A Curious Problem with Using the Colour Checker Dataset for Illuminant Estimation

Graham D. Finlayson[1]***, Ghalia Hemrit***[1]***, Arjan Gijsenij***[2]***, Peter Gehler***[3]
[1] ***School of Computing Sciences, University of East Anglia; Norwich, UK***
[2] ***Informatics Institute, Faculty of Science, University of Amsterdam; Amsterdam, The Netherlands***
[3] ***Bernstein Centre of Computational Neuroscience, University of Tübingen; Tübingen, Germany***

Abstract

In illuminant estimation, we attempt to estimate the RGB of the light. We then use this estimate on an image to correct for the light's colour bias. Illuminant estimation is an essential component of all camera reproduction pipelines. How well an illuminant estimation algorithm works is determined by how well it predicts the ground truth illuminant colour. Typically, the ground truth is the RGB of a white surface placed in a scene. Over a large set of images an estimation error is calculated and different algorithms are then ranked according to their average estimation performance. Perhaps the most widely used publically available dataset used in illuminant estimation is Gehler's Colour Checker set that was reprocessed by Shi and Funt. This image set comprises 568 images of typical everyday scenes.

Curiously, we have found three different ground truths for the Shi-Funt Colour Checker image set. In this paper, we investigate whether adopting one ground truth over another results in different rankings of illuminant estimation algorithms. We find that, depending on the ground truth used, the ranking of different algorithms can change, and sometimes dramatically. Indeed, it is entirely possible that much of the recent 'advances' made in illuminant estimation were achieved because authors have switched to using a ground truth where better estimation performance is possible.

Introduction

In Figure 1, we show the same scene rendered with respect to 4 hypothetical coloured lights. Of course a human observer placed in the scene would not see such a large variation in colour. The human visual system is capable of adapting to the changes to the colour of the light. In Figure 2, we show an image where the colours are biased by the illuminant colour (top) alongside the reproduction when the illuminant colour is estimated and then 'divided out'.

Illuminant estimation is a very active field of research with scores of algorithms being proposed each year. Each researcher asks and then proposes a solution to the question of 'how can the image content be analysed to infer the colour of the prevailing light'. The majority of the algorithms (and most of the recent progress) are for 'low level' approaches. Here the content of the image is viewed as an unstructured bag of pixels or bag of features (e.g. derivatives [3]).

As the field of illuminant estimation has burgeoned, so the need to benchmark one algorithm against another has become a critical concern. Crucially, we need to make sure that any new algorithm is evaluated relative to the same images and the same

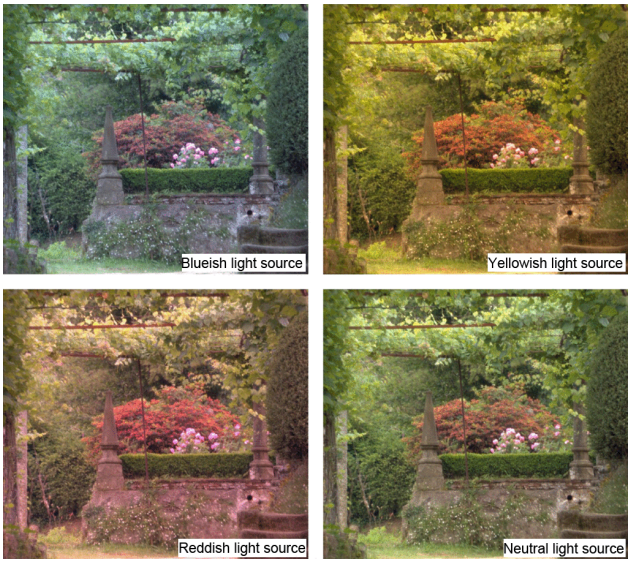

Figure 1. *Illustration of the influence of differently coloured light sources on the measured image values. These images are adapted from [1] and show the same scene, rendered under four different light sources [2]*

'ground truth' correct answers. Regarding the image set, perhaps the most widely used repository is Gehler's 'Colour Checker' dataset (so-named as the eponymous colour chart appears in every scene). The Colour Checker dataset comprises 568 images (from 2 cameras) of typical indoor and outdoor photographic scenes (in and around Microsoft research offices in Cambridge). Gehler's original dataset was reprocessed resulting in the Shi-Funt linear (raw) version [4]. And, it is the Shi-Funt dataset which is most widely used in illuminant estimation. For the Shi-Funt dataset the ground truth correct answer for illumination estimation is "the median of the RGB digital counts (i.e., median R, median G, median B) from the brightest achromatic square (ranked by average of each square) containing no RGB digital count $\succ$ 3300" [4].

