View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of East Anglia digital repository

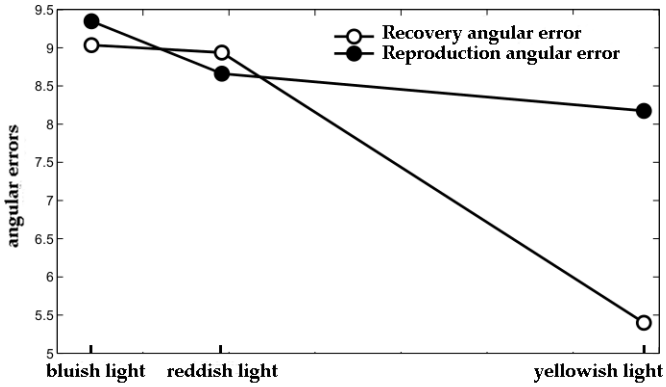
The Reproduction Angular Error for Evaluating the Performance of Illuminant Estimation Algorithms

Graham D. Finlayson, Roshanak Zakizadeh and Arjan Gijsenij

Abstract—The angle between the RGBs of the measured illuminant and estimated illuminant colors - the recovery angular error - has been used to evaluate the performance of the illuminant estimation algorithms. However we noticed that this metric is not in line with how the illuminant estimates are used. Normally, the illuminant estimates are 'divided out' from the image to, hopefully, provide image colors that are not confounded by the color of the light. However, even though the same reproduction results the same scene might have a large range of recovery errors. In this work the scale of the problem with the recovery error is quantified. Next we propose a new metric for evaluating illuminant estimation algorithms, called the reproduction angular error, which is defined as the angle between the RGB of a white surface when the actual and estimated illuminations are 'divided out'. Our new metric ties algorithm performance to how the illuminant estimates are used. For a given algorithm, adopting the new reproduction angular error leads to different optimal parameters. Further the ranked list of best to worst algorithms changes when the reproduction angular is used. The importance of using an appropriate performance metric is established.

Index Terms—Illuminant estimation, color constancy, performance evaluation, error metric.

(a)



(b)

Fig. 1. An example of similar color corrected images with varying recovery angular error. (a) First row: images of the same scene captured under chromatic illuminants (from SFU dataset [1]). Second row: corrected images using grey-world algorithm [8]. (b) Recovery versus Reproduction angular errors.

1 Introduction

Wherever colors are used as stable cues for a vision task, we wish to avoid any color bias due to illumination. To mitigate this problem, illuminant estimation algorithms infer the color of the light. Then, at a second stage the light color is removed (divided out) from the image. If an illuminant estimate is accurate then any color bias due to illumination is removed. The question of which algorithm works best is a key concern not only for those designing the algorithms, but also for those using them.

To measure the performance of an illuminant estimation algorithm, usually a set of images is agreed on as a benchmark (e.g. SFU Lab [1], Gehler-Shi colorchecker [2], [3], greyball [4], NUS [5] datasets, etc.). The RGB of the estimated light is then compared with a ground-truth measured illuminant. The *recovery* angular error - the angle between the RGBs of the actual and estimated lights - is often used to quantify the illuminant estimation error [6], [7]:

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

where $\underline{\rho}^{E}$ denotes the RGB of the actual measured light, $\underline{\rho}^{Est}$ denotes the RGB estimated by an illuminant estimation algorithm and '.' denotes the vector dot product. Over a data set summary statistics such as the average, median and quantile angular errors are calculated and algorithms are ranked according to these statistics.

In this paper, we show that the recovery angular error has a fundamental weakness. A visual illustration of the problem with recovery angular error is shown in Fig. 1. In the top row of Fig. 1, three images of the same scene form the SFU Lab dataset [1] are shown, which were captured under different chromatic illuminations. The RGB color of the illuminant for each scene is then estimated using the simple grey-world algorithm [8] and then we divide the image RGBs by this estimate to remove the color bias due to the illumination. The results of 'dividing out' are shown in the second row of the same figure. In part (b) of Fig. 1 we plot the recovery angular errors (open circles). Counter-intuitively, even though the output reproductions are similar the recovery error ranges from 5.5 to 9.5 degrees.

In this paper, we introduce a new illuminant estimation error metric, which we call **reproduction angular error**. Reproduction angular error measures the angle between the reproduction of a true achromatic surface under a white light ($[1\ 1\ 1]^t$) with the actual reproduction of an achromatic surface when an estimated illuminant color is divided out. The reproduction error is tied to how illuminant estimations are used and by design gives a similar error for the same scene reproduction, regardless of the illuminant color. In Fig. 1 (b) we show the reproduction errors (filled circles) and see they are much more stable than the recovery errors.

Our paper begins by calculating how large and small

G. D. Finlayson and R. Zakizadeh are with the School of Computing Sciences, University of East Anglia, Norwich, UK. (e-mail: g.finlayson@uea.ac.uk; r.zakizadeh@uea.ac.uk).

Arjan Gijsenij is with Akzo Nobel Decorative Coatings, Sassenheim, Netherlands. (e-mail: arjan.gijsenij@gmail.com)

Manuscript received 29 Oct. 2015; revised 27 May 2015.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2016.2582171

the mismatch between recovery errors and the images reproduced can be. We adopt the so-called diagonal model of illuminant change [9] and then, relative to this assumption, we solve for the illuminants that respectively induce the maximum and minimum recovery angular errors. We show that red, green and blue 'pure' lights lead to 0 errors. *Cyan*, *yellow* and *magenta* lights induce maximum error.

In order to observe the effect that the choice of error metric has on the ranking of algorithms and on their evaluation, we re-evaluated a large number of illuminant estimation algorithms over multiple benchmark datasets such as: SFU Lab [1], Gehler-Shi colorchecker [2], [3] and Grey-ball [4], using both recovery and the proposed reproduction angular errors. We have also evaluated a set of algorithms on the National University of Singapore [5] dataset.

