Re-evaluating Colour Constancy Algorithms

S. D. Hordley and G. D. Finlayson
School of Computing Sciences
University of East Anglia
Norwich NR4 7TJ, UK
{steve,graham}@cmp.uea.ac.uk

Abstract

We present a re-evaluation of previous experimental data for five different colour constancy algorithms, based on experiments on real and synthetic images. Our work is motivated by the observation that previous analysis of algorithm performance is flawed because it uses inappropriate statistical measures of performance. We discuss these flaws in detail and suggest more appropriate statistical tests. We show that using these tests conclusions as to the relative performance of algorithms are significantly changed as compared to the original analysis of the data. In particular we conclude that the performance of two algorithms: Gamut Mapping and Color by Correlation is statistically equivalent and significantly better than the three other algorithms tested (Max-RGB and two versions of Grey-world).

1. Introduction

The *colour constancy problem*: estimating the colour of the scene illuminant in an image and correcting the image to account for its effect, is an important problem in computer vision. There is a long history of colour constancy research and many algorithms have been published which solve the problem with varying degrees of success. In this paper we set out to re-evaluate the performance of some of these existing methods [8, 3, 6, 4] because, as we will show, previous algorithm evaluation [1, 2] is flawed. To this end we propose appropriate methods for judging algorithm performance and re-evaluate algorithms with respect to these performance measures.

The colour constancy problem can be simply stated as how, given an image of a scene taken under a single, arbitrary, unknown scene illuminant can we recover an estimate of that scene light? For the purposes of this work we assume that an image consists of a collection of RGB triplets representing a camera's response to light from a discrete set of sample spatial locations in a scene. Colour constancy algorithms vary in how they define "an estimate of the scene illuminant" but in this work we will focus on the degree to which algorithms are able to recover an estimate of the scene illuminant *white-point*. That is, the RGB value which represents the camera's response to a maximally and uniformly reflective surface viewed under the scene illuminant. We measure the accuracy of a colour constancy algorithm by measuring the error between an algorithms estimate of the white-point ($\hat{\underline{p}}_w^o$) and the actual scene illuminant white-point: $\underline{p}_w^o$.

Of course, algorithm accuracy varies from image to image and to obtain a robust estimate of algorithm performance we look at this error measure "averaged" over lots of different images. It is the shortcomings of this assessment of algorithm accuracy which we aim to address in this paper. In the next section we review previous methods of algorithm assessment and show that they suffer from a number of weaknesses. We then introduce statistical measures which are more appropriate for judging the relative performance of algorithms. In Section 3 we re-evaluate the performance of a number of existing colour constancy algorithms using the proposed methods using a previous experimental paradigm [1, 2]. We conclude the paper in Section 4 with a brief summary.

2. Evaluating Colour Constancy Algorithms

First we define our measure of algorithm accuracy. The illuminant white-point and an algorithm's estimate of it are RGB triplets: points in a 3-d space so one way to measure their difference is by the Euclidean distance between them:

$$\|\underline{p}_w^o - \hat{\underline{p}}_w^o\| \tag{1}$$

In estimating the scene illuminant however, accurately estimating the overall intensity of the illumination is of less importance than estimating its "colour". Thus, algorithms are most commonly assessed using an intensity independent error measure. Here (in common with previous work [1, 2])

we use the angle between the two RGB as our error measure:

$$e_{Ang} = \text{acos}\left(\frac{\underline{p}_w^{o\,t}\hat{\underline{p}}_w^o}{\|\underline{p}_w^o\|\|\hat{\underline{p}}_w^o\|}\right) \quad (2)$$

This error measure tells us the accuracy of a particular algorithm's performance and allow us to easily compare the relative performance of two or more algorithms on a single image. More generally we are interested in comparing algorithm performance over a large set of images and it is in this regard that current analysis of algorithm performance is lacking. When assessing performance over large sets of images authors typically compare algorithms using a single summary statistic such as the mean [4] (or root mean square [1]) angular error averaged over a set of images. If the mean error for algorithm A is found to be lower than the mean error for algorithm B then the conclusion is drawn that algorithm A is better than algorithm B. There are two potential problems with this assessment. First, a single summary statistic such as the mean does not always adequately summarise the underlying distribution. Second, the fact that one algorithm has a lower mean value than another is not sufficient information to draw the conclusion that one algorithm is better than the other.