On Gijsenij et al.'s [5] site *colorconstancy.com* various algorithms are benchmarked relative to the Shi-Funt dataset. Indeed, when a new algorithm is published, it is common to use the data on this website for comparison. If one wishes to claim that a new algorithm advances the state of the art, then the figures shown on the *colorconstancy.com* evaluation site need to be bettered.

The genesis for this paper is the curious - and frankly unsettling - observation that there are two different ground truths on the Gijsenij site and a third on

www.cs.sfu.ca/~colour/data/shi_gehler/. In this paper, we ask the obvious question: 'does the fact that there are three ground truths - three different sets of correct answers - affect the ranking of different illuminant estimation algorithms?'. The answer to this question is a resounding yes. Indeed, we posit that much of the recent 'advance' in illuminant estimation performance is because authors have adopted one of the three ground truths (in favour of another that had previously been commonly used). And, adopting this new ground truth, we argue, provides more favourable conditions for illuminant estimation.

The paper is organized as follows. We first give an overview of how illuminant estimation algorithms are evaluated. In particular, we review the recovery and reproduction angular errors. Then we consider the question of ranking algorithm performance. Here we show the large first-order changes that can result when different ground truths are employed. The paper finishes with a short conclusion.

Figure 2. *Example results on an image from the Colour Checker dataset. The upper image is taken with a Canon 1D in auto-white balance mode. The lower image was corrected using the Bayesian algorithm [6]*

Evaluating illuminant estimation algorithms

The Colour Checker dataset was introduced by Gehler et al. [6] and consists of 568 RGB images of indoor and outdoor real scenes, 86 images are taken with a Canon 1D and 482 images were taken with a Canon 5D. Compared to the other benchmark RGB datasets, the colour checker images are high quality images. The set is of a medium size and medium variety [2]. Samples of this dataset are shown in Figure 3.

Figure 3. *Examples of the colour checker dataset reprocessed by Shi [6][4] both captured with Canon 5D, with a 2.22 gamma correction applied*

The methodology for evaluating a given illuminant estimation algorithm is as follows. First, per image the error between the actual illuminant colour (the RGB from a physical white surface placed in the scene) and the estimate is calculated. Errors used include Euclidean distance and the more common recovery and reproduction errors (see next section). Then, over a dataset, aggregate summary statistics are calculated. Algorithm A is deemed better than algorithm B if its mean (or other some other summary measure) performance is better.

So what summary measure should we use to rank the performance of different illuminant estimation algorithms? The median gives a good overview of the whole distribution of errors and describes well the performance of one algorithm on all the images [7]. It is arguably a more appropriate summary statistic than the mean error [8]. The trimean captures the extreme values of the distribution [8], and the quantile error expresses the strength of the algorithm in being successfully applied to p% of the data. The 95% quantile returns the 95% highest error and is particularly interesting because the outliers (where illuminant estimation fails) perforce drive the development of illuminant estimation algorithms. Most algorithms (simple or complex) appear to deliver good performance for many images.

Angular error

Almost all research in illuminant estimation uses an angular metric to measure the error between the actual and estimated RGBs of the colours of the light. Angular measures are employed as it is not in general possible to recover the absolute magnitude of the illumination. Two different angular errors appear in the literature. The recovery angular error is defined as:

$$err_{recovery} = \cos^{-1}\left(\frac{\rho^E \cdot \rho^{Est}}{\|\rho^E\|\,\|\rho^{Est}\|}\right) \quad (1)$$

