Re-evaluation of Illuminant Estimation Algorithms in Terms of Reproduction Results and Failure Cases

by

Roshanak Zakizadeh

A thesis submitted for the degree of
Doctor of Philosophy

in the
University of East Anglia
School of Computing Sciences

February 2017

Acknowledgements

First, my sincere gratitude goes to my first supervisor, Professor Graham Finlayson. With his love of science, experience and great attention to detail, he has been a true mentor and motivator during these three years. I am very glad to know him as a researcher and a good friend. Also, I would like to thank Dr. Mark Fisher for being on my supervisory team and always being ready to help when I needed it. Many thanks to Professor Michael Brown who provided the opportunity for me to work with him at the National University of Singapore for some time. It was a pleasure working with him and his group. I am also thankful to Dr. Arjan Gijsenij who kindly offered his help and critique of some parts of this work. I thank Professor Theo Gevers who kindly accepted to travel from Amesterdam to examine my viva. Also, thank you to Professor Richard Harvey for being the internal examiner.

I am grateful to the University of East Anglia for funding my research. Apart from the educational experience, I enjoyed my time at the UEA a great deal.

To every member of the colour group at the UEA in past or present who has given me good memories and a helping hand along the way. In particular, Christopher Powell for proofreading parts of this thesis and Mike Harris whose words of wisdom have been a great help when I was just at the beginning of this work.

To complete, I thank the constant sources of encouragement in my life. My mother, Mina, who is the image of a successful strong woman to me. My father, Mohsen, who is an inspiration with his joy of life. My lovely sister, Nazanin, who is the brightest soul I know. And finally, the love of my life, Hamed, who not only has been the source of emotional support through all these years but whose patience and hard working attitude have been an example to me.

Abstract

Illuminant estimation algorithms are usually evaluated by measuring the recovery angular error, the angle between the RGB vectors of the estimated and ground-truth illuminants. However, this metric reports a wide range of errors for an algorithm-scene pair viewed under multiple lights. In this thesis, a new metric, “Reproduction Angular Error”, is introduced which is an improvement over the old metric and enables us to evaluate the performance of the algorithms based on the reproduced white surface by the estimated illuminant rather than the estimated illuminant itself. Adopting new reproduction error is shown to both effect the overall ranking of algorithms as well as the choice of optimal parameters for particular approaches.

A psychovisual image preference experiment is carried out to investigate whether human observers prefer colour balanced images predicted by, respectively, the reproduction or recovery error metric. Human observers rank algorithms mostly according to the reproduction angular error in comparison with the recovery angular error.

Whether recovery or reproduction error is used, the common approach to measuring algorithm performance is to calculate accurate summary statistics over a dataset. Mean, median and percentile summary errors are often employed. However, these aggregate statistics, by definition, make it hard to predict performance for individual images or to discover whether there are certain “hard images” where some illuminant estimation algorithms commonly fail. Not only do we find that such hard images exist, based only on the outputs of simple algorithms we provide an algorithm for identifying these hard images (which can then be assessed using more computationally complex advanced algorithms).

Contents

List of Figures

List of Tables

Publications

The following are publications by the author related to this work:

- G. Finlayson, R. Zakizadeh and A. Gijsenij. The Reproduction Angular Error for Evaluating the Performance of Illuminant Estimation Algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence (tPAMI)*, PP(99), July 2016

- R. Zakizadeh, M. S. Brown and G. D. Finlayson. A Hybrid Strategy for Illuminant Estimation Targeting Hard Images. In *Proceedings of The IEEE International Conference on Computer Vision (ICCV) Workshop on Color and Photometry in Computer Vision*, Santiago, Chile, 2015.

- R. Zakizadeh and G. D. Finlayson. The Correlation of Reproduction and Recovery Angular Errors for Similar and Diverse Scenes. In *Proceedings of IS&T's Twenty Third Color and Imaging Conference (CIC)*, pages 196-200, Darmstadt, Germany, 2015

- G. Finlayson and R. Zakizadeh. The Generalised Reproduction Error for Illuminant Estimation. In *Proceedings of International AIC (Association Internationale de la Couleur) Midterm Meeting*, Tokyo, Japan, 2015.

- G. Finlayson and R. Zakizadeh. Reproduction Angular Error: An Improved Performance Metric for Illuminant Estimation. *Proceedings of the British Machine Vision Conference (BMVC)*, Nottingham, UK, 2014.

Glossary

2D	2 **D**imensional
3D	3 **D**imensional
CCD	**C**harged **C**oupled **D**evices
CCI	**C**olour **C**onstancy **I**ndex
CIE	**C**ommision **I**nternationale de I’´**E**clairage (International Commission on Illumination)
CMF	**C**olour **M**atching **F**unction
DCRAW	An open-source computer program which is able to read numerous raw image formats [Coffin, 2008]
DSLR	**D**igital **S**ingle **L**ens **R**eflex
HDR	**H**igh **D**ynamic **R**ange
ISO	**I**nternational **O**rganization of **S**tandardization
JPG	**J**oint **P**hotographic **G**roup
JND	**J**ust **N**oticeable **D**ifference
LCD	**L**iquid **C**rystal **D**isplay
LED	**L**ight **E**mitting **D**iode
LUT	**L**ook **U**p **T**able
nm	**n**ano **m**etres
NUS	**N**ational **U**niversity of **S**ingapore
PCA	**P**rinciple **C**omponent **A**nalysis
PNG	**P**ortable **N**etwrok **G**raphics
RGB	**R**ed, **G**reen, **B**lue
SFU	**S**imon **F**raser **U**niversity
SPD	**S**pectral **P**ower **D**istribution
sRGB	Standard Default Color Space for the Internet [Stokes et al., 2012]
SVM	**S**upport **V**ector **M**achines

Illuminant Estimation Algorithms

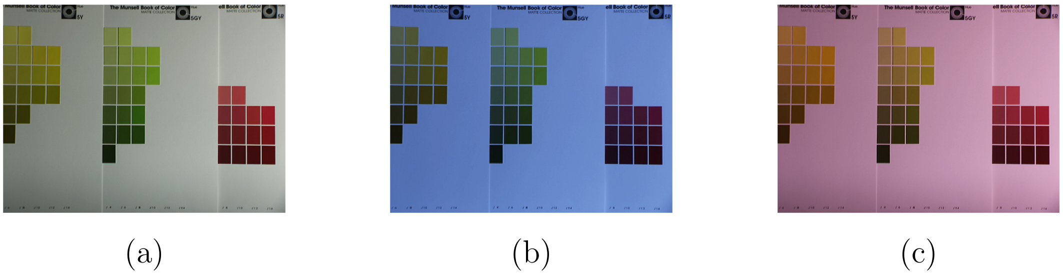
GP	**G**amut **P**ixel-based mapping
GE1	1^{st} order **G**rey **E**dge
GE2	2^{nd} order **G**rey **E**dge
IICS	**I**nverse **I**ntensity **C**hromaticity **S**pace
SOG	**S**hades **O**f **G**rey

The readers should notice that colour images could look different depending on the colour calibration of the printer or the device used for their display.

Chapter 1

Introduction

Colours of a scene captured by an imaging device such as a digital camera are subject to change due to the prevailing illumination (or illuminations). An example of this phenomenon can be seen in Figure 1.1. In the unprocessed images in this figure (from SFU dataset [Barnard et al., 2002c]), the Munsell colour chart [Munsell, 1950] is captured under three different lights, from left to right: Philips Ultralume fluorescent, Solux 4700 (which has a colour similar to daylight) and Sylvania warm white fluorescent. We can see the white colour of the paper changes under different illuminations. Also, some colours of the Munsell chart look more similar depend on the light's colour striking the chart.



(a) (b) (c)

FIGURE 1.1: The Munsell colour chart under: a) Philips Ultralume fluorescent, b) Solux 4700 and c) Sylvania warm white fluorescent lights. The charts are printed on white papers but here we see that different lights make the white paper appears yellowish, bluish and pinkish. Colours of the chart also look different under different illuminants. This is an example of how illuminant can affect the colours of a scene captured by the camera.

Unlike the human visual system, which has a relatively constant perception of colours, an imaging device initially lacks such a capability. In analogy to how we see, we would like digital cameras to be colour constant. In other words, we seek to develop computational approaches to recover the actual colour of surface objects which for instance in the case of Figure 1.1 are the colours of Munsell chart. Often colour constancy is posed as rerendering an image so white looks right and this is called white balancing in photography. The first step to white balancing is estimating the colour of illuminant. The estimate of colour is 'divided' out and if the estimate is correct a white surface should look white. Apart from the goal of creating images pleasant to the human eye with colours looking as natural as possible, colour constancy is essential for many computer vision applications such as image retrieval, colour reproduction and object recognition [Gevers and Smeulders, 1999; Abdel-Hakim and Farag, 2006; Slater and Healey, 1996; Van De Sande et al., 2010].

Over the years a variety of algorithms have been proposed to estimate the illuminant (We will review many of these algorithms in Chapter 2). A key concern of this thesis is how to evaluate which - of the many - illumination estimation algorithms works best.

The most popular way to evaluate the performance of a given algorithm for a given scene is to calculate the angle between the two RGB vectors of true illuminant of the scene and the one estimated by the algorithm[Finlayson et al., 1995]. This metric is generally known as angular error (In this thesis, it is called *recovery angular error.*). The error is usually calculated for a set of images from a benchmark dataset and eventually a summary of statistics (such as mean or median) is reported which decides the rank of an algorithm in comparison to other algorithms. Given the diversity of the existing illuminant estimation algorithms and the importance of removing the colour bias due to the illumination for many vision tasks, it is essential to know whether or not we can rely on the error reported by the used metric as well as the way the error data is analysed. Some work [Hordley and Finlayson, 2006; Gijsenij et al., 2009a] have also highlighted the importance of choice of error metric and the analysis of error data.

1.1 About This Thesis

The recovery angular error metric for illuminant estimation is problematic because it does not reflect how illuminant estimation algorithms are used. In practice the estimated light is divided out from the image. So, we propose, it is more useful to focus on how the reproduced image appears. This thesis addresses the problems with the existing workflow of evaluation which includes measuring the accuracy of the estimated illuminants and not their resulting reproduction. Further, this thesis also examines the other drawback of evaluation workflow that an aggregate of the error data can not quite reveal the relationship between the errors introduced by algorithms for individual images.

The major contributions of this thesis include:

- A part of this thesis is dedicated to studying the accuracy of the widely used angular error. We examine the angular error - or as we call it in this thesis *recovery angular error* - to derive its maximum value for a given scene. We show that for a given scene and a given illuminant estimation algorithm, there are specific lights that result in the maximum and minimum recovery angular errors. This analysis is based on both a set of theoretical feasible lights and a set of real lights. In both cases, we show that the range of recovery angular error could be very large for a given algorithm and a given scene.

- To mitigate this problem, we propose a new angular error which we call **"Reproduction Angular Error"**. The reproduction angular error measures the angle between the estimated reproduced white and 'true' white.

We show the stability of reproduction angular error against changes in the illumination.

- We study how reproduction angular error might change our judgement about the relative performance of illuminant estimation algorithms. Different statistical tests are performed to analyse this question. We also investigate the correlation between the two metrics: recovery and reproduction angular errors.

- The reproduction of white is at the heart of reproduction angular error. We generalise this idea and develop a novel framework to evaluate the accuracy of a range of reproduced colours.

- Some prior work [Gijsenij et al., 2009a; Banic and Loncaric, 2015; Vazquez-Corral et al., 2009] evaluate the performance of illuminant estimation algorithms psychophysically. In this work also a psychophysics experiment is conducted to study the correlation of perceptual judgements and recovery or reproduction angular error.

- As mentioned before, a summarised evaluation of an illuminant estimation algorithm's performance is often reported in the form of an aggregate (such as mean, median, etc.) over the whole benchmark dataset. However, the relationship of these summary statistics across different methods is unclear. The final part of this thesis investigates the relationship between the performance of several algorithms and motivated by this relationship a hybrid strategy for finding images which are commonly hard for many illuminant estimation algorithms is proposed.

1.1.1 Thesis Outline

The Outline of the thesis is as follows: In Chapter 2 we will review a selection of illuminant estimation algorithms. The evaluation techniques for illuminant estimation algorithms will be discussed in Chapter 3. Chapter 4 discusses a problem with the well-known metric of illuminant estimation, recovery angular error. Specifically that for the same scene illuminated under many lights - which when the estimated light is divided out produces the scene reproduction - results in a very large range of recovery angular error. A new metric, *Reproduction Angular Error*, is introduced and its theoretical foundation is explored. Reproduction angular error is very stable for a given algorithm and the scene viewed under multiple lights. Chapter 5 discusses the ranking of illuminant estimation algorithms in terms of recovery and reproduction angular errors. We show that the ranking of algorithms depends on the metric used. Using different statistical tests the significance of switches between the ranking of algorithms using the two metrics is discussed in this chapter. Also, the correlation of the two metrics is investigated in the same chapter. An evaluation framework based on other reproduced colours rather than only the reproduced white surface is performed in Chapter 6. In Chapter 7, a psychophysical experiment is carried out to investigate the choice of metric and image preference. Reproduction angular error accounts for observers' preference. In Chapter 8, a hybrid strategy for detecting commonly hard images for multiple illuminant estimation algorithm is proposed. The final chapter, Chapter 9, concludes the work in this thesis and shows the future avenues.

Chapter 2

Background: Image Formation & Illuminant Estimation

2.1 Introduction

To discuss illuminant estimation it is essential to understand the foundation of image formation in digital cameras. We start this chapter by giving the background on colour image formation in Section 2.2. The rest of this chapter is organised as follows: In Section 2.3, we discuss how the illuminant estimation problem is formulated. In Section 2.4, the illuminant estimation algorithms based on statistical information and methods are reviewed. In Section 2.5, an overview of gamut mapping methods for colour constancy is given. Learning-based methods are discussed in Section 2.6. In Section 2.7, the works that have used a combination of illuminant estimation algorithms or have optimised the results of one or more algorithms are introduced. The last section will conclude the chapter.

2.2 Colour Image Formation

The colour signal (e.g. [Wandell, 1987]) received by human eye is the product of spectral power distribution (SPD) of the ambient light and surface spectral reflectance of the object. This process is pictured in Figure 2.1. The colour signal ($C^x(\lambda_n)$) in a small region of the image is defined as:

$$C^x(\lambda_n) = E(\lambda_n)S^x(\lambda_n), \tag{2.1}$$

where $E(\lambda_n)$ is the SPD of the ambient light and $S^x(\lambda_n)$ is the surface reflectance at point x, both at sample wavelength λ_n.

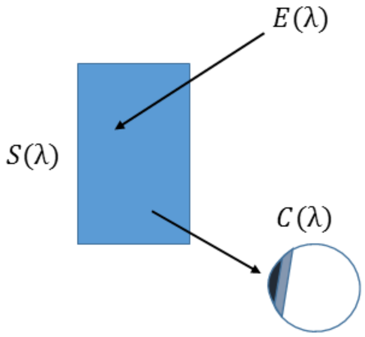


Figure 2.1: Colour signal received by the human eye (or a camera's sensor) from a small region of the surface.

An image captured by a digital camera (or seen by the eye) is a result of the sensor's (or human vision system's) response to this colour signal. In human visual system (illustrated in Figure 2.2), cone cells (indicated in red, green and blue colours in Figure 2.2) are the photoreceptor cells in the retina of the eye which are sensitive to light and are responsible for colour vision. The sensitivity of the cones is limited to a part of the electromagnetic spectrum (approx. 400-700 nanometres (nm)). There are three types of cone cells: S (short), M (medium) and L (long) for their relative spectral positions of their peak sensitivities. Figure 2.3 shows the spectral sensitivities of the human cone cells, S, M and L . The camera sensor sensitivity range could vary and cover more than the limited range of the human visual system, but it is ultimately filtered to capture the colour signal within the same range. An example of camera sensitivity functions is shown in Figure 2.4.

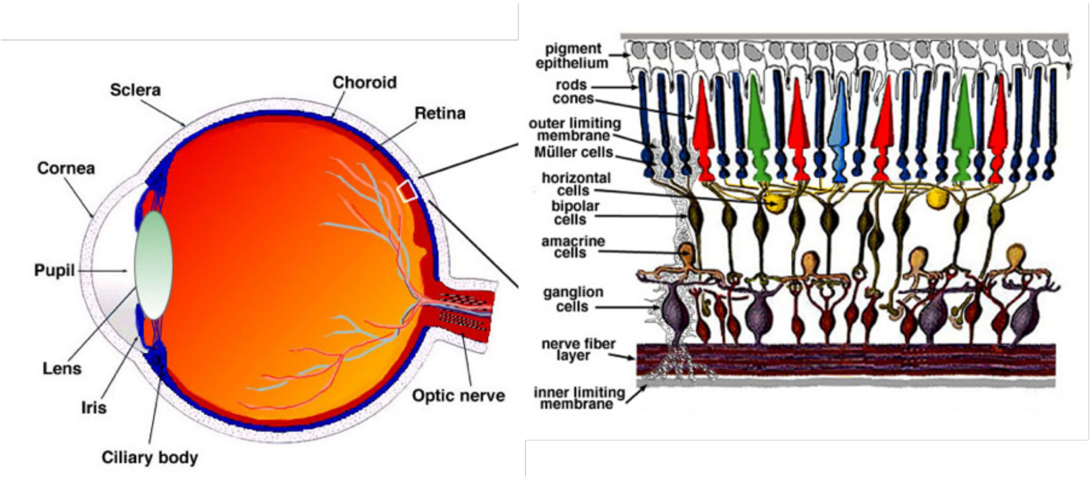


Figure 2.2: A close-up of the retinal cell layers. The cone cells are responsible for colour vision (illustration is taken from [Kolb, 2012]).

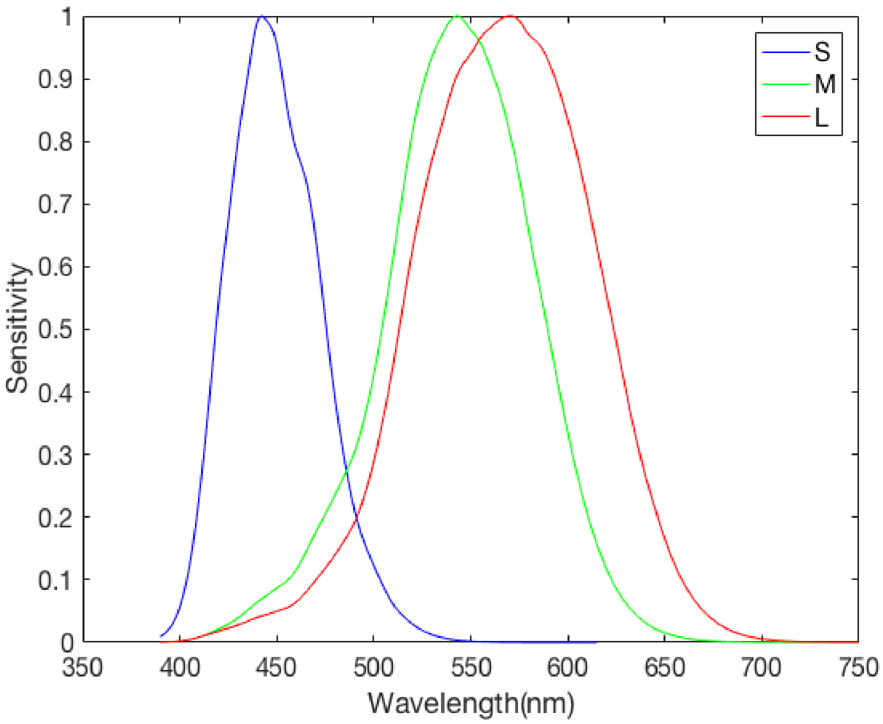


Figure 2.3: Spectral sensitivities of S, M and L cones (plotted from the data by [Stockman and Sharpe, 2000]).

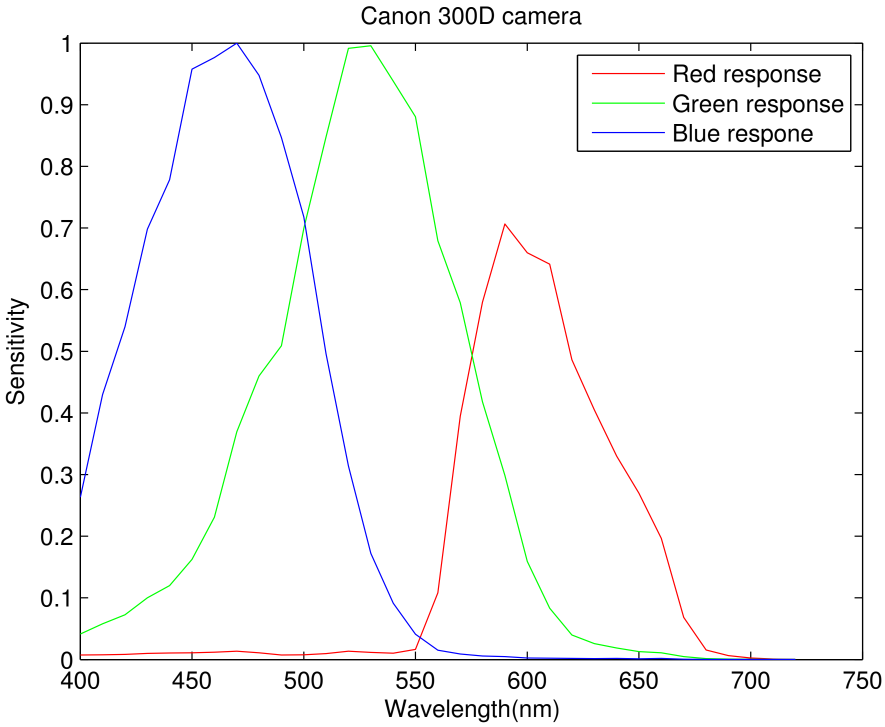


FIGURE 2.4: Spectral sensitivity functions of Canon 300D camera (measured by [Jiang et al., 2013]).

