Comparing objective and subjective error measures for color constancy

Marcel P. Lucassen, Arjan Gijsenij, Theo Gevers; Intelligent System Laboratory Amsterdam, University of Amsterdam; Amsterdam, The Netherlands

Abstract

We compare an objective and a subjective performance measure for color constancy algorithms. Eight hyper-spectral images were rendered under a neutral reference illuminant and four chromatic illuminants (Red, Green, Yellow, Blue). The scenes rendered under the chromatic illuminants were color corrected by 5 color constancy algorithms that are based on zero-, first- and second-order image statistics. The angular error is used as the objective performance measure for color constancy. It estimates the chromatic mismatch between the true and estimated illuminant vector in RGB space. A subjective performance measure was derived from a psychophysical experiment involving paired comparisons of the color corrected images shown on a calibrated monitor. Eight subjects indicated their preference with respect to color reproduction when comparing the two images (i.e. color constancy algorithms) against the reference image (the same scene under neutral illumination). Our results indicate a large negative correlation (-0.9 on average) between the objective and subjective color constancy measures. The data suggests the possibility for further improvement of the correlation between the two types of performance measures.

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 spectral composition of the illuminant. A key issue is how to disentangle the product of illumination and object reflection that is sampled by the visual system. In computer vision the usual way to approach the color constancy problem is by estimating the illuminant, so that reflectance can be recovered. Such color constancy algorithms may serve to correct the color balance of images for display or to support object recognition [1]. For objective evaluation of the effectiveness of color constancy algorithms the *angular error* is widely used [2]. The angular error ε is defined as the angular distance between the algorithm's estimate of the light source (e_e) and the true illuminant vector (e_l) in normalized RGB space:

$$\varepsilon = \cos^{-1}\left(\hat{e}_l \cdot \hat{e}_e\right) \tag{1}$$

Although the value of the angular error indicates how closely an original illuminant vector is approximated by the estimated one (after intensity normalization), it does not predict the color reproduction accuracy or color naturalness of color constancy algorithms. For that it is necessary to compare the color corrected images with the original images under reference illumination, which might be done by a computer algorithm or by visual inspection. Here we focus on the latter.

Recently, a framework of color constancy methods is proposed by van de Weijer et al. [3]. By varying one or more of the three framework parameters a set of color constancy methods can be generated. It incorporates some well-known algorithms based on zeroth-order statistics (i.e. pixel values) like the White-Patch algorithm [4], the Grey-World algorithm [5], but also new methods based on higher-order (e.g. first- and second-order) statistics. For the purpose of this study, we use five instantiations of the framework that represent a variety of algorithms: WP (White-Patch), GW (Grey-World), GGW (General Grey World), 1-GE (First-order Grey-Edge) and 2-GE (Second-order Grey-Edge). In this paper, we compare the angular errors of these five algorithms with visual judgments on the color fidelity and show that they are clearly correlated.

Methods

Eight hyperspectral images were selected that originate from Foster et al. [6]. We preferred these spectral images to normal RGB images because in this study we wanted to simulate a natural and colorimetrically correct interaction between illuminants and objects. Following Delahunt & Brainard [7] we selected one neutral reference illuminant (daylight CIE D65) and four chromatic test illuminants (Red, Green, Yellow, and Blue). The spectral power distributions of our illuminants were created with the CIE daylight basis functions [8] and were intensity scaled to ensure acceptable image quality when shown on an sRGB display. The four chromatic illuminants are perceptually equidistant from the neutral illuminant, at 28 ΔE_{ab} units. Figure 1 shows an example of the illumination effects.