The most thorough evaluation of colour constancy algorithms to-date has been given by Barnard *et al* [1, 2]. As part of their evaluation they looked at the distribution of the magnitude of chromaticity errors. That is, they calculated $r_w^o - \hat{r}_w^o$ where r_w^o and $\hat{r}_w^o$ are the actual and estimated r chromaticity value for the scene illuminant white-point. They found the distribution of these errors to be approximately normally distributed with a mean of zero. On this evidence they concluded that an appropriate error measure for assessing algorithm performance was the root mean square (RMS) error of a given error measure (e.g. angular error) since when an error measure is normally distributed with a mean of zero RMS error gives an estimate of the standard deviation of the error statistic. However, the fact that $r_w^o - \hat{r}_w^o$ is normally distributed does not imply that other error measures are also normally distributed and in the event that they are not, RMS error is not necessarily an appropriate measure.

The left-hand plot of Figure 1 shows the distribution of angular errors for a typical colour constancy algorithm (the *Max-RGB* algorithm) for 1000 randomly generated images each containing 8 surfaces. It is clear from this histogram plot that the angular error is not normally distributed. This fact is emphasised by the right-hand plot which plots quantiles of a standard normal distribution against the quantiles of the angular errors for the 1000 images. If the errors were normally distributed the points on this plot would fall along a straight line. This example illustrates the typical case for the algorithms we have tested on both real and synthetic images. On this evidence we should conclude that

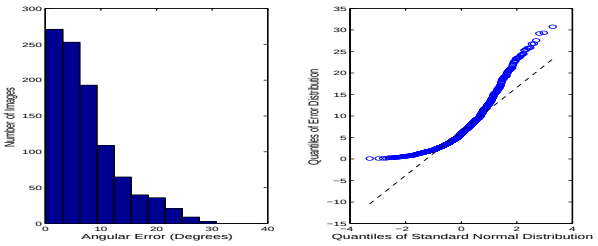

Figure 1. Left: Typical Distribution of Angular Error in white-point estimation. Right: Quantiles of this error distribution plotted against quantiles of a standard Normal distribution.

angular error is not normally distributed so that RMS error does not give an estimate of the standard deviation of the error measure.

So, if we want to look at a single summary statistic for this error distribution which should we choose? The mean error is often reported as a summary statistic however, for the type of distribution under investigation this statistic is not appropriate. It is well known [7] that the mean is a poor summary statistic for non-symmetric distributions: the distributions we are investigating are skewed as the example in Figure 1 illustrates. In these situations the median is a more reliable estimate of central tendency [7].

However, the median does not tell us everything about the distribution of errors. A more informative measure is the sample median together with a confidence interval for the statistic. In the case we are studying care must be taken when calculating a confidence interval because the underlying error distributions are not well modelled by standard statistical distributions. An appropriate method in this case is that of re-sampling [7]. In this method we take our 1000 error measures as an approximation of the underlying population distribution. We then re-sample this distribution many times. Each time we re-sample we obtain a new sample distribution whose median value we can calculate. This provides us with a set of estimates of the median statistic. A p% confidence interval for the median can be obtained from the $p/2$ and $(1-p/2)$ quantiles of this distribution.

An alternative to single summary statistic comparisons of algorithms is to somehow compare the whole error distribution of two algorithms. Since the underlying error distributions cannot be well modelled by a standard distribution we require a test which requires us to make no assumptions about the underlying error distributions. An appropriate test in this case is the Wilcoxon Sign Test [7]. Let A and

B be random variables representing the angular error in algorithm A and B's estimate of the scene illuminant. The Wilcoxon Test is used to test the hypothesis that the random variables A and B are such that $p = P(A > B) = 0.5$. That is, we hypothesise that algorithm A and B have the same performance. To test the hypothesis $H_0 : p = 0.5$ we consider independent pairs $(A_1, B_1) \ldots (A_N, B_N)$ of errors for N different images. We denote by W the number of images for which $A_i > B_i$. When H_0 is true W is binomially distributed ($b(N, 0.5)$) and the Wilcoxon test is based on this statistic. We can define an alternative hypothesis $H_1 : p < 0.5$ which if true implies that errors for algorithm A are lower than those for algorithm B. We accept or reject the null hypothesis at a given significance level α if the probability of observing the results we observe is less than or equal to α. The value of α we choose defines the error rate we accept when reject the null hypothesis. E.g. if we accept an error rate of 1% (we wrongly reject the null hypothesis in 1% of cases) we would choose $\alpha = 0.01$.