where ρ^E is the measured RGB triplet, $\rho^E st$ is the estimated RGB triplet (returned by an illuminant estimation algorithm), $'.'$ denotes the vector dot-product and $'\|\,\|'$ denotes the Euclidian norm. While (1) has served the community well it exhibits an unusual behaviour. Specifically, suppose we have the same scene viewed under lights A and B with respect to which we estimate the illuminant. After 'dividing out' the illuminant estimates let us further suppose that the same and identical reproduction is obtained. Despite the fact that the same endpoint is reached (the reproductions are the same), the recovery angular error can - and sometimes does - vary largely. The recovery error for the first and second lighting condition might be respectively 5 and 15 degrees (or even larger). The Reproduction angular error was introduced by Finlayson et al. [9] to mitigate this 'varying error with similar reproduction' problem. Here, the angle between a true achromatic surface under a white light $U = [111]^t$ and the actual reproduction of an achromatic surface when the estimated illuminant of one algorithm is 'divided out' is calculated. The reproduction error is defined as:

$$err_{reproduction} = \cos^{-1}\left(\frac{\left(\rho^E/\rho^{Est}\right)\cdot U}{\|\rho^E/\rho^{Est}\|\sqrt{3}}\right) \quad (2)$$

where '/' denotes the element-wise division and $\|U\| = \sqrt{3}$. The reproduction error stays stable with almost the same value when calculated regarding the same scene and different illuminant colours [9].

Experimental results

The research in this paper begins with the curious observation that - whether reproduction or recovery angular error is used - there are at least 3 ground truths for the Shi-Gehler Colour Checker dataset. That is, there are three different 568x3 matrices where each row is the putative white-point, correct answer, for each of the 568 Colour Checker images. The First two ground truths, which we call 'Gt1' and 'Gt2', appear in Gijsenij et al's *colorconstancy.com* website. The third, which we call SFU, appears on *www.cs.sfu.ca/~colour/data/shi_gehler/*. Crucially, we have found that the ground truth for each image differs by more than a scaling parameter which implies any angular errors that are calculated using the 3 ground truth datasets will be different from one another.

The chromaticities of the ground truth white-points for the Gt1 and SFU datasets are shown in Figure 4. The difference between Gt1 and Gt2 is small (but, as we shall see later, significant) and so is not plotted. Notice, the SFU ground truth is really quite separate from Gt1 and Gt2. Further, the SFU ground truth is much more compact which, if this is the 'true' set might indicate that illuminant estimation is easier than if Gt1 and Gt2 is used.

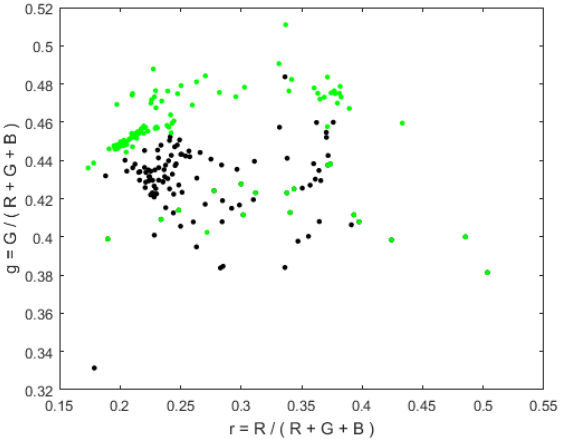

Figure 4. *2D chromaticity gamut (r,g) of the colour checker ground truths.(green) SFU ground truth [4], (black) Gt1 ground truth [5]. Gt2 (r,g) distribution is very close to Gt1,though slightly different (so, for clarity, is not plotted)*

Each of the 3 ground truths has been used, separately, to evaluate illuminant estimation algorithms. Though, the majority of methods (and all the methods on *colorconstancy.com*) are evaluated with Gt1 or Gt2. However, one of the latest and until recently the best algorithms by Barron [10] explicitly used the SFU ground truth [11]. We found that the errors calculated for [10] for Gt1 was much higher than that calculated for the SFU ground truth. This said, it seems possible that some of the performance increment presented in [10] is due to the SFU as oppose to the Gt1 ground truth being used.

Now, we calculate the mean recovery and reproduction angular errors for each of the 21 algorithms reported on Gijsenij's *colorconstancy.com* dataset. Figure 5 and Figure 6 show the change in the ranking of 6 algorithms chosen from the 21 that present the highest performance variation when we compare them. In Figure 5, mean recovery error, we see large changes in rank for the SFU ground truth compared with Gt1 or Gt2. But, there is no change in ranks when comparing Gt1 with Gt2. Significantly, for the 6 algorithms listed the simple grey-edge algorithm works best for the SFU ground truth but is second worst for Gt1 or Gt2. Figure 6 repeats this experiment for the mean reproduction error. Again there are large changes in ranking.