Figure 1. *Scene 7 (one of our 8 test scenes) rendered under the different illuminants. The central image is under reference illuminant D65. Starting at the top position, in clockwise rotation the images correspond to the Yellow, Red, Blue and Green illuminant, respectively. The four chromatic illuminants are perceptually equidistant from D65, at 28 ΔE_{ab} units. Prior to application of the spectral illuminants the calibration objects in the original images were removed, and resulting image gaps were filled in with the "inpainting" method of Criminisi et al. [9].*

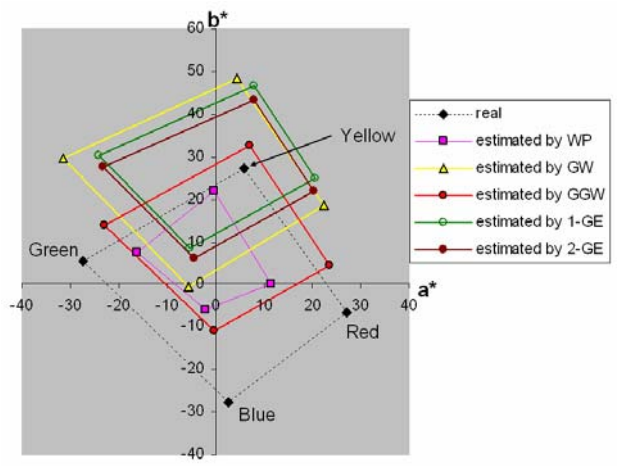

Figure 2. *Positions of the test illuminants Red, Yellow, Green, Blue and estimated counterparts in CIE a*,b* space. The neutral D65 reference illuminant is located at the origin. Dotted lines connect the a*b* coordinates of the real illuminants whereas solid lines connect those estimated by the color constancy algorithms. Illuminant coordinates shown here were obtained by averaging over the 8 test scenes.*

After "application" of the illuminants, the 5 different color constancy algorithms previously introduced were used to correct the color balance of the images. That is, they first estimate the illuminant from the scene and then correct the color balance using the von Kries diagonal transform [10], leaving the intensity of the images unchanged. The algorithms only differ in the way they estimate the illuminant from the scene. Here we do not further discuss the algorithm details, more information is presented in [3]. In Figure 2 we plot the a*b* coordinates of the (real) chromatic test illuminants Red, Green, Yellow and Blue and the estimated ones resulting from the five algorithms. Note that the estimates are shifted towards positive b* values, i.e. towards yellow. This is explained by the fact that the average color balance of the scenes under D65 is already shifted somewhat towards yellow, as shown in Figure 3 for test scene 7.

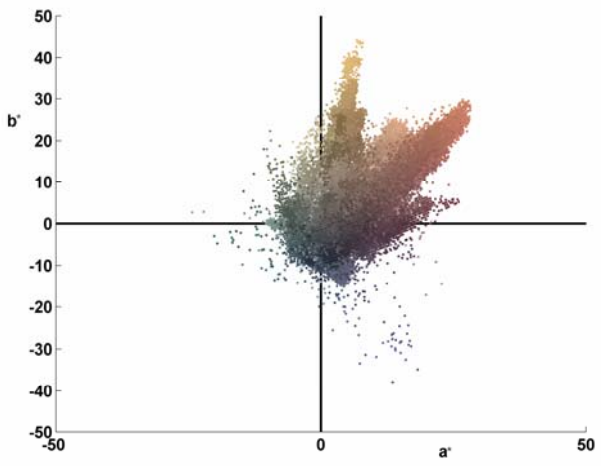

Figure 3. *Scatter plot of a*b* values for test scene 7. Each image pixel corresponds to one data point.*

In a dim room, images were presented on a calibrated LCD monitor (Eizo ColorEdge CG211). Before each experimental session, the self-adjusting monitor was calibrated to conform to the sRGB profile [11]. Eight subjects with normal color vision as confirmed by the HRR color vision test [12] and normal or corrected to normal visual acuity participated in the experiments. Images were shown on a neutral background (a*=b*=0) at an intensity level corresponding to L*=50 (Figure 4). Each experimental display was composed of four images. The upper two images serve as reference, representing the test scene under D65 illumination. The lower two images correspond to the test scene rendered under one of the chromatic test illuminants after which a color constancy algorithm was applied. Different algorithms are used on the lower left and right image. Subjects were instructed to compare the color reproduction of each of the lower images with the upper references. They then indicated which of the two lower images had the best color reproduction, but could also indicate equal performance (as good or as bad).