3. Colour Constancy Experiments

To re-evaluate algorithm performance in the light of the tests proposed above we analyse the performance of five algorithms on a set of synthetic and real image experiments following Barnard *et al* [1, 2]. The algorithms tested are Max-RGB *Mx*, Grey-world (*GW*), Database Grey-world (*DB*), a version of the Gamut Mapping algorithm (*GM*) and a version of Colour by Correlation (*CM*). The first three algorithms are implemented exactly as described in [1]. We used a linear programming implementation [5] of the Gamut Mapping algorithm which has been shown [5] to give near identical performance to the version tested in [1]. The version of *CM* we tested is also very similar to that tested in [1]. These five algorithms cover the major algorithm groups tested in the original work.

3.1. Synthetic Image Experiments

In the synthetic image experiment images are synthesised by first selecting n reflectances randomly from a collection of 1995 measured reflectance functions intended to be broadly representative of the world. The scene illuminant is selected randomly from a set of 287 measured illuminants. These reflectance functions and illuminant SPD are used together with the spectral sensitivities of a SONY DXC-900 digital video camera to generate synthetic sensor responses.

The n sensor response triplets form the input to the five tested algorithms each of which returns an estimate of the scene illuminant white-point. Algorithm performance is measured for images with number of surfaces n = 2, 4, 8, 16, 32, or 64 and for each value of n 1000 images are generated. Algorithm accuracy is measured using angular error defined in Eqn (2) above. The left-hand plot of Figure 2 shows the

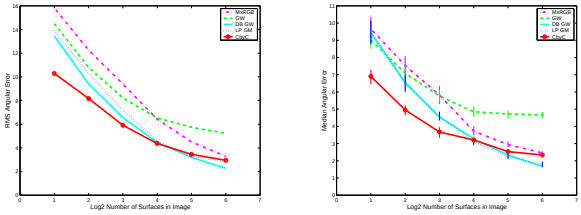

Figure 2. RMS (left) and Median (with 95% confidence intervals, right) angular error as a function of the number of surfaces in an image.

performance of the five algorithms as was reported in the original work: it shows RMS angular error as a function of the *log* (base 2) of the number of surfaces in an image. The right-hand plot shows results in terms of median error together with 95% confidence intervals (vertical bars) calculated using the re-sampling technique [7]. While the overall trends in these two plots are similar, if we assess algorithm performance based on only one of these summary statistics we will draw different conclusions depending on whether we look at RMS or median error. For small numbers of surfaces (up to 8), both plots suggest that *CM* performs best. Based on the RMS plot we might order the remaining algorithms, *Mx*, *GW*, *DB*, and *GM* in order of improving performance. However, if we look at median error statistics the picture is less clear: for example judged by median error *Mx* and *GW* are equal when we have 8 surfaces in an image. If we include 95% confidence intervals for the median in our assessment then the conclusions we draw change again: in this case, *CM* is still best for small numbers of surfaces, however, we cannot separate GW and Mx until we have at least sixteen surfaces in an image. Adding confidence inter-

	Mx	GW	DB	GM	CM
Mx		+	-	-	-
GW	-		-	-	-
DB	+	+			-
GM	+	+			-
CM	+	+	+	+	

Table 1. Results of Wilcoxon's Sign Test for all pairs of the five algorithms. See main text for interpretation of the table.

vals to a summary statistic provides more information than just the statistic and is a first step to determining the statistical significance of the results. To formally determine the statistical significance of the results we applied Wilcoxon's sign test to the error distributions. Table 1 summaries the results based on the errors for 6000 images: i.e. for 1000 images with 2, 4, ...64 surfaces. The table shows results for the 99% confidence level ($\alpha = 0.01$). A plus sign ($+$) in the ith row and jth column of the table means that algorithm i is statistically better than algorithm j when judged according to the Wilcoxon test. A minus ($-$) implies that it is worse while if the box is empty the two algorithms are statistically the same. On the basis of the results in Table 1 we would conclude that overall, *CM* is the best algorithm (better than all other algorithms), *GM* and *DB* are equally good and better than *Mx* and *GW* and that *Mx* is significantly better than *GW*.