In Section 2, we discuss illuminant estimation. In Section 3, the recovery angular error is presented and its range of variation is determined for a given illuminant estimation algorithm and a given scene. We present the reproduction angular error in Section 4. Section 5 discusses the evaluation of a large number of illuminant estimation algorithms. We summarize the paper in Section 6.

2 Illuminant Estimation

A simple model of image formation [10] that we often use when discussing illuminant estimation is given in (2).

$$\rho_k^{E,S} = \int_\omega R_k(\lambda)E(\lambda)S(\lambda)d\lambda \quad k \in \{R,G,B\}. \tag{2}$$

Here $\rho_k^{E,S}$ is the integrated response of a sensor to light and surface. There are R, G and B sensor channels. The spectral power distribution illuminating a scene is denoted as $E(\lambda)$, $S(\lambda)$ is the surface spectral reflectance and the light reflected is proportional to the multiplication of the two functions. The light is then sampled by a sensor with a spectral sensitivity $R(\lambda)$ and integrated over the visible spectrum ω.

Almost all illuminant estimation algorithms solve for the R, G and B responses for the illuminant which is defined as:

$$\rho_k^E = \int_\omega E(\lambda)R_k(\lambda)d\lambda. \tag{3}$$

Similarly we might write the surface response as:

$$\rho_k^S = \int_\omega S(\lambda)R_k(\lambda)d\lambda. \tag{4}$$

The response to light and surface together can be calculated as:

$$\rho_k^{E,S} = \rho_k^E \rho_k^S. \tag{5}$$

Assuming (5) holds and assuming an illuminant estimation algorithm provides a reasonable estimate of the light color ($\underline{\rho}^{Est}$), then we solve for ρ_k^S (remove color bias due to illumination), by dividing out:

$$\frac{\underline{\rho}^{E,S}}{\underline{\rho}^{Est}} \approx \underline{\rho}^S, \tag{6}$$

where the division of the vectors is component-wise.

An unknown light E' can be simulated by multiplying the actual light E by a 3-vector $\underline{d}$:

$$\underline{\rho}^{E',S} = \underline{d} * \underline{\rho}^{E,S} \qquad \underline{d} = [\alpha \ \ \beta \ \ \gamma]^t \quad \alpha,\beta,\gamma \geq 0 \tag{7}$$

Now let us assume that the illuminant of the scene is estimated as a statistical moment of the image RGB values for an N-pixel image. We write:

$$\underline{\rho}^{Est} = moment(\{\underline{\rho}^{E,S_1}, \underline{\rho}^{E,S_2}, ..., \underline{\rho}^{E,S_N}\}). \tag{8}$$

Combining (7) and (8):

$$\underline{d} * \underline{\rho}^{Est} = moment(\{\underline{\rho}^{E',S_1}, \underline{\rho}^{E',S_2}, ..., \underline{\rho}^{E',S_N}\}). \tag{9}$$

Equation (9) teaches that if two lights are related by 3 scaling factors $\underline{d}$ then the statistical moment estimates shift by the same scaling factors. Equation (9) is true for most illuminant estimation algorithms including all those that can be written in the Minkowski-framework [11]:

$$\left(\int |\frac{\delta^n \underline{\rho}(x)}{\delta x^n}|^p \mathrm{d}x\right)^{1/p} = k\underline{\rho}^{Est}_{n,p,\sigma}. \tag{10}$$

Here the 3-vector $\underline{\rho}(x)$ is the camera response at location x of an RGB image. The image can be smoothed with a Gaussian averaging filter with standard deviation σ pixels and then is differentiated with an order n differential operator. We then take the absolute Minkowski p-norm average [12] over the whole image. The unknown value k represents the fact that it is not possible to recover the true magnitude of the illuminants. The σ and p-norm are the tunable parameters which can be chosen so that the algorithms perform their best. The grey-world, MaxRGB [13] and grey-edge [11] algorithms are all instantiations of the Minkowski-framework.

For a full survey of illuminant estimation algorithms the reader is referred to [14].

3 The Range of Recovery Angular Error

Assuming the diagonal model of illumination change we show how to solve for the illuminant that results in the largest recovery angular error.

Theorem 1. Given a white reference light (the RGB of the light is $\underline{U} = [1 \ 1 \ 1]^t$) and denoting the illumination estimate made by a 'moment type' illuminant estimation algorithm as $\underline{\mu} = [\mu_r \ \mu_g \ \mu_b]^t$ then the illuminant that maximizes recovery angular error is an illuminant with 0 in exactly one of the either R, G or B channels.

From Theorem 1 and because the recovery error is intensity independent we can, without loss of generality, set one of the illuminant parameters to 1 and another to 0.

Lemma 1.1. Assuming $d_i = 1$ and $d_j = 0$ then $d_k = \mu_i/\mu_j$ (where $i \neq j \neq k$).

In other words, Theorem 1 and Lemma 1.1 combined state lights which are cyan, magenta and yellow maximize the recovery angular error. Conversely, pure red, green and blue lights result in the lowest angular error. In the limit lights which have two of the components tending towards 0 will, for all moment-type algorithms, result in a recovery angular error which tends towards 0. For a complete proof of Theorem 1 and Lemma 1.1 we refer the reader to [15].