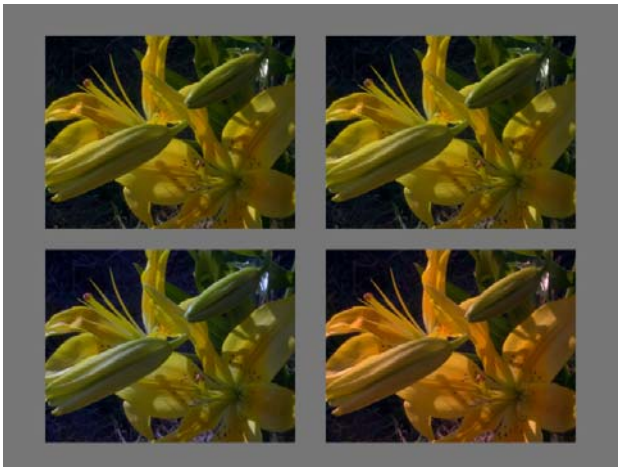

Figure 4. *Screen layout of the visual experiment. The upper two images are identical references and represent the test scene under neutral D65 illumination. The lower two images correspond to the test scene rendered under one of the test illuminants (Red, Green, Yellow or Blue) after which a color constancy algorithm was applied. Different algorithms are used on the lower left and right image. Subjects had to indicate which of the two lower images had the best color reproduction compared to the upper reference images.*

Results

Each of the 5 color constancy algorithms was tested against the other 4, implying that each subject took 320 trials (8 scenes x 4 test illuminants x 10 algorithm pairs). In each trial of our paired comparison experiment, the "winning" algorithm (visually having the best color fidelity) received 1 point, the "losing" algorithm received no points. In case of a tie, both algorithms received 0.5 point. In the latter case the two images have perceptually the same distance to the reference image, but this tells nothing about the absolute value of the distance. For each combination of test scene and test illuminant, the maximum score for an algorithm to gain is 4 since each algorithm "plays" against the four others exactly once.

In Table 1 we show the results as obtained by averaging over the 8 observers. Before we applied averaging however, we

analyzed the inter-observer agreement in different ways. First we calculated the individual differences from the mean observer scores. For each observer we computed the correlation coefficient of his/her average algorithm scores (averaged over scenes and illuminants) with the algorithm scores of the average observer. The latter are presented in the bottom row of Table 1. The correlation coefficients so obtained varied from 0.952 to 0.993, with an average of 0.983. Correlation coefficients between scores of the individual observers ranged from 0.895 to 0.998 and were all significant at the 95% confidence level (for 5 data points the critical value is 0.878). The average of 0.983 drops to about 0.4 when replacing the observer data with random data (as if observers randomly assigned their visual preference to the left or right image). Second, Root Mean Square analysis of the observer responses revealed that none of the observers is to be considered an outlier. Further analysis was therefore only performed on average observer data.

Table 1: Subjective scores per combination of scene and test illuminant, averaged over the 8 observers. The last column shows the correlation coefficient r between visual score and angular error. The bottom row shows the scores averaged over the eight scenes and four illuminants.

