Comments

Providing a Single Ground-Truth for Illuminant Estimation for the ColorChecker Dataset

Ghalia Hemrit, Graham D. Finlayson, Arjan Gijsenij, Peter Gehler, Simone Bianco, Mark S. Drew, Brian Funt, and Lilong Shi

Abstract—The ColorChecker dataset is one of the most widely used image sets for evaluating and ranking illuminant estimation algorithms. However, this single set of images has at least 3 different sets of ground-truth (i.e., correct answers) associated with it. In the literature it is often asserted that one algorithm is better than another when the algorithms in question have been tuned and tested with the different ground-truths. In this short correspondence we present some of the background as to why the 3 existing ground-truths are different and go on to make a new single and recommended set of correct answers. Experiments reinforce the importance of this work in that we show that the total ordering of a set of algorithms may be reversed depending on whether we use the new or legacy ground-truth data.

Index Terms—Color constancy, illuminant estimation, algorithms evaluation

1 Introduction

Color constancy is the ability of a visual system to correct the light-color bias in rendered image colors. For digital cameras, this is known as white balancing, and is one of the processing functions in a camera pipeline. It uses an estimate of the illuminant color (predominant light in the scene). Various illuminant estimation algorithms exist and when a new algorithm is introduced it is important to evaluate its performance compared to existing ones. This is done by referring to benchmark datasets of images.

The ColorChecker dataset is a widely used benchmark dataset for illuminant estimation, introduced in 2008 by Gehler et al. [1]. It has 568 RGB images of indoor and outdoor scenes taken with 2 widely used cameras: the Canon 1Ds (86 images) and the Canon 5D (482 images). All recent and state of the art algorithms have been evaluated using this dataset (there are 23 of them on colorconstancy.com [2], a widely used comparison site for illuminant estimation research, hosting data and results). The ColorChecker images are of typical photographic scenes (including, people, landscapes and typical tourist-type photos). The ground-truth—that is, the correct answer—for each image is defined as the RGB response from achromatic surfaces placed in the scene (actually, the achromatic patches in the eponymous Macbeth ColorChecker). The ground-truth is not only used for illuminant estimation algorithms evaluation but also for the training stage of learning methods. In Fig. 1, we show one image from the ColorChecker dataset.

In contrast, many of the other previously proposed datasets comprise lab-based or technical images that do not correspond to images that are normally captured. Since the year 2010 the vast majority of experiments based on the ColorChecker dataset have used a linearised variant [3] (linear raw images are used).

This article follows up on the rather worrying discovery in [4] that there are at least 3 different sets of ground-truths for the ColorChecker dataset. Further, the difference between at least two of the ground-truths was found to be large. This is a serious problem for the field. Indeed, in reading articles about illumination estimation, it is common for authors to rank their latest approach against previous work. But, this makes no sense if different ground-truths are used. In [4] it was shown that the rank-order of any algorithm depends strongly on the ground-truth employed.

2 The 3 Current Ground-Truths

In [4], the three ground-truth correct answers (a 568x3 matrix of 568 white values, one for each image in the ColorChecker dataset), were labelled Gt1, Gt2 and SFU/Gt3. The last, SFU/Gt3, was calculated by Shi and Funt and is still accessible from the SFU web-site [3], while Gt1 and Gt2 are alternative ground-truths, though purportedly also from [3], which are found on colorconstancy.com [2]. In Fig. 2, we plot as red squares the chromaticities of the Gt1 ground-truth and as green squares the SFU/Gt3 chromaticities. The convex hulls in the same figure show that the 3 ground-truths distributions have both overlapping and disjoint areas. In calculating the SFU/Gt3 results the 'black level' is subtracted from the raw images: the dark current signal image, captured with the camera lens cap on, is one of the noise factors in the rendered scene image [5]. The 'black level' was not subtracted in Gt1. However, as per [6], we note that when Gt1 was calculated the SFU web-site instructions indicated the 'black level' had already been subtracted. Post-facto, the web-site text was changed to indicate the converse.