The following tables show the 6 best ranked algorithms (from the 21 available on the *colorconstancy.com* website) in terms of median, trimean and 95% quantile reproduction error, respectively, for each of the 3 ground truths SFU, Gt1 and Gt2. In general the rank of algorithms using Gt1 and Gt2 are the same (for the mean and median errors). But, the ranks of Gt1 and Gt2 differ slightly for the 95% quantile error. Note often the same algorithm does not appear on the best 6 for a given set of ground truth. This is worrying because depending on the ground truth used, authors will claim that their particular algorithm is better than another.

On first glance it may appear as if the SFU ground truth is more challenging since the reported errors are higher. But, we actually argue the converse. Remember, that all the algorithms

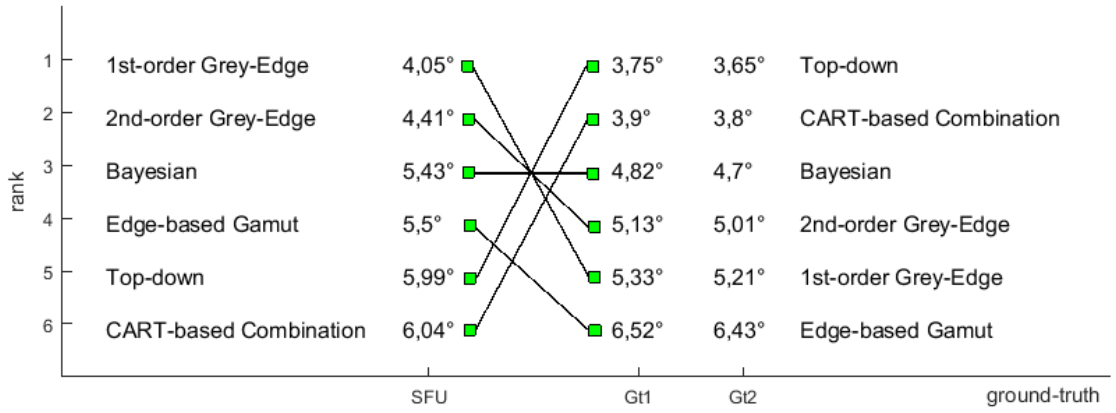

Figure 5. *Ranks of 6 algorithms in terms of mean recovery error for the 3 ground truths, SFU, Gt1 and Gt2 and values of the errors*

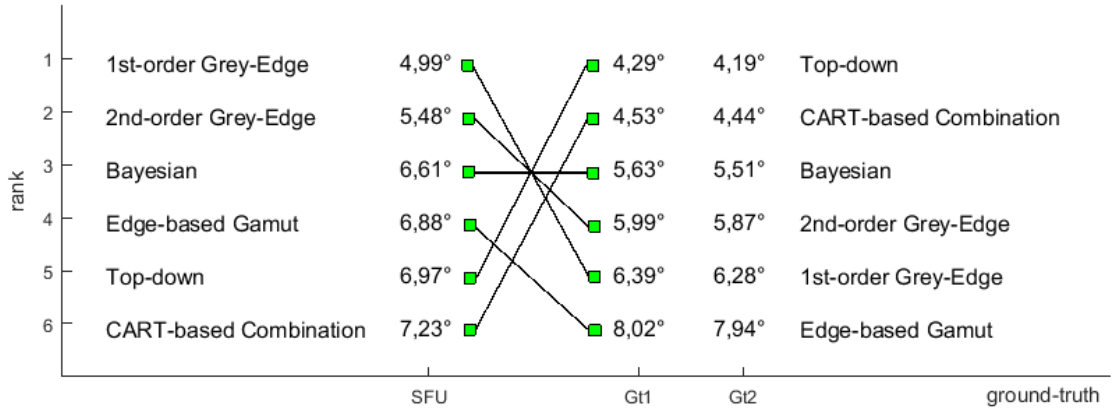

Figure 6. *Ranks of 6 algorithms in terms of mean reproduction error for the 3 ground truths, SFU, Gt1 and Gt2 and values of the errors*

Table 1. The 6 best algorithms in terms of median reproduction error for SFU vs Gt1 (best algorithms for Gt2 are the same), the Minkowski norm p and the smoothing value σ are the optimal parameters

