Colour Constancy: Past, Present and Future

S. D. Hordley
University of East Anglia
Norwich NR4 7TJ (UK)
Corresponding author: S. D. Hordley (steve@cmp.uea.ac.uk)

ABSTRACT

In this paper we focus on the problem of colour constancy - how a visual system is able to ensure that the colours it perceives remain stable, regardless of the prevailing scene illuminant - in the context of the more general computer vision problem. Our aim is firstly to summarise and review the most important theoretical advances that have been made in this field. Secondly, we present a summary of a comparative analysis of algorithm performance which we use as the basis of a discussion of the important questions for future research in this field. Finally, we highlight some areas of recent research which we believe are important in the context of improved colour constancy.

1. INTRODUCTION

In the context of computer vision, the problem of colour constancy is best understood by considering a simple physical model of image formation in which the "colour" of an object is controlled by the interaction of light, surface, and sensor. We assume that a scene is illuminated by a single light source with spectral power distribution (SPD) $E(\lambda)$, which illuminates surfaces characterised by their spectral reflectance functions $S(\lambda)$. The light reflected from a given surface is called the colour signal $C(\lambda) = E(\lambda)S(\lambda)$, and it is this light which enters an imaging device to produce a colour response q_k which is defined as

$$q_k = \int E(\lambda)S(\lambda)Q_k(\lambda)d\lambda \tag{1}$$

In (1), $Q_k(\lambda)$ characterises the spectral response function of a given class of light sensor on the imaging plane of the device: it determines the proportion of the colour signal the sensor absorbs on a per-wavelength basis. In most imaging devices there are 3 distinct classes of light sensor so that the response to light at a given pixel is defined by a triplet of responses $\underline{q}=(q_1, q_2, q_3)^t$, which are commonly referred to as RGB values. An inspection of (1) makes the colour constancy problem apparent. The colour response $\underline{q}$ of a device to a given surface depends on both the reflectance properties of the surface, and the SPD of the prevailing scene illuminant. So, when illumination changes, so too do the colours recorded in an image. From a computer vision perspective this illumination dependence is problematic since ideally, the colours recorded by a device would tell us something about the intrinsic properties of the imaged surfaces. That this is not true, implies that using colour as a cue to help solve fundamental vision tasks such as scene segmentation, object recognition, and tracking might run into problems if the scene illumination is changing. Indeed, there is evidence[14] to suggest that this is indeed the case. In light of these problems, solving the colour constancy problem is of fundamental importance in computer vision.

2. SOLVING FOR COLOUR CONSTANCY

In a theoretical sense, solving the colour constancy problem implies that we must invert Equation (1) to recover the scene illuminant SPD $E(\lambda)$. For a single surface we have only three measurements (the elements of $\underline{q}$), and both $E(\lambda)$ and $S(\lambda)$ are unknown. This implies that the problem is, in a strict sense, under-constrained. If we assume that $E(\lambda)$ is constant throughout an imaged scene, adding more surfaces provides us with additional constraints on the illuminant. However, it also introduces further unknowns: the reflectance functions of the added surfaces. Thus, we cannot in this way equalise the mis-match between knowns and unknowns. Given that the problem is fundamentally under-

constrained, all progress towards its solution is founded on enforcing one or more additional constraints. Computational colour constancy algorithms can be usefully classified as being *statistical* or *physics-based* approaches. Physics based approaches (e.g. [8,11,16]) usually adopt a more general model of image formation than Eqn. (1) and seek a solution by exploiting knowledge about the physical interaction between light and surfaces. For example, it is known that some surfaces reflect light in a mirror-like way and that as a result, these so-called *specular reflections* have the SPD of the illuminant. Identifying such reflections in an image would then determine the scene illuminant. Unfortunately, whilst physics-based approaches are theoretically sound, producing a robust algorithm based on these principles has proven difficult.