Following the model of colour signal Eq. 2.1, the model of image formation is written as:

$$\rho_k^x = \int_\omega R_k(\lambda) C^x(\lambda) d\lambda \quad k \in \{R, G, B\}, \tag{2.2}$$

where $R_k(\lambda)$ is the response function of the camera's k^{th} sensor and the integral is over the visible spectrum ω. ρ_k^x is the response of camera's k^{th} sensor to the colour signal at location x of sensor array. The sensitivity of the sensors in most digital cameras are concentrated in the Red (long), Green (medium) and Blue (short) parts of the visible spectrum of light. Therefore, k is denoted as R, G or B in (Eq. 2.2). Further, using the underscore notation to denote vector quantities we rewrite Eq. 2.2 as:

$$\underline{\rho} = \int_{\omega} \underline{R}(\lambda)C(\lambda)d\lambda \tag{2.3}$$

This model of image formation, despite the restrictions it imposes (e.g. surfaces are considered to be perfect Lambertian diffuser [Lee, 1986; Shafer, 1985]), is often used when discussing illuminant estimation. Making the role of light and surface explicit Eq. 2.2 and Eq. 2.3:

$$\rho_k^{E,S} = \int_{\omega} R_k(\lambda)E(\lambda)S(\lambda)d\lambda \quad k \in \{R, G, B\}. \tag{2.4}$$

Here $\rho_k^{E,S}$ is similar to ρ_k^x in Eq. 2.2, except we have dropped the x since there is a one-to-one relation between the scene point and colour signal. Then ρ_k^S is the surface colour which is observed under a reference light.

2.2.1 Discrete model of Image Formation and the Finite Basis Functions

The colour image integral (Eq. 2.4) can be written as a summation:

$$\underline{\rho} = \sum_{i=1}^{n} E(\lambda_i)S(\lambda_i)\underline{R}(\lambda_i)\Delta\lambda \tag{2.5}$$

where $\Delta\lambda$ accounts for the sampling interval (and is often set at $10nm$).

For further simplification, the spectral power distribution functions and surface spectral reflectance functions might be written as linear combinations of basis functions.

$$S(\lambda) \approx \sum_{i=1}^{d_S} S_i(\lambda)s_i, \tag{2.6}$$

where $S_i(\lambda)$ is the basis function and $\underline{s}$ is a vector of weights of d_S dimension. Similarly, illuminants can be modelled with a low dimension basis functions:

$$E(\lambda) \approx \sum_{j=1}^{d_E} E_j(\lambda)e_j \tag{2.7}$$

where $E_j(\lambda)$ is the basis function and $\underline{e}$ is a vector of weights of d_E dimensions.

The basis functions are decided by performing principal component analysis (PCA) on reflectances and illuminants [Maloney, 1986].

Following the finite basis models of the reflectance and illumination (Eq. 2.6 and Eq. 2.7), the image formation in Eq. 2.5 can be rewritten as a matrix transform. A lighting matrix $\Lambda(E(\lambda))$ maps reflectances defined by $\underline{s}$ onto the corresponding sensor responses $\underline{\rho}$ (colour observation by the sensors):

$$\underline{\rho} = \Lambda(E(\lambda))\underline{s} \tag{2.8}$$

where, $\Lambda(E(\lambda))$ is a $3 \times d_S$ matrix:

$$\Lambda(E(\lambda))_{ij} = \int_\omega R_i(\lambda)E(\lambda)S_j(\lambda)d\lambda \tag{2.9}$$

If $E(\lambda)$ is written as Eq. 2.7, then the lighting matrix is dependant only on the illuminant weighting vector $\underline{e}$.

In the next section, we will show how the colour constancy problem is formulated considering that only the colour observation by camera sensors (or $\underline{\rho}^{E,S}$ in Eq. 2.4) is known.

2.3 Illuminant Estimation

The aim of colour constancy is to derive an illuminant-independent colour observation ($\underline{\rho}^S$). Therefore, the task of colour constancy is formulated as:

$$\underline{\rho}^S = \psi(\underline{\rho}^{E,S}) \tag{2.10}$$

where $\underline{\rho}^{E,S}$ is a colour observation dependant on the surface reflectance and illumination. $\underline{\rho}^S$ is the colour of a surface under a reference (or canonical) light (e.g. [Forsyth, 1990] or [Maloney and Wandell, 1986]). The symbol ψ in Eq. 2.10 represents a transform function which maps all the colours of the formed image to the colours under reference lighting condition.

Forsyth [Forsyth, 1990] formally proved that the problem of colour constancy is *exactly* solvable if and only if the transformation in Eq. 2.10 is a 3 × 3 linear transform. So the problem summarises in solving for the nine parameters of matrix M:

$$\underline{\rho}^S \approx M\underline{\rho}^{E,S}. \tag{2.11}$$

Following Forsyth's formulation of illuminant estimation, in [Finlayson et al., 1994a] the authors demonstrate that a diagonal matrix transform suffices as a vehicle for illuminant estimation so long as the image RGBs are in a special basis. Specifically:

$$T\underline{\rho}^S \approx D\underline{\rho}^{E,S}, \tag{2.12}$$

where T is a fixed per camera 3 × 3 matrix.

Actually T for many cameras is simply the identity matrix and so can be removed [Barnard and Funt, 2002]. Henceforth, we simply assume:

$$\underline{\rho}^{S} \approx D\underline{\rho}^{E,S}. \tag{2.13}$$

This simple formulation of image formation is useful in simplifying the illuminant estimation problem. Suppose that $\underline{\rho}^{W}$ denotes the colour of a white surface (under a white reference light). Further, let's assume that $\underline{\rho}^{W} = [1\ 1\ 1]$. The colour of the same white surface under a second illuminant E is $\underline{\rho}^{W,E} = [d_1\ d_2\ d_3]^t$ and:

$$\underline{\rho}^{W} \approx D\underline{\rho}^{W,E}. \tag{2.14}$$

Clearly D is a diagonal matrix with components $[1/d_1\ 1/d_2\ 1/d_3]$. Remarkably, this D also models the physics of image formation for arbitrary surfaces (Eq. 2.13 holds).

In abstract form illuminant estimation can be posed equivalently as i) Finding an estimate of the illuminant colour $[d_1\ d_2\ d_3]^t$ or ii)finding the diagonal matrix D or iii) (In the gamut mapping formulation [Forsyth, 1990]) finding the diagonal matrix D^{-1}.

From Eq. 2.13 and assuming an illuminant estimation algorithm provides a reasonable estimate of the illuminant colour ($[d_1\ d_2\ d_3]^t$) - which we divide as $\underline{\rho}^{Est}$ - then we solve for $\underline{\rho}^{S}$, the colour of a surface under a reference illuminant, by dividing out:

$$\underline{\rho}^{S} \approx \frac{\underline{\rho}^{E,S}}{\underline{\rho}^{Est}}, \tag{2.15}$$

where the division of the vectors is component-wise. Equation (2.15) is simply a rewriting of Eq. 2.13. Figure 2.5 shows an example of illuminant estimation.

This is an image of an entrance to a handicraft shop in Ganjali Khan Complex in Kerman, Iran, captured by Canon EOS 100D camera. In Figure 2.5, the top image is the raw output by the camera with no corrections except for the gamma [1] (for display). In the image in the bottom, the colour of the illuminant for the raw image is estimated using a simple white balance algorithm (shades of grey [Finlayson and Trezzi, 2004]). Then the RGB of the estimated light is divided out from the raw image. Notice the white (19^{th}) patch of the colour checker which is corrected (it looks whiter) in the bottom image after the image is white balanced.

Forsyth [Forsyth, 1990] proposes colour constancy and illuminant estimation date back to 1878 when Von Kries's theory of chromatic adaptation [von Kries, 1878] was established and Judd [Judd, 1940] and later Land and MacCann [Land and McCann, 1971] associated colour constancy with it. Viewed as an algorithm, Von Kries is based on the coefficient rule [von Kries, 1878; West and Brill, 1982; Worthey and Brill, 1986]. Here the gain of each colour channel is adjusted independently to obtain the surface colour. Since then a variety of algorithms have been proposed to achieve a constant colour captured by an imaging device as it is observed by the human visual system. Some of the important surveys reviewing and evaluating illuminant estimation algorithms are [Barnard et al., 2002a], [Hordley, 2006], [Agarwal et al., 2006] and [Gijsenij et al., 2011].

Gijsenij et al. [Gijsenij et al., 2011] divide the state of art algorithms into three categories: 1) statistical methods, 2) gamut-based methods, and 3) learning-based methods. The following sections provide an overview of many illuminant estimation algorithms. We follow the same categorisation by Gijsenij et al. [Gijsenij et al., 2011] with an extension of a fourth category which is the combinational and optimisation methods for illumination estimation. Although as Gijsenij et al.

[1] Gamma correction (also known as gamma encoding or compression) is the process of applying a power function (usually 1/2) to the raw pixel values. Gamma correction is usually done in digital image processing to imitate the non linear human perception of luminance. Read more in [Plataniotis and Venetsanopoulos, 2013]

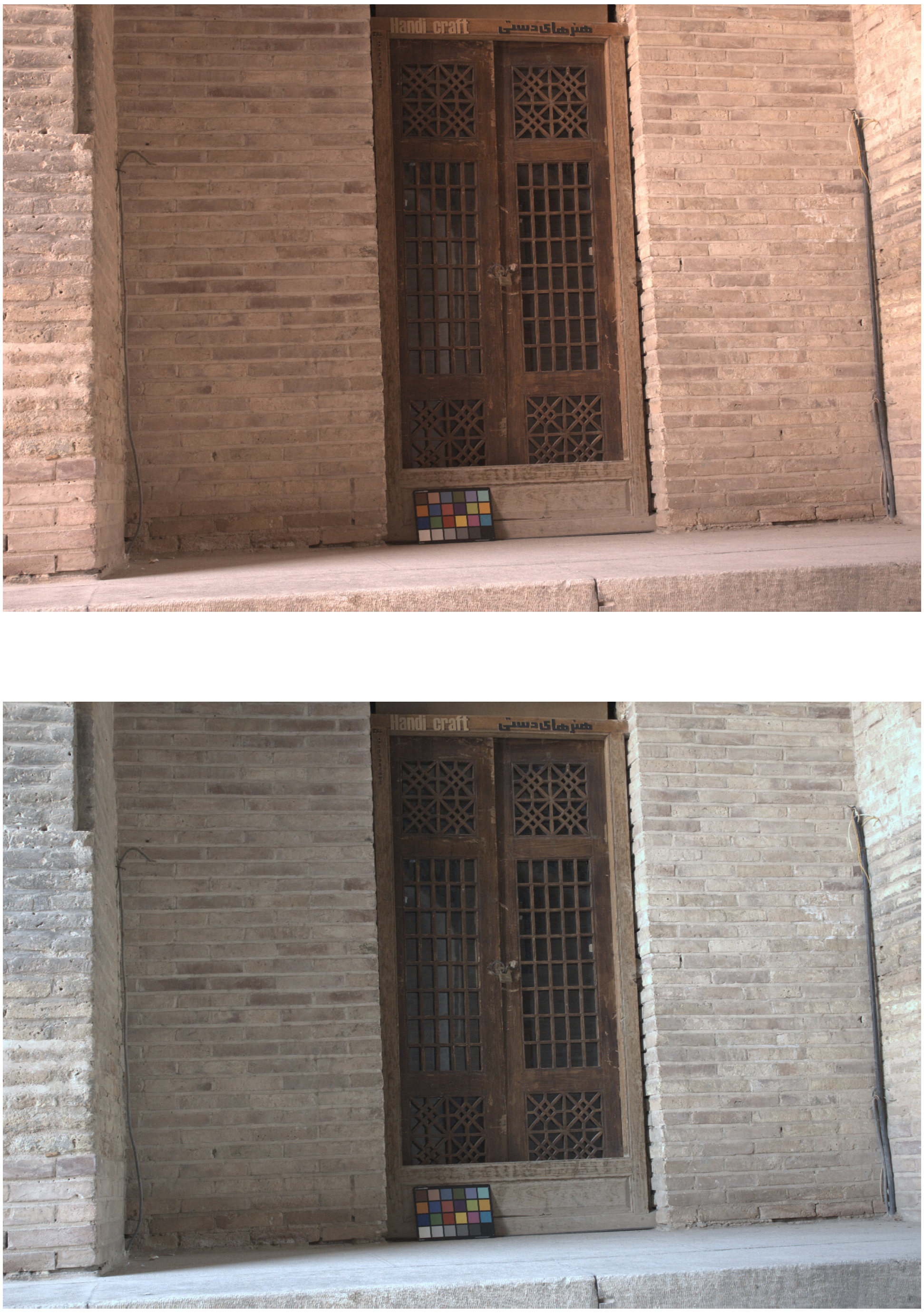


FIGURE 2.5: An entrance to a handicraft shop in Ganjali Khan Complex (Kerman, Iran), captured by Canon EOS 100D. An example of applying white balance to an image: The image in the top is the raw (unprocessed) camera output and the image in the bottom is white balanced by the shades of grey [Finlayson and Trezzi, 2004] algorithm.

mention, this categorisation is not absolute and some algorithms (like gamut-based techniques) might fall in more than one category.

2.4 Statistic-based Illuminant Estimation Algorithms

In this section we review the illuminant estimation methods based on the statistics of an image. These methods range from low-level statistics to high-level statistics and from pixel-based estimation to estimations based on the derivatives of an image.

2.4.1 MaxRGB

Incorporated in the early Retinex theory [Land et al., 1977], it is argued that the perceived white is associated with the maximum cone signals of the human visual system. Based on this hypothesis MaxRGB or White Patch algorithm assumes there is a white surface (or bright red, green and blue surfaces) in the scene that reflects the maximum brightness, which then the illuminant colour can be recovered. MaxRGB algorithm can be formulated as:

$$\max_x \underline{\rho}^x = k\underline{\rho}^{Est}. \tag{2.16}$$

The variable k represents the fact that the exact magnitude of light can never be recovered and the maximum value of all pixels in the image is calculated separately for each R, G and B channel:

$$\max_x \underline{\rho}^x = \left(\max_x R(x), \max_x G(x), \max_x B(x)\right). \tag{2.17}$$

The MaxRGB algorithm does not require the maximum of the three separate channels to be on the same location; hence, it also obtains correct estimated illuminant when the maximum reflectance is equal for the three channels [Van De Weijer et al., 2007a].

The MaxRGB algorithm imposes the restriction of estimating illuminant based on only the brightest pixel per channel in the scene. If the brightest response in an image - in all three channels - is from a yellow surface the MaxRGB algorithm will, wrongly, infer the light colour is yellow. Other researchers have proposed preprocessing steps which can improve the results of MaxRGB significantly. For instance, Ebner [Ebner, 2009] uses a local mean calculation as a preprocessing step or in [Funt and Shi, 2010] it is shown that capturing the full dynamic range of a scene and removal of clipped pixels (using a median filter and sub-sampling the image by bicubic interpolation as preprocessing steps) will improve the performance of MaxRGB algorithm. Joze et al. [Joze et al., 2012] extend MaxRGB hypothesis by not only considering the brightest pixel of the scene but the gamut of bright pixels and try to study the effect of bright pixels on MaxRGB and other colour constancy algorithms.

2.4.2 Grey-world

Grey-world algorithm is one of the simplest illuminant estimation algorithms and it is based on grey-world hypothesis [Buchsbaum, 1980] which states that the average reflectance in a well colour balanced scene under neutral light is grey. This means any deviation from grey is due to illumination. According to the grey-world assumption, the illumination prevailing the scene can be estimated by calculating the mean sensor response:

$$\int \underline{\rho}(x)\mathrm{d}x \approx k\underline{\rho}^{Est} \tag{2.18}$$

Again, k is unknown and represents the fact that the true magnitude of light can not be recovered.

It is clear that the grey-world hypothesis does not hold for the scenes with a single dominant reflectance. There are alternatives to overcome the flaw of the grey-world algorithm, such as taking preprocessing steps prior to applying the algorithm. An example of such preprocessing steps is segmenting the image and averaging over the segments e.g. counting the colour of each segment [Gershon et al., 1987] - regardless of its spatial extent - which may improve the results.

2.4.3 Shades of Grey

Finlayson and Trezzi [Finlayson and Trezzi, 2004] made the interesting observation, that both grey-world and MaxRGB algorithms are instances of Minkowski norms. They call the group of algorithms shades of grey. Shades of grey extends the idea of grey-world algorithm by assuming that the average is calculated as a Minkowski norm. The Minkowski norm framework is written as:

$$\left(\int \underline{\rho}^{p}(x)\mathrm{d}x\right)^{1/p} = k\underline{\rho}_{p}^{Est}. \tag{2.19}$$

Notice that substituting $p = 1$ equates Eq. 2.19 to the grey-world algorithm and with $p \to \infty$ Eq. 2.19 finds the maximum value per channel of $\underline{\rho}$ (which is the MaxRGB algorithm). Further by tuning the value p, one can achieve the best possible result for a given set of images. Often $p = 4$,5, or 6 seems to work best i.e. the best method is a *compromise* between MaxRGB and grey-world.

2.4.4 Grey-edge

Van de Weijer et al.'s [Van De Weijer et al., 2007a] grey-edge hypothesis is proposed as an alternative to the grey-world hypothesis and it states: the average of the reflectance *differences* in a scene is achromatic. The grey-edge framework is written as:

$$\left(\int|\frac{\delta^n\underline{\rho}(x)}{\delta x^n}|^p \mathrm{d}x\right)^{1/p} = k\underline{\rho}^{Est}_{n,p,\sigma}. \tag{2.20}$$

The image can be smoothed with a Gaussian averaging filter with the standard deviation of σ pixels and then is differentiated with an order n differential operator. We then take the absolute Minkowski p-norm average over the whole image.

There three important variables (n, p , σ) in the grey-edge framework (Eq. 2.20):

- The spatial derivative order n which determines if the method is a grey-world algorithm or a grey-edge algorithm. If $n = 0$ it means that the calculation is carried out directly on RGB values and therefore it is the grey-world method. Whereas, grey-edge method is based on the higher orders of spatial derivatives n. Usually, the highest order of n is considered to be two.
- The Minkowski norm p which determines the relative contribution of the image values or differentiated values.
- The smoothing parameter σ of the applied filter. For zero-order grey-world algorithm, the Gaussian filter with the smoothing parameter σ can be applied. For grey-edge algorithms, applying the filter is followed by a differentiation operation.

Further, Gijsenij et al. observed that different types of edges might contain various amounts of information such as shadow, geometry, material, etc. As a result they proposed weighted grey-edge algorithm [Gijsenij et al., 2009b, 2012] which assigned higher weights to specific types of edges. The weighting scheme - which leads to modest improvements in estimation performance - introduces a higher level of complexity.

In an attempt to further generalise Van de Weijer et al.'s work, Chakrabarti *et al.* [Chakrabarti et al., 2008, 2012] employed an explicit statistical model to capture the spatial dependencies between the pixels. This method takes a training step where a statistical model is learned from the cropped overlapping patches from an image observed under canonical illuminant. The model is then applied to a set of test images. Generally speaking, the idea is to suppress the smooth portion of data and keep the spatial high frequency components in the image.

Further, Cheng *et al.* [Cheng et al., 2014] investigate through multiple experiments why the spatial domain methods actually work. They observe that large colour differences which introduce higher gradient in an image are the key to the better (yet sometimes still false) results for illuminant estimation. On the other hand, by cutting the image into pieces and shuffling the pieces they find that relying on the content of an image to provide the colour differences is not the best way for illuminant estimation. Since, just shuffling the pieces and introducing artificial gradients resulted in lower errors for spatial-based methods.

2.5 Gamut Mapping Illuminant Estimation

One of the most powerful approaches to illuminant estimation is Gamut mapping algorithm which was first introduced by Forsyth [Forsyth, 1990]. The core idea of gamut mapping algorithm is that the set of feasible colours under a reference canonical illuminant is bounded by a convex canonical gamut $\mathcal{C}$. The canonical gamut is obtained by observing as many colours as possible under a canonical illuminant (known light source) during a training phase. The gamut of the unknown light source is assumed to be represented by the colours of the input image. Therefore, the input gamut $\mathcal{I}$ is constructed from all the colours of the input image. Forsyth's algorithm follows the diagonal model of illumination change (Eq. 2.14) and solves for all the feasible mappings $\mathcal{D}$ from the gamut of input image ($\mathcal{I}$) to the canonical gamut ($\mathcal{C}$):

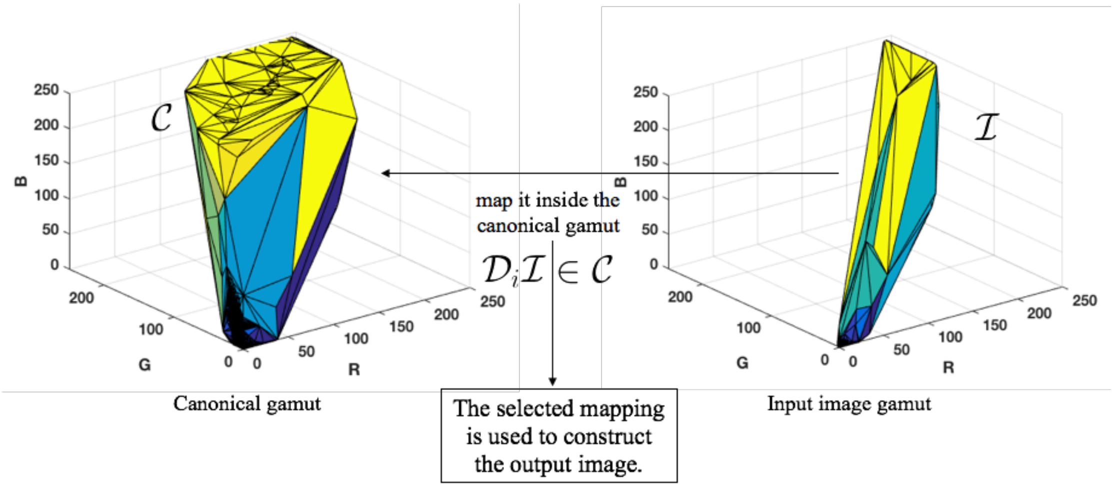


FIGURE 2.6: General overview of gamut mapping algorithms.

$$\mathcal{D}_i\mathcal{I} \in \mathcal{C}. \tag{2.21}$$

The feasible mappings $\mathcal{D}$ are all the mappings that can be applied to the gamut of input image $\mathcal{I}$ and result in a gamut that completely locates inside the canonical gamut $\mathcal{C}$. Ultimately, it chooses one mapping (the one which maximises the volume of the gamut) from all the diagonal maps as a proxy for the estimated illuminant, i.e. it applies the chosen mapping to the image gamut to obtain an estimate of the illuminant for the input image. The general framework of gamut mapping algorithm is pictured in Figure 2.6.