Clearly, the two sets of ground-truths are quite different from one another. This is a serious issue because although Gt1 is the most widely used ground-truth, the SFU/Gt3 ground-truth is more accurate. We also remark that the algorithms that use the SFU/Gt3 ground-truth are among the most recent and putatively among the best performers.

The ground-truth Gt2 is due to Bianco et al. [7], who comment that they found errors in how the ground-truth Gt1 was calculated: "*Using the one generated by Shi and Funt, we noticed that for some images the Macbeth ColorChecker coordinates (both the bounding box and the corners of each patch) were wrong and thus the illuminant ground-truth was wrong.* " We note that Gt1 and Gt2 are very close to each other (save for a few images) and so Gt2 is not plotted in Fig. 2.

3 Re-Calculating the Ground-Truth

We re-calculate the ground-truth using the methodology set forth by Shi and Funt (as described on their web-site [3] in 2018). We call this new ground-truth REC (for recommended). However, we get slightly different results than reported in [3]. The chromaticities of our ground-truth are shown in black in Fig. 2 and compared with

- *G. Hemrit and G. D. Finlayson are with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom. E-mail: {g.hemrit, g.finlayson}@uea.ac.uk.*
- *A. Gijsenij is with AkzoNobel, Amsterdam 1077, Netherlands. E-mail: arjan.gijsenij@gmail.com.*
- *P. Gehler is with Amazon, Seattle, WA 98109. E-mail: pgehler@amazon.com.*
- *S. Bianco is with University of Milan-Bicocca, Milano 20126, Italy. E-mail: simone.bianco@disco.unimib.it.*
- *M. S. Drew and B. Funt are with Simon Fraser University, Burnaby, BC V5A 1S6, Canada. E-mail: mark@cs.sfu.ca, funt@sfu.ca.*
- *L. Shi is with Samsung Semiconductor Inc., San Jose, CA 95134. E-mail: lilong10@gmail.com.*

Manuscript received 4 Dec. 2018; revised 3 Apr. 2019; accepted 17 May 2019. Date of publication 1 July 2019; date of current version 1 Apr. 2020.
(Corresponding author: Ghalia Hemrit.)
Recommended for acceptance by K. Nishino.
Digital Object Identifier no. 10.1109/TPAMI.2019.2919824

Fig. 1. An image from the Macbeth ColorChecker dataset.

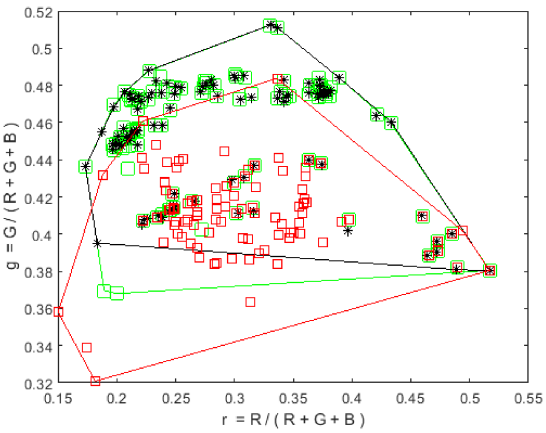

Fig. 2. New RECommended ground-truth chromaticities are plotted as black asterisks. The SFU/Gt3 ground-truth is shown as green squares and Gt1 as red squares. The distributions convex hulls are plotted. 100 of the 568 chromaticities are shown.

SFU/Gt3 in green. The differences are few but, visually, might be described as like SFU/Gt3 but with a few outliers removed.

The question of why our results differ is an important one. The reasons are partially supplied by Bianco et al. [7] and their derivation of Gt2 (see earlier quotation). Rather than visualising the differences as chromaticities they chose instead to measure the 3-vector angular difference between each of the 568 illuminant RGBs in Gt1 and Gt2. They add to their comment above,"*To have an idea of the differences, the maximum angular difference between Gt1 and Gt2 is around 20°, and the median one is around 0.03°*", which, again, speaks to a few 'outliers' being detected. We recall that Gt1 is derived from the same data as SFU/Gt3 without 'black level' subtracted (the same for Gt2).