In terms of practical algorithms, statistical based methods have, in general, proven more successful. Early statistical approaches (e.g.[12]) focused on attempting to reduce the discrepancy between the number of knowns and unknowns in Eqn (1) by adopting linear model representations of lights and surfaces. Such approaches clearly set out the theoretical conditions under which the colour constancy problem can be solved. Unfortunately, for the case of a trichromatic device, the required conditions are not satisfied in most typical images so that again, this avenue does not lead to practically applicable algorithms. Most algorithms which have found practical application render the colour constancy problem tractable by adopting (either implicitly or explicitly) a *diagonal model* of illumination change. Here RGB responses $\underline{q}^o$ and $\underline{q}^c$ of a device to the same surface viewed under two different lights denoted *o* and *c,* are assumed to be related by a diagonal matrix

$$\begin{pmatrix} q_1^c \\ q_2^c \\ q_3^c \end{pmatrix} = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} \begin{pmatrix} q_1^o \\ q_2^o \\ q_3^o \end{pmatrix} \tag{2}$$

In adopting this model two points are important to note. First, it is strictly valid only for a restricted class of lights and/or sensors[5]. However, for most scenes imaged under typical viewing illuminants, the model holds to a good approximation. Second, in essence the model assumes that light and surface can both be described by 3 parameters however, even in this case, we have fewer knowns (3*m* pixel values) than unknowns (3*m*+3) so that further constraints are required to produce a solution.

Given such a model, solving the colour constancy problem amounts to estimating the RGB response of a device to a white surface (we call this the white point of the scene light). Two simple algorithms which are sometimes used in practice are the *max RGB* and *Grey-World* methods. In the first case it is assumed that all images contain a white (neutrally reflecting) surface, which is brighter than all other surfaces in the scene, so that the white point of the scene illuminant can be found by locating the maximum pixel value in an image. The Grey-world approach is founded on a similar constraint. In this case it is assumed that the average of all surface reflectances in a scene is a neutral (grey) reflectance, so that an estimate of the scene illuminant can be found by calculating the mean sensor response in an image. Both these simple approaches can work well for some images, but their statistical assumptions are somewhat naive, and it is easy to imagine examples of real images where both methods can (and do) fail badly. More successful algorithms can be designed by adopting better founded assumptions about images. A neural network solution[3] attempts to learn the relationship between observed image RGBs and the scene illuminant by training a network on a large set of images for which the scene illuminant is known. The trained network can be seen as representing information about the statistical structure of images. In practice it is difficult to train a neural network such that it generalises well: i.e. so that it is able to accurately predict the scene illuminant in previously unseen images. In addition, the approach is unattractive with regard to the fact that it is essentially a black box solution to the problem.

Gamut mapping algorithms[1,6,7] are different from other algorithms in the sense that they do not place additional constraints on the problem to obtain a solution. Rather, they begin with the observation that the problem is inherently under-constrained and set out to solve for the set of all possible solutions. In

the context of Eqn. (2) the set of all possible solutions is a set of diagonal matrices, each of which corresponds to a plausible scene illuminant. For a given image, an illuminant is said to be plausible if its corresponding diagonal matrix *D* maps all the image RGBs into a pre-determined gamut of RGBs called the *canonical gamut*. The canonical gamut represents the set of all possible RGBs observable under some reference illuminant. Gamut mapping solutions differ in their implementation, but central to them all is the idea of the canonical gamut which encodes the statistical information about images they use. Once a set of valid diagonal matrices has been determined for an image, an estimate of the scene illuminant is selected from this set, usually by adopting some heuristic approach[1,6,7].

Bayesian approaches[2,4] share in common the fact that they attempt to capture information about the likelihood of observing a given RGB response under a given light, in the form of a statistical prior. Then, given an RGB image whose illuminant is to be classified, they calculate a measure of the likelihood that each possible scene illuminant was the illuminant in the given image. Typically, the illuminant with the highest likelihood is selected as the estimate of the scene illuminant. The simplest implementation of a Bayesian approach to colour constancy is the so-called *Color By Correlation*[4] method. In this algorithm prior information about plausible scene lights is encoded by restricting scene illuminants to one of a discrete finite set of possible lights, each of which is considered to have an equal probability of being observed. Prior information about surfaces is encoded by defining a distribution of image chromaticities for each of the possible scene lights. This chromaticity distribution can either capture information about the gamut of possible surfaces (by declaring that a given chromaticity either is or is not observable under a given light), or it can encode a measure defining how likely it is that a given chromaticity value will occur under each light. In the first case, the algorithm is quite similar to a gamut mapping approach, but with the addition of a restriction on the set of possible scene lights.