Sce Ne	Illu min	Color constancy algorithm					r
		1	2	3	4	5	
1	B	2.0	4.0	3.0	0.4	0.6	-0.99
	G	2.0	3.8	3.2	0.6	0.4	-0.93
	R	2.0	4.0	3.0	0.4	0.6	-0.98
	Y	2.1	3.8	3.1	0.2	0.8	-0.93
2	B	4.0	0.9	2.1	0.9	2.1	-0.95
	G	4.0	0.1	1.5	1.6	2.8	-0.99
	R	4.0	0.4	2.3	1.0	2.3	-0.96
	Y	3.9	0.0	1.3	2.3	2.5	-0.95
3	B	1.8	0.7	1.5	2.5	3.6	-0.98
	G	0.0	1.0	2.6	2.8	3.7	-0.80
	R	0.3	0.8	3.5	2.3	3.2	-0.79
	Y	0.8	0.3	2.9	2.5	3.6	-0.97
4	B	3.4	0.0	3.5	1.1	2.0	-0.97
	G	1.1	0.0	3.9	2.3	2.8	-0.68
	R	0.9	0.2	4.0	1.9	2.9	-0.67
	Y	2.8	0.0	3.9	1.2	2.2	-0.97
5	B	4.0	2.6	2.4	0.5	0.5	-0.99
	G	4.0	2.1	2.6	0.4	0.9	-0.99
	R	3.9	2.6	2.5	0.3	0.7	-0.98
	Y	4.0	0.4	3.0	0.7	1.9	-0.88
6	B	2.1	3.9	3.0	0.5	0.6	-0.99
	G	1.9	3.8	3.1	0.4	0.8	-0.99
	R	2.3	3.6	3.0	0.4	0.7	-0.99
	Y	2.7	3.4	2.8	0.9	0.3	-0.95
7	B	0.8	3.9	3.1	1.5	0.8	-0.94
	G	2.8	3.6	2.6	0.7	0.3	-0.99
	R	0.0	3.6	3.4	1.4	1.6	-0.80
	Y	1.1	3.7	3.0	1.2	1.1	-0.88
8	B	0.1	1.4	2.6	3.1	2.8	-0.95
	G	0.5	1.6	2.3	2.8	2.9	-0.01
	R	0.0	1.1	2.9	3.1	3.0	-0.69
	Y	1.0	0.4	2.3	2.9	3.4	-0.77
Average		2.1	1.9	2.8	1.4	1.8	-0.88

In Table 1, the maximum score per cell is 4 which is obtained when all 8 observers assigned the particular algorithm (1=WP, 2=GW, 3=GGW, 4=1-GE, 5=2-GE) as the "winner" in the visual image comparison. Table 1 shows that on average the General Grey World color constancy algorithm has the highest visual score for this particular set of scenes. However, this is an average result and it is clear from the Table that each scene has its "own" best performing color constancy algorithm. So it makes sense to compare the subjective scores with the objective scores (the angular error) per scene, or even per combination of scene and illuminant (i.e. per row in Table 1). Also, one may compare the scores per test illuminant, as presented in Figure 5 where data are averaged over the 8 observers and the 8 test scenes. Correlation coefficients between objective and subjective scores in Figure 5 are as follows: r=-0.98 for Blue, -0.80 for Green, -0.77 for Red and -0.85 for Yellow illumination. In terms of percentage explained variance this is R^2=0.95 for Blue, 0.64 for Green, 0.59 for Red, and 0.73 for Yellow illumination.

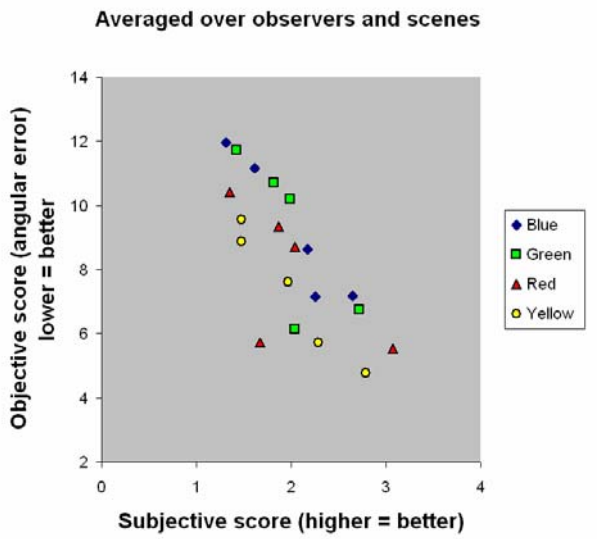

Figure 5. *Comparison of objective and subjective color constancy performance (averages over observers and scenes). Per set of 5 data points (belonging to the 5 color constancy algorithms) a simple linear regression can be performed that already accounts for reasonably high percentages explained variance.*

Note that the correlation coefficients are negative. This is due to the fact that a low value of the angular error represents a good illuminant estimation which will result in a good color correction of the image, and hence will give rise to a good visual judgment (a high score).