Rank	SFU		Gt1	
	algorithm	median	algorithm	median
1	1st order Grey-Edge (p=1, σ=9) [12]	3,77°	Deep colour constancy using CNNs [13]	2,24°
2	Edge-based Gamut (σ=3) [14]	4,54°	Exemplar-based colour constancy [15]	2,64°
3	2nd order Grey-Edge (p=1, σ=1) [12]	4,59°	Pixel-based Gamut (σ=4) [14]	2,73°
4	Bayesian [16][6]	4,62°	Intersection-based Gamut (σ=4) [17]	2,74°
5	Deep colour constancy using CNNs [13]	4,75°	Bottom-up+Top-down [18]	2,75°
6	Pixel-based Gamut (σ=4) [14]	5,21°	Bottom-up [18]	2,98°

shown on Gijseij's *colorconstancy.com* site have been tuned for Gt1 or Gt2. Thus, it is not surprising they do less well on the SFU dataset (to be clear, it is not possible to retune the 21 algorithms to new ground truth, as the illuminant estimation code is not available). Moreover, as seen in Figure 4, the SFU ground truth chromaticities actually, on average, appear in a much more compact region of chromaticity space than Gt1 or Gt2. Thus, an optimization method such as [10] tuned to this dataset should in principle return smaller errors. This is in fact the case: the median angular error with SFU is $0,86°$ [10] versus $3,53°$ with Gt1. From communication with the author of [10] it was confirmed that the SFU ground truth was used [11]. Strikingly, [10] demonstrated a

Table 2. The 6 best algorithms in terms of trimean reproduction error SFU vs Gt1 (best algorithms for Gt2 are the same), the Minkowski norm p and the smoothing value σ are the optimal parameters

Rank	SFU		Gt1	
	algorithm	trimean	algorithm	trimean
1	1st order Grey-Edge (p=1, σ=9) [12]	4.11°	Deep colour constancy using CNNs [13]	2,49°
2	2nd order Grey-Edge (p=1, σ=1) [12]	4.86°	Exemplar-based colour constancy [15]	2,87°
3	Deep colour constancy using CNNs [13]	5.05°	Bottom-up+Top-down [18]	2,94°
4	Edge-based Gamut (σ=3) [14]	5.19°	Bottom-up [18]	3,15°
5	Bayesian [16][6]	5.20°	Top-down [18]	3,25°
6	Exemplar-based colour constancy [15]	5.46°	Pixel-based Gamut (σ=4) [14]	3,36°

Table 3. The 6 best algorithms in terms of quantile 95% reproduction error SFU vs Gt1, the Minkowski norm p and the smoothing value σ are the optimal parameters

Rank	SFU		Gt1	
	algorithm	quantile 95%	algorithm	quantile 95%
1	Deep colour constancy using CNNs [13]	12,17°	Exemplar-based colour constancy [15]	8,32°
2	2nd order Grey-Edge (p=1, σ=1) [12]	12,77°	Deep colour constancy using CNNs [13]	9,36°
3	Exemplar-based colour constancy [15]	13,19°	HeavyTailed-based spatial correlations [19]	9,89°
4	1st order Grey-Edge (p=1, σ=9) [12]	13,7°	CART-based Combination [20]	11,43°
5	CART-based Combination [20]	14,26°	Bottom-up [18]	11,57°
6	Bottom-up [18]	14,78°	Grey-World [21]	12,41°

Table 4. The 6 best algorithms in terms of quantile 95% reproduction error for Gt1 vs Gt2, the Minkowski norm p and the smoothing value σ are the optimal parameters

	Gt1		Gt2	
Rank	algorithm	quantile 95%	algorithm	quantile 95%
1	Exemplar-based colour constancy [15]	8,32°	Exemplar-based colour constancy [15]	7,68°
2	Deep colour constancy using CNNs [13]	9,36°	Deep colour constancy using CNNs [13]	8,95°
3	HeavyTailed-based spatial correlations [19]	9,89°	HeavyTailed-based spatial correlations [19]	9,31°
4	CART-based Combination [20]	11,43°	Bottom-up [18]	10,32°
5	Bottom-up [18]	11,57°	CART-based Combination [20]	11,1°
6	Grey-World [21]	12,41°	Grey-World [21]	12,17°

large step change in illuminant estimation performance compared to all antecedent algorithms. Actually, we propose (and are investigating) that the greater part of this step is due to the adoption of the SFU set and comparing results against algorithms evaluated on the much more spread-out ground truths of Gt1 or Gt2 i.e. apples are being compared with oranges.