A drawback of the gamut mapping method proposed by Forsyth is that if the diagonal model fails (there are no maps that are feasible) the algorithm results in a null-solution. A solution to this problem is proposed by extending the size of the canonical gamut to find a feasible mapping [Finlayson, 1996; Barnard et al., 2002a].

Different publications have suggested different modifications to the method. For instance Finlayson [Finlayson, 1996] suggests computing gamut mapping in chromaticity space $(\frac{R}{B}, \frac{G}{B})$ to only recover the illuminant chromaticity and not its intensity as it is impossible to do so. However during this transformation from 3D to 2D space we lose some information which impacts on the performance of this version of gamut mapping. Nevertheless, other works [Finlayson and Hordley, 2000, 1999] have suggested to move back to 3D from 2D before selecting the best possible mapping which improves the results.

Forsyth's gamut mapping algorithm and most of its extended versions are based on the pixel values. In an extension of gamut mapping algorithm, Gijsenij *et al.* [Gijsenij et al., 2010] proposed using the derivative structure of the images. Gamut mapping on derivatives can improve estimation performance.

2.6 Learning-based Illuminant Estimation

Some methods of illuminant estimation use a model that is learned based on a training set of white-balanced images and then that model is used to estimate the illuminant for new images. Of course on that basis, gamut mapping algorithms can also be considered to be in this group but being very popular they are often categorised separately. In this section some of the learning-based algorithms which has gained more attention are reviewed.

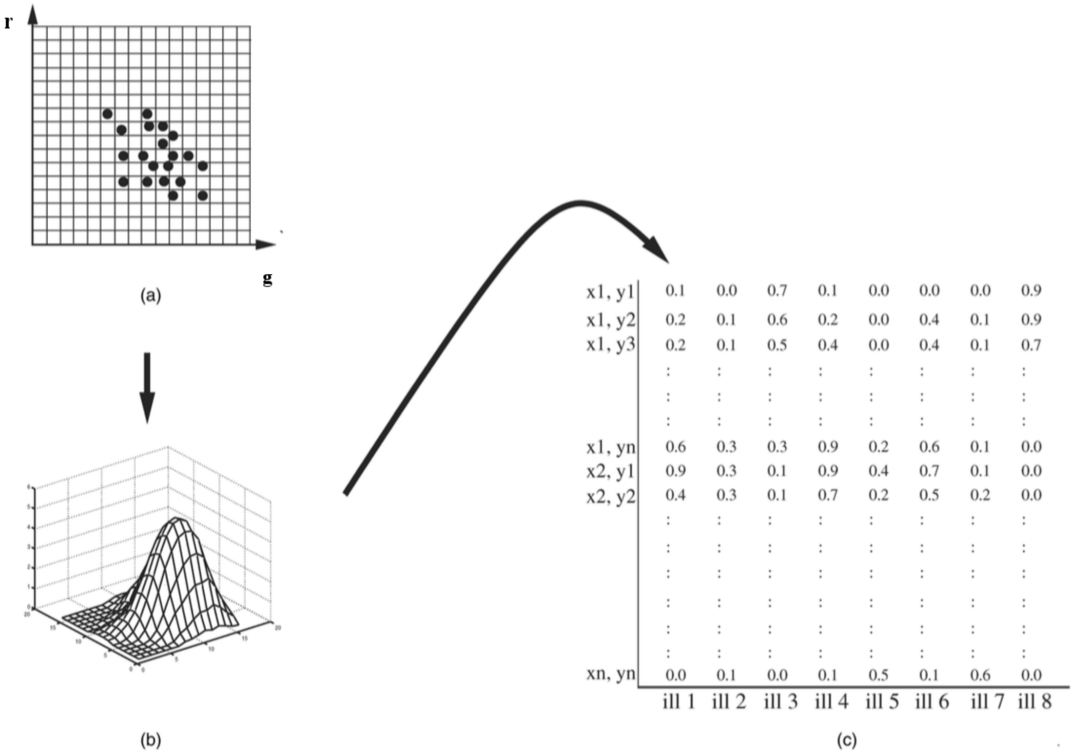
2.6.1 Probabilistic Methods

In Bayesian colour constancy ([DZmura et al., 1995], [Brainard and Freeman, 1997], [Sapiro, 1999],[Rosenberg et al., 2003], [Gehler et al., 2008]) the variability of reflectance and of illuminant is modelled as random variables, then the colour of illuminant is estimated from the posterior distribution conditioned on the power of light and surface reflectance in each channel. The formulation for Bayes Theorem is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.22}$$

where $P(A)$ is the probability of A or more specifically prior probability of A. $P(A|B)$ is the conditional probability of A after taking into account the new piece of evidence B. Often referring to $P(A|B)$ as a posterior priority. $P(B|A)$ is the likelihood of B happening given that A is true. In Bayesian illuminant estimation, B is assumed to be the observed sensor responses, and A contains parameters describing the illuminant, then $P(A)$ is estimated with minimum cost from $P(A|B)$. To calculate the likelihood ($P(B|A)$) of the observed image data B for a given illuminant A, a distribution for all reflectances is needed.

Colour by Correlation [Finlayson et al., 1997, 2001], another probabilistic method, is considered to be a discrete implementation of the 2D gamut mapping with more improvements. In Colour by Correlation the canonical gamut is replaced by a correlation matrix which contains the probabilities of occurrence of a certain coordinates in the rg chromaticity space ($r = R/(R+G+B)$ and $g = G/(R+G+B)$, see Appendix B for the complete diagram). These are possible image colours. In Figure 2.7 (a), the range of possible image colours (chromaticities) that can be observed under each reference illuminant are characterised. Then, this information is used to build a probability distribution (Figure 2.7 (b)) which gives the likelihood of observing an image colour under each scene illuminant (in simple words, what is the likelihood of observing an image colour under a given light?).The probability distributions for each light form the columns of a correlation matrix (Figure 2.7 (c)). Given an input image and the calculated correlation matrix: first, it is determined which image colours (chromaticities) are present in the image. For this step again the histogram of chromaticities in the image is calculated (see Figure 2.7 (a)). Then, this histogram (vector of values) is correlated with each column of the correlation matrix in Figure 2.7 (c) to obtain a probability of every considered light source. The unknown illuminant can be estimated from this measure of correlation. Ultimately, one light source can be selected as the scene illuminant by the maximum likelihood of these probabilities [Finlayson et al., 2001] or using Kullback-Leibler divergence [Rosenberg et al., 2001].



	ill 1	ill 2	ill 3	ill 4	ill 5	ill 6	ill 7	ill 8
x1, y1	0.1	0.0	0.7	0.1	0.0	0.0	0.0	0.9
x1, y2	0.2	0.1	0.6	0.2	0.0	0.4	0.1	0.9
x1, y3	0.2	0.1	0.5	0.4	0.0	0.4	0.1	0.7
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
x1, yn	0.6	0.3	0.3	0.9	0.2	0.6	0.1	0.0
x2, y1	0.9	0.3	0.1	0.9	0.4	0.7	0.1	0.0
x2, y2	0.4	0.3	0.1	0.7	0.2	0.5	0.2	0.0
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
xn, yn	0.0	0.1	0.0	0.1	0.5	0.1	0.6	0.0

FIGURE 2.7: Colour by correlation algorithm: The steps in building a correlation matrix (a): Characterising which image colours (chromaticites) are possible under each reference illuminant. (b) Build a probability distribution for each light. (c) The distributions are encoded in the columns of the matrix. (image from [Finlayson et al., 2001]).

2.6.2 Machine learning techniques

Early research using machine learning for solving colour constancy problem used neural networks to estimate the colour of illuminant [Funt et al., 1996; Cardei et al., 2002]. Here the rg chromaticity diagram (like the one in Figure 2.7) is partitioned into several bins and the inputs to the neural network are binary values indicating the presence of a pixel in the image falling in the corresponding bin. The trained network will then be able to estimate the chromaticity of the illuminant of an input image. Complementary approaches apply support vector regression [Funt and Xiong, 2004] or linear regression techniques [Agarwal et al., 2007] to the same type of input data.

With the emergence of deep learning more developed versions of neural networks (e.g. [Bianco et al., 2015; Barron, 2015; lou]) have been proposed. The most recent and successful of all these methods is Convolutional Colour Constancy (CCC) method by [Barron, 2015] which treats colour constancy as a discriminative learning problem, i.e. instead of training a generative model based on high likelihoods to white-balanced images he trains a model to distinguish between white-balanced images and non-white-balanced images. Barron transforms the pixel's RGB values into a log-chromaninace space [Berwick and Lee, 1998; Hubel et al., 2007]. Where, if I is the input image, the uv log-chrominance values of the image are defined as $I_u = log(I_g/I_r)$ and $I_v = log(I_g/I_b)$. He makes the interesting observation that scaling the colour channels of an image (generating a tinted version of the image) induces a translation in the log-chromaticity histogram of that image. This is pictured in Figure 2.8. In this figure, different tinted versions of the same image and their log-chrominance histograms (with their axis be horizontal $= u$ and vertical $= v$) are shown. Changing the illuminant can result in a simple translation of the log-chromaticity histogram (see second row of Figure 2.8). In Figure 2.8 the log-chromaticity histogram of the middle image (labeled as true image) is the correct white-balanced image. So, the algorithm would hopefully be able to discriminate between this and the wrong tinted ones. Following this observation, Barron frames the colour constancy problem as a discriminative learning problem and using tools such as convolutional neural network the algorithm learns to localise a histogram in this 2D space (to read more please refer to [Barron, 2015]). Barron's method shows significant improvement over the state of the art methods.

Other learning-based approaches include: Exemplar-based [Joze and Drew, 2012], predicting chromaticity from luminance [Chakrabarti, 2015], colour constancy by classification [Oh and Kim, 2017], a real-time neural system designed for colour constancy [Moore et al., 1991], another method using neural network [Funt et al., 1997], etc.

Recent attempts in solving colour constancy using machine learning techniques have shown encouraging results. However, using machine learning as a solution to

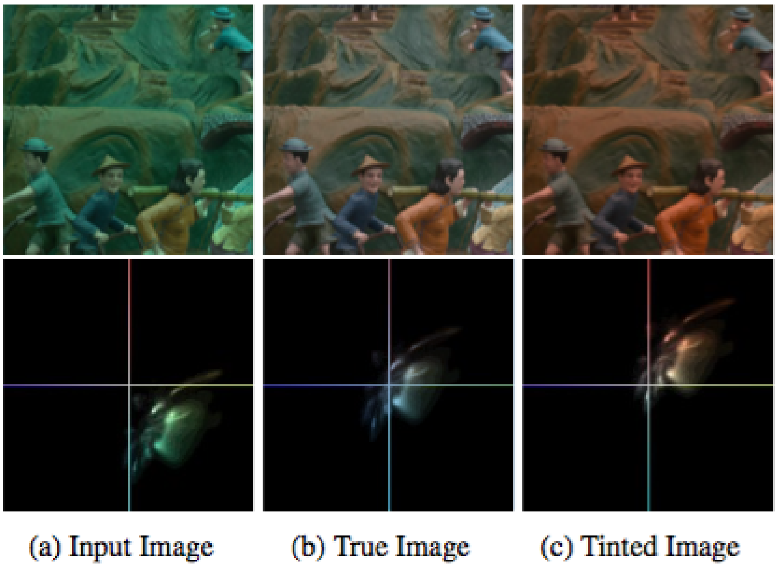


Figure 2.8: Log-chrominance histogram of different tinted versions of an image. Barron makes the interesting observation that scaling the colour channels of an image (generating a tinted version of the image) induces a translation in the log-chromaticity histogram of that image. This observation enables taking the convolutional approach to colour correction, in which the algorithm learns to localise a histogram in this 2D space (image taken from [Barron, 2015]).

illuminant estimation one always has to take into consideration the heavy computational burden it imposes. Also success of such techniques is very much dependant on the diversity of the training data, i.e. the more examples seen by the method the more accurate will be the estimated illuminant for an input image.

2.7 A Selection or Combination of Algorithms

Different Illuminant estimation algorithms can be combined to obtain good estimate of illuminant (e.g. [Cardei and Funt, 1999; Finlayson, 2013; Schaefer et al., 2005]).

In the committee-based method of [Cardei and Funt, 1999] it is shown that a weighted average of illuminant estimates of three algorithms outperforms the individual methods. In the following, [Cardei and Funt, 1999] calculates the weighted average of the rg chromaticity estimates by three algorithms (A simple neural network approach (NN) [Funt et al., 1996], grey-world method (GW) and white patch or MaxRGB (WP)):

$$[r_{NN} \;\; g_{NN} \;\; r_{GW} \;\; g_{GW} \;\; r_{WP} \;\; g_{WP}] \cdot C_{2\times 6}^{T} = [r_C \;\; g_C], \tag{2.23}$$

where $C_{2\times 6}$ is the weighted average matrix which is optimised in a least squares sense in the training step. Or, a non-linear combination (via a neural network based model) of the estimates can be found.

Finlayson [Finlayson, 2013] proposed the corrected-moment illuminant estimation algorithm. This work proposes a scheme to correct the results of moment-based algorithms, i.e. different instantiations of Minkowski framework (see Eq. 2.19) and grey-edge (see Eq. 2.20). The corrected-moment approach differs from the committee-based method in that it is not combining the results from two or more illuminant estimation algorithms. Rather, it corrects the estimated illuminants of m moments (such as 2^{nd} or higher moments) including the cross colour channel terms:

$$[\underline{\hat{\rho}}^{Est}]^t = \underline{\rho}_m^t C_{m\times 3}, \tag{2.24}$$

where $\underline{\rho}_m^t$ is a row vector compromising of m moments:

$$[E(R^2)^{0.5} \;\; E(G^2)^{0.5} \;\; E(B^2)^{0.5} \;\; ... \;\; E(RG)^{0.5} \;\; E(RB)^{0.5} \;\; E(GB)^{0.5}], \tag{2.25}$$

and $C_{m\times 3}$ denotes a $m \times 3$ regression matrix. The matrix C is learned through a training phase based on a set of known illuminant and their estimates. Unlike committee-based colour constancy, corrected-moment performs on the R, G and B estimates of the illuminant rather than the rg chromaticity values. A significant advantage of the corrected-moment method over the committee-based colour constancy is that it is exposure invariance, i.e. as the image data scales - e.g. due to a change of exposure or light brightness - so do the moments (that is if $\underline{\rho}^t_m \to \alpha\underline{\rho}^t_m$ then the estimated illuminant is $\alpha\hat{\underline{\rho}}^{Est}$). However, to find the best $C_{m\times 3}$ requires an iterative minimization where compensation for exposure is part of the formulation (details can be found in [Finlayson, 2013]). Corrected-moment algorithm while simple to use has a good performance compared to the state of the art methods.

Instead of combining the output of multiple algorithms into a more accurate estimate, the best algorithm can be selected based on the characteristic of an image. For instance, in [Gijsenij and Gevers, 2011] the intrinsic properties of natural images are used to select the most appropriate colour constancy method for every input image. Characteristics of natural images in terms of texture and contrast are captured using the Weibull parameterisation which captures the distribution of image derivatives. The Weibull statistics index the algorithm that should be used. Others [Bianco et al., 2010; Bianco and Schettini, 2014; Wu et al., 2010; Cheng et al., 2015c] have proposed using features other than Weibull parameterisation to select an algorithm for illuminant estimation.

In [Bianco et al., 2008] the selection (or combination) of the best algorithms or tuning of the algorithm is based on whether an image belongs to an indoor or outdoor scene. Further, Lu *et al.* [Lu et al., 2009] uses 3D geometry models to determine which colour constancy method to use for different geometrical regions found in images. In another approach [Van De Weijer et al., 2007b] use of high level visual information (semantic content) of an image is proposed as a clue for selection of the best illuminant estimation algorithm for each content.

2.8 Conclusion

Illuminant estimation is important as a preprocessing step for many computer vision tasks as well as being one of the major steps in camera pipeline to create colour images free of any casts created by colour of light.There are variety of algorithms proposed to solve the illuminant estimation. In this chapter we reviewed several algorithms which take different approaches to illuminant estimation, from simple statistics to more complicated techniques. We categorised them in four groups: 1) statistical-based algorithms, 2) gamut mapping algorithms, 3) learning-based algorithms and 4) selection of combination of algorithms. However, this categorisation is not abstract and some algorithms might fall in multiple groups.

The next chapter, will discuss different methods of evaluation of illuminant estimation algorithms as well as the existing bench-mark datasets in colour constancy.

Chapter 3

Background: Evaluation of Illuminant Estimation Algorithms

3.1 Introduction

Given the large body of illuminant estimation algorithms, it is important to agree on a framework for the evaluation and comparison of the performance of these algorithms. Figure 3.1 shows the common workflow usually followed for evaluating the performance of an illuminant estimation algorithm.

Most of the literature in colour constancy provide a relative comparison of the new proposed method with the state of the art algorithms using a summary of errors over a set of images form a benchmark dataset. Some work [Hordley and Finlayson, 2006; Gijsenij et al., 2009a] has both investigated the importance of choosing a proper error metric for evaluating the performance of illuminant estimation algorithms and also how the discovered errors should be analysed.

This chapter starts with an overview of the most popular benchmark datasets used in colour constancy research (Section 3.2). In Section 3.3, different metrics used for evaluation of the performance of illuminant estimation algorithms are

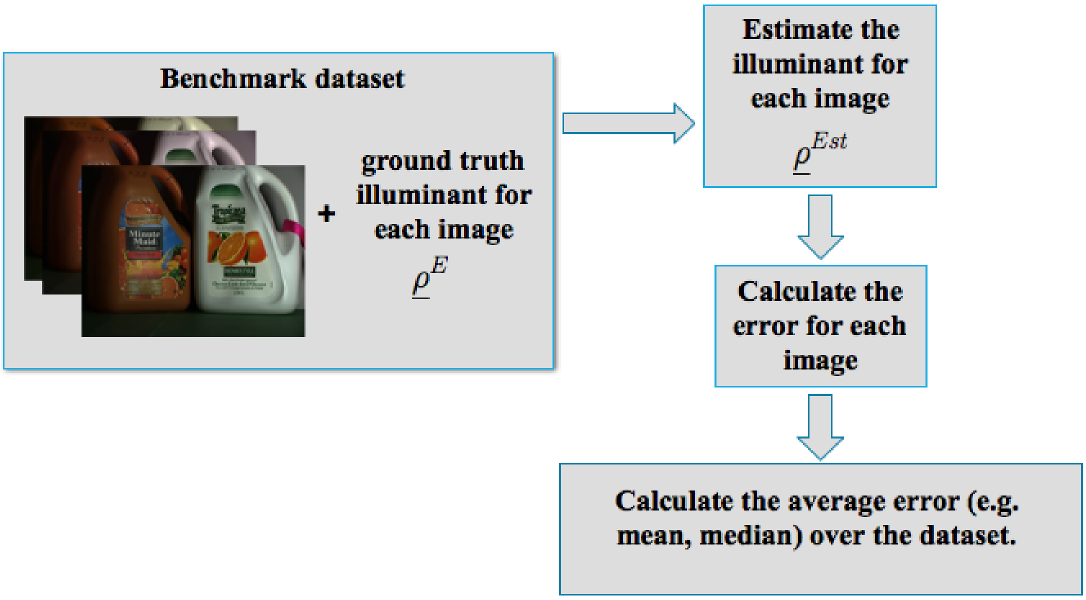

Benchmark dataset
+
ground truth illuminant for each image
$\underline{\rho}^E$
Estimate the illuminant for each image
$\underline{\rho}^{Est}$
Calculate the error for each image
Calculate the average error (e.g. mean, median) over the dataset.

FIGURE 3.1: General framework of performance evaluation of illuminant estimation algorithms.

presented. The way the error data is often reported in the literature is discussed in Section 3.4. The psychophysics experiments for acquiring the observers' point of view are introduced in Section 3.5. Also, since the new proposed methods are always compared against the existing methods in Section 3.6 different statistical tests performed to rank the algorithms within an acceptable significant difference are reviewed. Section 3.7 concludes the chapter.

