduced from 4.9° to 4.5°. The just noticeable difference is $0.06 \times 6.1° = 0.4°$ for indoor images and $0.06 \times 4.9° = 0.3°$ for outdoor images.

Using indoor–outdoor classification. Finally, Bianco *et al.* [41] propose to apply different illuminant estimation algorithms to indoor and outdoor images. For indoor images they propose to use the Shades-of-Gray algorithm [3], while the second-order Gray-Edge [4] is proposed for outdoor images. Without classification, the median angular error is 4.18°, so the just noticeable difference is $0.06 \times 4.18° = 0.25°$. Adding the classification step can reduce the median error to 3.78°, so this improvement can be considered to be perceptually significant.

8. DISCUSSION

In this paper, a taxonomy of different distance (performance) measures for color constancy algorithms is presented. Correlation between the observed quality of the output images and the different distance measures for illuminant estimates has been analyzed. It has been investigated to what extent distance measures mimic differences in color naturalness of images as obtained by human obervations.

Based on experimental results for two data sets, it can be concluded that the correlation between the angular error and the perceptual quality of the output of color constancy algorithms is not perfect, but quite high nonetheless. This means that the angular error is a reasonably good indicator of the perceptual performance of color constancy algorithms. The same conclusion holds for the Euclidean distance, but the correlation of this measure can be increased by using the perceptual Euclidean distance, optimizing the weights for a specific data set. A significant improvement can be obtained with respect to the angular error and the unweighted Euclidean distance. Using this optimized weight combination may change the ranking of color constancy algorithms, resulting in different conclusions on the performance of these algorithms. Note that the optimal weight combination depends on the data set that is used, which means that a psychophysical experiment is needed that uses a small subset of the complete data set to obtain the optimal weights. However, using these optimal weights can yield a significantly higher correlation with human observers, which means that the results of color constancy algorithms can be interpreted more reliably with respect to the perceptual quality of the output.

In addition to the correlation between subjective (human observers) and objective performance measures, the just noticeable difference is analyzed in this paper. It is shown that, independent of the distance measure that is used, performance improvements up to 5%–6% are not noticeable to human observers. This finding is in line with the values for the Weber fraction typically found in visual perception (e.g., [48]). Note that this implies that the summarizing statistic that is used to indicate the performance of a color constancy algorithm for a set of images is representative for the whole set. Previously, it was proposed that the median is more suited than the mean [7]. While this conclusion is not challenged here, it is suggested that other summarizing statistics can be used as well. For instance, the trimean is robust to outliers, like the median, but still has attention to the extreme values in the distribution [36,37]. Using the trimean instead of the median reveals small variations in the ranking of color constancy algorithms, indicating that some color constancy algorithms (e.g., second-order Gray-Edge) have a wider distribution of illuminant estimate errors than others (e.g., first-order Gray-Edge).

REFERENCES

1. E. H. Land, "The retinex theory of color vision," Sci. Am. **237**, 108–128 (1977).
2. G. Buchsbaum, "A spatial processor model for object colour perception," J. Franklin Inst. **310**, 1–26 (1980).
3. G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Twelfth Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications* (Society for Imaging Science and Technology, 2004), pp. 37–41.
4. J. van de Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," IEEE Trans. Image Process. **16**, 2207–2214 (2007).
5. F. Ciurea and B. V. Funt, "A large image database for color constancy research," in *Eleventh Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications* (Society for Imaging Science and Technology, 2003), pp. 160–164.
6. K. Barnard, L. Martin, B. V. Funt, and A. Coath, "A data set for color research," Color Res. Appl. **27**, 147–151 (2002).
7. S. D. Hordley and G. D. Finlayson, "Reevaluation of color constancy algorithm performance," J. Opt. Soc. Am. A **23**, 1008–1020 (2006).
8. D. A. Forsyth, "A novel algorithm for color constancy," Int. J. Comput. Vis. **5**, 5–36 (1990).
9. G. D. Finlayson, S. D. Hordley, and P. M. Hubel, "Color by correlation: a simple, unifying framework for color

