Color Constancy algorithms: psychophysical evaluation on a new dataset

Javier Vazquez, C.Alejandro Párraga, Maria Vanrell, and Ramon Baldrich; Centre de Visió per Computador, Computer Science Department, Universitat Autònoma de Barcelona, Edifíci O, Campus UAB (Bellaterra), C.P.08193, Barcelona,Spain {javier.vazquez, alejandro.parraga, maria.vanrell, ramon.baldrich}@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 we 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 a 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 grey sphere, which better copes with the illuminant variations of the scene.

Keywords: Color Constancy evaluation, Psychophysics, Computational Color.

Color Constancy algorithms: psychophysical evaluation on a new dataset

Javier Vazquez, C.Alejandro Párraga, Maria Vanrell, and Ramon Baldrich; Centre de Visió per Computador, Computer Science Department, Universitat Autònoma de Barcelona, Edifíci O, Campus UAB (Bellaterra), C.P.08193, Barcelona,Spain {javier.vazquez, alejandro.parraga, maria.vanrell, ramon.baldrich}@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 we 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 grey sphere, which better copes with the illuminant variations of the scene.

Keywords: Color Constancy evaluation, Psychophysics, Computational Color.

1. 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 twenty 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 in 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 towards a combination of different methods, since the performance of a method usually depends on the type of scene it deals with[11]. Recently, some interesting studies have pointed out towards 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.

In summary, the goal of this paper is to show the results of a new psychophysical experiment following the lines of that presented in[15]. The previous results were confirmed, that is, humans do not chose the minimum angular error solution as the more natural. Furthermore, in this paper we propose a new measure to reduce the gap between the error measure and the Humans 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-grey[16] algorithm instead of CRule[17]. This decision has been taken on the basis of CRule is calibrated whereas the other algorithms are not. This paper is divided as follows. In section 2 we present how the experiment has been driven. Afterwards, in section 3 we show the results. Later on, in section 4 a new perceptual measure to deal with the evaluation of color constancy algorithms is presented. Finally, in section 5, we sum up the conclusions.

2. 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.

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

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 in section 2.2). The camera calibration allows us to obtain the CIE1931 XYZ values for each pixel and consequently, we converted 83 images from CIE XYZ space to CIE sRGB. Following this, we replaced the original illuminant by D65 using the chromaticity values of the grey sphere that was present in all image scenes.

From the original images, 5 new pictures were created by re-illuminating the scene with 5 different illuminants. To this end we have used the chromatic values of each illuminant (3 Plankians: 4000K, 7000K, 10000K, and two arbitrary illuminants: Greenish (x = 0.3026, y = 0.3547) and Purplish (x = 0.2724, y = 0.2458), totaling 415 images. Afterwards, the three color constancy algorithms (Grey-World[2], Shades-of-Grey[16] and MaxName[15]) explained in section 2.2 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 (i.e. the only light present in the room came from the monitor itself).