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2016.2582171

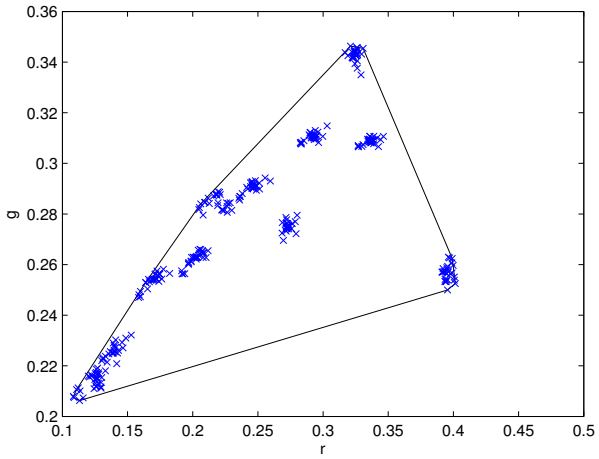


Fig. 2. 2D chromaticity gamut (solid line) bounding the set of SFU Lab dataset's measured illuminants [1].

3.1 Maximum recovery angular error for real lights

In reality, lights that induce a 0-response in the R, G or B channels are almost never encountered. This raises the question of whether we can revise Theorem 1 to cover more likely illuminants. Given that real lights are bounded to a restricted gamut area, what can we say about the range of recovery angular error? In Fig. 2 we plot on a rg chromaticity diagram the chromaticities of the lights from the SFU Lab dataset [1] (where [r,g,1-r-g] has the same orientation as the RGB of the light). Notice that the range of lights is really quite restricted and is far from allowing either pure red, green and blue lights or pure cyan, magenta or yellow. Our second theorem teaches where local maxima should lie when lights lie in a bounded region of chromaticity space.

Theorem 2. The maximum recovery angular error for a convex combination of a set of measured lights, belongs to a light which falls on the border of the convex set.

Proof: According to Theorem 1, for a given image and a given illuminant estimation algorithm, there are - when there are no restrictions on the color of the illuminant - three possible lights that result in local error maxima (one of which induces the overall maximum error). Further all three local maxima have one of R, G or B equal to 0. Let us assume now that for the restricted illuminant case - lights must lie within a convex region - that the light that induces the maximum error does not lie on the boundary of the convex set. As a consequence this light must be a local maximum. Further because this is an interior point of the set of illuminants all three components, R, G and B must be non-zero. It also follows that this illuminant must also be a local maximum even when the constraint on where the illuminant can lie is removed. By Theorem 1 this cannot be the case because all local maxima for the unrestricted case have one component of the RGB vector equal to 0. We have a contradiction and so the maximum error for a constrained convex set of lights must be on the boundary of the set. □

Theorem 2 is important because it teaches that we can find the light resulting in the maximum recovery angular error, belonging to a set of feasible lights, by searching the boundary of the feasible set.

4 Reproduction Angular Error: An Improvement over Recovery Angular Error

In very simple words, reproduction angular error is the angle between true white and estimated white (white surface under unknown light mapped to reference light using an illuminant estimate.). Remembering we cannot recover the absolute brightness of the light, we define the **Reproduction Angular Error** [15] - our new metric for assessing illuminant estimation algorithms - as:

$$err_{reproduction} = \cos^{-1}\left(\underline{w}^{Est}.\underline{w}\right), \tag{11}$$

where $\underline{w}^{Est} = \frac{\underline{\rho}^E/\underline{\rho}^{Est}}{|\underline{\rho}^E/\underline{\rho}^{Est}|}$ (reproduced color of white surface) and $\underline{w} = \frac{\underline{\rho}^E/\underline{\rho}^E}{\sqrt{3}}$ (true color of white surface).
According to the RGB model of image formation in Section 2, the RGB values in the image are scaled by the same three weighting factors as the illumination changes [16]. The reproduced image after color correction, is the image from which the estimated illuminant is 'divided out', so that the color bias due to illumination is removed. The color bias is removed from the images as is explained by (6):

$$\frac{\underline{\rho}^E}{\underline{\rho}^{Est}} \approx \underline{U} = \frac{\underline{\rho}^E}{\underline{\rho}^E}. \tag{12}$$

Theorem 3. Given a single scene viewed, separately, under two lights. The reproduction error of the estimated light by a 'moment type' illuminant estimation algorithm is the same.

Proof: For a chromatic light defined with $\underline{d} = [\alpha\ \beta\ \gamma]^t$ [see (7)], using the fact presented in (8), the reproduction angular error (11) can be written as:

$$err_{reproduction} = \cos^{-1}\frac{\left(\frac{\alpha}{\alpha\mu_r} + \frac{\beta}{\beta\mu_g} + \frac{\gamma}{\gamma\mu_b}\right)}{\sqrt{\left(\frac{\alpha}{\alpha\mu_r}\right)^2 + \left(\frac{\beta}{\beta\mu_g}\right)^2 + \left(\frac{\gamma}{\gamma\mu_b}\right)^2}} * \frac{1}{\sqrt{3}}. \tag{13}$$

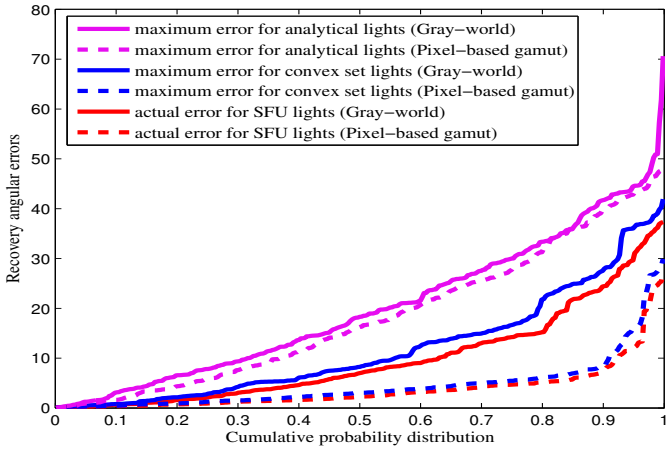
It can be seen easily in (13), that the scaling factors α, β and γ (which caused the illumination changes) cancel. The reproduction error is stable regardless of the color of the light. □
In Fig. 3(a), the two purple curves are the cumulative probability distribution functions of the analytical maximum recovery errors for the two algorithms: grey-world [8] (solid line) and pixel-based gamut mapping [17] (dashed line) algorithms for 321 images of SFU Lab dataset.