constancy," IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1209–1221 (2001).
10. G. D. Finlayson, S. D. Hordley, and I. Tastl, "Gamut constrained illuminant estimation," Int. J. Comput. Vis. **67**, 93–109 (2006).
11. D. H. Brainard and W. T. Freeman, "Bayesian color constancy," J. Opt. Soc. Am. A **14**, 1393–1411 (1997).
12. M. D'Zmura, G. Iverson, and B. Singer, "Probabilistic color constancy," in *Geometric Representations of Perceptual Phenomena* (Lawrence Erlbaum, 1995), pp. 187–202.
13. P. V. Gehler, C. Rother, A. Blake, T. P. Minka, and T. Sharp, "Bayesian color constancy revisited," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2008), pp. 1–8.
14. M. Ebner, "Evolving color constancy," Pattern Recogn. Lett. **27**, 1220–1229 (2006).
15. A. Gijsenij, T. Gevers, and J. van de Weijer, "Generalized gamut mapping using image derivative structures for color constancy," Int. J. Comput. Vis. (to be published), http://www.springerlink.com/content/q598825t7654648n/?p=155b2db7234942feaacfcf6d88a50b2c&pi=0. (September 2009).
16. Color constancy demonstration (Mathematica), http://cat.cvc.uab.es/~joost/code/ColorConstancy.zip.
17. J. von Kries, "Die gesichtsempfindungen," in *Handbuch der Physiologie des Menschen* (1904), Vol. 3, pp. 109–282.
18. G. West and M. H. Brill, "Necessary and sufficient conditions for von Kries chromatic adaptation to give color constancy," J. Math. Biol. **15**, 249–258 (1982).
19. G. D. Finlayson, M. S. Drew, and B. V. Funt, "Color constancy: generalized diagonal transforms suffice," J. Opt. Soc. Am. A **11**, 3011–3019 (1994).
20. B. V. Funt and B. C. Lewis, "Diagonal versus affine transformations for color correction," J. Opt. Soc. Am. A **17**, 2108–2112 (2000).
21. Commission Internationale de L'Eclairage (CIE), "Colorimetry," CIE Publ. no. 15.2, 2nd ed. (CIE, 1986).
22. Commission Internationale de L'Eclairage (CIE), "Improvement to industrial colour-difference evaluation," CIE Publ. no. 142-2001 (CIE, 2001).
23. M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, "A standard default color space for the Internet—sRGB," version 1.10 (1996) www.w3.org/Graphics/Color/sRGB.html.
24. G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae* (Wiley, 2000).
25. J. Slater, *Modern Television Systems to HDTV and Beyond* (Taylor & Francis, 2004).
26. L. E. Arend, A. Reeves, J. Schirillo, and R. Goldstein, "Simultaneous color constancy: papers with diverse Munsell values," J. Opt. Soc. Am. A **8**, 661–672 (1991).
27. E. Brunswik, "Zur Entwicklung der Albedowahrnehmung," Z. Psychol. **109**, 40–115 (1928).
28. P. B. Delahunt and D. H. Brainard, "Does human color constancy incorporate the statistical regularity of natural daylight?" J. Vision **4**, 57–81 (2004).
29. D. H. Foster, S. M. C. Nascimento, and K. Amano, "Information limits on neural identification of colored surfaces in natural scenes," Visual Neurosci. **21**, 331–336 (2004).
30. J. E. Bailey, M. Neitz, D. Tait, and J. Neitz, "Evaluation of an updated hrr color vision test," Visual Neurosci. **22**, 431–436 (2004).
31. H. A. David, "Ranking from unbalanced paired-comparison data," Biometrika **74**, 432–436 (1987).
32. R. L. Alfvin and M. D. Fairchild, "Observer variability in metameric color matches using color reproduction media," Color Res. Appl. **22**, 174–188 (1997).
33. E. Kirchner, G. J. van den Kieboom, L. Njo, R. Supèr, and R. Gottenbos, "Observation of visual texture of metallic and pearlescent materials," Color Res. Appl. **32**, 256–266 (2007).
34. Bruce Lindbloom's web site, http://www.brucelindbloom.com.
35. R. V. Hogg and E. A. Tanis, *Probability and Statistical Inference* (Prentice Hall, 2001).
36. J. W. Tukey, *Exploratory Data Analysis* (Addison-Wesley, 1977).
37. H. F. Weisberg, *Central Tendency and Variability* (Sage Publications, 1992).
38. A. Gijsenij and T. Gevers, "Color constancy using natural image statistics," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2007), pp. 1–8.
39. J. van de Weijer, C. Schmid, and J. J. Verbeek, "Using high-level visual information for color constancy," in *IEEE International Conference on Computer Vision* (IEEE, 2007), pp. 1–8.
40. S. Bianco, F. Gasparini, and R. Schettini, "Consensus-based framework for illuminant chromaticity estimation," J. Electron. Imaging **17**, 023013 (2008).
41. S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Improving color constancy using indoor-outdoor image classification," IEEE Trans. Image Process. **17**, 2381–2392 (2008).
42. A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color constancy beyond bags of pixels," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2008), pp. 1–8.
43. B. V. Funt, K. Barnard, and L. Martin, "Is machine colour constancy good enough?" in *Computer Vision—ECCV'98: 5th European Conference on Computer Vision* (Springer, 1998), pp. 445–459.
44. G. D. Finlayson, S. D. Hordley, and P. Morovic, "Colour constancy using the chromagenic constraint," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2005), pp. 1079–1086.
45. C. Fredembach and G. D. Finlayson, "The bright-chromagenic algorithm for illuminant estimation," J. Imaging Sci. Technol. **52**, 040906 (2008).
46. S. D. Hordley, "Scene illuminant estimation: past, present, and future," Color Res. Appl. **31**, 303–314 (2006).
47. E. H. Weber, "Der Tastinn und das Gemeingfühl," in *Handwörterbüch der Physiologie* (1846), Vol. 3, pp. 481–588.
48. T. N. Cornsweet, *Visual Perception* (Academic, 1970).