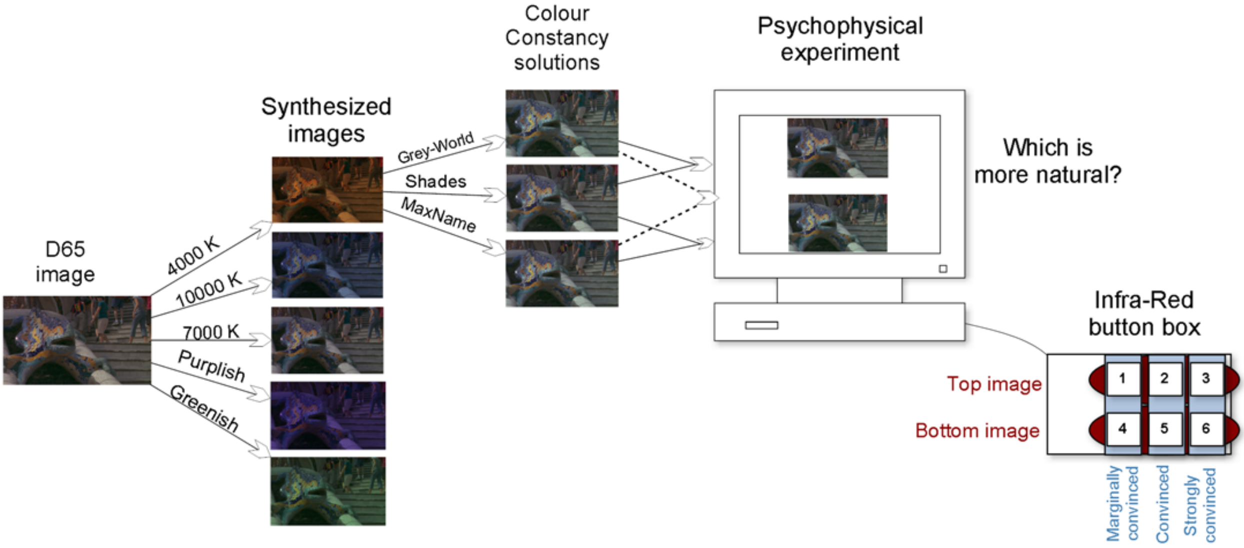

Figure 2: Experiment Schedule.

The experiment was conducted on 10 naïve 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 Test (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 grey background (31 Cd/m^2). The order and position of the picture pairs was random. Each picture subtended 10.5 x 5.5 degrees to the observer and was viewed from 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 set up (six buttons) allowed observers to register how convinced they were of their choice (e.g. strongly convinced, convinced, and marginally convinced). For example if an observer was strongly convinced that the top image was more natural that the bottom one, it would press button 3 (see Figure 2), if it was marginally convinced that the bottom picture was the most natural it would press button 4 and so on. There was no time limit but observers took an average of 2.5 seconds to respond to each choice. The total experiment lasted 90 minutes approximately (divided in three sessions of 30 minutes each)

2.1. 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, to reduce metameric effects. However, hyperspectral outdoor natural scenes are difficult to acquire since the exposure times needed are long and its capture implies control over small movements or changes in the scene, (not to talk of the financial cost of the equipment). There are currently good quality images databases available (such as the hyperspectral dataset built by Foster *et al* [18] and Brelstaff *et al*[19]), but they either contain specialised (i.e. non-general) 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 colour camera (Sigma Foveon D10) calibrated to produce CIEXYZ pixel representations.

The camera was calibrated at Bristol University (UK) Experimental Psychology lab by measuring its color sensors' spectral sensitivities using a set of 31 spectrally narrowband interference filters, a constant-current incandescent light source and a TopCon SR1 telespectroradiometer (a process similar to that 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 Barcelona city at different times of the day and in three different days in July 2008. The weather was mostly sunny with a few clouds. We mounted a grey ball in front of the camera (see Figure 4), following the ideas of Ciurea *et al*[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 grey 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 grey ball) were discarded.

Figure 3: Image dataset under D65 illuminant.

Figure 4: Camera and grey sphere setup.

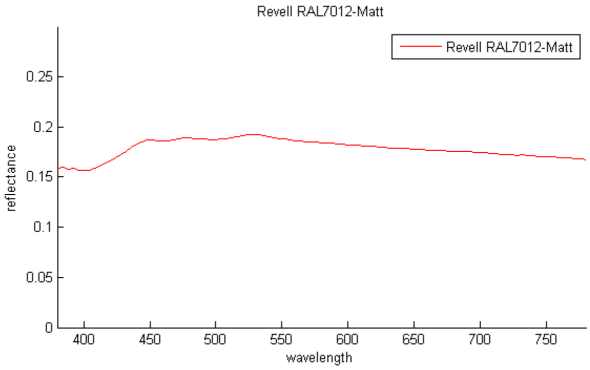

Figure 5: Reflectance of the paint used on the ball.

2.2. 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, Grey-World[2] and Shades-of-Grey[16], and a more recent method, the MaxName algorithm[15]. The Grey-World algorithm (an uncalibrated method based on a strong assumption about the scene) was selected because of its popularity in the literature. The Shades-of-Grey algorithm (another uncalibrated algorithm) was selected because it considerably improves performance with respect to Grey-World (another uncalibrated algorithm such as Grey-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. Grey-World. It was proposed by Bunschbaum[2] and it is based on the hypothesis that mean chromaticity of the scene corresponds to grey. 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 Grey-World hypothesis if

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

where k is a constant.

