Journal of Imaging Science and Technology® 53(3): 031105–031105-9, 2009.

Color Constancy Algorithms: Psychophysical Evaluation on a New Dataset

Javier Vazquez-Corral, C. Alejandro Párraga, Maria Vanrell▲ and Ramon Baldrich
Department of Computer Science, Centre de Visió per Computador, Universitat Autònoma de Barcelona, Edifici O, Campus UAB (Bellaterra), C.P. 08193 Barcelona, Spain
E-mail: javier.vazquez@cvc.uab.es

Abstract. *The estimation of the illuminant of a scene from a digital image has been the goal of a large amount of research in computer vision. Color constancy algorithms have dealt with this problem by defining different heuristics to select a unique solution from within the feasible set. The performance of these algorithms has shown that there is still a long way to go to globally solve this problem as a preliminary step in computer vision. In general, performance evaluation has been done by comparing the angular error between the estimated chromaticity and the chromaticity of a canonical illuminant, which is highly dependent on the image dataset. Recently, some workers have used high-level constraints to estimate illuminants; in this case selection is based on increasing the performance on the subsequent steps of the systems. In this paper the authors propose a new performance measure, the perceptual angular error. It evaluates the performance of a color constancy algorithm according to the perceptual preferences of humans, or naturalness (instead of the actual optimal solution) and is independent of the visual task. We show the results of a new psychophysical experiment comparing solutions from three different color constancy algorithms. Our results show that in more than half of the judgments the preferred solution is not the one closest to the optimal solution. Our experiments were performed on a new dataset of images acquired with a calibrated camera with an attached neutral gray sphere, which better copes with the illuminant variations of the scene.*

[DOI: 10.2352/J.ImagingSci.Technol.2009.53.3.031105]

INTRODUCTION

Color constancy is the ability of the human visual system to perceive a stable representation of color despite illumination changes. Like other perceptual constancy capabilities of the visual system, color constancy is crucial for succeeding in many ecologically relevant visual tasks such as food collection, detection of predators, etc. The importance of color constancy in biological vision is mirrored in computer vision applications, where success in a wide range of visual tasks relies on achieving a high degree of illuminant invariance. In the last 20 years, research in computational color constancy has tried to recover the illuminant of a scene from an acquired image.

This has been shown to be a mathematically ill-posed problem, which therefore does not have a unique solution. A common computational approach to illuminant recovery (and color constancy in general) is to produce a list of possible illuminants (feasible solutions) and then use some assumptions, based on the interactions of scene surfaces and illuminants to select the most appropriate solution among all possible illuminants. A recent extended review of computational color constancy methods was provided by Hordley.[1] In this review, computational algorithms were classified in five different groups according to how they approach the problem. These were (a) simple statistical methods,[2] (b) neural networks,[3] (c) gamut mapping,[4,5] (d) probabilistic methods,[6] and (e) physics-based methods.[7] Comparison studies[8,9] have ranked the performance of these algorithms, which usually depend on the properties of the image dataset and the statistical measures used for the evaluation. It is generally agreed that, although some algorithms may perform well on average, they may also perform poorly for specific images. This is the reason why some authors[10] have proposed a one-to-one evaluation of the algorithms on individual images. In this way, comparisons become more independent of the chosen image dataset. However, the general conclusion is that more research should be directed toward a combination of different methods, since the performance of a method usually depends on the type of scene with which it deals.[11] Recently, some interesting studies have pointed toward this direction,[12] i.e., trying to find which statistical properties of the scenes determine the best color constancy method to use. In all these approaches, the evaluation of the performance of the algorithms has been based on computing the *angular error* between the selected solution and the actual solution that is provided by the acquisition method.

Other recent proposals[13,14] turn away from the usual approach and deal instead with multiple solutions delegating the selection of a unique solution to a subsequent step that depends on high-level, task-related interpretations, such as the ability to annotate the image content. In this example, the best solution would be the one giving the best semantic annotation of the image content. It is in this kind of approach where the need for a different evaluation emerges, since the performance depends on the visual task and this can lead to an inability to compare different methods. Hence, to be able to evaluate this performance and to compare it with other high-level methods, we propose to explore a new evaluation procedure.

▲IS&T Member.

Received Aug. 25, 2008; accepted for publication Feb. 23, 2009; published online Apr. 28, 2009.
1062-3701/2009/53(3)/031105/9/$20.00.

Figure 1. Images regularly selected in the experiment as natural (left) vs images hardly ever selected (right).

In summary, the goal of this paper is to show the results of a new psychophysical experiment following the lines of that presented by Vazquez et al.[15] The previous results were confirmed, that is, humans do not choose the minimum angular error solution as the more natural one. Furthermore, in this paper we propose a new measure to reduce the gap between the error measure and the human preference. Our new experiment represents an improvement over the old one in that it considers the uncertainty level of the observer responses and it uses a new, improved image dataset. This new dataset has been built by using a neutral gray sphere attached to the calibrated camera to better estimate the illuminant of the scene. We have worked with the Shades-of-Gray[16] algorithm instead of CRule.[17] This decision was made on the basis that CRule is calibrated whereas the other algorithms are not.