3.2 Benchmark datasets

A benchmark dataset for colour constancy usually includes a number of images captured under a variety of lighting conditions such as indoor and outdoor situations. Any benchmark dataset also provides the ground-truth illuminant colour for every image. Generally, the colour of the light is defined to be the RGB of a physical white or achromatic surface placed in a scene. Often the Macbeth colour checker [xri] (a standard reference chart with 24 colours is used). In [Ciurea and Funt, 2003] a grey ball is placed in every scene.

Commonly used evaluation datasets include:

- **Synthetic dataset:** Barnard et al. [Barnard et al., 2002a] provide a large corpus of 287 illuminant and 1995 surface reflectance spectra. Random selections of surfaces for a given light are numerically integrated to make RGBs. These RGBs in turn can be used as the input to illumination estimation algorithms. The advantage of the synthetic test method is that the data is 'clean'. There is no image noise and the world is perfectly Lambertian.

- **Hyperspectral dataset:** There are a few measured hypersectral data sets (e.g. Foster et al.'s dataset [Foster et al., 2006; Nascimento et al., 2002]). Here again numerical integration can be used to form RGB images. An advantage of this dataset is that the researcher explicitly knows the spectral sensitivity of the camera under investigation.

- **SFU (Simon Fraser University) dataset:** This dataset [Barnard et al., 2002c] consists of different scenes captured using Sony DXC-930 3CCD camera under 11 different lights in laboratory environment. The light sources include three fluorescent lights (Sylvania warm white, Sylvania cool white, and Philips Ultralume), four different incandescent lights, and these four used in conjunction with a blue filter (Roscolux 3202). The spectrum of one of the incandescent sources (Sylvania 50MR16Q) is very similar to a regular incandescent lamp. The other three have spectra which are similar to daylight of three different color temperatures (Solux 3500K, Solux 4100K, Solux 4700K). When used in conjunction with the blue filter, these bulbs provide a reasonable coverage of the range of outdoor illumination. The dataset is captured under 11 lights. There are 321 images in total. There are 21 scenes which are mainly Lambertian and 10 more that contain specular objects. The correct illuminant RGB for this dataset is measured by placing a white tile in each scene and finding the average RGB response for the tile. The SFU images are linear (natural camera raw).

- **Greyball (videoframes) dataset:**Ciurea and Funt [Ciurea and Funt, 2003] placed a grey sphere within the field of the view of a digital camera and produced a dataset consisting of 11000 frames of video. The grey ball is used as a reference for calculating the groundtruth illuminant colour. Using a grey sphere instead of the standard grey card facilitates measurement of the variation in illumination as a function of incident angle. The greyball images are rendered (post-the camera processing pipeline).

- **Barcelona dataset:** The Barcelona dataset [Vazquez-Corral et al., 2009] is made using a Sigma Foveon D10 camera which is a DSLR camera. The major difference between this dataset and others is by calibrating the camera they manage to recover the values for each pixel in CIE1931 XYZ coordinates.

- **Gehler dataset:** Perhaps the first dataset for colour constancy with typical photography images was provided by [Gehler et al., 2008]. It contains 568 images of a variety of indoor and outdoor shots taken around Cambridge, England by two DSLR cameras (Canon 1D with 86 images and Canon 5D with 482 images) with all settings in auto mode. The reference for calculating the ground-truth illuminant colour is a Macbeth colour-checker chart [xri] located in every scene. The last six neutral patches of the colour-checker often count as a clue for the ground-truth illuminant. However, when applying a colour constancy algorithm on the images the colour-checker needs to be occluded to create a real photographic situation. For this purpose the coordinates of the colour-checker is given with the dataset. Shi-Gehler [Shi and Funt, 2010] is a linear raw 12-bit Portable Network Graphics (PNG) version of Gehler dataset. In raw spaces all the images have a cyan tint. Later, Lynch *et al.* [Lynch et al., 2013] updated the dataset by re-rendering the raw images and allowing DCRAW [1] to apply a D65 Colour Correction

[1] An open-source computer program which is able to read numerous raw image formats [Coffin, 2008]

matrix to all images. As a result the strong Cyan tint on the images is removed but the data is still linear.

- **NUS (National University of Singapore) dataset:** A recent dataset for colour constancy resembling real-life images is NUS (National University of Singapore) [Cheng et al., 2014]. This dataset is made using eight different cameras each captured around 217 images in average. The ground-truth illuminant colour is recovered using the same method as [Gehler et al., 2008] (the neutral patches Macbeth colour-checker). Again, the linear raw data is provided using DCRAW for research purposes.

- **HDR images:** The high dynamic dataset by SFU [Funt and Shi, 2010] consists of 105 scenes captured by Nikon D700 digital still camera. Each scene is captured up to nine exposures and the raw 16-bit Portable Network Graphics (PNG) format (lossless compression) images are created from the NEF data using DCRAW. After aligning the base images the hdr images were created using matlab built-in function *makehdr*. Every scene is captured twice, once with four GretagMacbeth mini Colorcheckers positioned at different angles with respect to one another and once without them.

  Figure 3.2 shows a couple of examples of benchmark datasets mentioned in this section.

3.3 Evaluation metrics

Different research [Barnard et al., 2002a,b; Hordley and Finlayson, 2004, 2006; Funt et al., 1998; Li et al., 2011] have studied the relative performance of illuminant estimation algorithms and explored the existing problems in the evaluation techniques.

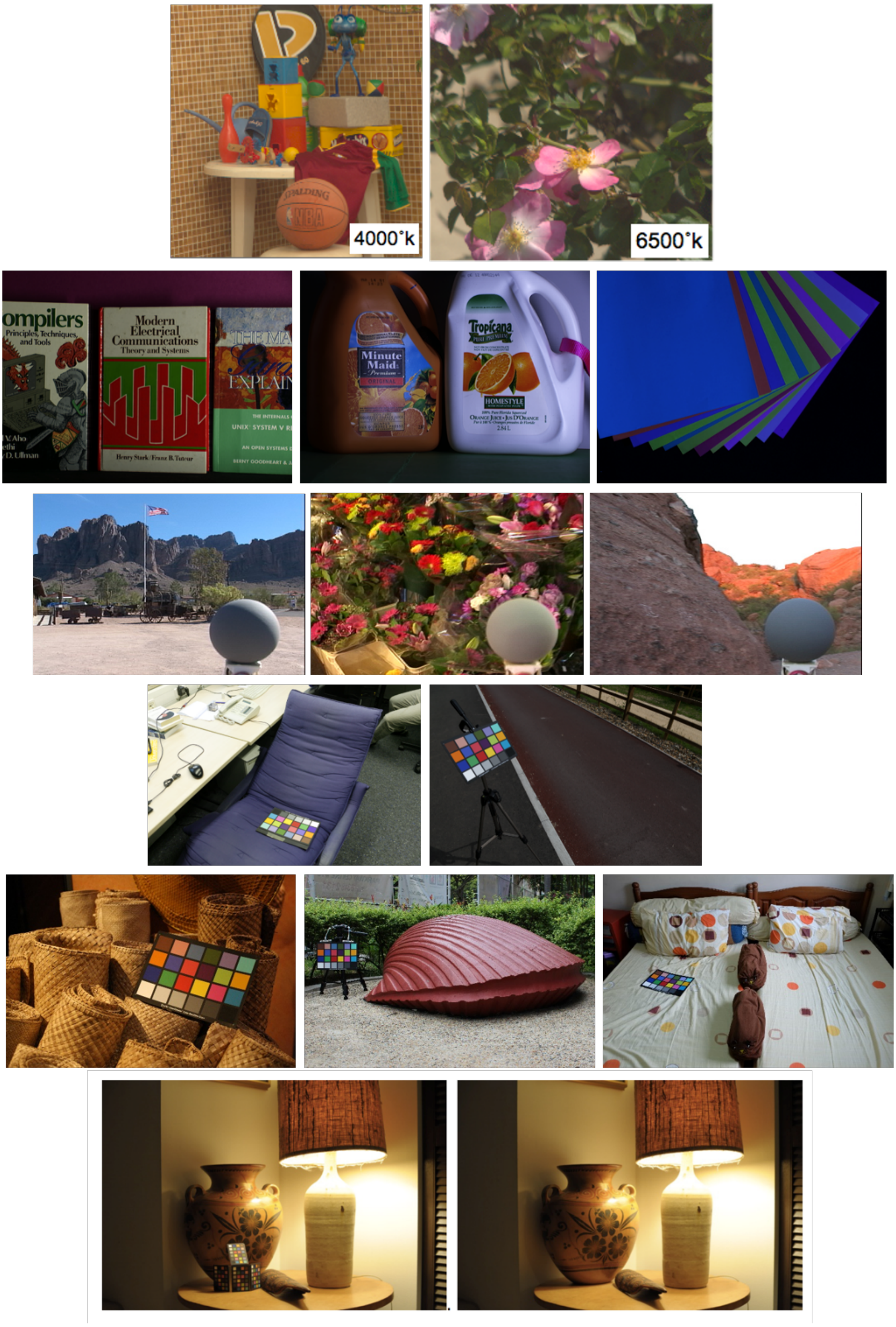


Figure 3.2: Example of different datasets: row 1- Synthesised hyperspectral dataset [Foster et al., 2006]. row 2- SFU dataset [Barnard et al., 2002c]. row 3- Greyball dataset [Ciurea and Funt, 2003]. row 4- Shi-Gehler dataset [Gehler et al., 2008; Shi and Funt, 2010]. row 5- NUS dataset [Cheng et al., 2014]. row 6- HDR dataset [Funt and Shi, 2010] (with and without the mini colour checker).

When a new illuminant estimation algorithm is proposed apart from the white-balanced images represented in the literature, a great part of evaluation is done by reporting the average error of the algorithm over a benchmark dataset. As discussed before in Section 3.2, the datasets vary from real images of different types to synthetic images. Barnard *et al.* [Barnard et al., 2002a,b] suggest an empirical framework in which algorithms are tested on sets of synthetic and real test data. In [Hordley and Finlayson, 2006] it is shown that the empirical framework and the choice of error metric has a significant effect on the judgment of algorithms.

3.3.1 Euclidean distance

The *Euclidean distance* between the chromaticity vectors ($r = R/(R+G+B)$, $g = G/(R+G+B)$) is calculated as:

$$err_{Euc}(\underline{c}^E, \underline{c}^{Est}) = \sqrt{(c_r^E - c_r^{Est})^2 + (c_g^E - c_g^{Est})^2}, \tag{3.1}$$

where (c_r^E, c_g^E) and (c_r^{Est}, c_g^{Est}) are chromaticity coordinates of true and estimated illuminant respectively.

Some [Gijsenij et al., 2009a] have considered calculating the Euclidean distance with taking into account the third chromaticity vector, i.e. $b = B/(R+G+B)$:

$$err_{Euc}(\underline{c}^E, \underline{c}^{Est}) = \sqrt{(c_r^E - c_r^{Est})^2 + (c_g^E - c_g^{Est})^2 + (c_b^E - c_b^{Est})^2}. \tag{3.2}$$

Finally, the weighted Euclidean distance or perceptual Euclidean distance (PED) is introduced by [Gijsenij et al., 2009a] which associates weights to different colour channels:

$$\mathrm{PED}(\underline{c}^E, \underline{c}^{Est}) = \sqrt{w_R(c_r^E - c_r^{Est})^2 + w_G(c_g^E - c_g^{Est})^2 + w_B(c_b^E - c_b^{Est})^2}, \tag{3.3}$$

where $w_R + w_G + w_B = 1$. The associated weights are based on the property of human vision system which states a deviation in one colour channel might have a stronger effect on the perceived difference between two images than a deviation in another channel.

3.3.2 Brunswick Ratio

The Brunswick Ratio [Leibowitz, 1956] also known as colour constancy index CCI [Arend et al., 1991] is often used to measure perceptual colour constancy [Delahunt and Brainard, 2004] and is defined as follows:

$$r = \frac{||D - S||}{||P - S||}, \tag{3.4}$$

where D denotes the estimated light source , P a white reference light, S the true (measured) light source (in a human vision referenced chromaticity space), and $||x - y||$ the distance between x and y in a chromaticity space. Usually during the evaluation different colour spaces can be used to compute the absolute difference between the lights. The index value is typically between zero and one.

3.3.3 Recovery Angular Error

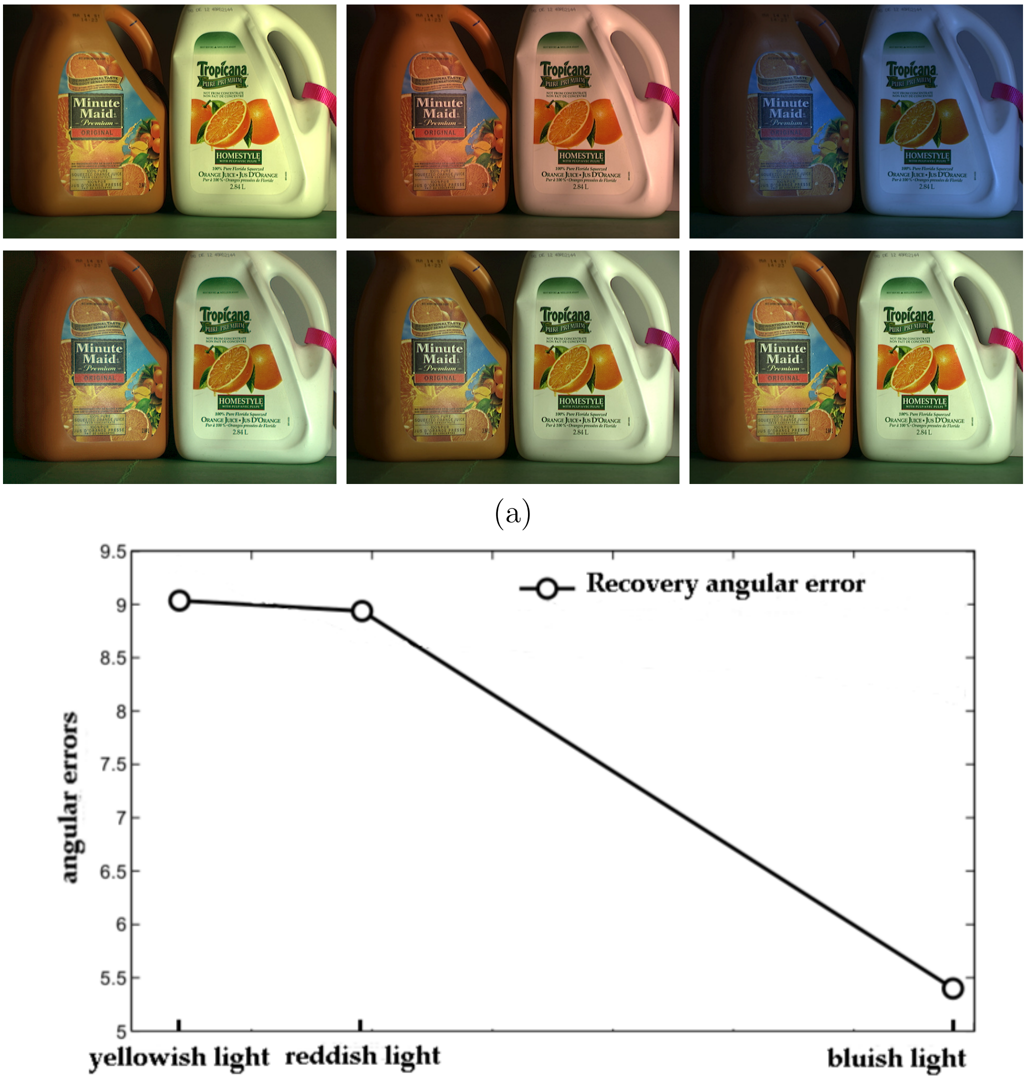
The *angular error* is the most popular metric used for evaluating the performance of illuminant estimation algorithms and it is calculated [Finlayson et al., 1995] as:

$$err_{recovery} = \cos^{-1}\left(\frac{(\underline{\rho}^{E} \cdot \underline{\rho}^{Est})}{\|\underline{\rho}^{E}\|\|\underline{\rho}^{Est}\|}\right), \tag{3.5}$$

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. Throughout this thesis the traditional angular error is called *recovery angular error*.

The recovery angular error is widely used for evaluating the illuminant estimation algorithms. However, recovery angular error (as well as Euclidean distance) assesses only the accuracy of the estimated illuminant colour and not the quality of reproduced images. That this is a problem is illustrated in Figure 3.3. The first row in Figure 3.3 (a) shows three images from SFU dataset [Barnard et al., 2002c] of the same scene captured under three different illuminants (from left to right: Philips Ultralume fluorescent, Sylvania warm white fluorescent and Solux- 4700K+blue filter). The second row shows same images with their colours corrected using the estimate by grey-world algorithm [Buchsbaum, 1980]. The recovery angular error of grey-world algorithm for the three images can be see in Figure 3.3 (b). Although recovery angular error provides a reasonable prediction of the error of illuminant estimation for the three images, the range of error from 5.5° to 9° is relatively high. Just a change in the colour of illuminant results in a 3.5 degrees decrease in angular error for the same algorithm. Naturally while evaluating an algorithm such a change in the error will have a significant affect on our judgment about the algorithm.

Of course it needs to be mentioned that changes in the illuminant are only due to the exposure variances when they follow the simplest model of illuminant changes. But in reality there could be many other changes in light, such as a shift in the colour values [Van De Sande et al., 2010].



(a)

(b)

Figure 3.3: An example of similar colour corrected images with varying recovery angular error. (a) First row: images of the same scene captured under chromatic illuminants (from SFU Lab dataset [Barnard et al., 2002c]). Second row: Corrected images using grey-world algorithm [Buchsbaum, 1980]. (b) The Recovery angular errors.

An error measure for illuminant estimation algorithms needs to be ideally simple to calculate, correlates with the human perception of colour reproduction and not to be very sensitive to changes in the colour of illuminant. Angular error is a simple error measure which according to [Gijsenij et al., 2009a] correlates with human perception of the performance quality of illuminant estimation algorithms more than other error measures. However, as it was shown in Figure 3.3, as much as it does a reasonable evaluation of the algorithm's performance, just a change in the colour of light due to the exposure variance results in a relatively wide range of errors. An improvement to the recovery angular error is one of the main contributions in this thesis (chapter 3 and 4). The new error measure is less effected by changes in the colour of light due to exposure. Whether the proposed error measure is robust towards other changes in the illuminant (apart from exposure changes) could be an avenue for future research.

3.4 Aggregate Error Values

It is common when comparing algorithms' performance to look at the "average" performance of the tested methods over a set of images [Hordley and Finlayson, 2006]. It is shown that the average performance in the form of mean or root mean square - e.g. of the recovery angular error - do not give an accurate summary of the underlying distribution of error data. Since the distribution of error data could be skewed and mean value of data could result in poor analysis, it is argued [Hordley and Finlayson, 2006; Gijsenij et al., 2009a] that the median and trimean are more appropriate measures for summarising the data. The trimean of a distribution is defined as the weighted average of the first, second, and third quantile (25%, 50% [median] and 75%) errors. Recently, researchers have also presented a wider range of statistics, e.g. worst 25% and best 25% errors [Cheng et al., 2014].

Whether summarising results in the form of an aggregate over the whole dataset or analysing the distribution of data using a box plot, the relationship of these statistics across different methods is unclear. It is also interesting to analyse the performance of algorithms on individual images and whether there are algorithms that commonly fail for certain outlier images. A part of this thesis (Chapter 8) proposes a framework which enables us to detect hard images in colour constancy (as these images are not well represented in the summary statistics).

3.5 Perceptual analysis

3.5.1 Just Noticeable Difference

Regarding the noticeable error by human observers in colour-corrected images, Funt *et al.* [Funt et al., 1998] state that the minimum root mean squared Euclidean error of the estimated chromaticity for accurate colour-based object recognition is 0.04. In terms of angular error, a deviation of 1° with respect to the ground truth was found to be not noticeable, while an angular error of 3° was found noticeable but acceptable [Finlayson et al., 2005; Fredembach and Finlayson, 2008]. Further, Hordley [Hordley, 2006] derives that an angular error of 2° represents good enough color constancy for complex images. Also, another important outcome of experiment by Gijsenij *et al.* [Gijsenij et al., 2009a] is to indicate whether an observer is sensitive to the difference between the reproduction results of two illuminant estimation algorithm. They concluded that the difference in terms of angular error between two methods A and B should be at least $0.06 \times err_{max}$ to be noticeable, where $err_{max} = \max(err_A, err_B)$. This means that for instance if method A has an angular error of 10°, then an improvement of at least 0.6° is necessary; otherwise the improvement will not be visible to a human observer. They state that this finding is in line with the values for the Weber fraction found in visual perception [Cornsweet, 1970]. Although this JND (Just Noticeable Difference) is based on their experiments on the hyperspectral data and might vary depending on the scene content. Later, it is suggested by [Banic and Loncaric, 2015] that if the angular error is more than one, instead the natural logarithm of the angular error be used (their suggestion is based on the Weber's law [Thurstone, 1927] the just noticeable difference increases linearly with the absolute error).

3.5.2 Perceptual Distances

In [Gijsenij et al., 2009a, 2008], Gijsenij *et al.* proposed a perceptual distance for colour constancy. First, they convert the RGB values of the colour of ground-truth and estimated illuminants to a human vision colour space such as CIElab (See Appendix A). After which, they compare the two illuminants. However, the conversion between RGB to CIElab requires a few assumptions such as the reference white point. In [Gijsenij et al., 2009a, 2008] the error measurements and experiments are designed for the reference white point of D65 and the sRGB colour profile [Commission et al., 1999].

In the same study [Gijsenij et al., 2009a], Gijsenij *et al.* also conducted a psychophysics study to reveal the correlation between the human perception of white balanced images and many distance measures including recovery angular error.