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

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

where k is a constant. Actually, this is a family of methods where $p=1$ is Grey-World method, and $p=\infty$ is Max-RGB 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

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

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[28]). In addition to the basic color terms, we added a set of skin colored reflectances. In Equation 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 Equation 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:

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

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 3-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:

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

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

3. 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 (Method A and B are labeled as 1 and 0, respectively). The confidence of the decision is considered at three different levels (the three buttons that the subject was allowed to press –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's consider an exemplary layout). If the subject thinks that the top picture is more natural it will press one of the top buttons in Figure 2 according to how much he/she is 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 1). If, on the contrary the subject thinks the bottom picture (Method B) is more natural, then he/she will press a button from the lower row (Figure 2). If he/she is marginally convinced, will pick button 4 (bottom-left) and the response will be coded as 0.4 according to Table 1. Similarly if he/she is strongly convinced, 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. Observer's 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].

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

Table 1: Buttons codification.

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 Figure 6). Table 2 shows this measure. The idea behind this analysis is to detect outliers (subjects with a distribution of results significantly different to 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 2). Both measures were calculated for the whole 1245 observations (3 combinations of color constancy solutions x 415 observations per combination).

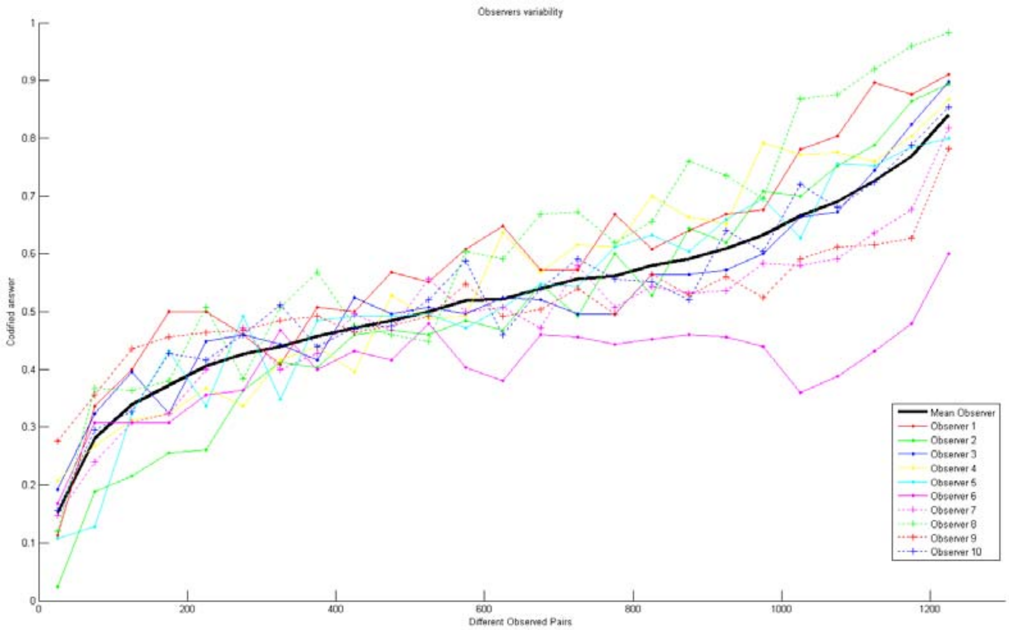

Figure 6: Comparison to the mean observer (black line).

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%

Table 2: Correlation between each observer and mean observer.

From this table, and from the distribution of the plots in Figure 6, we decided to omit data from observer 6 (very low correlation coefficient and highest coefficient of variation) in all subsequent analysis.

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 9 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 3. For each pair of methods, we show the percentage of cases where it has been selected against the others. Thus, results in Table 3 can be interpreted as follows: each method (in rows) is preferred a certain percentage of trials over the method in the columns. For example, Shades-of-Grey is preferred in 68.1% of the trials against Grey-world.

vs. Method Selected method	Shades-of-Grey	Grey-World	MaxName
Shades-of-Grey	-	68.1%	50.6%
Grey-World	31.9%	-	37.6%
MaxName	49.4%	62.4%	-

Table 3: Results of the experiment in the 1-to-1 comparison.

The percentages in Table 3 show that the images produced by Shades-of-Grey and MaxName are preferred to those produced by Grey-World (68,1% and 62,4%). However, there is no clear preference when compared against each other (50.6% Shades-of-Grey preference vs. MaxName).