The blue curves represent the cumulative probability functions of the maximum recovery angular errors for an example of the real lights (see Theorem 2.) (in this case these lights are within the convex combination of the measured illuminants of SFU Lab dataset [1]). The red curves in the same figure are the actual recovery angular errors of the estimated illuminant using the two grey-world [8] (solid line) and pixel-based gamut mapping [17] (dashed line) algorithms applied on SFU Lab dataset.

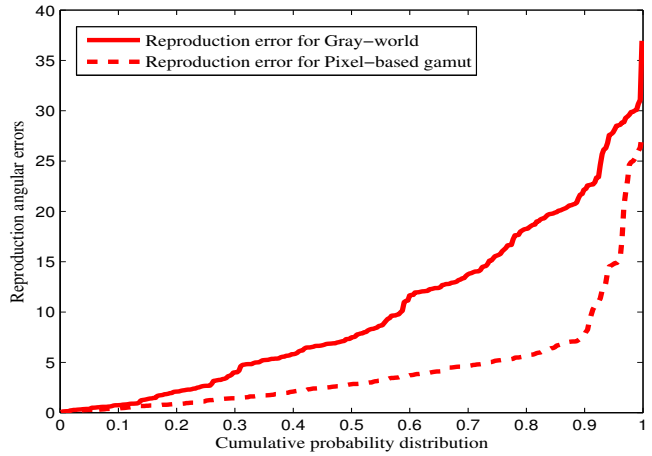
In terms of the maximum angular error Fig. 3 (a) teaches that grey-world, in the worst case, performs about the same as gamut mapping. This is a surprising result as gamut mapping is a much more complex algorithm and is assumed to perform better. Note also that for real lights the worst case error is still worse for grey-world but the worst-case

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2016.2582171



(a)



(b)

Fig. 3. (a) Cumulative probability distribution function of analytical maximum recovery angular errors (in magenta), maximum error of real lights within the convex of SFU Lab dataset's [1] measured illuminants (in blue) and the recovery angular errors of the estimated lights of 321 SFU Lab images using the two algorithms (in red). (b) Cumulative probability distribution function of maximum reproduction angular errors [15]

for pixel based gamut mapping is similar to the actual performance (though, still significantly different especially for the higher quantile errors).

In Fig. 3 (b) we show the reproduction angular error for grey-world and pixel-based gamut mapping. This error is stable across illumination changes. Fig. 3 (b) informs us - what we knew - that for all lights pixel-based gamut mapping works better than grey-world.

Another way of articulating the benefits of reproduction error is that it is 'skew' invariant. That is under the diagonal (or indeed linear) model of illuminant change the colors 'skew' from one light to another. Other formulations 'diagonal' skew invariants can be made e.g. $\underline{\rho}^{Est}/\underline{\rho}^{E}$ or $norm(log(\underline{\rho}^{Est}/\underline{\rho}^{E}))$. However, the former effectively measures the reproduction error assuming the illuminant is actually what we estimated and is corrected with - for the purposes of this example - the 'wrong' actual light. The normal - as oppose to this inverse - reproduction error makes more sense. The latter skew invariant measure is derived from the normal reproduction error. We remark that any function of the reproduction error will also be skew invariant. And as it follows from the derivation of reproduction angular error (see Theorem 3) that the same scene+algorithm pair will return the same reproduction error for all lights if the algorithm is skew invariant or so-called 'moment-based' i.e. if the illuminant change is modeled by the diagonal matrix $\underline{d}$ the moment type estimate also maps by the same i.e. $\underline{d} * \underline{\rho}$. An avenue for future research is to assess how well reproduction error - and other skew invariants - correlate to judgements made by human observers.

The relative performance of different algorithms based on reproduction and recovery angular error with a more realistic case study has also been discussed in [18].

4.1 The Reproduction Error for a non-diagonal illuminant model

The efficacy of a diagonal model of illuminant change is strongly related to the spectral shape of the sensors. The more bandlimited, or narrow, the sensitivities the more applicable the diagonal model. The majority of commercial photographic cameras have narrow band sensors and, to our knowledge, the illuminant is discounted by applying the diagonal model. However, there are exceptions such as the Sigma range of sensors where their X3 sensing technology [19] results in broad sensitivities. Thus, it is an interesting question to consider whether reproduction angular error can be applied more widely.

First we note that even when a diagonal model of illuminant change does not hold it can often be made to hold via a change in sensor basis. With respect to this new sensor basis [20], [21] the reproduction error can be used directly.

More generally, an illuminant estimate can be used to parametrize a 3×3 correction matrix [22]. For example, given finite dimensional approximation of light and surfaces when given estimated RGB of light $\underline{\rho}^{Est}$ the function $\mathcal{M}(\underline{\rho}^{Est})$ returns a 3×3 matrix which maps image colors - where the illuminant is $\underline{\rho}^{Est}$ - to a reference $[1\ 1\ 1]$ e.g. [10]. That is we substitute $\underline{w}^{Est} = \mathcal{M}(\underline{\rho}^{Est})\underline{\rho}^{E}$ into (11). In fact we can be more general still. In [23], Forsyth introduces the function $\psi(\underline{\rho};\underline{\rho}^{Est})$ the meaning of which is the RGB $\underline{\rho}$ mapped to a reference lighting condition using the light estimation ρ^{Est}. Adopting this idea we can substitute $\underline{w}^{Est} = \psi(\underline{\rho}^{E};\underline{\rho}^{Est})$ into (11) and so arrive at even more general form of reproduction error.

Reproduction error is generalized to encompass more reflectances in [24], [25]. Importantly, [24] found that simple reproduction angular error could be used as a proxy for calculation based on many reflectance.