Of the approaches we have briefly discussed, the Gamut Mapping and Bayesian methods are the most attractive from a theoretical perspective. Both methods address the fundamental fact that the problem is under-constrained and that therefore the solution is, in general, non-unique. In the ideal case, it might be considered that a Bayesian approach should be capable of better performance, since it encodes more statistical information than does Gamut Mapping. However, in practice the success of both methods depends on how accurately the statistical information they encode is matched to the statistics of real images. This point is important and can only be resolved by an evaluation of algorithm performance on real images.

3. ALGORITHM EVALUATION

There are two main approaches to algorithm evaluation. Either algorithms are tested on synthetic images, rendered according to a model of image formation such as that in Eqn. (1), and using sets of measured reflectances and illuminants, or they are tested on real images, captured under controlled illumination conditions. Testing on synthetic images means that algorithm performance can be evaluated over many thousands of images, and also ensures that the statistical information encoded by the various algorithms precisely matches the statistics of the test images. In this sense, synthetic image testing gives a best-case measure of algorithm performance. In theory, testing on real images should give a more realistic view of an algorithm's typical performance, and it should allow for the evaluation of how robust an algorithm is to noise, and perhaps most importantly, to a mis-match between an algorithm's training data and the image data it is tested on.

Figure 1 summarises the performance of 5 algorithms in a typical synthetic image experiment. Algorithm performance is plotted as a function of the number of surfaces in an image for *Max RGB*, *Grey-World*, a *Neural Network* implementation, an implementation of the *Gamut Mapping* algorithm and *Color By Correlation*: a Bayesian approach. Algorithm performance is measured in terms of *angular error*, that is, the angle between the actual white point of the scene illuminant $\underline{p}^w$ and algorithm's estimate of that white point $\hat{\underline{p}}^w$. Performance is measured over many different images and a summary statistic (the median angular error) is plotted in Figure 1. This approach to evaluation is

typical of the evaluations that appear in the literature (e.g. [1]), except for the fact that often, summary statistics other than the median are used to compare algorithms.

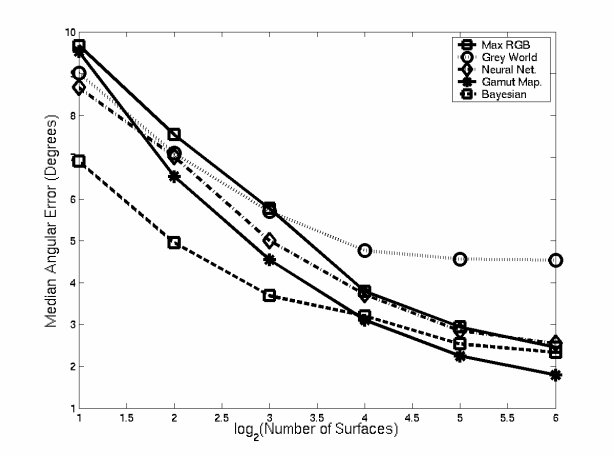

Figure 1 Median Angular Error as a function of the number of surfaces per image.

Table 1 ranks algorithm performance over all 6000 test images in terms of Root Mean Square Error, Median Error, Mean Error, and when judged according to a non-parametric Sign Test[9]. For the five algorithms shown there is good correlation (with the exception of RMS error), but in general, the choice of summary statistic can have a strong effect on the conclusions about relative algorithm performance. A more in-depth discussion can be found in[10] where it is concluded that the nature of the error distributions under study implies that the median error is the single best summary statistic, and that non-parametric tests are most appropriate for determining the statistical significance of the results.

In terms of what these results tell us about the performance of different algorithms, the ranking of algorithms is largely in line with what we might intuitively expect given what we know about their design. Essentially, the results support the theory that incorporating more (and more realistic) statistical information about image content leads to better algorithm performance. So, for example, a Bayesian approach to colour constancy gives the most accurate performance, with the Neural Network and Gamut Mapping approaches also performing reasonably well. As might be expected, all algorithms (except for Grey-World) tend towards very good performance as we increase the number of surfaces in an image. It is also noticeable that the Bayesian approach significantly out performs all other methods for images with few surfaces.

Table 1 Ranking of the five algorithms according to four different error criteria (synthetic images).