EXPERIMENTAL SETUP

Subjects were presented with a pair of images (each one a different color constancy solution) on a CRT monitor and asked to select the image that seems "most natural." The term "natural" was chosen not because it refers to natural objects but because it refers to natural viewing conditions, implying the least amount of digital manipulation or global perception of an illuminant. Figure 1 shows some exemplary pictures from the database. The pictures on the left are examples of images selected as natural most of the time, while those on the right are examples of images hardly ever selected as natural.

The global schematics of the experiment are shown in Figure 2. We used a set of 83 images from a new image dataset that was built for this experiment (the image gathering details are explained below). The camera calibration allows us to obtain the Commission Internationale de l'Eclairage (CIE) 1931 XYZ values for each pixel and consequently, we converted 83 images from CIE XYZ space to CIE standard red, green blue (sRGB). Following this, we replaced the original illuminant by D65 using the chromaticity values of the gray sphere that was present in all image scenes.

From the original images, five new pictures were created by reilluminating the scene with five different illuminants. To this end we have used the chromatic values of each illuminant (three Plankians: 4000, 7000, and 10,000 K, and two arbitrary illuminants: greenish ($x=0.3026$, $y=0.3547$) and purplish ($x=0.2724$, $y=0.2458$), totaling 415 images. Afterward, the three color constancy algorithms (Gray-World,[2] Shades-of-Gray,[16] and MaxName[15]) explained below were applied to the newly created images. Consequently, we obtain one solution per test image per algorithm, totaling 1245 different solutions. These solutions were converted back to CIE XYZ to be displayed on a calibrated CRT monitor (Viewsonic P227f, which was tested to confirm its uniformity across the screen surface) using a visual stimulus generator (Cambridge Research Systems ViSaGe). The monitor's white point chromaticity was ($x=0.315$, $y=0.341$), and its maximum luminance was 123.78 Cd/m^2. The experiment was conducted in a dark room in which the only light present in the room came from the monitor itself.

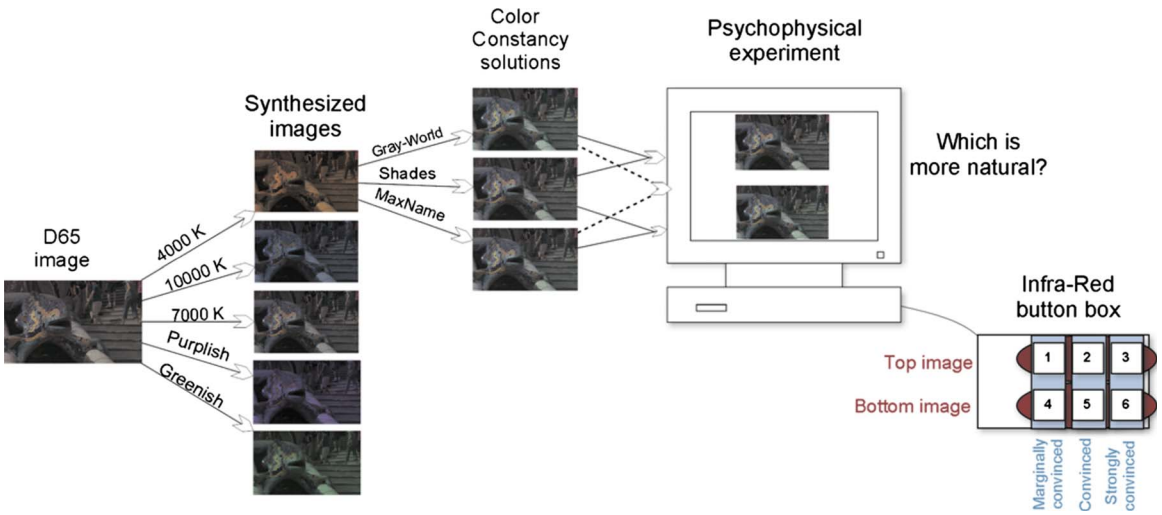

Figure 2. Experiment schedule.

Figure 3. Image dataset under D65 illuminant.

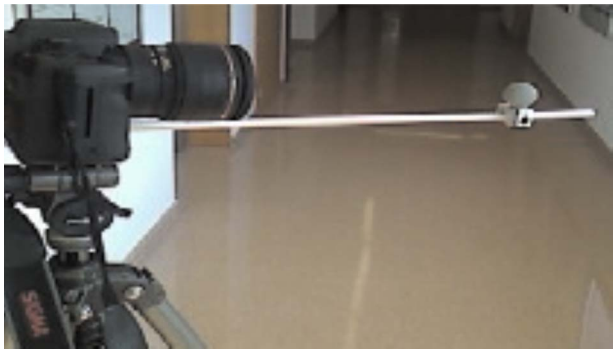

Figure 4. Camera and gray sphere setup.

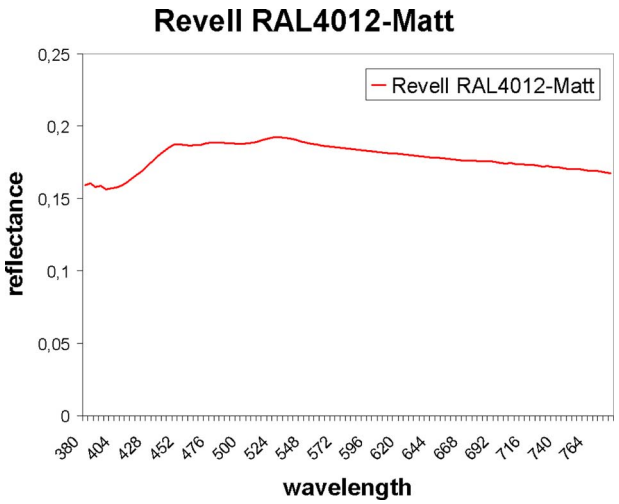

Figure 5. Reflectance of the paint used on the ball.