5 EXPERIMENTS

Gijsenij *et al.* [14] carried out a comprehensive evaluation of a large selection of illuminant estimation algorithms using recovery angular error. In this section we revisit their experiments for the SFU Lab dataset [1]. The SFU data has 30 objects under up to 11 lights. This makes it ideal for our purpose because for these lights reproduction error should be similar but recovery error will vary. We also wish to consider illuminant estimation performance for the recent NUS dataset [5] which comprises a large set of typical photographic pictures captured with a wide range

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2016.2582171

TABLE 1
Recovery and Reproduction errors in terms of median and 95% quantile for several color constancy algorithms applied on SFU dataset [1]. The ranks for some algorithms have changed based on the two error calculations. There are also changes in the optimal parameters.

	Recovery error								Reproduction Error							
Method	p	σ	Median	Rank	p	σ	95%	Rank	p	σ	Median	Rank	p	σ	95%	Rank
Grey-world	-	-	7°	11	-	-	30.3°	11	-	-	7.5°	11	-	-	28°	11
MaxRGB	-	-	6.5°	10	-	-	27.2°	10	-	-	7.4°	10	-	-	27.2°	10
Shades of grey	7	-	3.7°	**9**	4	-	18.7°	**9**	7	-	3.9°	**8**	3	-	19°	**8**
1^{st} order grey-edge	7	4	3.2°	**7**	2	1	14.3°	6	14	4	3.58°	**6**	2	1	15.6°	6
2^{nd} order grey-edge	14	10	2.7°	4	2	2	14.2°	5	15	10	3°	4	2	2	15.1°	5
Pixel-based gamut [17]	-	4	2.26°	**2**	-	6	9.8°	**1**	-	4	2.8°	**3**	-	7	11.1°	**1**
Edge-based gamut	-	2	2.27°	**3**	-	2	12.6°	**3**	-	2	2.7°	**2**	-	2	14.3°	**4**
Inter-based gamut	-	4	2.1°	1	-	6	9.8°	**1**	-	3	2.5°	1	-	7	11.2°	**2**
Union-based gamut	-	2	3°	5	-	3	12.8°	**4**	-	2	3.4°	5	-	3	13.2°	**3**
Heavy tailed-based [26]	-	-	3.5°	**8**	-	-	15.9°	7	-	-	4.1°	**9**	-	-	16.6°	7
Weighted grey-edge	2	1	3.1°	**6**	2	1	18°	**8**	2	1	3.62°	**7**	2	1	19.3°	**9**

TABLE 2
Recovery and Reproduction errors in terms of max and 95% quantile for several algorithms applied on Canon1D camera from NUS dataset [5].

	Recovery error								Reproduction Error							
Method	p	σ	Max	Rank	p	σ	95%	Rank	p	σ	Max	Rank	p	σ	95%	Rank
Grey-world	-	-	22.37°	**5**	-	-	12.78°	4	-	-	24.69°	**4**	-	-	16.19°	4
MaxRGB	-	-	39.12°	**7**	-	-	17.28°	**7**	-	-	33.76°	**6**	-	-	18.14°	**6**
Shades of grey	5	-	14.62°	**2**	5	-	9.01°	**1**	5	-	18.41°	**3**	8	-	11.71°	**2**
1^{st} order grey-edge	7	9	14.08°	1	7	2	9.09°	**2**	5	3	17.35°	1	9	2	11.50°	**1**
2^{nd} order grey-edge	4	10	15.00°	**3**	3	5	9.12°	3	5	4	17.91°	**2**	1	2	12.09°	3
Pixel-based gamut	-	0	38.60°	**6**	-	0	16.64°	**6**	-	0	35.52°	**7**	-	0	18.45°	**7**
Edge-based gamut	-	5	21.64°	**4**	-	3	13.01°	5	-	5	27.60°	**5**	-	3	16.37°	5

of commercial cameras. Here we do not have access to the whole set of algorithms used in the original study [14] by Gijsenij et al. (indeed, the performances supplied there for the datasets were contributed by many authors (including the recent methods [27], [28], [29], [30]) i.e. there is not a complete code repository). So, for the NUS dataset we evaluate the Minkowski family of algorithms (10) as well as pixel-based and edge-based gamut mapping [17], [23].

Table 1 reports the recovery and reproduction median and 95% quantile angular errors for the SFU Lab dataset [1]. The SFU Lab dataset comprises a set of 321 images captured under relatively chromatic lights (we adopt the same algorithm naming conventions used by Gijsenij [14]). The p and σ values shown in the table (the tunable parameters (see (10))) provide the lowest angular error for an illuminant estimation algorithm and the error statistic being used (e.g. median or 95% quantile). Notice that using recovery vs reproduction angular error and median vs 95% quantile we end up with different optimal p and σ values.

For each of the four test scenarios (Recovery vs Angular error for the median and 95% quantile statistic) we also show the rank of the different algorithms. We remark that it is possible for two algorithms, to the precision tested, to have the same performance (according to the median or 95% quantile) and so these algorithms will have the same rank. In bold and underlined we highlight the algorithms whose ranks change. Here we compare the performance measured according to the same statistical measure but for the recovery vs reproduction angular error. That is, we compare the ranks of the 1^{st} and 3^{rd} columns and the 2^{nd} and 4^{th} columns (respectively, the median angular error and 95% quantile). These highlighted rank changes also include the case where two algorithms have delivered the same performance for one error metric (and are assigned the same rank) but different for the other metric. We highlight one occasion where it happens below.

Table 2 reports the recovery and reproduction max and 95% quantile angular errors for NUS dataset [5] which consists of 1736 images from 8 different cameras. Here we are reporting the results for one of the cameras, Canon1D.