In their experiment, the observers were shown four images at once. Two identical images at the top which are the reference images white balanced by the ground-truth illuminant and two images at the bottom which are corrected using the estimates of two different illuminant estimation algorithms. The observers are then asked which of the reproduced images in the bottom row they prefer. They compare five illuminant estimation algorithms on a set of images and score them based on observers' preferences. The results are then compared with different distance measures.

Regarding recovery angular error, they concluded that the correlation between this metric and the perception of the human observer is reasonably high. However, for some images the correlation is very low. In a closer analysis, they find that for these images the observers judge the results of white balance to be much worse than indicated by the recovery angular error (meaning that human observers do not agree with the angular error.).

Another research proposing perceptual-based performance measurement of colour constancy algorithm is [Vazquez-Corral et al., 2009]. Vazques *et al.* [Vazquez-Corral et al., 2009] follow more or less the same pair-wise comparison method as in [Gijsenij et al., 2009a], except in their experiment the observers are shown two images and are asked to choose the "most natural" one. Their experiment only involves three illuminant estimation algorithms; Grey-world, Shades of Grey and Maxname [Vazquez-Corral et al., 2009]. Vazques *et al.* study results in defining a new measure of an algorithm's accuracy which is the angle between the perceived white point of a scene and the estimated illuminant. The perceived white point has to be measured from the chosen colour-corrected images of the scene during their proposed psychophysical experiment.

3.6 Ranking and comparison of algorithms

Ultimately, we wish to develop a way to conclude that one algorithm is better than the other. However, Hordley and Finlayson [Hordley and Finlayson, 2006] noted that a single summary statistic - such as the mean - does not adequately summarise the underlying distribution and further the fact that one algorithm has a lower mean value than another does not necessarily indicate that one algorithm is better than the other. Hordley and Finlayson [Hordley and Finlayson, 2006] rank algorithms according to statistical significance. They hypothesise that one algorithm is better than another and then test this hypothesis using appropriate statistical tools and the error distributions of each algorithm over a large set of sample images. Since the error distributions are not well described by standard statistical distributions (e.g. a normal distribution) non-parametric tests which are independent of the underlying distribution are more suitable for this purpose.

- **Wilcoxon Signed-Ranks test:** Suppose one wish to compare the relative performance of two algorithms in terms of their median angular error. Let

A and B be random variables representing the error in algorithm A and B's estimate of the scene illuminant. The Wilcoxon Sign Test[Conover, 1999] can be used to test the hypothesis that the random variables A and B are such that $p = P(A > B) = 0.5$. In other words the hypothesis says that algorithm A and B have the same median:

$H_0 \ : \ p = 0.5$, the medians of the two distributions are the same

An alternative hypothesis can also be defined as:

$H_1 \ : \ p < 0.5$, algorithm A has a lower median than algorithm B.

For independent pairs $(A_1, B_1) \ . \ . \ . \ (A_N, B_N)$ of errors for N different images, W is denoted as the number of images for which $A_i > B_i$. When H_0 is true, any particular observed value of W belongs to a sampling distribution whose mean is equal to zero (i.e. $p = 0.5$). We then compare W against its critical value ω (from the Wilcoxon Signed-Ranks Table) for N samples at α significance level (e.g. $\alpha = 0.05$). If $W > \omega$, we can't reject the null hypothesis (i.e. $P(W > \omega) \geq 0.05$) and so it is concluded that the medians of the two errors are the same. If $P(W \leq \omega) < \alpha$ we reject the null hypothesis H_0 and accept the alternative hypothesis H_1 at the significance level α.

The value of α determines the probability with which the null hypothesis is rejected when it is in fact true. So, for example if $\alpha = 0.05$ and the calculated probability is 0.04 then the null hypothesis is rejected at the 0.05 significance level. In this case we will be correct in rejecting the null hypothesis 95% of the time. To be more sure that we are correct the significance level can be decreased.

When $N > 30$ the Wilcoxon Signed-Ranks statistic follows the z distribution (standard normal distribution). Then W is used to calculate the z score:

$$z = \frac{W - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}, \tag{3.6}$$

We then find the area above z if z is positive or beyond z if it is negative (using a table of areas under the normal distribution curve). This value shows the probability of occurrence of W (that if it is very likely to happen, the null hypothesis will be accepted). For instance, if the area above the calculated z is 0.0099, the value of W is likely to occur by chance with a probability of 0.0099.

Using Wilcoxon Sign Test it is implicitly assumed that the median is a good summary statistic for the distributions.

Introducing a new algorithm or applying a different error metric might change the ranking of algorithms for a bechmark dataset. To study whether the change in the ranking of algorithm is significant or not, Kendall's test can be performed.

- **Kendall's test:** Kendall test [Sprent and Smeeton, 2007; Conover, 1999]. A change in the ranking of a selection of illuminant estimation algorithms can be considered as a permutation problem. Kendall test is a method to compare two permutations and it correlates to the number of exchanges needed in a bubble sort to convert one permutation to the other [Fagin et al., 2003].

  The Kendall's test statistic T 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 agree, otherwise the pair are concordant. T is defined as:

  $$T = C - D, \tag{3.7}$$

  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 the case of a tie the pair is counted as 1/2 concordant and 1/2 discordant, although as it is obvious by Eq. 3.7 this makes no difference in our final Kendall's T value.

To study the discordance in ranking of the algorithms, the Lower-Tailed Kendall's Test [Conover, 1999] is performed which is defined as follows:

Lower-Tailed Test

H_0 : X and Y are independent.

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 [Conover, 1999]. □

3.7 Conclusions

Considering the diversity of illuminant estimation algorithms, it is important to agree on a common workflow for measuring the accuracy of algorithms and analyse the error data to compare the algorithms on a selection of images. In this chapter we reviewed metrics and workflows commonly used for evaluating the performance of illuminant estimation algorithms. The most popular way of measuring the error for an algorithm is to calculate the angle between the RGB vectors of the true and estimated illuminant. However, with an example in Section 3.3.3 we briefly showed how the angular error, or as it is called in this thesis *"recovery angular error"*, has a weakness. Specifically that when the same image reproduction is produced that different error is calculated.

Chapter 4

Reproduction Angular Error

The angle between the RGBs of the measured and estimated illuminant colours - the recovery angular error - has been widely used to evaluate the performance of illuminant estimation algorithms. However, this metric is not in line with how illuminant estimates are used. Normally, illuminant estimates are 'divided out' from the image to, hopefully, provide image colours that are not confounded by the colour of the light. However, even though the same reproduction results, the same scene might have a large range of recovery errors. In this chapter, the scale of the problem with the recovery error is quantified. Further, a new metric for evaluating the performance of illuminant estimation algorithms; 'Reproduction Angular Error'; [Finlayson and Zakizadeh, 2014; Finlayson et al., 2016] is introduced which is more in line with the application of the estimated illuminants. We will demonstrate that the new metric shows much more stability towards changes in the colour of illuminant compared to the recovery angular error.

4.1 Introduction

To measure the performance of an illuminant estimation algorithm, usually the RGB of the estimated light is compared with the RGB of a ground-truth measured illuminant by calculating the angle between the two vectors. This metric is known as angular error, or as we call it in this thesis *recovery angular error*. As mentioned before in Section 3.3.3, the recovery angular error is the most common metric used to quantify illuminant estimation error [Hordley and Finlayson, 2006; Gijsenij et al., 2009a]. Although this metric has been previously introduced in Section 3.3.3, its definition is repeated below (since it will be referred to very often in this chapter):

$$err_{recovery} = \cos^{-1}(\frac{(\underline{\rho}^E \cdot \underline{\rho}^{Est})}{\|\underline{\rho}^E\|\|\underline{\rho}^{Est}\|}), \tag{4.1}$$

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. The final error for an illuminant estimation algorithm reported, is usually an average error (e.g. mean, median, quantiles) over the whole dataset. The algorithms are ranked according to their reported errors. Gijsenij et al. [Gijsenij et al., 2011] have done a comprehensive study of several illuminant estimation algorithms and have provided the ranking for these algorithms for many benchmark datasets.

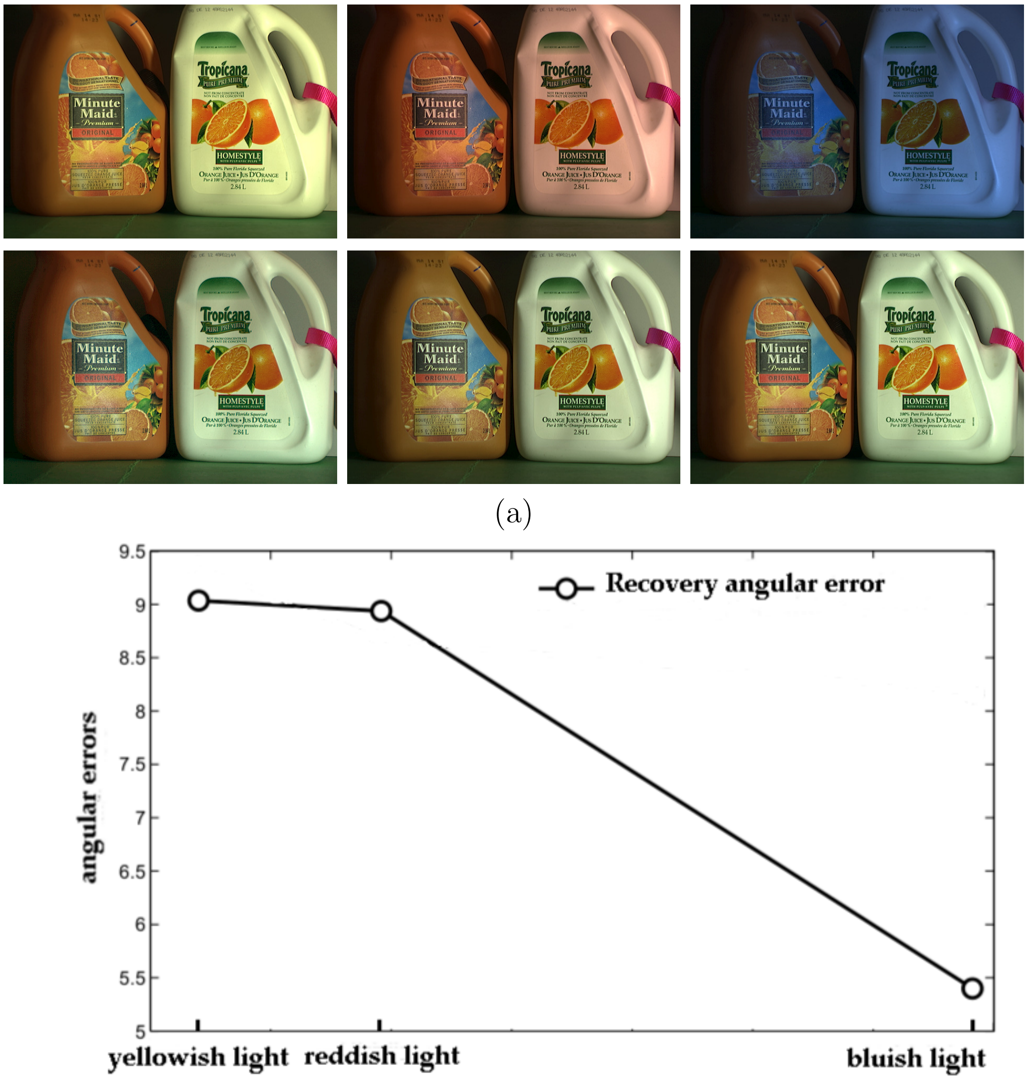
In this chapter, we show that recovery angular error has a fundamental weakness. Further, we introduce a new metric for evaluating the performance of illuminant estimation algorithms and we discuss its stability over the changes in illumination.

The organisation of this chapter is as follows: Section 4.2 discusses the problem with the recovery angular error. In Section 4.3, the new metric for evaluating the performance of illuminant estimation algorithms, which we call **Reproduction Angular Error**, is introduced. This chapter is concluded in Section 4.4.

4.2 The Problem with Recovery Angular Error

According to the diagonal model of illuminant change (see Eq. 2.13), changes in the colour of light can be simulated by multiplying the R, G and B values of the light. Interestingly, it can be noticed that the colours of an image are also corrected following the same principle, i.e. they are divided by the colour of the estimated illuminant to almost look as if they have been captured under the white reference illuminant ($\underline{U} = [1\ 1\ 1]^t$). Following this observation, imagine an object in the exact same environment being captured under different colours of light. Now if a simple algorithm like grey-world [Buchsbaum, 1980] which estimates the illuminant by averaging the colours of the scene is used, we are expecting the same colour-corrected images by the algorithm regardless of the colour of light. Although other factors such as specularity (which are often ignored while addressing the colour constancy problem) might affect the performance of an illuminant estimation algorithm, colour of the light (when it is the only changing element) is not expected to have a significant affect on the algorithm's performance. The question is whether recovery angular error is robust enough against the changes in the colour of light. If it can provide a reasonable range of error for the same algorithm and the same scene when only the colour of light is changing from one image to another.

Figure 4.1, which was previously shown in Section 3.3, is a good demonstration of this problem. In Figure 4.1, the weakness of recovery angular error in evaluating the performance of an algorithm is illustrated. In the top row of Figure 4.1, three images of the same scene from the SFU Lab dataset [Barnard et al., 2002c] are shown, which were captured under different chromatic illuminations (From left to right: solux-4700K+blue filter, Sylvania warm white fluorescent and Philips Ultralume fluorescent). The RGB colour of the illuminant for each scene is then estimated using the simple grey-world algorithm [Buchsbaum, 1980] and then we divide the R, G and B values at each pixel of the image by this estimate to remove



(a)

(b)

Figure 4.1: An example of similar colour corrected images with varying recovery angular error. (a) First row: images of the same scene captured under chromatic illuminants (from SFU Lab dataset [Barnard et al., 2002c]). Second row: Corrected images using grey-world algorithm [Buchsbaum, 1980]. (b) The Recovery angular error.

the colour bias due to illumination. The results of ‘dividing out’ are shown in the second row of the same figure. Notice that the reproduced images look better as the colours of the objects are not biased by the illuminant colour. In this case the grey-world algorithm has delivered good illuminant estimation equally for the three images. In part (b) of Figure 4.1 we plot the recovery angular errors for the given algorithm. Even though the output reproductions are similar the recovery errors are quite different. The recovery errors range from 5.5 to 9.5 degrees. This can be very misleading and the performance of the algorithm (in this case the grey-wrold algorithm) might be interpreted wrongly.

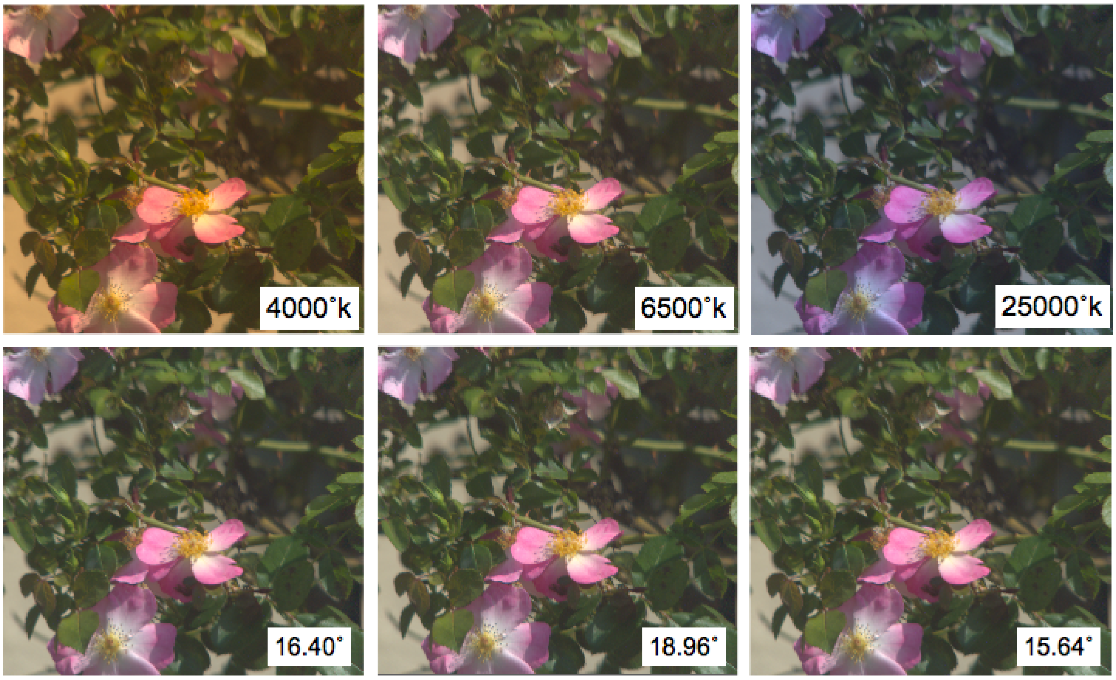


FIGURE 4.2: An example of similar colour-corrected images with varying recovery angular error. First row: hyperspectral images [Foster et al., 2006] rendered in sRGB under three lights of different temperature. Second row: the images are white balanced using the shades of grey algorithm in hyperspectral space before converting to sRGB (The recovery angular error can be seen on each image).

Another example of the different range of recovery angular errors for a similar scene and algorithm can be seen in Figure 4.2. The images in the top row of Figure 4.2 are the captured data from a hyperspectral camera [Foster et al., 2006] rendered under three lights with different spectra, from left to right: 4000°k, 6500°k and 25000°k illuminants. To display the images, first CIE 1931 colour matching functions [Wyszecki and Stiles, 1982b] are used to get the X, Y, Z values at each pixel and then the X, Y, Z values are converted to their corresponding RGB colour representation in sRGB ($IEC_6 1966 - 2 - 1$). The light with a temperature of 6500°k is similar to daylight and falls in the central white point of the CIE 1931 chromaticity diagram (see Appendix B). Moving away from 6500°k light results in colourfull lights like the bluish light with the temperature of 25000°k or the yellowish light with the temperature of 4000°k. In the second row of the figure, the white-balanced images by shades of grey algorithm [Finlayson and Trezzi, 2004] ($p = 2$) are shown, along with the recovery angular error of the algorithm for each image. It can be seen in the second row of Figure 4.2 that the images reproduced by dividing out the estimates of light made by shade of grey algorithm are exactly the same. However, this observation is not supported by the error values reported by recovery angular error.

With the synthetic images we have control over the factors involved in the image formation. To bypass the effects of XYZ to RGB conversion as well as the CIE 1931 colour matching functions, we performed the shades of grey algorithm and calculated the recovery angular errors in hyperspectral space before any conversions. In other words, the illuminant estimation is done on the radiance data which is the product of the scene reflectances and the illuminant spectra. This enables us to study the effect of the lights with different spectra on the range of recovery error for a single algorithm. It can be seen in Figure 4.2 that different lights result in different recovery angular errors for the same image and the same algorithm.

4.2.1 The Range of Recovery Angular Error

In this section, we address the problem of the mismatch between the recovery angular error and the reproduced images (as illustrated in Figure 4.1 (b) and Figure 4.2). We calculate how large and small the mismatch between recovery errors and images reproduced can be.

Recalling the diagonal model of illumination change [von Kries, 1902] (Eq. 2.13), where the RGB response of a device to the same surface viewed under two different lights are related by three factors of a diagonal matrix:

$$\begin{pmatrix} \rho_R^{E',S} \\ \rho_G^{E',S} \\ \rho_B^{E',S} \end{pmatrix} = \text{diag}\,(\underline{d}) * \begin{pmatrix} \rho_R^{E,S} \\ \rho_G^{E,S} \\ \rho_B^{E,S} \end{pmatrix} \quad \underline{d} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \quad \alpha, \beta, \gamma \geq 0 \tag{4.2}$$

and assuming that the illuminant estimate $\underline{\rho}^{Est}$ can be viewed as some statistical moment of the RGB values of an image with N pixels (towards the end of this thesis, *moments* refer to statistical moments such as average, maximum, etc. which do not change by scaling the data, similar to those in Eq. 2.25):

$$\underline{\rho}^{Est} = moment(\underline{\rho}^{E,S_1}, \underline{\rho}^{E,S_2}, ..., \underline{\rho}^{E,S_N}) \tag{4.3}$$

The estimated illuminant for the second light E' based on the first light E can be written as:

$$\text{diag}\,(\underline{d}) * \underline{\rho}^{Est} = moment(\underline{\rho}^{E',S_1}, \underline{\rho}^{E',S_2}, ..., \underline{\rho}^{E',S_N}) \tag{4.4}$$

We notice again that Eq. 4.4 teaches that if two lights are related by three scaling factors $\underline{d}$ then the statistical moment estimates shift by the same scaling factor as well. Equation 4.4 is true for most illuminant estimation algorithms including all that can be written in the Minkowski-framework[Finlayson and Trezzi, 2004] (see Eq. 2.19). Considering Eq. 4.2 and Eq. 4.4, we seek the illuminants that result in the largest and the smallest recovery angular errors.