The experiment was conducted on ten naive observers recruited among university students and staff (none of the observers had previously seen the picture database). All observers were tested for normal color vision using the Ishihara and the Farnsworth dichotomous tests (D-15). Pairs of pictures, each obtained using one of two different color constancy algorithms, were presented one on top of the other on a gray background (31 Cd/m^2). The order and position of the picture pairs were random. Each picture subtended $10.5° \times 5.5°$ to the observer and was viewed from a distance of 146 cm. This brings us to 1245 pairs of observations per observer. No influence on picture (top or bottom) position in the observers' decision was found.

For each presentation, observers were asked to select the picture that seemed most natural and to rate their selection by pressing a button on an IR button box. The setup (six buttons) allowed observers to register how convinced they were of their choice (e.g., strongly convinced, convinced, and marginally convinced). For example, observers who were strongly convinced that the top image was more natural than the bottom one would press button 3 (see Fig. 2), if they were marginally convinced that the bottom picture was the most natural they would press button 4 and so on. There was no time limit, but observers took an average of 2.5 s to respond to each choice. The total experiment lasted approximately 90 min (divided in three sessions of 30 min each).

A New Image Dataset

To test the models we need a large image dataset of good quality natural scenes. From a colorimetric point of view, the obvious choice is to produce hyperspectral imagery in order to reduce metameric effects. However, hyperspectral outdoor natural scenes are difficult to acquire since the exposure times needed are long, and their capture implies control over small movements or changes in the scene, (not to mention the financial cost of the equipment). There are currently good quality image databases available (such as the hyperspectral dataset built by Foster et al.[18] and Brelstaff et al.[19]), but they either contain specialized (i.e., nongeneral) imagery, or the number of scenes is not large enough for our purposes. For this reason, and because metamerism is relatively rare in natural scenes,[20,21] we decided to acquire our own dataset of 83 images (see Figure 3) using a trichromatic digital color camera (Sigma Foveon D10) calibrated to produce CIE XYZ pixel representations.

The camera was calibrated at Bristol University (UK) Experimental Psychology laboratary by measuring its color sensors' spectral sensitivities using a set of 31 narrow band interference filters, a constant-current incandescent light source, and a TopCon SR1 telespectroradiometer (a process similar to that used by others[22,23]). The calibrated camera allows us to obtain a measure of the CIE XYZ values for every pixel in the image. Images were acquired around the city of Barcelona at different times of the day and on three different days in July 2008. The weather was mostly sunny with a few clouds. We mounted a gray ball in front of the

camera (see Figure 4) following the ideas of Ciurea and Funt.[24] The ball was uniformly painted using several thin layers of spray paint (Revell RAL7012-Matt, whose reflectance was approximately constant across the camera's response spectrum, and its reflective properties were nearly Lambertian—see Figure 5). The presence of the gray ball (originally located at the bottom-left corner of every picture and subsequently cropped out) allows us to measure and manipulate the color of the illuminant. Images whose chromaticity distribution was not spatially uniform (as measured on the gray ball) were discarded.

Selected Color Constancy Algorithms

In this section we briefly summarize the three methods we have selected for our analysis. We have chosen two well-known methods, Gray-World[2] and Shades-of-Gray,[16] and amore recent method, the MaxName algorithm.[15] The Gray-World algorithm (an uncalibrated method based on a strong assumption about the scene) was selected because of its popularity in literature. The Shades-of-Gray algorithm (another uncalibrated algorithm) was selected because it considerably improves performance with respect to Gray-World (another uncalibrated algorithm such as Gray-Edge[25] could also have been used). Finally, MaxName[15] was selected because it uses high-level knowledge to correct the illuminant. We give a brief outline of these methods below.

(1) *Gray-World*. It was proposed by Buchsbaum,[2] and it is based on the hypothesis that mean chromaticity of the scene corresponds to gray. Given an image $f=(R,G,B)^T$ as a function of RGB values, and adopting the diagonal model of illuminant change,[26] then an illuminant (α,β,γ) accomplishes the Gray-World hypothesis if

$$\frac{\int f\partial x}{\int \partial x} = k\cdot(\alpha,\beta,\gamma), \tag{1}$$

where k is a constant.

(2) *Shades-of-Gray*. It was proposed by Finlayson and Trezzi.[16] This algorithm is a statistical extension of the Gray-World and MaxRGB[27] algorithms. It is based on the Minkowski norm of images. An illuminant (α,β,γ) is considered as the scene illuminant if it accomplishes

$$\left(\frac{\int f^p\partial x}{\int \partial x}\right)^{1/p} = k\cdot(\alpha,\beta,\gamma), \tag{2}$$

where k is a constant. Actually, this is a family of methods where $p=1$ is the Gray-World method and $p=\infty$ is the MaxRGB algorithm. In this case we have used $p=12$, since it is the best solution for our dataset.