Looking at Table 1 and Table 2, we make two observations. Firstly, using reproduction angular error there are clearly changes in the ranking of algorithms. Although the overall ranking of illuminant estimation algorithms remains similar (e.g. gamut mapping algorithms still perform the best for the SFU dataset), but the local rank of different algorithms can swap. For example, based on median errors, the pixel-based gamut-mapping algorithm is better than the derivative-based counterpart for the SFU dataset for the recovery angular error but the converse is true when the reproduction angular error is used. We also notice the tunable parameters for an algorithm can change if the reproduction angular error is used for evaluation of the algorithm.

The Kendall's test statistic T [31] can give us a measure of correlation between pairs of ranks. A pair of unique observations (x_1, y_1) and (x_2, y_2) are said to be discordant if the ranks of the two elements (x_1, x_2) and (y_1, y_2) do not

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2016.2582171

TABLE 3
Changes in ranking of algorithms for SFU Lab dataset [1] (based on median errors).

Method	Median			
	Reproduction Rank	Recovery Rank	C	D
Edge-based gamut	1	2	4	1
Pixel-based gamut	2	1	4	0
1^{st} order grey-edge	3	4	2	1
Weighted grey-edge	4	3	2	0
shades of grey	5	6	0	1
Heavy tailed-based	6	5	0	0
T quantile for 6 samples at 99.5% confidence = 13			>(T = 9)	

TABLE 4
Changes in ranking of algorithms for SFU Lab dataset [1] (based on 95% quantile errors).

Method	95% quantile			
	Reproduction Rank	Recovery Rank	C	D
Pixel-based gamut	1	1	4.5	0.5
Inter-based gamut	2	1	4	0
Union-based gamut	3	4	2	1
Edge-based gamut	4	3	2	0
shades of grey	5	6	0	1
Weighted grey-edge	6	5	0	0
T quantile for 6 samples at 99.5% confidence = 13			>(T = 10)	

agree, otherwise the pair are concordant. T is defined as:

$$T = C - D, \tag{14}$$

where C is the number of concordant pairs and D is the number of discordant pairs. If $y_1 = y_2$ while $x_1 \neq x_2$ we call it a tie. In case of a tie the pair is counted as $1/2$ concordant and $1/2$ discordant, although as it is obvious by (14), this makes no difference in our final Kendall's T value.

To study the discordance in ranking of the algorithms, we perform the Lower-Tailed Kendall's Test [31], which is defined as follows:

Lower-Tailed Test

H_0 : X and Y are independent. This means the pairs of data are neither discordant nor concordant.

H_1 : Pairs of data tend to be discordant.

Reject null hypothesis (H_0) at $\alpha\%$ confidence level if T is less than its quantile at this confidence level in the null distribution. The T quantile at different confidence levels for $n \leq 60$ can be looked up in table of the quantiles for the Kendall's test in [31]. For instance, if the null hypothesis (H_0) is rejected at 95%, this means we can say that the pairs of data tend to be discordant with 95% confidence. □

We are interested in measuring the discordancy (or otherwise) for the algorithms whose ranks change. The number of algorithms where the ranks change depends both on the error measure used (median or 95% quantile) and the dataset (SFU Lab and NUS). Thus we measure concordant and discordant pairs for 6, 6, 6 and 4 algorithms for respectively the error measure and dataset pairs: (median, SFU Lab), (95% quantile, SFU Lab), (max, NUS) and (95% quantile, NUS). Respectively, the data for these pairs are recorded in Tables 3 through 6. Breaking down the calculations, for instance in Table 3 (median error and for the SFU Lab dataset), in total there are 12 concordant and 3 discordant pairs of ranking which result in $T = 12-3 = 9$. This T value is then compared with its quantile, which in this case is 13 at 99.5 $\%$ confidence level. Based on the comparison made, the null hypothesis (H_0) in the Lower-Tailed Kendall's test is rejected and it concludes that the pairs tend to be discordant. That is the ranks of the algorithms are significantly different. Similarly, for Table 4 (SFU Lab dataset and the 95% quantile error) we find the 6 algorithms (whose ranks change) are ranked differently (at a 99.5% confidence level). In this table, the algorithms with the same rank given based on 95% quantile recovery error are also included.

TABLE 5
Changes in ranking of algorithms for Canon1D camera from NUS dataset [5] (based on max errors).

Method	Max			
	Reproduction Rank	Recovery Rank	C	D
2^{nd} order grey-edge	1	2	4	1
Shades of grey	2	1	4	0
Grey-world	3	4	2	1
Edge-based gamut	4	3	2	0
MaxRGB	5	6	0	1
Pixel-based gamut	6	5	0	0
T quantile for 6 samples at 99.5% confidence = 13			>(T = 9)	

TABLE 6
Changes in ranking of algorithms for Canon1D camera from NUS dataset [5] (based on 95% quantile errors).

Method	95% quantile			
	Reproduction Rank	Recovery Rank	C	D
1^{st} order grey-edge	1	2	2	1
shades of grey	2	1	2	0
MaxRGB	3	4	0	1
Pixel-based gamut	4	3	0	0
T quantile for 4 samples at 99.5% confidence = 6			>(T = 2)	

Tables 5 and 6 report the ranking performance for the NUS Canon1D dataset [4] from Table 2 where again we focus only on the algorithms whose ranks change. We wish to measure how much the ranks change. Again the algorithms in these two tables have changed in their ranking orders when they were ranked using median and 95% quantile reproduction angular errors respectively (see Table 2).

It can be seen that the null hypothesis (H_0) in Lower-Tailed Kendall's test is rejected for all pairs of algorithms in Tables 3 to 6, showing the fact that the ranking of these algorithms using recovery and reproduction angular errors are strongly discordant. A pictorial scheme of Kendall's test in Table 3 is shown in Fig. 4. It is interesting to notice that according to recovery errors in this case edge-based gamut mapping algorithm is followed immediately by weighted grey-edge. Whereas, based on reproduction errors they are two steps apart in the ranking table.