In Table 4 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 3 (that is, a greater than a 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 non-transitive 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 closely followed by Shades-of-Grey (39.28% and 35.18% respectively). We can also state that Grey-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.

Method	Wins
Shades-of-Grey	35.18%
Grey-World	16.63%
MaxName	39.28%
3-equally selected	8,92%

Table 4: Experiment results in a general comparison.

We have also calculated the Thurstone's Law of Comparative Judgement[36] coefficients from our data (Table 5), obtained from the ordinal pairwise comparisons. Using this measure, results are not very different (Shades-of-Grey and Maxname are clearly better than Grey-World although the ranking changes) and images with minimal angular error are only selected in 45% of the cases.

Method	Wins
Shades-of-grey	42.65 %
MaxName	36.39 %
Grey-World	20.96 %

Table 5: Results using Thurstone's Law of Comparative Judgement

Finally, we have computed two overall analyses (considering all scenes as one) in order to extract a global ranking for our color constancy methods: the Thurstone's Law of Comparative Judgement[36] and the Bradley-Terry[37] analysis. Table 6 shows the results of the Bradley and Terry's cumulative logit model for pairwise evaluations extended to ordinal comparisons[29]. These results are shown on the "estimate" column where the estimate reference has been set to 0 for the smallest value (Grey-World model). The standard error of this ranking measure shows that the two best models (Shades-of-Grey and MaxName) are better than Grey-World and arguably close to each other. Table 7 shows a similar analysis using Thurstone's Law of Comparative Judgement[36] and considering all scenes as one.

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr>ChiSq
Shades-of-grey	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
Grey-World	0	0	0	0	0	.	.

Table 6: Results using Bradley-Terry ordinal pairwise comparison analysis

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr>ChiSq
Shades-of-grey	1	0.196	0.0031	0.19	0.2021	4040.2	<.0001
MaxName	1	0.1283	0.0031	0.1223	0.1343	1743.22	<.0001
Grey-World	0	0	0	0	0	.	.

Table 7: Results using Thurston law of comparative judgment binary pairwise comparison analysis

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 of 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.

4. 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 an alternative line research 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 before, the most common performance evaluation for color constancy algorithms consists in measuring how close their proposed solution is to the physical solution, independently of the other concerns. This has been computed as

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

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

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

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 Equation 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 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 Figure 7. In the figure we can see that both Shades-of-Grey and MaxName work quite similarly, while Grey-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.

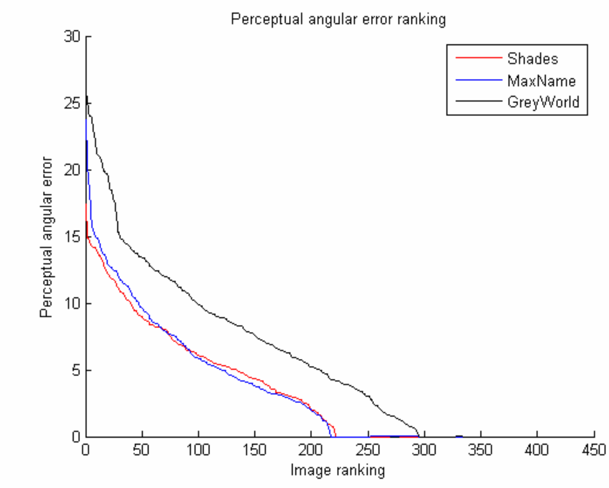

Figure 7: Estimated Perceptual Angular error (between method estimations and preferred illuminants).

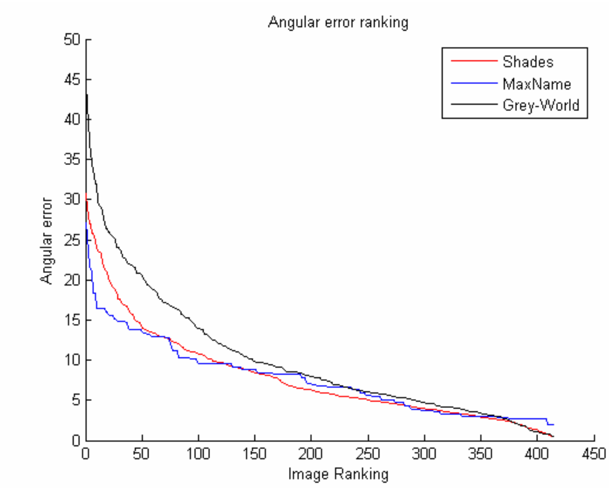

Figure 8: Angular error between methods estimations and canonical illuminant.