Conclusion

Our paper begins with the curious observation that the Shi-Gehler variant of the Colour Checker image set - used extensively in evaluating illuminant estimation algorithms - has 3 sets of ground truth. Further, the difference in reported ground truth can be surprisingly large. One consequence of there being more than one ground truth dataset is that the ranking of algorithms (which algorithm works best) changes with the ground truth being used. Further, the change in rankings can be very large. We posit that a large part of the putative recent progress in illuminant estimation is because one particular set of ground truth has been selected whereas, previously, different ground truth had been used. Of course, an interesting question is how do we extricate ourselves

from this circumstance. Because, right now, the field is comparing apples with oranges when performance statistics are compared because which ground truth was used is often not known nor is their a recommendation on which ground truth should be used. We are are working with Peter Gehler (who acquired the initial data) and Arjan Gijsenij (who maintains the illuminant estimation benchmarking site, *colorconstancy.com*) to define a single ground truth and a single methodology for using the Colour Checker dataset for illuminant estimation.

References

[1] D. H. Foster, K. Amano, and S. M. C. Nascimento, "Color constancy in natural scenes explained by global image statistics," *Vis. Neurosci.*, vol. 23, no. 3-4, pp. 341–349, 2006.

[2] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Computational color constancy: Survey and experiments," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2475–2489, 2011.

[3] B. Funt and G. D. Finlayson, "Color constant color indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 5, pp. 522–529, 1995.

[4] L. Shi and B. Funt, "Re-processed Version of the Gehler Color Constancy Dataset of 568 Images." www.cs.sfu.ca/∼colour/data/shi_gehler/,[accessed 10-February-2017].

[5] A. Gijsenij and T. Gevers, "Datasets and Results per Datasets." www.colorconstancy.com,[accessed 10-February-2017].

[6] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, 2008.

[7] A. Gijsenij, T. Gevers, and M. P. Lucassen, "Perceptual analysis of distance measures for color constancy algorithms," *J. Opt. Soc. Am. A*, vol. 26, no. 10, pp. 2243–2256, 2009.

[8] S. D. Hordley and G. D. Finlayson, "Reevaluation of color constancy algorithm performance," *J. Opt. Soc. Am. A*, vol. 23, no. 5, pp. 1008–1020, 2006.

[9] G. D. Finlayson, R. Zakizadeh, and A. Gijsenij, "The Reproduction Angular Error for Evaluating the Performance of Illuminant Estimation Algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1482 – 1488, 2016.

[10] J. T. Barron and Y. Tsai, "Fast Fourier Color Constancy," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[11] J. T. Barron, "Personal Communication." 2017.

[12] J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2207–2214, 2007.

[13] S. Bianco, C. Cusano, and R. Schettini, "Color constancy using CNNs," in *IEEE Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 81–89, 2015.

[14] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Generalized gamut mapping using image derivative structures for color constancy," *Int. J. Comput. Vis.*, vol. 86, no. 2, pp. 127–139, 2010.

[15] M. S. Drew and H. R. V. Joze, "Exemplar-Based Colour Constancy," in *Procedings Br. Mach. Vis. Conf.*, vol. 26, pp. 1–12, 2012.

[16] C. Rosenberg, T. Minka, and A. Ladsariya, "Bayesian Color Constancy with Non-Gaussian Models," *Adv. Neu-ral Inf. Process. Syst.*, vol. 16, 2003.

[17] K. Barnard, "Improvements to gamut mapping colour constancy algorithms," *Eur. Conf. Comput. Vis.*, vol. 1842, pp. 390–403, 2000.

[18] J. Van De Weijer, C. Schmid, and J. Verbeek, "Using high-level visual information for color constancy," in *IEEE Proc. Int. Conf. Comput. Vis.*, pp. 1–8, 2007.

[19] A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color constancy with spatio-spectral statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1509–1519, 2012.

[20] S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Automatic color constancy algorithm selection and combination," *Pattern Recognit.*, vol. 43, no. 3, pp. 695–705, 2010.

[21] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Inst.*, vol. 310, no. 1, pp. 1–26, 1980.