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 maximises recovery* *angular error is an illuminant with 0 in exactly one of the either R, G or B channels.*

Proof. From the diagonal model of illumination change (Eq. 4.2) and without the loss of generality we assume that the reference illuminant is $\underline{U}$ (if it is not, we can map the illuminant to $\underline{U}$ using 3 scaling factors). For a given scene under the reference light, $\underline{\mu}$ is a moment type illuminant estimate. Under a second light $\underline{d}$ and remembering Eq. 4.3 and Eq. 4.4 we have the new illuminant estimate $\underline{d} * \underline{\mu}$ and the recovery angular error (Eq. 4.1) can be written as:

$$err_{recovery}(\underline{d}, \underline{d} \times \underline{\mu}) = \cos^{-1}\left(\frac{(\alpha^2\mu_r + \beta^2\mu_g + \gamma^2\mu_b)}{\sqrt{\alpha^2+\beta^2+\gamma^2}\sqrt{(\alpha\mu_r)^2 + (\beta\mu_g)^2 + (\gamma\mu_b)^2}}\right) \tag{4.5}$$

where $\underline{d} = [\alpha\ \beta\ \gamma]$. Since in illuminant estimation we are only interested in the orientation of $\underline{d}$, let us set $\alpha = 1$. Assume we are given β and γ and we would like to know whether the error varies if γ is fixed and we solve for the optimal β. We now maximise Eq. 4.5 by minimising $f(\beta)$ (If the cosine of an angle is minimised the angle is maximised):

$$f(\beta) = \left(\frac{(\mu_r + \beta^2\mu_g + \gamma^2\mu_b)}{\sqrt{1+\beta^2+\gamma^2}\sqrt{(\mu_r)^2 + (\beta\mu_g)^2 + (\gamma\mu_b)^2}}\right). \tag{4.6}$$

To find the stationary points of $f(\beta)$, its derivative is computed:

$$\begin{aligned}\frac{\partial f}{\partial \beta} &= \frac{-\beta\cdot(\mu_r + \mu_g\beta^2 + \mu_b\gamma^2)\cdot({\mu_r}^2 + {\mu_g}^2\beta^2 + {\mu_b}^2\gamma^2)}{(1+\beta^2+\gamma^2)^{\frac{3}{2}}\cdot({\mu_r}^2 + {\mu_g}^2\beta^2 + {\mu_b}^2\gamma^2)^{\frac{3}{2}}}\\ &+ \frac{2\mu_g\beta\cdot(1+\beta^2+\gamma^2)\cdot({\mu_r}^2 + {\mu_g}^2\beta^2 + {\mu_b}^2\gamma^2)}{(1+\beta^2+\gamma^2)^{\frac{3}{2}}\cdot({\mu_r}^2 + {\mu_g}^2\beta^2 + {\mu_b}^2\gamma^2)^{\frac{3}{2}}}\\ &- \frac{{\mu_g}^2\beta\cdot(1+\beta^2+\gamma^2)\cdot(\mu_r + \mu_g\beta^2 + \mu_b\gamma^2)}{(1+\beta^2+\gamma^2)^{\frac{3}{2}}\cdot({\mu_r}^2 + {\mu_g}^2\beta^2 + {\mu_b}^2\gamma^2)^{\frac{3}{2}}}. \end{aligned} \tag{4.7}$$

f is maximised when equating Eq. 4.7 to zero. β is the common factor in all three fractions. Therefore, $\beta = 0$ and this is true for all γ including the γ that maximises Eq. 4.6. □

Similarly, Theorem 1 can be proved by fixing β in Eq. 4.5) and minimising $f(\gamma)$ or setting $\gamma = 1$ and minimising $f(\alpha)$. We have proved that the light with the maximum recovery angular error is the one with 0 in exactly one of the three R, G or B channels.

Lemma 1.1. *Assuming $\alpha = 1$ and $\beta = 0$, the recovery angular function has at most three stationary values.*

Proof. Since $\alpha = 1$ and $\beta = 0$, $f(\gamma)$ is written as:

$$f(\gamma) = \frac{(\mu_r + \gamma^2\mu_b)}{\sqrt{1+\gamma^2}\sqrt{(\mu_r)^2 + (\gamma\mu_b)^2}}. \tag{4.8}$$

$f(\gamma)$ is again the cosine of the angle we seek to maximise. The derivative of $f(\gamma)$ is calculated as:

$$\frac{\partial f}{\partial \gamma} = \frac{(\mu_r - \mu_b)^2 \cdot \gamma \cdot (\mu_b \gamma^2 - \mu_r)}{(\gamma^2+1)^{\frac{3}{2}} \cdot ({\mu_b}^2 \gamma^2 + {\mu_r}^2)^{\frac{3}{2}}}. \tag{4.9}$$

Setting it to zero implies:

$$\gamma = \pm\sqrt{\mu_r/\mu_b} \quad \gamma = 0, \tag{4.10}$$

When $\gamma = 0$ the angle is a global minimum (0 degrees). We know that real lights are all positive. So $\sqrt{\mu_r/\mu_b}$ is the other solution to Eq. 4.9. We apply the standard second derivative test:

$$\frac{\partial^2 f}{\partial^2 \gamma} = \frac{-(\mu_b - \mu_r)^2}{(\gamma^2+1)^{\frac{5}{2}} \cdot ({\mu_b}^2 \gamma^2 + {\mu_r}^2)^{\frac{5}{2}}} \cdot \left(3{\mu_b}^3\gamma^6 - 5\mu_r{\mu_b}^2\gamma^4 - 2\mu_r{\mu_b}^2\gamma^2 - 3{\mu_r}^2\mu_b\gamma^2 - 2{\mu_r}^3r^2 + {\mu_r}^3\right) \tag{4.11}$$

Substituting $\gamma = \sqrt{\mu_r/\mu_b}$:

$$\frac{\partial^2 f(\sqrt{\mu_r/\mu_b})}{\partial^2 \gamma} = (\mu_b - \mu_r)^2.(4\mu_r^3 + 2\mu_r^2\mu_b + 2\frac{\mu_r^3}{\mu_b}), \tag{4.12}$$

which is a positive value, since $\mu_r > 0$ and $\mu_b > 0$. Therefore, $f(\gamma)$ Eq. 4.8 is a local minimum at $\gamma = \sqrt{\mu_r/\mu_b}$ and this means $cos^{-1}(f(\gamma))$ at this point is a local maximum. □

Alternatively setting $\alpha = 0$ and $\gamma = 0$, the above argument can be repeated which results in respectively $\beta = \sqrt{\mu_b/\mu_g}$ and $\alpha = \sqrt{\mu_g/\mu_r}$. Thus there are three possible local maximums, one of which is the global maximum. One might wonder if there are six local maximums for Eq. 4.5 (i.e. while $[1\ 0\ \sqrt{\mu_r/\mu_b}]$ is a local maximum, $[\sqrt{\mu_b/\mu_r}\ 0\ 1]$ might also be a possible local maximum). That is actually true. Lemma 1.1 can be repeated by assuming $\gamma = 1$ instead of $\alpha = 1$, which results in $\alpha = \sqrt{\mu_b/\mu_r}$. But we have to mention that they both result in the same output once substituted in Eq. 4.5. The same applies for $[0\ 1\ \sqrt{\mu_g/\mu_b}]$ and $[0\ \sqrt{\mu_b/\mu_g}\ 1]$, as well as $[\sqrt{\mu_g/\mu_r}\ 1\ 0]$ and $[1\ \sqrt{\mu_r/\mu_g}\ 0]$.

Theorem 1 and consequently Lemma 1.1 follow that lights with one wavelength set to zero (e.g. $[1\ 0\ \sqrt{\mu_r/\mu_b}]$) result in maximum recovery angular errors for a given illuminant estimation algorithm applied on a given scene. In other words, Theorem 1 states lights which are cyan, purple and yellow maximise the recovery angular error. Conversely, pure red, green and blue lights result in the lowest angular error.

4.2.2 Maximum Recovery Angular Error for Real Lights

Theorem 1 suggests lights with 1 and 0 in two channels (e.g. $[1\ 0\ \sqrt{\mu_r/\mu_b}]$) result in the maximum recovery angular error. Nevertheless, we have to take into consideration that the majority of lights do not satisfy this property. This raises the question of whether we can revise Theorem 1 to cover more likely illuminants. This leads us to Theorem 2. Given that real lights are bounded to a restricted gamut area (such as the one in Figure 4.3), Theorem 2 answers this question: for a given set of real lights can we solve for the maximum error light? In Figure 4.3 we plot on a rg ($r = R/(R+G+B)$ and $G = G/(R+G+B)$) chromaticity diagram, the chromaticities of the lights from the SFU Lab dataset [Barnard et al., 2002c] (where [r,g,1-r-g] defines the corresponding 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 either. Our second theorem teaches where local maxima should lie when lights lie in a bounded region of colour space.

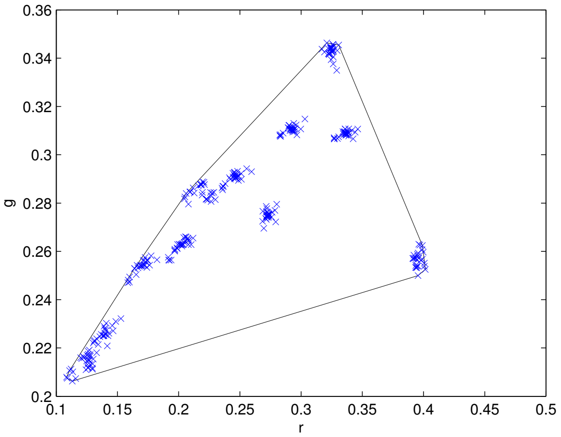


FIGURE 4.3: 2D Gamut of SFU Lab dataset's measured illuminants [Barnard et al., 2002c].

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 colour of 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 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 (as we move away from the light in any direction the error must decrease). 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 as it enables us to find the light resulting in the maximum recovery error, belonging to a set of feasible lights, by searching the boundary of the feasible set.

4.3 Reproduction Angular Error

Here we introduce a new metric for evaluating illuminant estimation algorithms. We call our new error measure, which is an improvement over the conventional recovery error, **Reproduction Angular Error** [Finlayson and Zakizadeh, 2014; Finlayson et al., 2016]. We prove that, reproduction angular error by design gives the same error for the scene reproduction where the difference is only in the colour of illumination prevailing the scene. Reproduction angular error is tied to the application of illuminant estimation which is discarding the estimated illuminant from the scene by dividing it out from the image. Further, by design it is as simple to compute as the legacy recovery angular error.

4.3.1 Introducing Reproduction Angular Error

According to the RGB model of image formation (Eq. 2.4) in Chapter 1, the RGB values in the image are scaled by the same three weighting factors as the illumination changes [Finlayson, 2013]. The reproduced image after colour correction, is the image from which the estimated illuminant is ‘divided out’, so that the colour bias due to illumination is removed. The colour bias is removed from the images as follows:

$$\frac{\underline{\rho}^{E,S}}{\underline{\rho}^{Est}} \approx \underline{\rho}^{S}. \tag{4.13}$$

where the division of the vectors is component-wise. Considering that the colour of a white surface under a certain illuminant would represent the colour of the illuminant, we rewrite Eq. 4.13 for the specific example of a white surface $\underline{\rho}^{E,W}$, where its colour is similar to the colour of the light $\underline{\rho}^{E}$ (i.e. $\underline{\rho}^{E,W} = \underline{\rho}^{E}$):

$$\frac{\underline{\rho}^{E,W}}{\underline{\rho}^{Est}} \approx \underline{U} = \frac{\underline{\rho}^{E,W}}{\underline{\rho}^{E}}, \quad \underline{U} = [1\ 1\ 1]^t. \tag{4.14}$$

The above equation states that the colour of a white surface will be recovered as $[1\ 1\ 1]^t$ if we knew the ground-truth illuminant $\underline{\rho}^{E}$. But in reality an illuminant estimation algorithm, in the best-case scenario, will recover only an estimate ($\underline{\rho}^{Est}$) close to $\underline{\rho}^{E}$ which will not give us the exact white ($[1\ 1\ 1]^t$).

Remembering that we cannot recover the absolute brightness of the light, we define the **Reproduction Angular Error** [Finlayson and Zakizadeh, 2014] - our new metric for assessing illuminant estimation algorithms - as:

$$\boxed{err_{reproduction} = \cos^{-1}\left(\frac{(\underline{\rho}^{E}/\underline{\rho}^{Est}).\underline{U}}{|(\underline{\rho}^{E}/\underline{\rho}^{Est})|\sqrt(3)}\right).} \tag{4.15}$$

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.).

4.3.2 Stability of Reproduction Angular Error

In the last section, Figures 4.1 and 4.2 showed that for the same scene and the same illuminant estimation algorithms different recovery angular errors can occur as a result of change in illumination. Here reproduction angular errors are calculated for the same set of images in Figures 4.1 and 4.2.

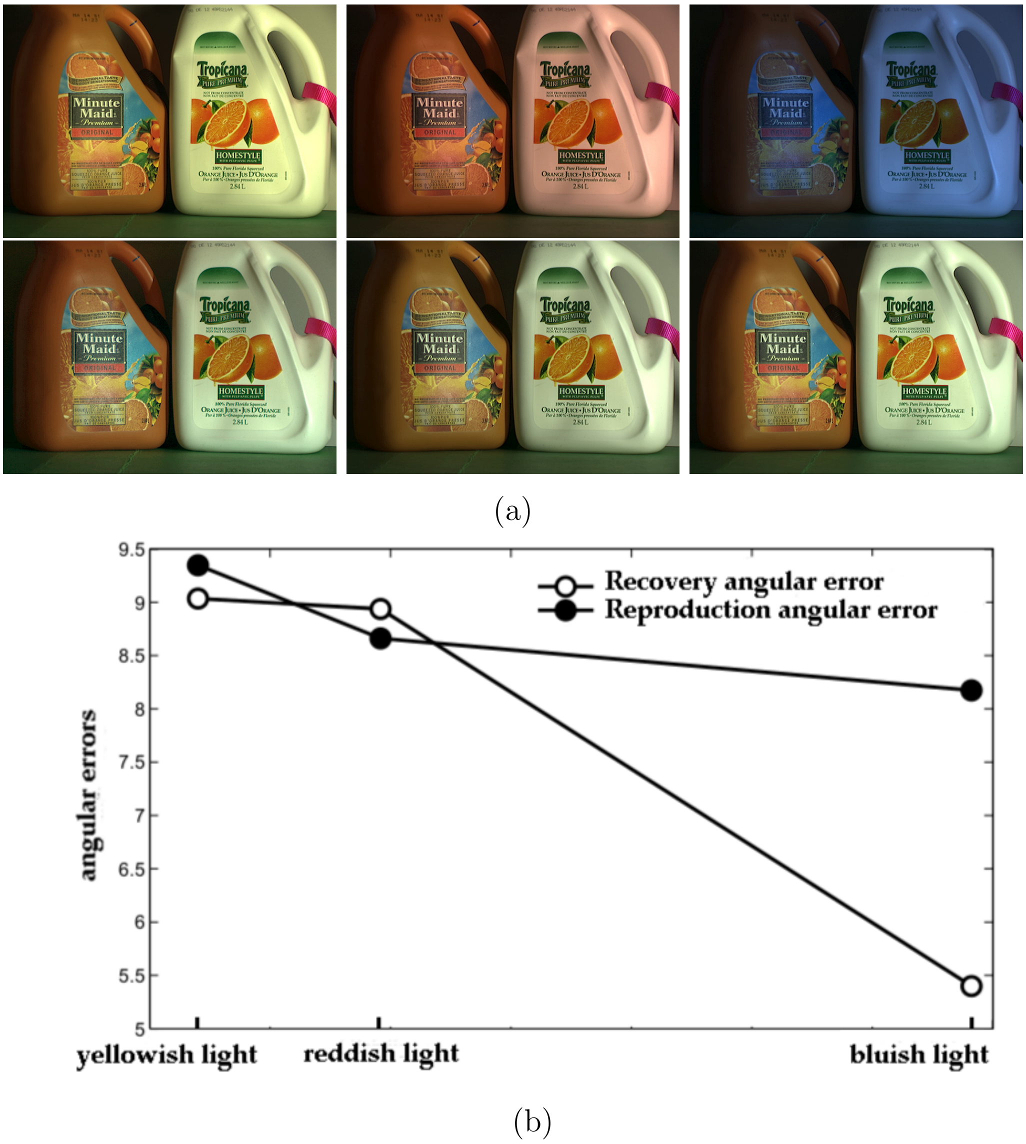


FIGURE 4.4: An example of similar colour corrected images with varying recovery angular error. (a) First row: images of the same scene captured under chromatic illuminants (from SFU Lab dataset [Barnard et al., 2002c]). Second row: Corrected images using grey-world algorithm. (b) The Recovery angular error (conventional metric, open circles) versus the Reproduction angular error (proposed metric, filled circles).

In Figure 4.4 (b), the recovery angular errors are shown with open circles and reproduction angular errors with filled circles. There is hardly any difference observed between the corrected images in the second row of Figure 4.4 (a). Reproduction angular errors are very in line with this observation and are much more stable than recovery angular errors.

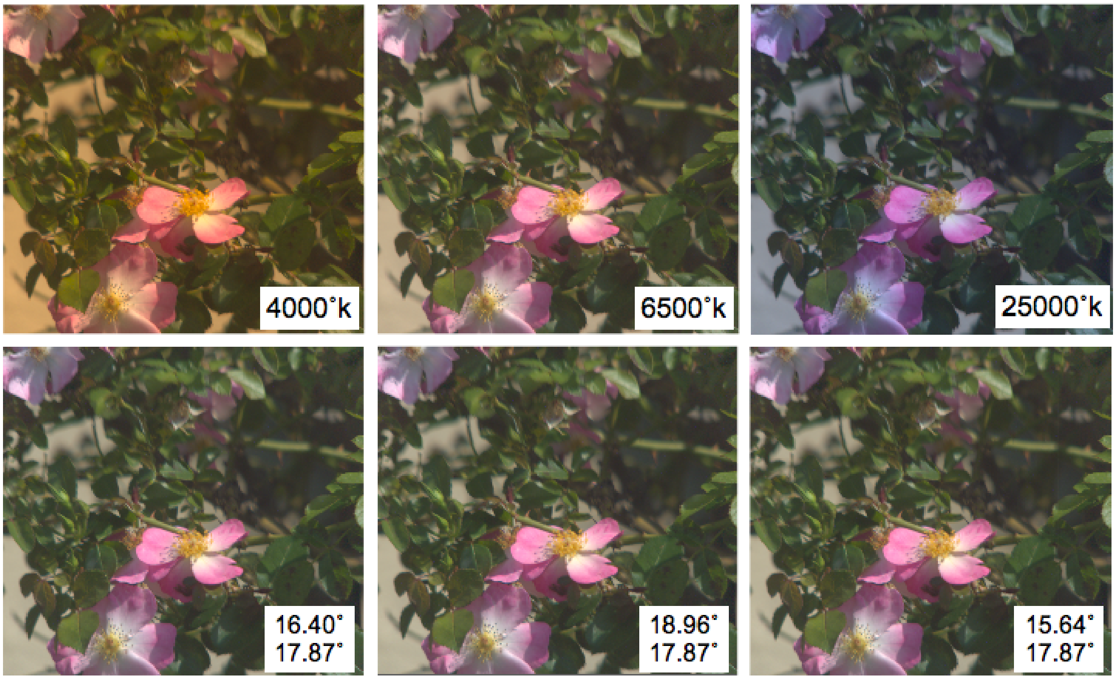


FIGURE 4.5: An example of similar colour corrected images with varying recovery angular error. First row: hyperspectral images [Foster et al., 2006] rendered in sRGB under three light of different temperature. Second row: the images are white balanced using the shades of grey algorithm in hyperspectral space before converting to sRGB (The recovery and **reproduction** angular errors (bottom errors) can be seen on each image).

In another example, in Figure 4.5, with the images rendered from hyperspectral data where there are no camera sensor sensitivity functions affecting the pixel values (same as Figure 4.2), we can see that reproduction angular error (the bottom errors on each image) is reporting the exact same errors for shades of grey algorithm (with $p = 2$, see Eq. 2.19 for more details) applied on the radiance images generated under three illuminants with different spectra. As mentioned before, illuminant estimation and error calculation is performed in hyperspectral space on the radiance values at each pixel which is the result of the product of reflectance data and illuminant spectra at each pixel.

Having seen the visual examples of stability of reproduction angular error in the last two figures, here we state and prove the theory of stability of reproduction angular error.

Theorem 3. *Given a single scene viewed 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 Eq. 4.2], using the fact presented in Eq. 4.4, the reproduction angular error (Eq. 4.15) can be written as:

$$err_{reproduction} = \cos^{-1} \frac{(\frac{\alpha}{\alpha\mu_r} + \frac{\beta}{\beta\mu_g} + \frac{\gamma}{\gamma\mu_b})}{\sqrt{(\frac{\alpha}{\alpha\mu_r})^2 + (\frac{\beta}{\beta\mu_g})^2 + (\frac{\gamma}{\gamma\mu_b})^2}\sqrt{(3)}}. \tag{4.16}$$

It can be seen easily in Eq. 4.16, that the scaling factors α, β and γ cancel. The reproduction error is stable regardless of the colour of the light. □

In Figure 4.6 (a), the two purple curves are the cumulative probability distribution functions of the analytical maximum recovery errors ($[1\ 0\ \sqrt{\mu_r/\mu_b}]$, see Lemma 1.1) for the two algorithms: gray-world [Buchsbaum, 1980] (solid line) and pixel-based gamut mapping [Gijsenij et al., 2010] (dashed line) algorithms for 321 images of the 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 [Barnard et al., 2002c]). The red curves in the same figure are the actual recovery angular errors of the estimated illuminant using the gray-world [Buchsbaum, 1980] (solid line) and pixel-based gamut mapping [Gijsenij et al., 2010] (dashed line) algorithms applied on the SFU Lab dataset.