(3) *MaxName*. This algorithm is a particular case of the one presented by Vazquez et al.[15] It is based on giving more weight to those illuminants that maximize the number of color names in the scene. That is, MaxName builds a weighted feasible set by considering *nameable* colors; this is prior knowledge given by

$$\mu_k = \int_\omega S(\lambda)E(\lambda)R_k(\lambda)\partial\lambda, \quad k=R,G,B, \tag{3}$$

where, $S(\lambda)$ are the surface reflectances having maximum probability of being labeled with a basic color term, also called focal reflectances from the work of Benavente et al.[28] In addition to the basic color terms, we added a set of skin colored reflectances. In Eq. (3), $E(\lambda)$ is the power distribution of a D65 illuminant, and $R_k(\lambda)$ are the CIE RGB 1955 color matching functions.

We define μ as the set of all k-dimensional nameable colors obtained from Eq. (3). The number of elements of μ depends on the number of reflectances used. Following this, we compute the *semantic matrix*, denoted as (SM), which is a binary representation of the color space as a matrix, where a point is set to 1 if it represents a nameable color, that is, it belongs to μ and 0 otherwise. Then, for a given input image, I, we compute all possible illuminant changes $I_{\alpha,\beta,\gamma}$. For each one, we calculate its nameability value. This is done by counting how many points of the mapped image are nameable colors in SM and can be computed by a correlation in log space:

$$Nval_{\alpha,\beta,\gamma} = \log(H_{bin}(I)) * \log(\mathrm{SM}). \tag{4}$$

In the previous equation, H_{bin} is the binarized histogram of the image, *Nval* at the position (α,β,γ) is the number of coincidences between the SM and $I_{\alpha,\beta,\gamma}$. *Nval* is a three-dimensional matrix, depending on all the feasible maps, (α,β,γ). From this matrix, we select the most feasible illuminant as the one that accomplishes

$$(\alpha,\beta,\gamma) = \arg\max_{(\alpha,\beta,\gamma)} Nval, \tag{5}$$

that is, the one giving the maximum number of nameable colors.

RESULTS

The results of the experiment validate those presented by Vazquez et al.[15] with a different image dataset and a different set of algorithms. The main finding is that preferred solutions, namely, the more natural in the psychophysical experiment, do not always coincide with solutions of minimum angular error. In fact, this agreement only happened in 43% of the observations, independently of the degree of certainty of the observers when making the decision.

Since the experimental procedure allows us to define a partition in the interval [0,1] to encode the subject selection and each observation represents a decision between two images, then for each observation we label one image as the result from Method A and the other as the result from Method B (Methods A and B are labeled as 1 and 0, respectively). The confidence of the decision is considered at three

Table I. Button codification.

Image at the bottom is more natural than image at the top			Image at the top is more natural than image at the bottom		
Button 6	Button 5	Button 4	Button 1	Button 2	Button 3
Definitely more natural	Sufficiently more natural	Marginally more natural	Marginally more natural	Sufficiently more natural	Definitely more natural
0	0.2	0.4	0.6	0.8	1

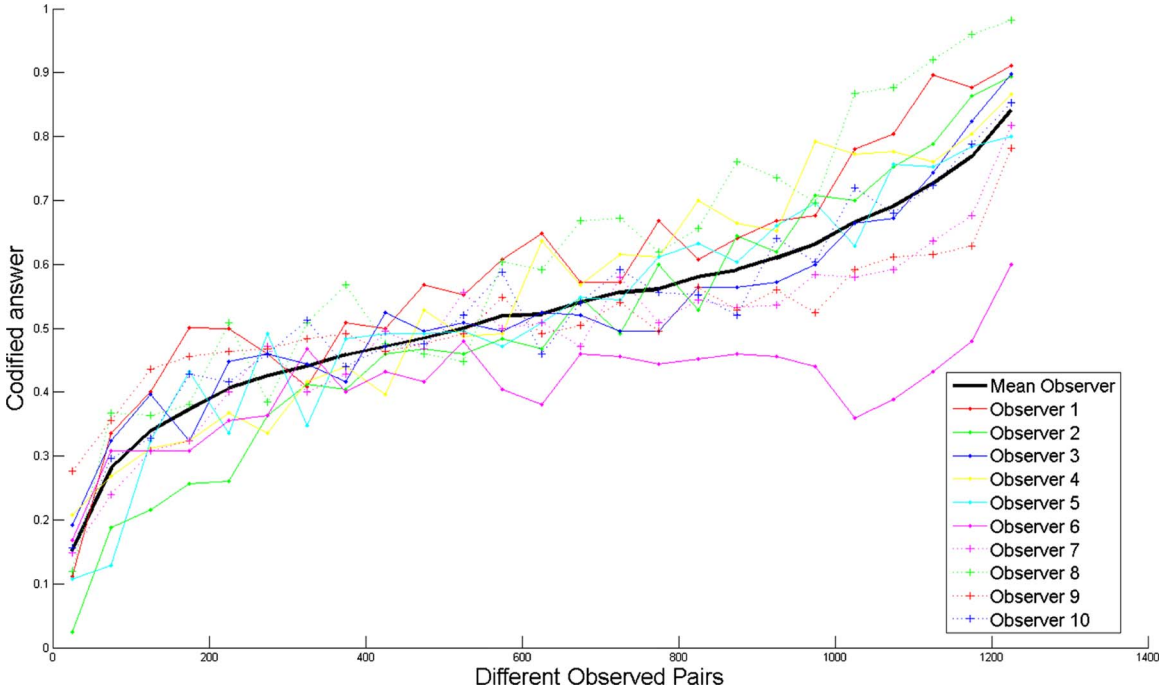

Figure 6. Comparison to the mean observer (black line).