In Tables 8 and 9 we show the different statistics on the computed angular errors. In Table 8, the angular error between the estimated illuminant and the canonical illuminant are shown. In this case, MaxName and Shades-of-Grey present better results than

Grey-World. In Table 9 equal statistics are computed for the estimated perceptual angular error. The results on this table confirm the conclusions we obtained from Figure 7.

	Mean	RMS	Median
MaxName	7.64º	8.84º	6.78º
Shades-of-Grey	7.84º	9.70º	5.95º
Grey-World	10.05º	12.70º	7.75º

Table 8: Angular error for the different methods on 415 images of the dataset.

	Mean	RMS	Median
MaxName	3.86º	6.02º	2.61º
Shades-of-Grey	3.79º	5.66º	2.86º
Grey-World	6.70º	9.01º	5.85º

Table 9: Estimated perceptual angular error for the different methods on 415 images of the dataset.

5. Conclusion

This paper explores a new research line, the psychophysical evaluation of color constancy algorithms. Previous research point out 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-Grey, Grey-World and MaxName. The results of the experiment show Shades-of-grey and MaxName methods have quite similar results which are better than those obtained by the Grey-World method and that in almost half of 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[39], must be done in order to obtain the perceptual solution of the images, independently of the algorithms being judged.

Acknowledgements

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

References

1. S. Hordley, "Scene illuminant estimation: Past, present, and future", Color Research and Application, 31: 303, (2006).
2. G. Buchsbaum, "A spatial precessor model for object color perception", Journal of the Franklin Institute-Engineering and Applied Mathematics, 310: 1, (1980).
3. V. C. Cardei, B. Funt, & K. Barnard, "Estimating the scene illumination chromaticity by using a neural network", J Opt Soc Am A Opt Image Sci Vis, 19, (2002).
4. G. Finlayson, S. Hordley, & R. Xu, *Convex programming colour constancy with a diagonal-offset model*. International Conference on Image Processing (ICIP). (IEEE Computer Society Press 2005) 2617-2620.
5. K. Barnard, *Improvements to gamut mapping colour constancy algorithms*. European Conference on Computer Vision (ECCV). (Springer 2000) 390-403.
6. G. Finlayson, P. Hubel, & S. Hordley, *Color by correlation*. 5th Color Imaging Conference: Color Science, Systems, and Applications. (IS&T - The Society for Imaging Science and Technology 1997) 6-11.
7. B. Funt, M. Drew, & J. Ho, "Color constancy from mutual reflection", International Journal of Computer Vision, 6: 5, (1991).
8. K. Barnard, V. Cardei, & B. Funt, "A comparison of computational color constancy algorithms - part i: Methodology and experiments with synthesized data", IEEE Transactions on Image Processing, 11, (2002).
9. K. Barnard, L. Martin, A. Coath, & B. Funt, "A comparison of computational color constancy algorithms - part ii: Experiments with image data", IEEE Transactions on Image Processing, 11: 985, (2002).
10. S. Hordley, & G. Finlayson, *Re-evaluating colour constancy algorithms*. 17th International Conference on Pattern recognition. (IEEE Computer Society 2004) 76-79.
11. V. Cardei, & B. Funt, *Committee-based color constancy*. 7th Color Imaging Conference: Color Science, Systems and Applications. (IS&T - The Society for Imaging Science and Technology 1999) 311-313.
12. A. Gijsenij, & T. Gevers, *Color constancy using natural image statistics*. 2007 IEEE Conference on Computer Vision and Pattern Recognition, Vols 1-8. (IEEE Computer Society Press 2007) 1806-1813.
13. F. Tous, *Computational framework for the white point interpretation base on color matching.* Unpublished PhD. Thesis, Universitat Autònoma de Barcelona, Barcelona (2006).
14. J. V. van de Weijer, C. Schmid, & J. Verbeek, *Using high-level visual information for color constancy*. International Conference on Computer Vision. (IEEE Computer Society Press 2007)