	RMS	Median	Mean	Sign Test
Max-RGB	5	4	4	4
Grey-World	4	5	5	5
Gamut Map.	3	2	2	2
Bayesian	1	1	1	1
Neural Net.	2	3	3	3

Real image performance of algorithms is most usefully evaluated using the set of 321 real images gathered at Simon Fraser University[1]. Thirty different scenes were captured under up to 11 different lights, using a well-calibrated digital camera. The performance of the same five algorithms on this test set is summarised in Table 2. For the most part the trend of the real image results follows that for the synthetic images: algorithm performance is correlated with the level of statistical information incorporated into each method. There are some notable exceptions however. First, the Bayesian approach is no longer the best performing algorithm: judged using the Sign Test, its performance is equivalent to Gamut Mapping. Second, the *Max-RGB* algorithm performs better on real images than it does in synthetic image tests. We explore the significance of these results and discuss what conclusions can be drawn from them in the next section.

4. FUTURE RESEARCH DIRECTIONS

Algorithm evaluations such as the one given above are informative in the sense that they enable a comparison of the relative performance of different algorithms to be made. It is clear from the evaluation that no algorithm affords perfect colour constancy. However, it is not easy to draw conclusions about the whether algorithm performance is “good enough”. One reason for this is the fact that a given angular error measure of algorithm accuracy does not easily translate into an assessment of whether the corresponding colour constancy performance is good or bad. A bigger

problem however, is the fact that algorithm evaluation to date is weak in the sense that it is based on a relatively small set of test images. In effect, algorithms are tested on at most 30 different scenes (albeit under a range of illuminants) and moreover, the content of these scenes is quite restricted, consisting for the most part of scenes containing just a few highly coloured objects. While in some respects this represents quite a hard test set, more conclusive evidence as to the relative performance of algorithms could be obtained by testing on a much larger set of images. In addition it is important to include scenes captured outdoors as well as indoors, and which contain, for example, faces and other natural phenomena, since these are the type of images which a real computer vision system will process.

	RMS	Median	Mean	Sign Test Rank
Max RGB	8.88	4.05	6.38	3
Grey-World	14.52	8.94	11.48	5
Gamut Map.	6.85	3.71	5.00	1
Bayesian	10.09	3.19	6.56	1
Neural Net.	11.04	7.78	9.18	4

Table 2 Errors for the 321 real images.

The issue of test scene content is indicative of a more fundamental problem in algorithm evaluation: the fact that "good enough" colour constancy performance depends on the context in which the algorithm is applied. In computer vision terms, algorithm performance is good enough, so long as the illuminant estimate, or equivalently, the illuminant independent description of a surface it provides, is of sufficient accuracy to enable subsequent vision tasks such as object recognition or tracking, to be performed successfully. Attempts to link algorithm performance to real vision tasks have been made. However, such studies are on a very small scale, and it is difficult to draw meaningful conclusions from them. Further investigation as to the relation between algorithm performance and the accuracy of subsequent visual tasks is therefore necessary, both to properly evaluate the performance of existing algorithms, and - by determining the situations in which existing algorithms fail – to guide the development of future algorithms.

The evaluation above does provide some useful information with regard to the performance of existing algorithms. In particular it suggests that on average, both the Gamut Mapping and the Bayesian algorithms provide good colour constancy (an angular error of 3-4 degrees represents quite good accuracy – for example an image with errors of this order would in many cases be acceptable to a human observer). The evaluation is ambiguous however, as to whether exploiting gamut information, or likelihood information, leads to better performance. The synthetic results support the intuitive view that a Bayesian approach is more powerful (since it in effect encodes more information). However, the fact that the performance of this algorithm on real images degrades, suggests that it is difficult to obtain accurate information about the distributions of image colours in real images. Compiling a large test set of images would also be useful in this regard since it would enable a more accurate picture of the statistical structure of real images to be obtained.

The evaluation also suggests that algorithm performance is not good enough when estimates are derived based on only a few surfaces. This point is important since in practice many scenes will be lit by multiple, spatially varying illuminants, so that illuminant independent images can only be obtained by estimating the scene illuminant locally, and as a result, on information from just a few surfaces. Accepting that the training of existing algorithms is currently imperfect, the results neverthess suggest that the current state-of-the-art in algorithm development is not sufficient to enable local processing and that further theoretical advances are required. Further information about the scene illuminant might be obtained in one or more of several ways. It may be for example, that different algorithms perform well or badly on different images, so that by combining the output of two or more algorithms (e.g. Gamut Mapping and Bayesian), we obtain a more robust illuminant estimate. There is some work in this area[15,17] which suggests that this may indeed be the case. For example, it has been found[15] that combining the output of a Bayesian algorithm with the output from an approach based on exploiting specular highlights, leads to improved illuminant estimation. An alternative approach might be to look for higher-level image cues to help determine the scene content: for example, knowing that an image is of an outdoor scene, significantly restricts the set of possible illuminants.