Table II. Correlation between each observer and mean observer.

Observer	1	2	3	4	5	6	7	8	9	10
Correlation	0.54	0.57	0.59	0.55	0.52	0.23	0.48	0.63	0.61	0.55
CV	52,49%	57,96%	37,65%	52,28%	52,69%	59,85%	47,12%	51,13%	25,36%	42,81%

different levels (the three buttons that the subject was allowed to press yield an ordinal paired comparison[29]). For example, suppose that a scene processed by Method A is presented on top of the screen and a second scene processed by Method B is presented at the bottom (the physical position of the scenes was randomized in each trial, but let us consider an exemplary layout). If subjects think that the top picture is more natural they will press one of the top buttons in Fig. 2, according to how strongly they are convinced. Suppose the subject presses button 3 (top-right: definitely more natural), then the response is coded as 1. If the choice is button 2 (top-center: sufficiently more natural) the response is coded as 0.8, etc. (see Table I). If, on the contrary subjects think the bottom picture (Method B) is more natural, then they will press a button from the lower row (Fig. 2). If they are marginally convinced, they will pick button 4 (bottom-left), and the response will be coded as 0.4 according to Table I. Similarly if they are strongly convinced, they will press button 6 (bottom-right), and the response will be coded as 0. In this way we collect not only the direction of the response but its certainty. Observers' certainty was found to be correlated (corr. coef. 0.726) to a simple measure of image difference (the angular error between each image pair). This technique is similar to that used by other researchers.[30–33]

We have computed two different measures of observer variability. The first measure is the correlation coefficient between individual subjects and the average (in black in Fig-

Table III. Results of the experiment in the one-to-one comparison.

Selected method vs method	Shades-of-Gray (%)	Gray-World (%)	MaxName (%)
Shades-of-Gray	—	68.1	50.6
Gray-World	31.9	—	37.6
MaxName	49.4	62.4	—

Table IV. Experiment results in a general comparison.

Method	Wins (%)
Shades-of-Gray	35.18
Gray-World	16.63
MaxName	39.28
Three-equally selected	8.92

Table V. Results using Thurstone's law of comparative judgment.

Method	Wins (%)
Shades-of-Gray	42.65
MaxName	36.39
Gray-World	20.96

ure 6). Table II shows this measure. The idea behind this analysis is to detect outliers (subjects with a distribution of results significantly different from the rest of the observers, i.e., low correlation). Our second measure is the coefficient of variation (CV),[34,35] which computes the difference between two statistical samples (see Table II). Both measures were calculated for the whole 1245 observations (three combinations of color constancy solutions $\times$ 415 observations per combination). From the table, and from the distribution of the plots in Fig. 6, we decided to omit data from observer 6 (very low correlation coefficient and highest coefficient of variation) in all subsequent analyses.

As a first approach to analyze our results we computed the mean of the observers' responses for each pairwise comparison. We considered that a method was selected if the mean of the encoded decisions, computed for all nine observers, is greater than 0.5 (when the method was encoded as 1) or lower than 0.5 (when the method was encoded as 0). The performance does not vary significantly if we do not consider the cases where the average value is too close to the chance rate (e.g., averages between 0.45 and 0.55). The results of these pairwise comparisons are given in Table III. For each pair of methods, we show the percentage of cases where it has been selected against the others. Thus, results in Table III can be interpreted as follows: each method (in rows) is preferred to a certain percentage of trials over the method in the columns. For example, Shades-of-Gray is preferred in 68.1% of the trials against Gray-World.

The percentages in Table III show that the images produced by Shades-of-Gray and MaxName are preferred to those produced by Gray-World (68.1% and 62.4%). However, there is no clear preference when compared against each other (50.6% Shades-of-Gray preference versus MaxName).

In Table IV we show a global comparison of all algorithms (the percentages are computed for all 415 images). A method was considered a "winner" for a given image if it was selected in two of the three comparisons. Methods were evaluated in the same way as we did for results in Table III (that is, a greater than 0.5 mean value from all observers is encoded as 1). Evaluating this way, there are some cases where the three methods are equally selected (this happens in 8.92% of the images). This analysis was formulated in order to remove nontransitive comparisons (e.g., Method A beats Method B, Method B beats Method C, and Method C beats Method A). Hence, we can conclude from these straightforward analyses that solutions from MaxName are preferred in general but are closely followed by Shades-of-Gray (39.28% and 35.18%, respectively). We can also state that Gray-World solutions are the least preferred in general (with a low percentage of 16.63%). Moreover, the best angular error solution is selected in 42.96% of the cases.

We have also calculated Thurstone's law of comparative judgment[36] coefficients from our data (Table V), obtained from the ordinal pairwise comparisons. Using this measure, results are not very different (Shades-of-Gray and MaxName are clearly better than Gray-World although the ranking changes), and images with minimal angular error are only selected in 45% of the cases.

Finally, we have computed two overall analyses (considering all scenes as one) in order to extract a global ranking for our color constancy methods: Thurstone's law of comparative judgment[36] and the Bradley–Terry[37] analysis. Table

Table VI. Results using Bradley–Terry ordinal pairwise comparison analysis.