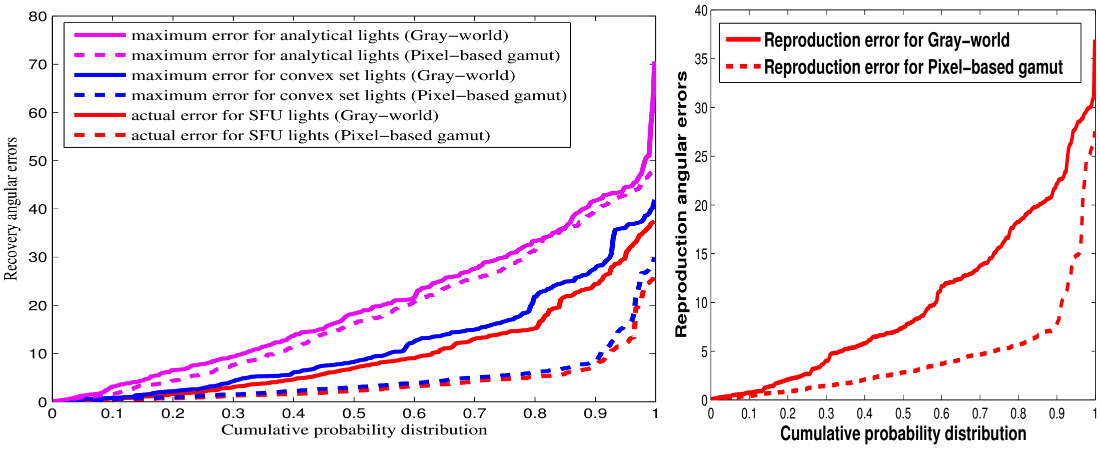


Figure 4.6: (a) Cumulative probability distribution function of analytical maximum recovery angular errors (in magenta), maximum error of real lights within the convex of SFU Lab dataset's [Barnard et al., 2002c] measured illuminants (in blue) and the recovery angular errors of the estimated lights of 321 SFU Lab images using the two algorithms (in red). (b) Cumulative probability distribution function of maximum reproduction angular errors [Finlayson and Zakizadeh, 2014]

In terms of the maximum angular error Figure 4.6 (a) teaches that gray-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.

In Figure 4.6 (b) we show the reproduction angular error for gray-world and pixel-based gamut mapping. This error is stable across illumination changes. Figure 4.6 (a) informs us - what we knew - that for all lights pixel-based gamut mapping works better than gray-world.

We note that the worst case performance is not just a mathematical curiosity, rather with the advent of LED lights it is possible to encounter lights that might invoke the worst case performance of recovery angular error.

4.3.3 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 [Hubel, 2005] 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 [Finlayson et al., 1994b; Chong et al., 2007] the reproduction error can be used directly.

More generally, an illuminant estimate can be used to parametrize a 3×3 correction matrix [Maloney and Wandell, 1986]. 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. [Wandell, 1987]. That is we substitute $\underline{w}^{Est} = \mathcal{M}(\underline{\rho}^{Est})\underline{\rho}^{E}$ into Eq. 4.15. In fact we can be more general still. In [Forsyth, 1990], 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 $\underline{\rho}^{Est}$. Adopting this idea we can substitute $\underline{w}^{Est} = \psi(\underline{\rho}^{E}; \underline{\rho}^{Est})$ into Eq. 4.15 and so arrive at even more general form of reproduction error.

Reproduction error is generalized to encompass more reflectances in [Finlayson and Zakizadeh, 2015; Cheng et al., 2015b]. Importantly, [Finlayson and Zakizadeh, 2015] found that simple reproduction angular error could be used as a proxy for calculation based on many reflectances. Chapter 6 will discuss the work in [Finlayson and Zakizadeh, 2015].

4.4 Conclusions

The most efficient illuminant estimation algorithm is often chosen based on its performance over a benchmark dataset. The performance of the algorithms is often evaluated using recovery angular error or simply angular error. In this chapter, this widely used metric is re-studied. We argue the conventional metric can report a huge range of errors for the same scene and algorithm pair (where the difference is only in the lighting condition under which the images are captured). That is, even though the images reproduced by dividing out the estimated illuminant using the same algorithm look very much the same a large range of estimation errors are reported. One of the contributions of this thesis is to solve for the range of recovery angular error for a given illuminant estimation algorithm and a given scene. We show that the maximum recovery angular error is for the cyan, yellow and magenta lights. The minimum recovery error is close to 0 for pure red, green and blue lights (all 'moment-type' algorithms can produce close to zero error for these lights). Although the same image is reproduced when the illuminant colour bias is removed, the angular recovery error can range from 0 to 40 degrees (or more).

In this chapter, we proposed the **Reproduction Angular Error** as an improvement over the recovery angular error. We prove that reproduction angular error is not very dependent on the illumination colour which prevails the scene in the sense that the same scene and algorithm pair will generate the same image reproduction and so the same reproduction angular error. Indeed, the new reproduction angular error is defined as the angle between a true white surface and the estimated reproduced white when an algorithm's estimate is used

After all, if we wish to recognise colourful content independent of the illuminant colour (i.e. we first remove the colour bias due to illumination by dividing out the illuminant colour [Funt et al., 1998]) then we need to adopt the new reproduction angular error to measure the performance. More generally, if illuminant estimates are used to discount colour casts - this is by far the main reason for estimating the illumination - from images due to the prevailing illuminant colour (for recognition, tracking or navigation) then the new metric should be used.

In the next chapter we re-evaluate a large selection of illuminant estimation algorithms for most well-known benchmark datasets in colour constancy. We will study the effect of using reproduction angular error on the rank order of illuminant estimation algorithms, as well as the way that they are used to give optimal performance.

Chapter 5

Rank Study of Illuminant Estimation Algorithms

In the last chapter, we showed that the traditional recovery angular error might introduce a wide range of error for the same algorithm applied on the same scene when only the colour of light is changing. We discussed that this instability of recovery angular error might lead to misjudgement about the performance of an illuminant estimation algorithm. Further, we introduced a new metric for evaluation of illuminant estimation algorithms, 'Reproduction Angular Error'. In this chapter, reproduction angular error is used to re-evaluate [Finlayson et al., 2016; Zakizadeh and Finlayson, 2015] most state of the art illuminant estimation algorithms for well-known benchmark datasets (including Simon Fraser University (SFU) [Barnard et al., 2002c], Gehler-Shi [Gehler et al., 2008; Shi and Funt, 2010], National University of Singapore (NUS) [Cheng et al., 2014], Greyball [Ciurea and Funt, 2003] datasets and a multispectral dataset by Foster et al. [Foster et al., 2006]). If there are algorithms for which the results are not provided in this chapter that is because the error data per image was not provided for public use.

When evaluating the performance of illuminant estimation algorithms, researchers are often interested in assigning a rank order to an algorithm. Of course the rank of an algorithm could be dependent on the scene content and the type of images for which the error of the algorithm is calculated. In this chapter, we will also study the effect of using the new metric on the rank order of the algorithms for different datasets. Whether two algorithms' positions in the ranking table is decided based on a slight or significant difference in their performance is also of great importance. Here this is examined using different non-parametrical statistical tests, which are usually run on a summary of data or the individuals to study the relation of data across different observations. Further, we analyse the effect that reproduction angular error has on choosing the optimal parameters for tunable algorithms. Also, the correlation of the two metrics (reproduction and recovery angular errors) and where it happens has also been studied.

The results of re-evaluation using reproduction angular error are available on the colour constancy website [1].

[1] http://colorconstancy.com/?page_id=703

5.1 Introduction

When evaluating an illuminant estimation algorithm, often a set of images is agreed on as a benchmark dataset, arguably the most well-known colour constancy datasets are Simon Fraser University (SFU) Lab [Barnard et al., 2002c], Gehler-Shi colour-checker [Shi and Funt, 2010], Greyball [Ciurea and Funt, 2003] datasets and the recent National University of Singapore (NUS) [Cheng et al., 2014] dataset (see Section 3.2 for a summary of colour constancy datasets). The RGB of the ground-truth illuminant of each image is also provided with each dataset. Eventually, a summary of the error data (such as mean, median or quantiles) is reported over the whole dataset. As mentioned in Section 3.4 whether or not such aggregates give an accurate summary of the underlying distribution of the error data is of some debate and different researchers prefer some over the others.

When evaluating the performance of illuminant estimation algorithms, researchers are also interested in analysing the performance of the algorithms in relation with each other, i.e. assigning rank orders to the algorithms. Understanding whether or not the algorithms in the ranking table are significantly different in terms of their performance, requires utilising appropriate statistical tests (some of which were introduced in Section 3.6).

Gijsenij *et al.* [Gijsenij et al., 2011] did a comprehensive evaluation of a great selection of illuminant estimation algorithms using recovery angular error. The evaluation is done for multiple benchmark datasets and the results are available online. Where the algorithms need to be tuned to perform their best, the optimal parameters for each algorithm are also reported. Over time, the evaluation results of some of state of the art algorithms using recovery angular error for different datasets are added to the website.

Here we re-evaluate most of these algorithms for the following datasets: SFU Lab [Barnard et al., 2002c], Shi colour-checker [Shi and Funt, 2010], NUS [Cheng et al., 2014] and Grey-ball [Ciurea and Funt, 2003] datasets as well as the hyperspectral dataset by Foster *et al.* [Foster et al., 2006] . The effect of using the new metric, reproduction angular error on ranking the algorithms has also been studied in this chapter. The two non-parametric statistical tests for studying the relation of data across different observations (Kendall rank correlation test and Wilcoxon signed-rank test [Conover, 1999; Sprent and Smeeton, 2007]) have been used to analyse whether the changes in the ranking of algorithms have been significant or not. Moreover, we show that using reproduction angular error, the algorithms (where applicable) might be tuned differently. Also, considering the long time use of recovery angular error in colour constancy research, we need to study the correlation between the two metrics, recovery and reproduction angular errors. Here, we investigate this correlation for different sets of images.

The organisation of this chapter is as follows: In Section 5.2, the results of re-evaluation of illuminant estimation algorithms for different benchmark datasets using reproduction angular error are provided. Also the effect of using the new metric in choosing the optimal parameter for tunable algorithms are discussed in the same section. In Section 5.3, the significance of rank switches of the algorithms when evaluated by reproduction angular error is analysed using two statistical tests: Kendall rank correlation and Wilcoxon signed-rank tests. In Section 5.4, we investigate the correlation of reproduction and recovery angular errors for similar and diverse scenes.

5.2 Re-evaluation of the Algorithms by Reproduction Angular Error on Several Benchmark datasets

In this section, the results of evaluation of several illuminant estimation algorithms for well-known benchmark datasets including SFU Lab, Shi colour-checker, NUS, Grey-ball datasets and the hyperspectral dataset by Foster *et al.* [Foster et al., 2006] using **Reproduction Angular Error** are given. The results are reported in the form of a summary of error data (e.g. median) along with the optimal parameters for each algorithm (where applicable) concluded based on both reproduction and recovery angular errors . For each dataset, the results of at least one statistical moment is presented here. In each table the recovery angular errors are also provided for comparison. We will observe that some algorithms might be ranked differently if evaluated by reproduction angular error in comparison to recovery angular error. We will also see that in some cases the optimal parameters for tunable algorithms such as grey-edge or gamut mapping might be chosen differently when selected based on the errors reported by reproduction angular error.

5.2.1 Simon Fraser University Dataset

The SFU dataset, introduced in Section 3.2, is linear and is useful to examine the performance of illuminant estimation algorithms assuming *raw* capture. Indeed, all images captured at the sensor level are linear. When these images are shown on a display (without additional processing) they appear dark. This is because of the inherent non-linearity of displays and because images are processed by a camera pipeline to make visually appealing images. Using our new metric, reproduction angular error, in this section we are presenting the results of our evaluation for the SFU Lab [Barnard et al., 2002c].

Table 5.1 and 5.2 contain median and 95% quantile of the recovery and reproduction angular errors for the linear SFU Lab dataset [Barnard et al., 2002c].

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

Table 5.1: Median Recovery and Reproduction errors for several colour constancy algorithms applied on SFU dataset [Barnard et al., 2002c]. 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	Median	Rank
Grey-world	-	-	7°	11	-	-	7.5°	11
MaxRGB	-	-	6.5°	10	-	-	7.4°	10
Shades of grey	7	-	3.7°	**9**	7	-	3.9°	**8**
1st order Grey-edge	7	4	3.2°	**7**	14	4	3.58°	**6**
2nd order Grey-edge	14	10	2.7°	4	15	10	3°	4
Pixel-based gamut	-	4	2.26°	**2**	-	4	2.8°	**3**
Edge-based gamut	-	2	2.27°	**3**	-	2	2.7°	**2**
Inter-based gamut	-	4	2.1°	1	-	3	2.5°	1
Union-based gamut	-	2	3°	5	-	2	3.4°	5
Heavy tailed-based [Chakrabarti et al., 2012]	-	-	3.5°	**8**	-	-	4.1°	**9**
Weighted grey-edge	2	1	3.1°	**6**	2	1	3.62°	**7**

Table 5.2: 95% quantile Recovery and Reproduction errors for several colour constancy algorithms applied on SFU dataset [Barnard et al., 2002c].

	Recovery error				Reproduction Error			
Method	p	σ	95%	Rank	p	σ	95%	Rank
Grey-world	-	-	30.3°	11	-	-	28°	11
MaxRGB	-	-	27.2°	10	-	-	27.2°	10
Shades of grey	4	-	18.7°	**9**	3	-	19°	**8**
1st order Grey-edge	2	1	14.3°	6	2	1	15.6°	6
2nd order Grey-edge	2	2	14.2°	5	2	2	15.1°	5
Pixel-based gamut	-	6	9.8°	**1**	-	7	11.1°	**1**
Edge-based gamut	-	2	12.6°	**3**	-	2	14.3°	**4**
Inter-based gamut	-	6	9.8°	**1**	-	7	11.2°	**2**
Union-based gamut	-	3	12.8°	**4**	-	3	13.2°	**3**
Heavy tailed-based	-	-	15.9°	7	-	-	16.6°	7
Weighted grey-edge	2	1	18°	**8**	2	1	19.3°	**9**

same performance (according to the median of 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. 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; in this case pixel-based and intersection-based gamut mapping algorithms in Table 5.2. Later, in Section 5.4, the significance of changes in the ranking of algorithms will be discussed. p and σ in Tables 5.1 and 5.2 are the parameters which are tuned for some algorithms to give the minimum errors. We notice, that these parameters could be chosen differently based on reproduction angular error, compared to when tuned based on recovery angular error. This will be discussed in more detail in Section 5.4.

Looking at Table 5.1 and Table 5.2, we can conclude that using reproduction angular error there are changes in the ranking of algorithms. Although the overall ranking of illuminant estimation algorithms remains the same (e.g. gamut mapping algorithms still performing the best for the SFU dataset); local rank switches can be still observed. 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 that in many cases grey-world and MaxRGB are not performing well and when that is the case they perform poorly with a noticeable distance from other algorithms. This is true regardless of the choice of evaluation technique and we can see that using reproduction angular error, they are still ranked similarly with respect to each other and the rest of algorithms.

5.2.2 Colour-Checker Dataset (by Shi)

Colour-Checker dataset (introduced in Section 3.2) by Gehler *et al.* [Gehler et al., 2008] is a wide selection of indoor and outdoor scenes captured in a real-life photographic sense. As mentioned in Section 3.2 the dataset was later reprocessed by Shi *et al.* [Shi and Funt, 2010] to create almost raw images and avoid the post-processing steps such as clipping or tone-curve. Here we re-evaluated several illuminant estimation algorithms by reproduction angular error for the Colour-Checker dataset by Shi.

Tables 5.3 and 5.4 report the mean and 95% quantile recovery and reproduction angular errors for Shi-Gehler dataset.

Again, there are changes in ranking of algorithms when using reproduction angular error (changes are highlighted in bold and underlined). In the case of Colour-Checker dataset, there are few rank switches according to the mean errors (Table 5.3). Whereas, looking at 95% quantile errors, we notice there are many switches between the rank orders. The 95% quantile of images represent those for which illuminant estimation algorithms have a very poor performance. Whether, there is any commonality between the images for which a certain number of algorithms fail to estimate the illuminant could be interesting. This needs a detailed investigation and is a stand alone topic, which is examined in Chapter 7.

Interestingly, we notice Exemplar-based algorithm [Joze and Drew, 2012] is outperforming the rest of the algorithms with an almost significant error difference, even when looking at the 95% quantile error. We need to point out that there might be other algorithms proposed during the very recent years outperforming Exemplar-based method; however, the error data or a suitable code to reproduce the results for those algorithms were not available to be included here.

Table 5.3: Mean Recovery and Reproduction errors for several algorithms applied on Shi Colour-checker dataset [Shi and Funt, 2010].

	Recovery error				Reproduction Error			
Method	p	σ	mean	Rank	p	σ	mean	Rank
Grey-world	-	-	6.4°	14	-	-	7.1°	14
MaxRGB	-	-	7.5°	16	-	-	8.1°	16
Shades of grey	3	-	4.9°	**11**	3	-	5.5°	**10**
1^{st} order grey-edge	1	9	5.3°	13	1	1	6.2°	13
2^{nd} order grey-edge	1	1	5.1°	12	1	1	6.0°	12
Pixel-based gamut	-	5	4.1°	7	-	5	4.7°	7
Edge-based gamut	-	4	6.5°	15	-	4	7.8°	15
Bayesian	-	-	4.82°	**10**	-	-	5.63°	**11**
Heavy-tailed based	-	-	3.67°	4	-	-	4.42°	4
Bottom-up [Van De Weijer et al., 2007b]	-	-	3.43°	**2**	-	-	3.98°	**2**
Top-down [Van De Weijer et al., 2007b]	-	-	3.75°	5	-	-	4.29°	5
Bottom-up + Top-down	-	-	3.48°	**3**	-	-	3.98°	**2**
Natural image statistics [Gijsenij and Gevers, 2011]	-	-	4.19°	8	-	-	4.83°	8
CART-based selection [Bianco et al., 2010]	-	-	4.49°	9	-	-	5.16°	9
CART-based combination [Bianco et al., 2010]	-	-	3.9°	6	-	-	4.53°	6
Examplar-based	-	-	2.89°	1	-	-	3.4°	1

Table 5.4: 95% quantile Recovery and Reproduction errors for several algorithms applied on Shi Colour-checker dataset [Shi and Funt, 2010].

	Recovery error				Reproduction Error			
Method	p	σ	95% quantile	Rank	p	σ	95% quantile	Rank
Grey-world	-	-	11.25°	**7**	-	-	12.41°	**6**
MaxRGB	-	-	19.01°	18	-	-	20.05°	18
Shades of grey	2	-	10.56°	5	2	-	11.97°	5
1st order grey-edge	1	1	11.33°	**8**	1	1	14.56°	**11**
2nd order grey-edge	1	1	11.01°	**6**	1	1	13.66°	**9**
Pixel-based gamut	-	5	13.60°	14	-	5	15.44°	14
Edge-based gamut	-	5	16.1°	**16**	-	5	19.93°	**17**
Intersection-based gamut	-	5	13.6°	15	-	5	15.47°	15
Regression (SVR) [Agarwal et al., 2007]	-	-	17.25°	**17**	-	-	18.89°	**16**
Bayesian	-	-	12.60°	13	-	-	15.39°	13
Heavy-tailed based	-	-	8.68°	2	-	-	9.89°	2
Bottom-up	-	-	9.53°	**3**	-	-	11.57°	**4**
Top-down	-	-	12.13°	**11**	-	-	13.81°	**10**
Bottom-up + Top-down	-	-	11.55°	**9**	-	-	13.59°	**8**
Natural image statistics	-	-	11.69°	**10**	-	-	12.59°	**7**
CART-based selection	-	-	12.49°	12	-	-	14.63°	12
CART-based combination	-	-	10.14°	**4**	-	-	11.43°	**3**
Exemplar-based	-	-	6.95°	1	-	-	8.23°	1

5.2.3 National University of Singapore Dataset

The recently proposed NUS datasett [Cheng et al., 2014] consists of 1736 images from eight different cameras of indoor and outdoor scenes (see Section 3.2 for more details).

Tables 5.5 and 5.6 reports the max and 95% quantile errors for the Canon1D camera from NUS dataset. We performed a selection of popular illuminant estimation algorithms on all eight cameras of NUS dataset and more or less the same pattern can be observed for all cameras, so, here we are presenting the results of one cameras, Canon1D. Like other cameras in NUS dataset, there are around 220 images captured by Canon1D camera.

Table 5.5: Maximum Recovery and Reproduction errors for several algorithms applied on Canon1D camera from NUS dataset [Cheng et al., 2014].

	Recovery error				Reproduction Error			
Method	p	σ	Max	Rank	p	σ	Max	Rank
Grey-world	-	-	22.37°	**5**	-	-	24.69°	**4**
MaxRGB	-	-	39.12°	**7**	-	-	33.76°	**6**
Shades of grey	5	-	14.62°	**2**	5	-	18.41°	**3**
1^{st} order grey-edge	7	9	14.08°	1	5	3	17.35°	1
2^{nd} order grey-edge	4	10	15.00°	**3**	5	4	17.91°	**2**
Pixel-based gamut	-	0	38.60°	**6**	-	0	35.52°	**7**
Edge-based gamut	-	5	21.64°	**4**	-	5	27.60°	**5**

Table 5.6: 95% quantile Recovery and Reproduction errors for several algorithms applied on Canon1D camera from NUS dataset [Cheng et al., 2014].

	Recovery error				Reproduction Error			
Method	p	σ	95%	Rank	p	σ	95%	Rank
Grey-world	-	-	12.78°	4	-	-	16.19°	4
MaxRGB	-	-	17.28°	**7**	-	-	18.14°	**6**
Shades of grey	5	-	9.01°	**1**	8	-	11.71°	**2**
1^{st} order grey-edge	7	2	9.09°	**2**	9	2	11.50°	**1**
2^{nd} order grey-edge	3	5	9.12°	3	1	2	12.09°	3
Pixel-based gamut	-	0	16.64°	**6**	-	0	18.45°	**7**
Edge-based gamut	-	3	13.01°	5	-	3	16.37°	5

For NUS dataset, when looking at the maximum or 95% quantile recovery and reproduction errors, the rank order of the algorithms changes very frequently.