We also found a problem with the calculation itself of SFU/Gt3 (for a small number of images). Specifically, we found that a saturation problem occurred in 3 images where the R, G and B of the light come from two (or three) different achromatic patches, e.g., from the white patch and the lightest grey-patch. We propose that the sensible way to resolve this ambiguity is to take all three measurements (R,G and B) from the brightest achromatic patch that has no saturated channel at all (no digital count in any channel > 3300).

4 Re-Evaluation of Illuminant Estimation Algorithms

We re-evaluated the 23 illuminant estimation algorithms on the ColorChecker output results available to us on [2]. We compared the calculated angular errors for the new REC ground-truth with the results of Gt1, Gt2 and SFU/Gt3. The angular error is a measure of the angle between the ground-truth illuminant color RGB vector and the estimate vector.

TABLE 1
The Performance of 6 Algorithms (see colorconstancy.com) are in Reverse Order in Terms of their Median Angular Error when the New REC vs the Legacy Gt1 Ground-Truth is Used

Algorithm	REC	Gt1
Edge-based Gamut ($\sigma = 4$)	3.27°	5.04°
2nd order Grey-Edge (p=1, $\sigma = 1$)	3.57°	4.44°
Bayesian	3.85°	3.46°
Using Natural Image statistics	4.70°	3.13°
Heavy Tailed-based Spatial Correlations	4.76°	2.96°
Bottom-Up	4.90°	2.56°

Note the Minkowski norm p and the smoothing value σ are the optimal parameters.

For the REC vs Gt1 (Gt1 is the most widely used but incorrect ground-truth) we show, in Table 1, the ranking by median angular error of 6 algorithms. This rank-ordering is in exact reverse order compared with using the new RECommended ground-truth. On colorconstancy.com we now make available the new REC ground-truth and the processed raw images, which can be used to evaluate illuminant estimation algorithms.

5 Conclusion

The widely used ColorChecker dataset has 3 ground-truth versions and two of them are very different from one another. In this short correspondence we provide an explanation as to why the 3 ground-truths are different and also why none of the 3 is completely accurate. We then adopt the methodology of [3] (but using our own code) to make a new recommended set of ground-truth which we make available to the community.

Acknowledgments

This research was supported by EPSRC Grant M001768; UEA grant EP/P007910/1; and Apple Inc.

References

[1] P. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[2] A. Gijsenij and T. Gevers, "Datasets and results per datasets," Oct. 2011. [Online]. Available: www.colorconstancy.com/evaluation/, Accessed on: Jan. 2018.
[3] L. Shi and B. Funt, "Re-processed version of the gehler color constancy dataset of 568 images," Mar. 2016. [Online]. Available: www.cs.sfu.ca/_colour/data/shi_gehler/, Accessed on: Mar. 2018.
[4] G. D. Finlayson, G. Hemrit, A. Gijsenij, and P. Gehler, "A curious problem with using the colour checker dataset for illuminant estimation," in *Proc. Color Imaging Conf.*, 2017, pp. 64–69.
[5] R. Ramanath, W. Snyder, Y. Yoo, and M. S. Drew, "Color image processing pipeline in digital still cameras," *IEEE Signal Process.*, vol. 22, no. 1, pp. 34–43, Jan. 2005.
[6] G. Hemrit, G. D. Finlayson, A. Gijsenij, P. Gehler, S. Bianco, B. Funt, M. S. Drew, and L. Shi, "Rehabilitating the colorchecker dataset for illuminant estimation," in *Proc. Color Imag. Conf.*, 2018, pp. 350–353.
[7] S. Bianco, "A curious problem with using the colour checker dataset for illuminant estimation," *Personal Communication*, Jan. 12, 2018.