A Perceptual Comparison of Distance Measures for Color Constancy Algorithms

Arjan Gijsenij, Theo Gevers, and Marcel P. Lucassen

Intelligent Systems Laboratory Amsterdam
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

Abstract. Color constancy is the ability to measure image features independent of the color of the scene illuminant and is an important topic in color and computer vision. As many color constancy algorithms exist, different distance measures are used to compute their accuracy. In general, these distances measures are based on mathematical principles such as the angular error and Euclidean distance. However, it is unknown to what extent these distance measures correlate to human vision.

Therefore, in this paper, a taxonomy of different distance measures for color constancy algorithms is presented. The main goal is to analyze the correlation between the observed quality of the output images and the different distance measures for illuminant estimates. The output images are the resulting color corrected images using the illuminant estimates of the color constancy algorithms, and the quality of these images is determined by human observers. Distance measures are analyzed how they mimic differences in color naturalness of images as obtained by humans.

Based on the theoretical and experimental results on spectral and real-world data sets, it can be concluded that the perceptual Euclidean distance (PED) with weight-coefficients ($w_R = 0.26$, $w_G = 0.70$, $w_B = 0.04$) finds its roots in human vision and correlates significantly higher than all other distance measures including the angular error and Euclidean distance.

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. [1,2,3,4]. For benchmarking, the accuracy of color constancy algorithms is evaluated by computing a distance measure on the same data sets such as [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

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part I, LNCS 5302, pp. 208–221, 2008.
© Springer-Verlag Berlin Heidelberg 2008

distance and the angular error, of which the latter is probably more widely used than the first. In [7], an analysis is presented of the distribution of these measures, with the aim to find the best summarizing statistic over a large set of images. 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 humans, color corrected images will be 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.

The paper is organized as follows. In section 2, color constancy and image transformation is discussed. Further, a set of color constancy methods will be introduced. Then, the different distance measures will be 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. Thirdly, 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 given in section 5.

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:

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

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:

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

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

2.1 Color Constancy Algorithms

For the purpose of the experiments in this paper, the focus is on a number of simple algorithms. Recently, van de Weijer et al. [1] proposed a framework with which systematically many different algorithms can be constructed. Possible algorithms include methods using 0^{th}-order statistics (i.e. pixel values), like the White-Patch [2], the Grey-World [3] and the Shades-of-Grey algorithms [4], and methods using higher-order (e.g. 1^{st}- and 2^{nd}-order) statistics, like the Grey-Edge and 2^{nd}-order Grey-Edge algorithms. The framework is given by:

$$\left(\int \left| \frac{\partial^n \mathbf{f}_\sigma(\mathbf{x})}{\partial \mathbf{x}^n} \right|^p d\mathbf{x} \right)^{\frac{1}{p}} = k\mathbf{e}^{n,p,\sigma}, \tag{3}$$

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. For the purpose of this article, five instantiations are used, representing a wide variety of algorithms, being the White-Patch ($\mathbf{e}^{0,\infty,0}$), the Grey-World ($\mathbf{e}^{0,1,0}$), the General Grey-World ($\mathbf{e}^{0,13,2}$), the 1^{st}-order Grey-Edge ($\mathbf{e}^{1,1,6}$) and the 2^{nd}-order Grey-Edge algorithm ($\mathbf{e}^{2,1,5}$). Of course, many other algorithms can be generated, but for simplicity, the focus is on these five instantiations as they are derived from different orders of image statistics.