Parameter	Degrees of freedom	Estimate	Standard error	Wald 95% confidence limits		Chi-square	Pr > Chisq
Shades-of-Gray	1	1.609	1.2231	−0.7882	4.0063	1.73	0.1883
MaxName	1	1.0256	0.8435	−0.6278	2.6789	1.48	0.2241
Gray-World	0	0	0	0	0		

Table VII. Results using Thurston's law of comparative judgment binary pairwise comparison analysis.

Parameter	DF	Estimate	Standard error	Wald 95% confidence limits		Chi-square	Pr > Chisq
Shades-of-Gray	1	0.196	0.0031	0.19	0.2021	4040.2	<0.0001
MaxName	1	0.1283	0.0031	0.1223	0.1343	1743.22	<0.0001
Gray-World	0	0	0	0	0		

VI shows the results of Bradley and Terry's cumulative logit model for pairwise evaluations extended to ordinal comparisons.[29] These results are shown in the "estimate" column where the estimate reference has been set to 0 for the smallest value (Gray-World model). The standard error of this ranking measure shows that the two best models (Shades-of-Gray and MaxName) are better than Gray-World and arguably close to each other. Table VII shows a similar analysis using Thurstone's law of comparative judgment[36] and considering all scenes as one.

As we mentioned above, our experiment shows that images having minimum angular error with respect to the canonical solution are selected in less than half the observations (when we ask people for the most natural image, the response does not always correspond to the optimal physical solution). Moreover, this result is maintained even if we discard responses with low levels of certainty. In order to quantify this fact, in the next section we will introduce a new measure to complement the current performance evaluation of color constancy algorithms.

PERCEPTUAL PERFORMANCE EVALUATION

Assuming the ill-posed nature of the problem, the difficulty of finding an optimal solution and the results of the present experiment, we propose an approach to color constancy algorithms that involves human color constancy by trying to match computational solutions to perceived solutions. Hence, we propose a new evaluation measurement, the *perceptual angular error*, which is based on perceptual judgments of adequacy of a solution instead of the physical solution. The approach that we propose in this work does not try to give a line of research alternative to the current trends, which focus on classifying scene contents to efficiently combine different methods. Here we try to complement these efforts from a different point of view that we could consider as more "top-down," instead of the "bottom-up" nature of the usual research.

As mentioned above, the most common performance evaluation for color constancy algorithms consists of measuring how close their proposed solution is to the physical solution, independently of the other concerns. This has been computed as

$$e_{ang} = a\cos\left(\frac{\rho_w \hat{\rho}_w}{\|\rho_w\|\|\hat{\rho}_w\|}\right), \tag{6}$$

which represents the angle between the actual white point of the scene illuminant, ρ_w, and the estimation of this point

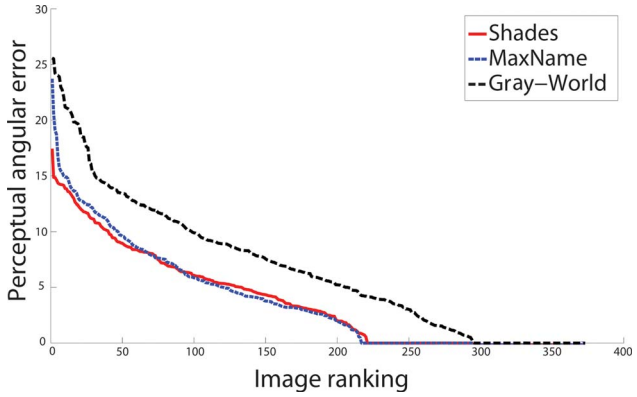

Figure 7. Estimated perceptual angular error (between method estimations and preferred illuminants).

given by the color constancy method, $\hat{\rho}_w$, which can be understood as a chromaticity distance between the physical solution and the estimate. The current consensus is that none of the current algorithms present a good performance on all the images,[38] and a combination of different algorithms offers a promising option for further research. Our proposal here is to introduce a new measure, the *perceptual angular error*, e^p_{ang}, that would be computed in a similar way:

$$e^p_{ang} = a\cos\left(\frac{\rho^p_w \hat{\rho}_w}{\|\rho^p_w\|\|\hat{\rho}_w\|}\right), \tag{7}$$

where ρ^p_w is the perceived white point of the scene (which should be measured psychophysically) and $\hat{\rho}_w$ is an estimation of this point, that is the result of any color constancy method, as in Eq. (6). The difficulty of this new measurement arises from the complexity of building a large image dataset, where ρ^p_w, the perceived white point of the images has been measured.

In this work we propose a simple estimation of this perceived white point by considering the images preferred in the previous experiment. Hence, the perceived white point is given by the images coming from the color constancy solutions that have been preferred by the observers. The preferred solutions, that is, the most natural solutions, can give us an approximation to the perceived image white point.

Making the above consideration, in Figure 7 we can see how the estimation of the perceptual angular error works for the three tested algorithms. In the abscissa we plot a ranking of the observations in order to get the perceptual errors in descending order. In the ordinate we show the estimated

Table VIII. Angular error for the different methods on 415 images of the dataset.

Method	Mean	rms	Median
MaxName	7.64°	8.84°	6.78°
Shades-of-Gray	7.84°	9.70°	5.95°
Gray-World	10.05°	12.70°	7.75°

Table IX. Estimated perceptual angular error for the different methods on 415 images of the dataset.