3.2. Real Image Experiments

We conducted a second experiment on 321 real images which follows the procedure detailed in [2]. As noted by Barnard *et al* [2] image pre-processing has a significant effect on algorithm performance. The results we report here are based on a pre-processing scheme which involves segmenting images according to the method outlined by Barnard *et al* in [2]. In the case of *CM* we found that significant improvement is obtained if we consider only "bright" segments of the image where a "bright" segment is considered to be any segment with an intensity greater than the 70th quantile of all image segments (ordered by intensity). Also, if multiple segments have the same RGB value, that value is counted multiple times: this differs from what was proposed in the original algorithm [4] in which each RGB is counted only once. Table 2 summarises the results for the real image experiments. If we judge algorithms according to RMS error (column 1) we would rank the algorithms: *GM, Mx, CM, DB, GW* in order of decreasing performance. However, when judged according to median error (column 2) the ranking changes. In this case *CM* is better than *Mx* and very similar to *GM*. Columns 3-7 summarise the results of Wilcoxon's sign test. According to this test *CM* and *GM* are equivalent and both are better than the three remaining algorithms. Of those three algorithms, *Mx* and *DB* are equivalent and both are better than *GW*. Once again, the statistical tests proposed in this paper significantly change the conclusions we draw about algorithm performance. Most significantly, previous analysis would suggest that *CM* performs worse on real images than it does on synthetic images. Our analysis offers less support for this view. Rather we should conclude that *CM* is significantly better than all algorithms apart from *GM* to which it is equivalent. This is similar to the trend observed on synthetic images.

	RMS Error	Median Errror	Mx	GW	DB	GM	CM
Mx	8.77	4.02		+		-	-
GW	14.32	8.85	-		-	-	-
DB	12.25	6.58		+		-	-
GM	5.46	2.92	+	+	+		
CM	9.93	2.93	+	+	+		

Table 2. Algorithm performance (real images).

4. Conclusions

The most important point raised by the re-evaluation of previous colour constancy experiments discussed above is that the relative performance of algorithms changes considerably depending on the criteria by which they are judged. The Wilcoxon test of statistical significance we have applied is strong in the sense that it makes minimal assumptions about the underlying error distributions. In summary we recommend that the future evaluation of colour constancy algorithms should follow the guidelines set out in this paper or at least pay attention to the underlying distributions of the error statistics used for evaluation and apply appropriate statistical tests.

References

[1] K. Barnard, V. Cardei, and B. Funt. A comparison of computational color constancy algorithms; part one: Methodology and experiments with synthetic images. *IEEE Transactions on Image Processing*, 11(9):972–984, 2002.

[2] K. Barnard, L. Martin, A. Coath, and B. Funt. A comparison of computational color constancy algorithms; part two: Experiments with image data. *IEEE Transactions on Image Processing*, 11(9):985–996, 2002.

[3] G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310:1–26, 1980.

[4] G. Finlayson, S. Hordley, and P. Hubel. Color by correlation: A simple, unifying framework for color constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1209–1221, 2001.

[5] G. D. Finlayson and R. Xu. Convex programming colour constancy. In *Workshop on Color and Photometric Methods in Computer Vision*, pages 1–7. IEEE, October 2003.

[6] D. Forsyth. A Novel Algorithm for Colour Constancy. *International Journal of Computer Vision*, 5(1):5–36, 1990.

[7] R. V. Hogg and E. A. Tanis. *Probability and Statistical Inference*. Prentice Hall, 2001.

[8] E. H. Land. The Retinex Theory of Color Vision. *Scientific American*, pages 108–129, 1977.