2.2 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* [8]. The diagonal mapping is given as follows:

$$\mathbf{f}^c = \mathcal{D}^{u,c}\mathbf{f}^u \Rightarrow \begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix}, \tag{4}$$

where $\mathbf{f}^u$ is the image taken under an unknown light source, $\mathbf{f}^c$ is the same image transformed, so it appears if it was taken under the reference light, and $\mathcal{D}^{u,c}$ is a diagonal matrix which 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 only recover the color of the light source 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:

$$r = \frac{R}{R+G+B}\ , \quad g = \frac{G}{R+G+B}\ , \quad b = \frac{B}{R+G+B}. \tag{5}$$

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 than the first. The Euclidean distance between the estimated light source $\mathbf{e}_e$ and the true, ground truth, light sources $\mathbf{e}_u$ is given by:

$$\mathcal{L}_2(\mathbf{e}_e, \mathbf{e}_u) = \sqrt{(R_e - R_u)^2 + (G_e - G_u)^2 + (B_e - B_u)^2}. \tag{6}$$

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

$$d_{\text{angle}}(\mathbf{e}_e, \mathbf{e}_u) = \cos^{-1}\left(\frac{\mathbf{e}_e \cdot \mathbf{e}_u}{||\mathbf{e}_e|| \cdot ||\mathbf{e}_u||}\right), \tag{7}$$

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 values correspond to human vision. 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 (section 3.1), to perceptual measures (section 3.2) and color constancy specific measures (section 3.3).

3.1 Minkowski Distance

A well-known measure is the Minkowski distance:

$$\mathcal{L}_p(\mathbf{e}_e, \mathbf{e}_u) = (|R_e - R_u|^p + |G_e - G_u|^p + |B_e - B_u|^p)^{\frac{1}{p}}, \tag{8}$$

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 ($\mathcal{L}_1$), the Euclidean distance ($\mathcal{L}_2$) and the Chebychev distance ($\mathcal{L}_\infty$).

3.2 Perceptual Distances

The goal of color constancy algorithms, in this paper, is to obtain perceptual distance to a reference image. 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 perceptually uniform color spaces $L^*a^*b^*$ and $L^*u^*v^*$ [9], as well as in the more intuitive color channels chroma C and hue H.

Most color constancy algorithms are restricted to estimating the chromaticity values of the illuminant. To evaluate the performance of an estimated light source in different color spaces, this (intensity normalized) estimate, as well as the ground truth light source, is applied to a perfect white reflectance. Hence, two (R, G, B)-values are obtained, representing the color of a white reflectance under the estimated and the true light source. These (R, G, B)-values can consequently be converted to different color spaces. Conversion from RGB to XYZ is done using the following linear transformation:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.4125\ 0.3576\ 0.1804 \\ 0.2127\ 0.7152\ 0.0722 \\ 0.0193\ 0.1192\ 0.9502 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{9}$$

Then, the two perceptual color models $L^*a^*b^*$ and $L^*u^*v^*$ are computed using $(X_w, Y_w, Z_w) = (0.9505, 1.0000, 1.0888)$ as reference white [9]. From these perceptual color spaces, different color channels can be computed, like chroma C and hue H. The transformation from $L^*a^*b^*$ to C and H is given by:

$$C_{ab} = \sqrt{(a^*)^2 + (b^*)^2}\ , \quad H_{ab} = \tan^{-1}\left(\frac{b^*}{a^*}\right), \tag{10}$$

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

Finally, it is known that the human eye is more sensitive to some colors than to others. This important property of the human visual system is used, for instance, in the conversion of RGB-images to luminance-images [10]:

$$Lum = 0.3R + 0.59G + 0.11B. \tag{11}$$

Hence, a change in the green-channel has a stronger effect on the perceived difference between two images than a change in the blue-channel, for instance. This leads us to 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 $w_R + w_G + w_B = 1$.

3.3 Color Constancy Distances

In this section, two color constancy specific distances are discussed. The first is the color constancy index CCI [11], also called Brunswik ratio [12], and is generally used to measure perceptual color constancy [13,14]. It is defined as the ratio of the amount of adaptation that is obtained by a human observer versus no adaptation at all:

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

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.