Finally, more robust illuminant estimation might be obtained by making more measurements of the light reflected from each scene point. One recent approach in this vein which shows promise is the so-called *Chromagenic Camera*[16]. Here, a conventional RGB imaging paradigm is adopted but two images of each scene are captured. The first is just a conventional RGB image, whilst the second is an RGB image, optically pre-filtered using a special *chromagenic* filter. It has been shown that by exploiting the relationship between corresponding pixels in the image pair (a relationship which is determined by the known filter), it is possible to derive a simple algorithm for estimating the scene illuminant. In synthetic experiments similar to those reported above, this algorithm was found to give significantly better performance than all other methods, and in particular, its performance on images with only a few surfaces is very good. Further algorithm development and testing on real images is ongoing, but initial results suggest that this is a rich area for future algorithm development.

References

1. K. Barnard. *Practical Colour Constancy.* PhD thesis, Simon Fraser Univ., School of Computing Science, 2000.
2. David H. Brainard and William T. Freeman. Bayesian Method for Recovering Surface and Illuminant Properties from Photosensor Responses. In *Proceedings of the IS&T/SPIE Symposium on Electronic Imaging Science & Technology*, volume 2179, pages 364-376, 1994.
3. Vlad C. Cardei, Brian Funt, and Kobus Barnard. Estimating the scene illuminant chromaticity by using a neural network. *Journal of the Optical Society of America, A*, 19(12):2374-2386, 2002.
4. G. D. Finlayson, S. D. Hordley, and P. M. 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. Graham D. Finlayson, Mark S. Drew, and Brian V. Funt. Color constancy: generalized diagonal transforms suffice. *Journal of the Optical Society of America, A*, 11(11):3011--3019, 1994.
6. Graham D. Finlayson and Ruixia Xu. Convex programming colour constancy. In *Workshop on Color and Photometric Methods in Computer Vision. IEEE*, October 2003.
7. D. A. Forsyth. A Novel Algorithm for Colour Constancy. *International Journal of Computer Vision*, 5(1):5-36, 1990.
8. B. V. Funt, M.S. Drew, and J.Ho. Color constancy from mutual reflection. *International Journal on Computer Vision*, 6:5-24, 1991.
9. Robert V. Hogg and Elliot A. Tanis. *Probability and Statistical Inference*. Prentice Hall, 2001.
10. S. D. Hordley and G. D. Finlayson. Re-evaluating colour constancy. In *Proceedings of the 17th International Conference on Pattern Recognition. IEEE*, August 2004.
11. Gudrun J. Klinker, Steven A. Shafer, and Takeo Kanade. A physical approach to color image understanding.
International Journal of Computer Vision, 4:7-38, 1990.
12. Laurence T. Maloney and Brian A. Wandell. Color constancy: a method for recovering surface spectral reflectance. *Journal of the Optical Society of America, A*, 3(1):29-33, 1986.
13. Shoji Tominaga and Brian A. Wandell. Standard surface-reflectance model and illuminant estimation. *Journal of the Optical Society of America, A*, 6(4):576-584, 1996.
14. Brian Funt, Kobus Barnard and Lindsay Martin. Is Machine Colour Constancy Good Enough?, In *Proceedings of 5th European Conference on Computer Vision*, June", pages 455-459, 1998.
15. G. Schaefer. *A combined physical and statistical approach to computational colour constancy.* PhD thesis, University of East Anglia, School of Computing Sciences, 2004.
16. G. D. Finlayson, S. D. Hordley, and P. M. Morovic. Gamut constrained Chromagenic Colour Constancy, In *Proceedings CVPR 2005*, to appear.
17. Vlad C. Cardei and Brian Funt. Committee-Based Colour Constancy, In *Proceedings IS&T\SID Seventh Color Imaging Conference*, pages 311-313, 1999.