To further study the behaviour of two metrics on individual images we performed the Wilcoxon sign test [31]

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2016.2582171

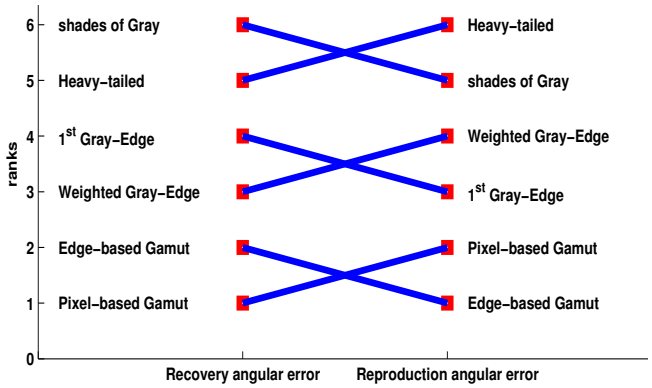


Fig. 4. The pictorial scheme of Kendall test for the changed rank algorithms in Table 3 [15].

TABLE 7
Wilcoxon sign test on SFU dataset for Recovery and Reproduction errors of the algorithms in Table 3.

	Recovery error						Reproduction error					
	1. Edge-based gamut	2.Pixel-based gamut	3.1st grey-edge	4.weighted grey-edge	5.shades of grey	6.Heavy tailed-based	1. Edge-based gamut	2.Pixel-based gamut	3.1st grey-edge	4.weighted grey-edge	5.shades of grey	6.Heavy tailed-based
1	0	-1	+1	+1	+1	+1	0	+1	+1	+1	+1	+1
2	+1	0	+1	+1	+1	+1	-1	0	+1	+1	+1	+1
3	-1	-1	0	-1	+1	0	-1	-1	0	+1	+1	+1
4	-1	-1	+1	0	+1	0	-1	-1	-1	0	+1	+1
5	-1	-1	-1	-1	0	0	-1	-1	-1	-1	0	0
6	-1	-1	0	0	0	0	-1	-1	-1	-1	0	0

which allows us to show the statistically significance of the difference between two algorithms [6]. In the Wilcoxon sign test we can test the hypothesis that the median of algorithm i is significantly lower than the median of algorithm j at some confidence level.

The Wilcoxon sign test results for the algorithms in Table 3 applied on SFU dataset are shown in Table 7. Here, a positive value (green color) at location (i, j) indicates that the median of algorithm i is significantly lower than the median of algorithm j at the 90% confidence level. For such a small set of objects (SFU set has 30 objects) 90% confidence level is reasonable. The value (-1) (red color) indicates the opposite and a zero (yellow color) shows there is no significant difference between the performance of two algorithms. As can be seen there are cases where reproduction angular error interprets the significance of difference between performance of two methods differently from recovery angular error. For instance based on recovery error there isn't much difference between the performance of Heavy tailed-based and 1^{st} grey-edge but for reproduction error they are different. Or in case of 1^{st} order grey-edge and weighted grey-edge methods there is a complete switch between the ranking of two algorithms. In summary, the Wilcoxon sign test demonstrates that for images where state of the art illuminant estimation algorithms performed reasonably the recovery and reproduction errors ranked these algorithms differently.

That the new reproduction angular error ranks algorithms differently is a matter of considerable importance. Indeed, not only do the *absolute* values change with respect to the currently used recovery angular error, the *relative* differences between the algorithms (the rank order of algorithms) change as well. Especially this latter observation is an important argument in favor of switching to the new reproduction error instead of keep using the legacy recovery error.

After all, if we wish to recognize colorful content independent of the illuminant color (i.e. we first remove the color bias due to illumination by dividing out the illuminant color [32]) then we need to adopt the new reproduction angular error to measure the performance. More generally, if illuminant estimates are used to discount color casts - this is by far the main reason for estimating the illumination - from images due to the prevailing illuminant color (for recognition, tracking or navigation) then the new metric should be used.

5.1 Multispectral Illuminant Estimation

Considering that illuminant estimation is the preprocessing step to many computer vision tasks which mostly make use of 3-band RGB images, most of our analysis have been done on such benchmark datasets. However, one might find the difference between recovery and reproduction angular errors on a set of multispectral data applicable. Here we repeat the same experiment on the images from Foster et al. dataset [33]. The dataset consists of eight scenes captured by a progressive-scanning monochrome digital camera. The data is provided between 410 and 710 nm with 10 nm intervals. We have assumed the lighting condition to be under 6500 k illuminant. The recovery and reproduction errors for four illuminant estimation algorithms applied on the five of these 31-band images are presented in Table 8.

TABLE 8
Changes in ranking of algorithms for Foster et al. dataset [33] (based on median errors).

	Recovery		Reproduction			
Method	Median error	Rank	Median error	Rank	C	D
1^{st} order grey-edge	7.18	1	7.50	2	2	1
2^{st} order grey-edge	7.23	2	7.73	4	0	2
General grey world	7.26	3	6.08	1	1	0
Shades of grey	7.85	4	7.52	3	0	0
T quantile for 4 samples at 99.5% confidence = 6					>(T = 0)	

In multispectral illuminant estimation, rather than the actual and estimated light being three vectors they are 31-vectors. Relative to this 31 vectors the recovery and reproduction errors are analogously defined. The reader will notice the errors are higher. Intuitively, this is to be expected as in 31-space there are more degrees of freedom. The discrepancy between the ranking of reproduction versus recovery error is even more marked for the multispectral case.