The second is a new measure, called the *gamut intersection*, that makes use of the gamuts of the colors that can occur under a given light source. It measures 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:

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

where $\mathcal{G}_i$ is the gamut of all possible colors under illuminant i and $\text{vol}(\mathcal{G}_i)$ 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 real-world scenes. The images are shown on a calibrated monitor, and observers are shown images in a round-robin schedule. For every pair of images, the observers have to specify which of the two results is closer to the ideal result. In this way, comparison of the distance measures (objective performance) is compared with visual judgment (subjective performance) by computing the correlation between the two performance measures.

4.1 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 real-world images and is used to analyze the results of the first experiments.

Hyperspectral data. The first data set, originating from [14] consists of eight hyperspectral images, of which four are shown in figure 1(a)-(d). These images were 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 [13], one neutral illuminant (CIE $D65$) and four chromatic illuminants (Red, Green, Yellow, Blue) were selected to create images under different light sources. The spectral power distributions of the selected illuminants are shown in figure 2(a) and were created with the use of the CIE daylight basis function, as described in [9]. In figure 2(b), images of scene 3 rendered under these four illuminants are shown.

Real-world data. The second data set consisted of real-world images and were a subset of 50, both indoor and outdoor, images taken from a data set that is widely

Fig. 1. Four examples of the hyperspectral scenes used in this study are shown in figures (a)-(d), rendered under the neutral $D65$ illuminant. In figures (e)-(h), four examples of the real-world scenes are shown.

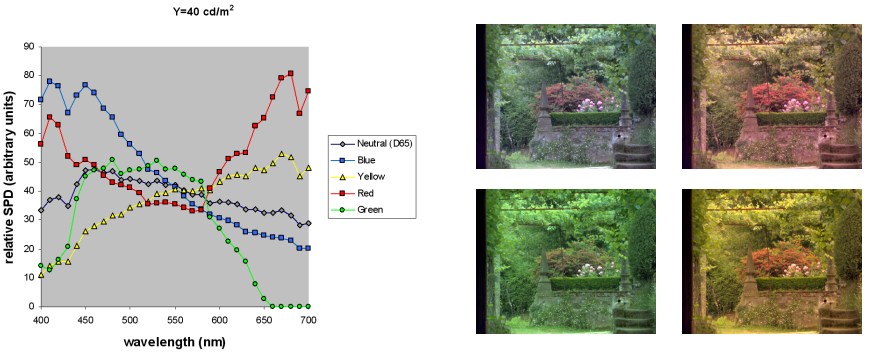
used for performance evaluation of color constancy methods [5]. The original data set consists of over $11,000$ images, and for all images, the ground truth of the color of the light source is known from a grey sphere that was mounted on top of the camera. This grey-sphere was cropped during the experiments. Some example images are shown in figure 1(e)-(h). Images from this data set are not as well calibrated as the previous set, and are therefore mostly used to confirm the results on the hyperspectral data.

4.2 Monitor

Images were viewed on a high-resolution (1600×1200 pixels, 0.27 mm dot pitch) calibrated LCD monitor, an Eizo ColorEdge CG211. The monitor was calibrated to a $D65$ 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 [15], which was used for rendering the spectral scenes under our illuminants. Spatial uniformity of the display, measured relative to the center of the monitor, was $\Delta\ E_{\mathrm{ab}} < 1.5$ according to the manufacturer's calibration certificates.

4.3 Observers

All observers that participated in the experiments had normal color vision and normal or corrected to normal visual acuity. Subjects were screened for color vision deficiencies with the HRR pseudo-isochromatic plates (4^{th} edition), allowing color vision testing along both the red-green and yellow-blue axes of color space [16]. After taking the color vision test, our subjects first adapted for about five



(a) Illuminant spectra (b) Illuminants applied to scene 3

Fig. 2. Relative spectral power distribution of the illuminants used in the experiments. The illuminants were 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.

minutes to the light level in a dim room that only received some daylight from a window that was covered with sunscreens (both inside and outside). In the meantime they were made familiar with the experimental procedure.