Method	Mean	rms	Median
MaxName	3.86°	6.02°	2.61°
Shades-of-Gray	3.79°	5.66°	2.86°
Gray-World	6.70°	9.01°	5.85°

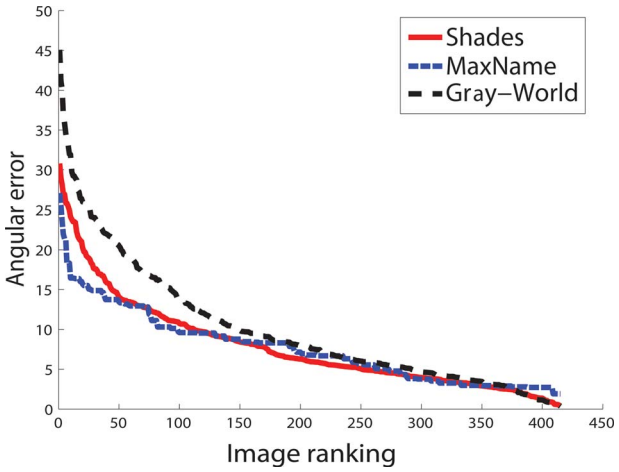

Figure 8. Angular error between methods estimations and canonical illuminant.

perceptual angular error for each created image (that is, 415 different inputs to the algorithms). A numerical estimation of the perceptual angular error could be the area under the curves plotted in Fig. 7. In the figure we can see that both Shades-of-Gray and MaxName work quite similarly, while Gray-World presents the highest perceptual error. This new measurement agrees with the conclusion we summarized in the previous section and provides a complementary measure to evaluate color constancy algorithms. In Figure 8 we show a similar plot for the usual angular error.

In Tables VIII and IX we show the different statistics on the computed angular errors. In Table VIII, the angular error between the estimated illuminant and the canonical illuminant are shown. In this case, MaxName and Shades-of-Gray present better results than Gray-World. In Table IX equal statistics are computed for the estimated perceptual angular error. The results in this table confirm the conclusions we obtained from Fig. 7.

CONCLUSIONS

This paper explores a new research line, the psychophysical evaluation of color constancy algorithms. Previous research points to the need to further explore the behavior of high-level constraints needed for the selection of a feasible solution (to avoid the dependency of current evaluations on the statistics of the image dataset). With this aim in mind, we have performed a psychophysical experiment in order to compare three computational color constancy algorithms: Shades-of-Gray, Gray-World, and MaxName. The results of the experiment show Shades-of-Gray and MaxName methods have quite similar results, which are better than those obtained by the Gray-World method and that in almost half the judgments, subjects have preferred solutions that are not the closest ones to the optimal solutions.

Considering that subjects do not prefer the optimal solutions in a large percentage of judgments, we have introduced a new measure based on the perceptual solutions to complement current evaluations: the perceptual angular error. It tries to measure the proximity of the computational solutions versus the human color constancy solutions. The current experiment allows computing an estimation of the perceptual angular error for the three explored algorithms. However, our main conclusion is that further work should be done in the line of building a large dataset of images linked to the perceptually preferred judgments.

To this end a new, more complex experiment, perhaps related to the one proposed in Ref. 39, must be done in order to obtain the perceptual solution of the images independently of the algorithms being judged.

Acknowledgments

This work has been partially supported by projects TIN2004-02970, TIN2007-64577, Consolider-Ingenio 2010 CSD2007-00018, and the Ramon y Cajal research programme (RYC-2007-00484) of the Spanish MEC (Ministry of Science). The authors thank Dr. J. van de Weijer for his insightful comments.

REFERENCES

[1] S. Hordley, "Scene illuminant estimation: Past, present, and future," Color Res. Appl. **31**, 303 (2006).
[2] G. Buchsbaum, "A spatial processor model for object colour perception," J. Franklin Inst. **310**, 1 (1980).
[3] V. C. Cardei, B. Funt, and K. Barnard, "Estimating the scene illumination chromaticity by using a neural network," J. Opt. Soc. Am. A **19**, 2374 (2002).
[4] G. Finlayson, S. Hordley, and R. Xu, "Convex programming colour constancy with a diagonal-offset model," *Proc. International Conference on Image Processing (ICIP)* (IEEE Computer Society, Los Alamitos, CA, 2005) pp. 2617–2620.
[5] K. Barnard, "Improvements to gamut mapping colour constancy algorithms," *Proc. European Conference on Computer Vision (ECCV)* (Springer, Berlin, 2000) pp. 390–403.
[6] G. Finlayson, P. Hubel, and S. Hordley, "Color by correlation," *Proc. IS&T/SID 5th Color Imaging Conference* (IS&T, Springfield, VA, 1997) pp. 6–11.
[7] B. Funt, M. Drew, and J. Ho, "Color constancy from mutual reflection," Int. J. Comput. Vis. **6**, 5 (1991).
[8] K. Barnard, V. Cardei, and B. Funt, "A comparison of computational color constancy algorithms. I: Methodology and experiments with synthesized data," IEEE Trans. Image Process. **11**, 972 (2002).
[9] K. Barnard, L. Martin, A. Coath, and B. Funt, "A comparison of