15. J. Vazquez, M. Vanrell, R. Baldrich, & C. A. Párraga, *Towards a psychophysical evaluation of colour constancy algorithms*. CGIV 2008 / MCS/08 - 4th European Conference on Colour in Graphics, Imaging, and Vision 10th International Symposium on Multispectral Colour Science, Terrassa – Barcelona, España. (Society for Imaging Science and Technology 2008) 372-377.
16. G. Finlayson, & E. Trezzi, *Shades of gray and colour constancy*. 12th Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications. (IS&T - The Society for Imaging Science and Technology 2004) 37-41.
17. D. A. Forsyth, "A novel algorithm for color constancy", International Journal of Computer Vision, 5: 5, (1990).
18. D. H. Foster, S. M. C. Nascimento, & K. Amano, "Information limits on neural identification of colored surfaces in natural scenes", Vis. Neurosci., 21: 331, (2004).
19. G. J. Brelstaff, C. A. Parraga, T. Troscianko, & D. Carr, *Hyperspectral camera system: Acquisition and analysis [2587-30]*. Proceedings- Spie the International Society For Optical Engineering. (SPIE Publishing 1995) 150-159.
20. D. H. Foster, K. Amano, S. M. Nascimento, & M. J. Foster, "Frequency of metamerism in natural scenes", J Opt Soc Am A Opt Image Sci Vis, 23: 2359, (2006).
21. M. G. A. Thomson, S. Westland, & J. Shaw, "Spatial resolution and metamerism in coloured natural scenes", Perception, 29: 123, (2000).
22. A. Olmos, & 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, & D. J. Tolhurst, "Spatiochromatic properties of natural images and human vision", Curr Biol, 12, (2002).
24. F. Ciurea, & B. Funt, *A large image database for color constancy research*. 11th Color Imaging Conference: Color Science and Engineering - Systems, Technologies, Applications. (IS&T - The Society for Imaging Science and Technology 2003) 160-164.
25. J. V. van de Weijer, T. Gevers, & A. Gijsenij, *Edge-based color constancy*. IEEE Transactions on Image Processing. (IEEE Computer Society Press 2007) 2207-2214.
26. G. Finlayson, M. Drew, & B. Funt, *Diagonal transforms suffice for color constancy*. 4th International Conference on Computer Vision. (IEEE Computer Society Press 1993) 164-171.
27. E. Land, "Retinex theory of color-vision", Scientific American, 237: 108, (1977).
28. R. Benavente, M. Vanrell, & R. Baldrich, "Estimation of fuzzy sets for computational colour categorization", Color Research and Application, 29: 342, (2004).
29. A. Agresti (1996). *An introduction to categorical data analysis* (Wiley, New York ; Chichester, 1996) 436-439.
30. P. Courcoux, & M. Semenou, "Preference data analysis using a paired comparison model", Food Quality and Preference, 8, (1997).
31. G. Gabrielsen, "Paired comparisons and designed experiments", Food Quality and Preference, 11, (2000).
32. J. Fleckenstein, R. A. Freund, & J. E. Jackson, "A paired comparison test of typewriter carbon papers", Tappi (Technical Association of the Pulp and Paper Industry), 41, (1958).
33. A. Agresti, "Analysis of ordinal paired comparison data", Applied Statistics - Journal of the Royal Statistical Society, Series C, 41, (1992).
34. M. Luo, A. Clarke, P. Rhodes, A. Schappo, S. Scrivener, & C. Tait, "Quantifying color appearance i. LUTCHI color appearance data", Color Research and Application, 16: 166, (1991).
35. M. Luo, A. Clarke, P. Rhodes, A. Schappo, S. Scrivener, & C. Tait, "Quantifying color appearance ii. Testing color models performance using LUTCHI color appearance data", Color Research and Application, 16: 181, (1991).
36. L. Thurstone, "A law of comparative judgment", Psychological Review, 34, (1927).
37. R. A. Bradley, & M. B. Terry, "Rank analysis of incomplete block designs: I the method of paired comparisons", Biometrika, 39: 22, (1952).

38. B. Funt, K. Barnard, & L. Martin, *Is machine colour constancy good enough?* 5th European Conference on Computer Vision, Freiburg, Germany. (Springer 1998) 445-459.

39. P. D. Pinto, J. M. Linhares, & S. M. Nascimento, "Correlated color temperature preferred by observers for illumination of artistic paintings", J. Opt. Soc. Am. A, 25: 623, (2008).