4.4 Experimental Procedure

The experimental procedure consists of a sequence of image comparisons. The subjects were shown 4 images at once, arranged in a square layout. The images were shown on a gray background having $L^* = 50$ and $a^* = b^* = 0$. 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 were 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 were to be addressed. Subjects then indicated (by pressing a key on the computer's keyboard) which of the two lower images had the best color reproduction. If the color reproduction of the two test images were identical (as good or as bad), the subjects had the possibility of indicating this. Subjects were told that response time would be measured, but that they were not under time pressure, they could use as much time as they needed 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 [17]. We applied a scoring mechanism

in which the color constancy algorithm underlying a win was awarded with 1 point and the algorithm underlying a loss with no points. In case of a tie, the competing algorithms both received 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 inter-observer variability will be 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 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 for the human observers, this time using the error values are criterion if one result is better than another.

5.1 Hyperspectral Data

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

Eight observers participated in this experiment, 4 men and 4 women, with ages ranging from 24 to 43 (an average of 34.6). At a viewing distance of about 60 cm, each of the four images subtended a visual angle of $16.6^\circ \times 12.7^\circ$. Horizontal and vertical separation between images was 2.1° and 0.9°, respectively.

Inter-observer variability. As a measure of the inter-observer variability, the individual differences from the mean observer scores are computed, a procedure that is often used in studies involving visual judgements, e.g. [18,19]. For each observer, the correlation coefficient of his/her average algorithm scores (averaged over scenes and illuminants) with the algorithm scores of the average observer is computed. The correlation coefficients so obtained varied from 0.974 to 0.999, with an average of 0.990. Correlation coefficients between scores of the individual observers ranged from 0.937 to 0.997. The significance of this result becomes clear when comparing these high values with the values that are 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 individual "observers" range from 0.074 to 0.948, with an average of 0.396. Correlation coefficients between individual observers in this case ranged 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 vs. 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(a), where the correlation coefficients on the spectral data set for all distance measures are summarized. Also shown in this table are the results 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. Comparing every distance measure to 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.

By zooming in on individual images, it can be seen that 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 observing the results of the images with such a low correlation, 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 the 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 ($\mathcal{L}_1$) and the Chebychev distance ($\mathcal{L}_\infty$), 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 vs. 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 can be compared 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 computing the correlation of the human observers with the Euclidean distance in

different color spaces, 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 like hue and chroma, the correlation is remarkably low; considering both chroma and hue, correlation is 0.646, which is significantly lower than the correlation of other color spaces. Considering chroma or hue alone, correlation drops even further to $\rho = 0.619$ and $\rho = 0.541$, respectively. In conclusion, using perceptual uniform spaces provide similar or lower correlation then rgb.

As was 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 for different color channels related to human vision (e.g. for Lum the coefficients are $R = 0.3$, $G = 0.59$ and $B = 0.11$). The question remains, however, which weights to use. For this purpose, an exhaustive search has been performed to find the optimal weighting scheme, denoted by $\text{PED}_{\text{hyperspectral}}$ in table 1(a). The weight-combination $(w_R, w_G, w_B) = (0.20, 0.79, 0.01)$ results in the highest correlation ($\rho = 0.963$), but differences with similar weighting combinations are very small such as Luminance $Lum = 0.3R + 0.59G + 0.11B$ which corresponds to the sensitivity of the human visual system. In conclusion, as the human eye is sensitive according to the well-known Lum sensitivity curve, incorporating this property yields a perceptual sound distance measure providing the highest correlation in the experiments on the spectral data.

Color constancy measures vs. 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 can be derived that the highest correlation with the human observers is obtained when measuring the color constancy index with $L^*a^*b^*$ ($\rho = 0.905$). However, differences between other distance measures are small. In conclusion, 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 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).