Averaged over all 32 cases in Table 1, the average correlation coefficient is -0.88. Remarkable exceptions to the high correlations are found for illuminants Red and Green in scenes 4 and 8. The value -0.01 for illuminant Green in scene 8 is really standing out. Leaving this data point out improves the correlation coefficient to an average of -0.92, but from inspection of the particular data point we could not find a reason to exclude it. We arrived at the same conclusion for the other deviant data points. Assuming that all 8 observers did their job correctly – which we do not doubt - this indicates that the angular error might be improved to better correlate with our perceptual error measure.

Conclusion

Comparison of the angular error (objective performance) with visual judgment (subjective performance) reveals a large correlation (about -0.9 on average) between the two performance measures, using color constancy methods with different orders of image statistics. Although a restricted set of test scenes was studied, the data already suggest that there is room for an improved correlation between the two types of performance measures.

References

[1] B. Funt, K. Barnard, L. Martin, Is machine colour constancy good enough? In: H. Burkhardt, B. Neumann (Eds.), Proc. ECCV, pg. 445–459 (1998).

[2] S.D. Hordley, G.D. Finlayson, Reevaluation of color constancy algorithm performance, J. Opt. Soc. Am. A, 23, 1008-1020 (2006).

[3] J. van de Weijer, T. Gevers, A. Gijsenij, Edge-based color constancy, IEEE T Image Process, 16, 2207–2214 (2007).

[4] E. Land, The retinex theory of color vision, Sci. Am., 237, 108–128 (1977).

[5] G. Buchsbaum, A spatial processor model for object colour perception, J. Franklin Inst., 310, 1–26 (1980).

[6] D.H. Foster, S.M.C. Nascimento, K. Amano, Information limits on neural identification of colored surfaces in natural scenes, Visual Neurosci., 21, 331-336 (2004).

[7] P.B. Delahunt, D.H. Brainard, Does human color constancy incorporate the statistical regularity of natural daylight? J. Vis., 4, 57-81 (2004).

[8] G. Wyszecki, W.S. Stiles, Color Science: Concepts and Methods, Quantitative Data and Formulae (Wiley, New York, 2nd edition, 2000), pg. 145-146 .

[9] A. Criminisi, P. Pérez, K. Toyama. Region filling and object removal by exemplar-based image inpainting, *IEEE Trans. On Image Processing*, 13, 1200–1212 (2004).

[10] J. von Kries, Influence of adaptation on the effects produced by luminous stimuli. In D. MacAdam (ed.): Sources of Color Vision, MIT Press, 109–119 (1970).

[11] M. Stokes, M. Anderson, S. Chandrasekar, R. Motta, A standard default color space for the Internet–sRGB, Nov 1996, http://www.w3.org/Graphics/Color/sRGB.html.

[12] J.E. Bailey, M. Neitz, D. Tait, J. Neitz, Evaluation of an updated HRR color vision test, Visual Neurosci., 21, 431-436 (2004).

Author Biography

Marcel Lucassen received an M.S. degree in Technical Physics from Twente University (The Netherlands) in 1988 and a Ph.D. in Biophysics (color constancy) from Utrecht University in 1993. In the period 1993-2007 he worked with Akzo Nobel Coatings and TNO Human Factors. He is now a freelance color scientist (Lucassen Colour Research) and holds a part-time position at the University of Amsterdam. His interests lie in basic and applied vision research, and color vision in particular. He is an associate editor for Color Research and Application.