6 Conclusion

In this paper, we propose the reproduction angular metric as an improvement over the recovery angular error. The

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2016.2582171

latter measure calculates the angle between the actual and estimated lights whereas the former calculates the angle between the actual true rgb for white surface and the estimated white. We showed that the recovery angular error has the property that it varies widely for the same scene viewed under different lights. This is surprising when we factor in how the illuminant estimates are used: they are used to balance a scene that has a colour cast due to the prevailing light so that the color bias is removed. The new reproduction error is stable for a fixed scene-algorithm pair.

The best 'tuning' parameters for different algorithms is found to depend on the error metric used. Further we show that the ranking of illuminant estimation algorithms while broadly the same for recovery or reproduction angular error can change for the local pairs of algorithms (e.g. pixel-based and edge-based gamut mapping). The change in the ranks is statistically significant.

Almost always, illuminant estimation algorithms provide estimates of the prevailing illuminant color which is then removed from the image. The resulting reproduction is the image used in computer vision processing for tasks ranging from recognition to tracking to navigation. Not only is our new reproduction angular error targeted towards how illuminant estimation algorithms are used (as a pre-processing step for other vision processing) but they rank algorithms differently from the current recovery angular error.

Acknowledgments

This research was supported by EPSRC grant H022236.

References

[1] K. Barnard, L. Martin, B. Funt, and A. Coath, "A data set for color research," *Color Research & Application*, vol. 27, no. 3, pp. 147–151, 2002.

[2] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[3] L. Shi and B. Funt, "Re-processed version of the gehler color constancy dataset of 568 images," *Simon Fraser University*, 2010.

[4] F. Ciurea and B. Funt, "A large image database for color constancy research," in *IS&T/SID Color and Imaging Conference*, 2003, pp. 160–164.

[5] D. Cheng, D. K. Prasad, and M. S. Brown, "Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution," *Journal of the Optical Society of America A*, vol. 31, no. 5, pp. 1049–1058, 2014.

[6] S. D. Hordley and G. D. Finlayson, "Reevaluation of color constancy algorithm performance," *Journal of the Optical Society of America A*, vol. 23, no. 5, pp. 1008–1020, 2006.

[7] A. Gijsenij, T. Gevers, and M. P. Lucassen, "Perceptual analysis of distance measures for color constancy algorithms," *Journal of the Optical Society of America A*, vol. 26, no. 10, pp. 2243–2256, 2009.

[8] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin institute*, vol. 310, no. 1, pp. 1–26, 1980.

[9] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Color constancy: generalized diagonal transforms suffice," *Journal of the Optical Society of America A*, vol. 11, no. 11, pp. 3011–3019, 1994.

[10] B. A. Wandell, "The synthesis and analysis of color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, pp. 2–13, 1987.

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

[12] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *IS&T/SID Color and Imaging Conference*, 2004, pp. 37–41.

[13] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–128, 1977.

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

[15] G. D. Finlayson and R. Zakizadeh, "Reproduction angular error: An improved performance metric for illuminant estimation." in *British Machine Vision Conference (BMVC)*, 2014.

[16] G. D. Finlayson, "Corrected-moment illuminant estimation," in *IEEE International Conference on Computer Vision*, 2013, pp. 1904–1911.

[17] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Generalized gamut mapping using image derivative structures for color constancy," *International Journal of Computer Vision*, vol. 86, no. 2-3, pp. 127–139, 2010.

[18] R. Zakizadeh and G. D. Finlayson, "The correlation of reproduction and recovery angular errors for similar and diverse scenes," in *IS&T/SID Color and Imaging Conference*, 2015, pp. 196–200.

[19] P. M. Hubel, "Foveon technology and the changing landscape of digital cameras," in *IS&T/SID Color and Imaging Conference*, 2005, pp. 314–317.

[20] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Spectral sharpening: sensor transformations for improved color constancy," *Journal of the Optical Society of America A*, vol. 11, no. 5, pp. 1553–1563, 1994.

[21] H. Y. Chong, S. J. Gortler, and T. Zickler, "The von Kries hypothesis and a basis for color constancy," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[22] L. T. Maloney and B. A. Wandell, "Color constancy: a method for recovering surface spectral reflectance," *Journal of the Optical Society of America A*, vol. 3, no. 1, pp. 29–33, 1986.

[23] D. A. Forsyth, "A novel algorithm for color constancy," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 5–36, 1990.

[24] G. Finlayson and R. Zakizadeh, "The generalised reproduction error for illuminant estimation," in *Proceedings of AIC 2015 Color and Image, Interim Meeting of the International Color Association*. Association Internationale de la Couleur, 2015.

[25] D. Cheng, B. Price, S. Cohen, and M. S. Brown, "Beyond white: Ground truth colors for color constancy correction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 298–306.

[26] A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color constancy with spatio-spectral statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1509–1519, 2012.

[27] S. Bianco, C. Cusano, and R. Schettini, "Color constancy using cnns," in *IEEE CVPR Workshops, Deep Vision: Deep Learning in Computer Vision*, 2015.

[28] J. T. Barron, "Convolutional color constancy," in *IEEE International Conference on Computer Vision*, 2015.

[29] N. Banić and S. Lončarić, "Color dog: Guiding the global illumination estimation to better accuracy," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2015.

[30] B. Li, W. Xiong, W. Hu, B. Funt, and J. Xing, "Multi-cue illumination estimation via a tree-structured group joint sparse representation," *International Journal of Computer Vision*, vol. 117, no. 1, pp. 21–47, 2015.

[31] W. Conover, *Practical nonparametric statistics, Third Edition*. John Wiley & Sons, New York, 1999.

[32] B. Funt, K. Barnard, and L. Martin, "Is machine colour constancy good enough?" in *Proceedings of the European conference on computer vision*. Springer, 1998, pp. 445–459.

[33] D. H. Foster, K. Amano, S. M. Nascimento, and M. J. Foster, "Frequency of metamerism in natural scenes," *Journal of the Optical Society of America A*, vol. 23, no. 10, pp. 2359–2372, 2006.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.