Table 1. An overview of the correlation coefficients ρ of several distance measures and using several color spaces, with respect to the human observers. Significance is shown using a Student's t-test (at the 95% confidence level). By comparing every distance measure with all others, a score is generated representing the number of times the null hypothesis (i.e. two distances measures have a similar mean correlation coefficient) is rejected. The results of the experiments on the hyperspectral data are shown in table (a), the results on the real-world data are shown in table (b).

(a) Hyperspectral data

Measure	ρ	T-test (#)
d_{angle}	0.895	3
$\mathcal{L}_1$	0.893	3
$\mathcal{L}_2$	0.890	3
$\mathcal{L}_\infty$	0.817	3
$\mathcal{L}_2 - L^*a^*b^*$	0.902	4
$\mathcal{L}_2 - L^*u^*v^*$	0.872	3
$\mathcal{L}_2 - C + H$	0.646	0
$\mathcal{L}_2 - C$	0.619	0
$\mathcal{L}_2 - H$	0.541	0
$\text{PED}_{\text{hyperspectral}}$	0.963	13
$\mathbf{PED_{proposed}}$	**0.960**	13
$\text{CCI}(d_{\text{angle}})$	0.895	3
$\text{CCI}(\mathcal{L}_{2,RGB})$	0.893	3
$\text{CCI}(\mathcal{L}_{2,L^*a^*b^*})$	0.905	4
$\text{CCI}(\mathcal{L}_{2,L^*u^*v^*})$	0.880	3
d_{gamut}	0.965	13

(b) Real-world data

Measure	ρ	T-test (#)
d_{angle}	0.926	3
$\mathcal{L}_1$	0.930	3
$\mathcal{L}_2$	0.928	3
$\mathcal{L}_\infty$	0.906	3
$\mathcal{L}_2 - L^*a^*b^*$	0.927	3
$\mathcal{L}_2 - L^*u^*v^*$	0.925	3
$\mathcal{L}_2 - C + H$	0.593	1
$\mathcal{L}_2 - C$	0.562	1
$\mathcal{L}_2 - H$	0.348	0
$\text{PED}_{\text{real-world}}$	0.961	14
$\mathbf{PED_{proposed}}$	**0.957**	14
$\text{CCI}(d_{\text{angle}})$	0.931	3
$\text{CCI}(\mathcal{L}_{2,RGB})$	0.929	3
$\text{CCI}(\mathcal{L}_{2,L^*a^*b^*})$	0.921	3
$\text{CCI}(\mathcal{L}_{2,L^*u^*v^*})$	0.927	3
d_{gamut}	0.908	3

Discussion. From table 1(a), 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 like $L^*a^*b^*$ and $L^*u^*v^*$ does not increase the correlation with human observers. Using chroma C and hue H significantly decrease the correlation with the human observers. The gamut intersection distance and the perceptual Euclidean distance (PED) 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 spectral data set.

5.2 Real-World Data

The experiments on the real-world data were run in three sessions, with the number of images equally divided in three parts. The sequence of the sets was randomized for every observer. In this experiment, seven observers participated (4 men and 3 women), with ages ranging from 24 to 43. The difference between the observers was analyzed similarly to the experiments on the hyper-spectral data, and again the agreement of the individual observers was found to be sufficiently high.

Objective vs. subjective scores. In general, the same trends on this data set as on the hyperspectral data are observed, see table 1(b). In general, 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 (PED) 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 $\text{PED}_{\text{real-world}}$ in table 1(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 on this data set.

Discussion. The results of the experiments on the real-world data set, see table 1(b), 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 used algorithms, but this was not investigated in this paper.

The most noticeable difference between the results on this data set and the results on the previous data set is the correlation of the gamut intersection distance. This distance has the highest correlation with the human observers on the hyperspectral data. However, on the real-world data set, the correlation is considerably lower, though not significant, than the other measures. The correlation of the perceptual Euclidean distance on the real-world data 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 real-world data are averaged. The optimal correlation is found for the weight-combination $(w_R, w_G, w_B) = (0.26, 0.7, 0.04)$, which is the weight-combination we propose to use to compute the PED. Using these weights, correlation of the perceptual Euclidean distance with human observers on the hyperspectral data is 0.960, and on the real-world data is 0.957, denoted PED_{proposed} in table 1(a) and (b), both still significantly higher (at the 95% confidence level) than all other distance measures.