computational color constancy algorithms. II: Experiments with image data," IEEE Trans. Image Process. **11**, 985 (2002).
[10] S. Hordley and G. Finlayson, "Re-evaluating colour constancy algorithm," *Proc. 17th International Conference on Pattern Recognition* (IEEE Computer Society, Los Alamitos, CA, 2004) pp. 76–79.
[11] V. Cardei and B. Funt, "Committee-based color constancy," *Proc. IS&T/SID 7th Color Imaging Conference* (IS&T, Springfield, VA, 1999) pp. 311–313.
[12] A. Gijsenij and T. Gevers, "Color constancy using natural image statistics," *Proc. 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Vols. 1–8 (IEEE Computer Society, Los Alamitos, CA, 2007) pp. 1806–1813.
[13] F. Tous, "Computational framework for the white point interpretation base on color matching," Ph.D. thesis, Universitat Autònoma de Barcelona, Barcelona (2006) (unpublished).
[14] J. V. van de Weijer, C. Schmid, and J. Verbeek, "Using high-level visual information for color constancy," *Proc. International Conference on Computer Vision* (IEEE Computer Society, Los Alamitos, CA, 2007).
[15] J. Vazquez, M. Vanrell, R. Baldrich, and C. A. Párraga, "Towards a psychophysical evaluation of colour constancy algorithms," *Proc. IS&TsCGIV 2008/MCS/08—4th European Conference on Colour in Graphics, Imaging, and Vision* (IS&T, Springfield, VA, 2008) pp. 372–377.
[16] G. Finlayson and E. Trezzi, "Shades of gray and colour constancy," *Proc. IS&T/SID 12th Color Imaging Conference* (IS&T, Springfield, VA, 2004) pp. 37–41.
[17] D. A. Forsyth, "A novel algorithm for color constancy," Int. J. Comput. Vis. **5**, 5 (1990).
[18] 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 (2004).
[19] G. J. Brelstaff, C. A. Parraga, T. Troscianko, and D. Carr, "Hyperspectral camera system: Acquisition and analysis," Proc. SPIE **2587**, 150–159 (1995).
[20] D. H. Foster, K. Amano, S. M. C. Nascimento, and M. J. Foster, "Frequency of metamerism in natural scenes," J. Opt. Soc. Am. A **23**, 2359 (2006).
[21] M. G. A. Thomson, S. Westland, and J. Shaw, "Spatial resolution and metamerism in coloured natural scenes," Perception **29**, 123 (2000).
[22] A. Olmos and F. A. A. Kingdom, "A biologically inspired algorithm for the recovery of shading and reflectance images," Perception **33**, 1463 (2004).
[23] C. A. Párraga, T. Troscianko, and D. J. Tolhurst, "Spatiochromatic properties of natural images and human vision," Curr. Biol. **12**, 483 (2002).
[24] F. Ciurea and B. Funt, "A large image database for color constancy research," *Proc. IS&T/SID 11th Color Imaging Conference* (IS&T, Springfield, VA, 2003) pp. 160–164.
[25] J. V. van de Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," IEEE Trans. Image Process. **16**, 2207–2214 (2007).
[26] G. Finlayson, M. Drew, and B. Funt, "Diagonal transforms suffice for color constancy," *Proc. 4th International Conference on Computer Vision* (IEEE Computer Society, Los Alamitos, CA, 1993) pp. 164–171.
[27] E. Land, "Retinex theory of color-vision," Sci. Am. **237**, 108 (1977).
[28] R. Benavente, M. Vanrell, and R. Baldrich, "Parametric fuzzy sets for automatic color naming," J. Opt. Soc. Am. A **25**, 2582 (2008).
[29] A. Agresti, *An Introduction to Categorical Data Analysis* (Wiley, New York and Chichester, 1996) pp. 436–439.
[30] P. Courcoux and M. Semenou, "Preference data analysis using a paired comparison model," Food Qual. Preference **8**, 353 (1997).
[31] G. Gabrielsen, "Paired comparisons and designed experiments," Food Qual. Preference **11**, 55 (2000).
[32] J. Fleckenstein, R. A. Freund, and J. E. Jackson, "A paired comparison test of typewriter carbon papers," Tappi J. **41**, 128 (1958).
[33] A. Agresti, "Analysis of ordinal paired comparison data," J. R. Stat. Soc., Ser. C, Appl. Stat. **41**, 287 (1992).
[34] M. Luo, A. Clarke, P. Rhodes, A. Schappo, S. Scrivener, and C. Tait, "Quantifying color appearance I. LUTCHI color appearance data," Color Res. Appl. **16**, 166 (1991).
[35] M. Luo, A. Clarke, P. Rhodes, A. Schappo, S. Scrivener, and C. Tait, "Quantifying color appearance II. Testing color models performance using LUTCHI color appearance data," Color Res. Appl. **16**, 181 (1991).
[36] L. Thurstone, "A law of comparative judgment," Psychol. Rev. **34**, 273 (1927).
[37] R. A. Bradley and M. B. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," Biometrika **39**, 22 (1952).
[38] B. Funt, K. Barnard, and L. Martin, "Is machine colour constancy good enough?" *Proc. 5th European Conference on Computer Vision* (Springer, Berlin, 1998) pp. 445–459.
[39] P. D. Pinto, J. M. Linhares, and S. M. Nascimento, "Correlated color temperature preferred by observers for illumination of artistic paintings," J. Opt. Soc. Am. A **25**, 623 (2008).