6 Conclusion

In this paper, a taxonomy of different distance measures for color constancy algorithms has been presented. Correlation has been analyzed between the observed quality of the output images and the different distance measures for illuminant estimates. Distance measures have been investigated to what extent they mimic differences in color naturalness of images as obtained by humans.

Based on the theoretical and experimental results on spectral and real-world data sets, it can be concluded that the perceptual Euclidean distance (PED) with weight-coefficients ($w_R = 0.26$, $w_G = 0.70$, $w_B = 0.04$) finds its roots in human vision and correlates significantly higher than all other distance measures including the angular error and Euclidean distance.

References

1. van de Weijer, J., Gevers, T., Gijsenij, A.: Edge-based color constancy. IEEE Transactions on Image Processing 16(9), 2207–2214 (2007)
2. Land, E.: The retinex theory of color vision. Scientific American 237(6), 108–128 (1977)
3. Buchsbaum, G.: A spatial processor model for object colour perception. J. Franklin Institute 310(1), 1–26 (1980)
4. Finlayson, G., Trezzi, E.: Shades of gray and colour constancy. In: Proc. CIC, pp. 37–41 (2004)
5. Ciurea, F., Funt, B.: A large image database for color constancy research. In: Proc. CIC, pp. 160–164 (2003)
6. Barnard, K., Martin, L., Funt, B., Coath, A.: A data set for color research. Color Research and Application 27(3), 147–151 (2002)
7. Hordley, S., Finlayson, G.: Reevaluation of color constancy algorithm performance. J. Opt. Soc. America A 23(5), 1008–1020 (2006)
8. von Kries, J.: Influence of adaptation on the effects produced by luminous stimuli. In: MacAdam, D. (ed.) Sources of Color Vision, pp. 109–119. MIT Press, Cambridge (1970)
9. Wyszecki, G., Stiles, W.: Color science: concepts and methods, quantitative data and formulae. John Wiley & sons, Chichester (2000)
10. Slater, J.: Modern television systems to HDTV and beyond. Taylor & Francis Group, Abington (2004)
11. Arend, L., Reeves, A., Schirillo, J., Goldstein, R.: Simultaneous color constancy: papers with diverse munsell values. J. Opt. Soc. America A 8(4), 661–672 (1991)
12. Brunswik, E.: Zur entwicklung der albedowahrnehmung. Zeitschrift fur Psychologie 109, 40–115 (1928)
13. Delahunt, P., Brainard, D.: Does human color constancy incorporate the statistical regularity of natural daylight? Journal of Vision 4(2), 57–81 (2004)
14. Foster, D., Nascimento, S., Amano, K.: Information limits on neural identification of colored surfaces in natural scenes. Visual Neuroscience 21, 331–336 (2004)
15. Stokes, M., Anderson, M., Chandrasekar, S., Motta, R.: A standard default color space for the internet-srgb (1996), `www.w3.org/Graphics/Color/sRGB.html`
16. Bailey, J., Neitz, M., Tait, D., Neitz, J.: Evaluation of an updated hrr color vision test. Visual Neuroscience 22, 431–436 (2004)
17. David, H.: Ranking from unbalanced paired-comparison data. Biometrika 74, 432–436 (1987)
18. Alfvin, R., Fairchild, M.: Observer variability in metameric color matches using color reproduction media. Color Research and Application 22, 174–188 (1997)
19. Kirchner, E., van den Kieboom, G., Njo, L., Supr, R., Gottenbos, R.: Observation of visual texture of metallic and pearlescent materials. Color Research and Application 32, 256–266 (2007)