5.2.4 Greyball (videoframes) dataset

Tables 5.7 and 5.8 contain median and 95% quantile recovery and reproduction errors for Grey-ball dataset [Ciurea and Funt, 2003] which consists of 11346 images (video frames) of a variety of indoor and outdoor scenes. Every image has a grey sphere in view in the bottom-right of the image. The average RGB over the sphere is taken to be the RGB of the light (read more in Section 3.2). (Inverse intensity chromaticity space algorithm [Tan et al., 2004] is denoted as IICS in Tables 5.7 and 5.8)

Table 5.7: Median Recovery and Reproduction errors for several colour constancy algorithms applied on Greyball dataset [Ciurea and Funt, 2003]. 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	Median	Rank
Grey-world	-	-	7°	11	-	-	7.6°	11
MaxRGB	-	-	5.3°	**6**	-	-	5.5°	**5**
Shades of grey	8	-	5.28°	**5**	14	-	5.6°	**6**
1^{st} Grey-edge	2	1	4.6°	3	2	1	4.8°	3
2^{nd} Grey-edge	1	2	4.8°	4	1	2	5°	4
Pixel-based gamut	-	2	5.67°	**9**	-	2	5.87°	**8**
Edge-based gamut	-	1	5.62°	**8**	-	1	5.85°	**7**
Inter-based gamut	-	6	5.7°	**10**	-	2	5.92°	**9**
IICS	-	-	5.6°	**7**	-	-	6°	**10**
Using natural image statistics	-	-	3.9°	2	-	-	4.3°	2
Exemplar-based	-	-	3.4°	1	-	-	3.67°	1

We point out that in Table 5.8 the ranks of 'Shades of Grey' and '2^{nd} order grey-edge' are the same for the 95% quantile error (they have the same rank 4) but different when the reproduction error is used. That is although 'shades of grey' has the same rank for both error measures we highlight a ranking difference because in one case there is a tie in the ranking and in the other there is no tie.

Table 5.8: 95% quantile Recovery and Reproduction errors for several colour constancy algorithms applied on Grey-ball dataset [Ciurea and Funt, 2003]. 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	95%	Rank	p	σ	95%	Rank
Grey-world	-	-	17.9°	11	-	-	20.9°	11
MaxRGB	-	-	17.4°	9	-	-	18°	9
Shades of grey	9	-	13.8°	4	8	-	14.5°	4
1^{st} Grey-edge	1	2	13.5°	3	1	2	14.3°	3
2^{nd} Grey-edge	1	3	13.8°	4	1	4	14.7°	5
Pixel-based gamut	-	5	17.8°	10	-	5	18.5°	10
Edge-based gamut	-	3	16.2°	7	-	4	16.6°	7
Inter-based gamut	-	9	16.2°	7	-	8	17°	8
IICS	-	-	15.2°	6	-	-	16°	6
Using natural image statistics	-	-	13.2°	2	-	-	13.7°	2
Exemplar-based	-	-	11.3°	1	-	-	12.5°	1

5.2.5 Foster et al. Hyperspectral dataset

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 is 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 [Foster et al., 2006]. 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 six of these 31-band images are presented in Table 5.9.

It can be seen that in the case of hyperspectral images, the ranking of algorithms might differ depending on which error metric is used.

Table 5.9: Maximum Recovery and Reproduction errors for several algorithms applied on six scenes from Foster et al. hyperspectral dataset [Foster et al., 2006].

	Recovery error				Reproduction Error			
Method	p	σ	Median	Rank	p	σ	Median	Rank
General Grey-world	6	10	7.25°	**3**	7	10	6.08°	**1**
Shades of grey	3	-	7.85°	**4**	3	-	7.51°	**3**
1st order grey-edge	4	2	7.18°	**1**	1	1	7.50°	**2**
2nd order grey-edge	10	6	7.23°	**2**	1	2	7.73°	**4**
MaxRGB	-	-	12.60°	6	-	-	12.99°	6
Grey-world	-	-	10.14°	5	-	-	8.80°	5

In multispectral illuminant estimation, rather than the actual and estimated light being three vectors they are 31-vectors. Relative to these 31 vectors the recovery and reproduction errors are analogously defined. It can be noticed that the errors are higher. Intuitively, this is to be expected as in 31-space there are more degrees of freedom.

5.2.6 The effect of Reproduction Angular Error on the Choice of Optimal Parameters

The parameters such as Mink-norm (p) in Eq. 2.19 and the parameter for the Gaussian filter (σ in Eq. 2.20) as well as in edge-based gamut mapping algorithm) can be tuned to achieve the best performance for these algorithms; or in other words has the lowest error on a given set of images. For instance, looking at the median errors of an algorithm for all its possible parameters calculated over the whole dataset, it is decided for which parameters the algorithm is more likely to have a minimum error. An algorithm might introduce a wide range of error for the same image with different assigned parameters. Therefore, choosing the correct parameter and consequently the proper metric is of great importance.

The second important outcome of Tables 5.1 to 5.9 is the changes in the optimal parameters for the algorithms. We notice the tunable parameters for an algorithm can change if the reproduction angular error is used for evaluation of the algorithm instead of recovery angular error.

For instance, for SFU dataset (Table 5.1) the mink-norm (p) resulting in the minimum median recovery angular error for 1^{st} order grey-edge algorithm is seven. Whereas, for the same algorithm, $p = 14$ results in the minimum median reproduction angular error. Or, for the maximum-error images of NUS dataset (Table 5.6), both p and σ are chosen differently for the 2^{nd} order grey-edge algorithm depending on whether they are selected based on the recovery error or the reproduction error. Similarly, in Table 5.7, according to median recovery angular error, shades of grey is performing best for Grey-ball dataset when the Minkowski norm [Finlayson and Trezzi, 2004] (p) equals eight but median reproduction angular error reports that the best performance of shades of grey is with $p = 14$.

The choice of the best parameters for an algorithm can have a great impact on the final evaluation of the algorithm and its rank in the table of illuminant estimation algorithms. Observing that this choice depends on the error metric used, emphasises the importance of which error metric is used for evaluation of the algorithms' performance.

5.3 Rank Switches by Recovery and Reproduction Angular Error

5.3.1 Kendall's Rank Correlation Test

To study to what extent the ranking of these algorithms has changed using our new metric, we performed the the Kendall test [Sprent and Smeeton, 2007; Conover, 1999] for all the algorithms in Tables 5.1 to 5.9 in Section 5.2 where their ranks changed once evaluated using reproduction angular error. Kendall's test, as discussed in Section 3.6, is an appropriate statistical test to study whether the change in the ranking of algorithm is significant or not.

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 (i.e. median, mean, max or 95% quantile) and the dataset (SFU Lab, Colour Checker (by Shi), NUS, Grey-ball or Foster *et al.* hyperspectral).

In Tables 5.10 and 5.11, Kendall's T is calculated for the *changed* rank algorithms for SFU Lab dataset [Barnard et al., 2002c] from Tables 5.1 and 5.2. Breaking down the calculations, in Table 5.10 (median error and for 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.

TABLE 5.10: Changes in ranking of algorithms for SFU Lab dataset [Barnard et al., 2002c] (based on median errors).

	Median			
Method	Reproduction Rank	Recovery Rank	C	D
Edge-based gamut	1	2	4	1
Pixel-based gamut	2	1	4	0
1^{nd} 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 5.11: Changes in ranking of algorithms for SFU Lab dataset [Barnard et al., 2002c] (based on 95% quantile errors).

	95% quantile			
Method	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)	

Similarly, Tables 5.12 and 5.13 contain the results of Kendall's test for the Colour-Checker dataset by Shi from Tables 5.3 and 5.5. Again, for the changed-rank algorithms based on mean (Table 5.12) and 95% quantile (5.13) errors, Kendall's test results shows the switches in the ranking of algorithms are significant.

Tables 5.14 and 5.15 report the ranking performance for the NUS Canon1D dataset [Ciurea and Funt, 2003] from Tables 5.5 and 5.6 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 order

Table 5.12: Changes in ranking of algorithms for the Colour-Checker dataset by Shi [Shi and Funt, 2010] (based on mean errors).

	Mean			
Method	Reproduction Rank	Recovery Rank	C	D
Bottom-up	1	1	2.5	0.5
Bottom-up + Top-down	1	2	2	0
Bayesian	4	3	0	1
shades of grey	3	4	0	0
T quantile for 4 samples at 99.5% confidence = 6			>(T = 3)	

Table 5.13: Changes in ranking of algorithms for Shi dataset [Shi and Funt, 2010] (based on 95% quantile errors).

	95% quantile			
Method	Reproduction Rank	Recovery Rank	C	D
Bottom-up	1	2	8	1
CART-based combination	2	1	7	0
2^{nd} order grey-edge	3	6	4	3
Grey-world	4	3	6	0
1^{st} order grey-edge	5	8	2	3
Bottom-up + Top-down	6	5	3	1
Natural image statistics	7	4	3	0
Top-down	8	7	2	0
Edge-based gamut	9	10	0	1
Regression(SVR)	10	9	0	0
T quantile for 10 samples at 99.5% confidence = 27			>(T = 26)	

when they were ranked using max and 95% quantile reproduction angular errors respectively.

Tables 5.16 and 5.17 contain the same information for Grey-ball dataset [Ciurea and Funt, 2003] from Tables 5.7 and 5.8 . 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 . The tied rank algorithms (i.e. the algorithms with the same rank given once evaluated based on 95% quantile

Table 5.14: Changes in ranking of algorithms for Canon1D camera from NUS dataset [Cheng et al., 2014] (based on max errors).

	Max			
Method	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 5.15: Changes in ranking of algorithms for Canon1D camera from NUS dataset [Cheng et al., 2014] (based on 95% quantile errors).

	95% quantile			
Method	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)	

recovery angular error) from Tables 5.7 and 5.8 are also included in Tables 5.16 and 5.17.

Table 5.18 shows the Kendall test results for the changed rank algorithms in Table 5.9 which contains the median recovery and reproduction errors for the Foster *et al.* hyperspectral dataset. The discrepancy between the ranking of reproduction versus recovery error is even more marked for the multispectral case.

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 5.10 to 5.18, showing the fact that the ranking of these algorithms using recovery and reproduction angular errors

Table 5.16: Changes in ranking of algorithms for Grey-ball dataset [Ciurea and Funt, 2003] (based on median errors).

	Median			
Method	Reproduction Rank	Recovery Rank	C	D
MaxRGB	1	2	4	1
Shades of grey	2	1	4	0
Edge-based gamut	3	4	2	1
Pixel-based gamut	4	5	1	1
Intersection-based gamut	5	6	0	1
IICS	6	3	0	0
T quantile for 6 samples at 99.5% confidence = 13			>(T = 7)	

Table 5.17: Changes in ranking of algorithms for Grey-ball dataset [Ciurea and Funt, 2003] (based on 95% quantile errors).

	95% quantile			
Method	Reproduction Rank	Recovery Rank	C	D
shades of grey	1	1	2.5	0.5
2^{nd} grey-edge	2	1	2	0
Edge-based gamut	3	3	0.5	0.5
Intersection-based gamut	4	3	0	0
T quantile for 4 samples at 99% confidence = 6			>(T = 4)	

Table 5.18: Changes in ranking of algorithms for Foster et al. hyperspectral dataset [Foster et al., 2006] (based on median errors).

	Median			
Method	Reproduction Rank	Recovery Rank	C	D
1^{st} order grey-edge	2	1	2	1
2^{nd} order grey-edge	4	2	0	2
General grey world	1	3	1	0
Shades of grey	3	4	0	0
T quantile for 4 samples at 99.5% confidence = 6			>(T = 0)	

are strongly discordant. This implies that indeed the ranking of algorithms in Tables 5.10 to 5.17 have changed significantly based on reproduction angular error.

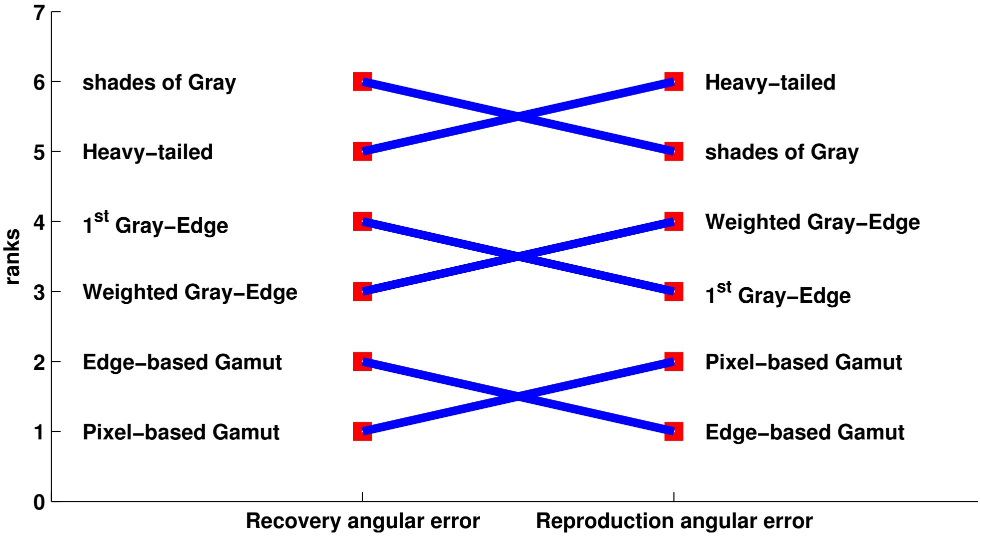


Figure 5.1: The pictorial scheme of Kendall test for the changed rank algorithms in Table 5.10 [Finlayson and Zakizadeh, 2014].

A pictorial scheme of Kendall's test in Table 5.10 is shown in Figure 5.1. 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.

Apart from changes observed in a coarse selection of the best algorithms applied on the five datasets which were represented here, there are many switches in the local ranking of algorithms (e.g. 1^{st} grey-edge algorithm with different σ and p-norm values applied on a set of images). The same trend can be observed with other datasets.

5.3.2 Wilcoxon Signed-Rank Test

To further study the behaviour of two metrics on individual images we performed the Wilcoxon signed-rank test [Conover, 1999] (previously explained in Section 3.6) which allows us to show the statistical significance of the difference between two algorithms [Hordley and Finlayson, 2006]. 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.

Here, we perform the Wilcoxon test for two of the datasets from Section 4.2: SFU dataset and Grey-ball dataset, which based on median of the errors for the algorithms in Table 5.10 and 5.16, Kendall's test results showed that there is a significant change in the ranking of algorithms. Using Wilcoxon signed-rank test, we want to investigate whether there is a significant difference between the median of errors of those algorithms .

The Wilcoxon sign test results for the algorithms in Table 5.10 applied on SFU dataset are shown in Table 5.19. Here, a positive value (green colour) at location (i, j) (i being the row and j the column) 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 colour) indicates the opposite and a zero (yellow colour) shows there is no significant difference between the performance of two algorithms. For example, at location $(1, 3)$ the positive value for recovery angular error indicates that algorithm 1. Edge-based gamut mapping has a significantly lower error than 3. 1^{st} grey-edge. In this case, looking at the median of recovery angular errors in Table 5.1 for the two algorithms, the same conclusion is drawn. 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 the 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.

Table 5.19: Wilcoxon sign test on SFU dataset for Recovery and Reproduction errors of the algorithms in Table 5.10.

	Recovery error						Reproduction error					
	1. Edge-based gamut	2. Pixel-based gamut	3. 1^{st} grey-edge	4. weighted grey-edge	5. shades of grey	6. Heavy tailed-based	1. Edge-based gamut	2. Pixel-based gamut	3. 1^{st} 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

Table 5.20: Wilcoxon sign test on Grey-ball dataset for Recovery and Reproduction errors of the algorithms in Table 5.16.

	Recovery error						Reproduction error					
	1. MaxRGB	2.shades of grey	Edge-based gamut	4.Pixel-based gamut	5.Intersection-based gamut	6.IICS	1. MaxRGB	2.shades of grey	Edge-based gamut	4.Pixel-based gamut	5.Intersection-based gamut	6.IICS
1	0	-1	+1	+1	+1	+1	0	+1	+1	+1	+1	0
2	+1	0	+1	+1	+1	+1	-1	0	+1	+1	+1	+1
3	-1	-1	0	+1	+1	-1	-1	-1	0	0	+1	+1
4	-1	-1	-1	0	+1	-1	-1	-1	0	0	+1	+1
5	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	0	+1
6	-1	-1	+1	+1	+1	0	0	-1	-1	-1	-1	0

Table 5.20 reports the results of Wilcoxon sign-rank test for the algorithms in Table 5.16 applied on the Grey-ball dataset. Similar to the results for SFU dataset,

Wilcoxon test confirms the significance of the difference between the algorithm's performance. Different colours for the same (i, j) location for recovery and reproduction angular errors show that for some algorithms there are switches in the ranking.

5.4 Correlation of Recovery and Reproduction Angular Errors

This section investigates the correlation between the two metrics, reproduction and recovery angular errors, for each individual algorithm applied on a set of images [Zakizadeh and Finlayson, 2015]. We consider two cases: first, where the images are from the same scene under different illuminations; second: when the same algorithm is applied on the images of diverse scenes. We will observe that the correlation of the two metrics differ considering the two scenarios.

We have used the grey-world estimations of 11 illuminants for 30 objects in the SFU data set to illustrate the degree of deviation of recovery errors from one illuminant to the other for a single object. As mentioned before, in the SFU dataset the same object is captured under 11 different lights. The box plots in Figure 5.2 show the range of reproduction and recovery angular errors for the 30 objects in SFU dataset. It can be seen that the range of errors according to recovery angular error (top box plot) is much wider than the range of reproduction angular errors (bottom box plot).

We also calculate the standard deviation of the recovery error per object and the per object standard deviation for the reproduction error. We plot (for all 30 objects) the standard deviation of recovery against reproduction standard deviations in Figure 5.3. Clearly, the reproduction error is much more stable than the recovery error.

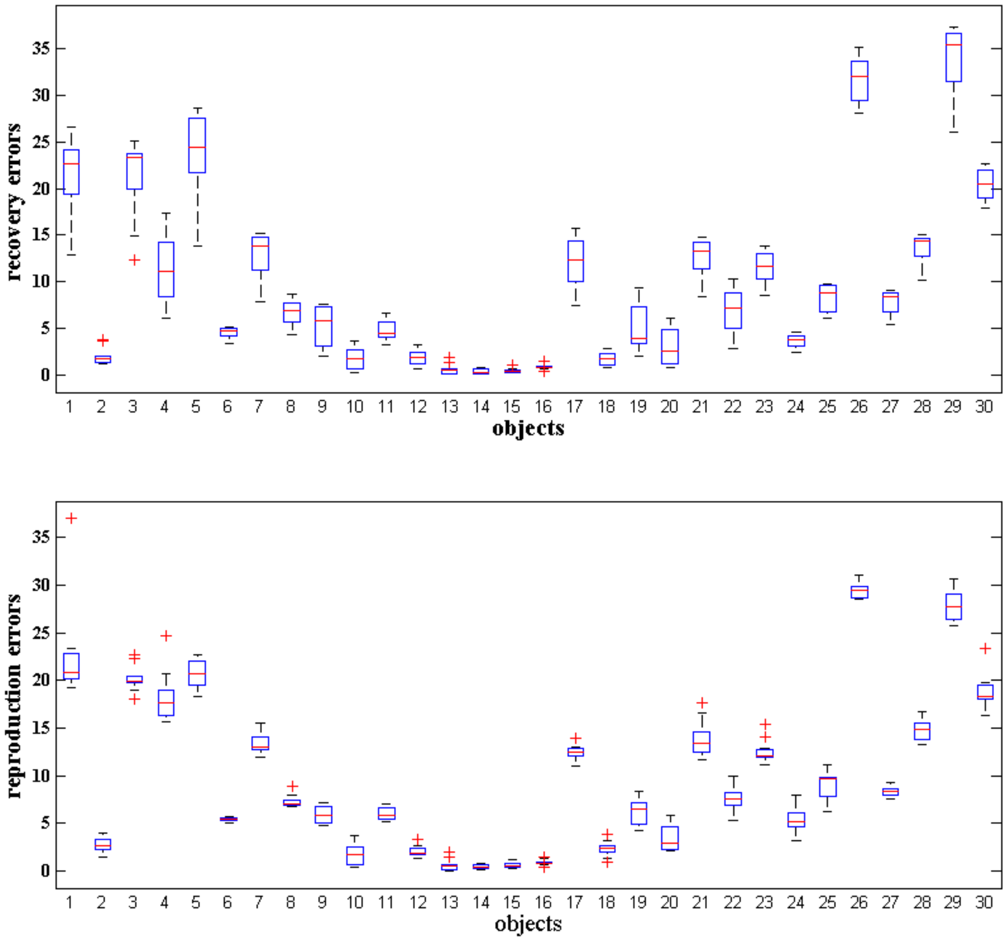


FIGURE 5.2: Box plots of recovery (top) and reproduction (bottom) angular errors for the 30 objects in SFU dataset.

To study the correlation of recovery and reproduction angular error for the two cases of the same and diverse scenes, two datasets are considered: 1) SFU data set (multiple objects each being viewed under multiple lights) for a range of algorithms. Our expectation here is that, recovery and reproduction errors, while correlated, the correlation will be less for a data set where the same object is viewed under multiple lights. 2) The Gehler-Shi colour checker data set which comprises a wide variety of scenes viewed under a single light.

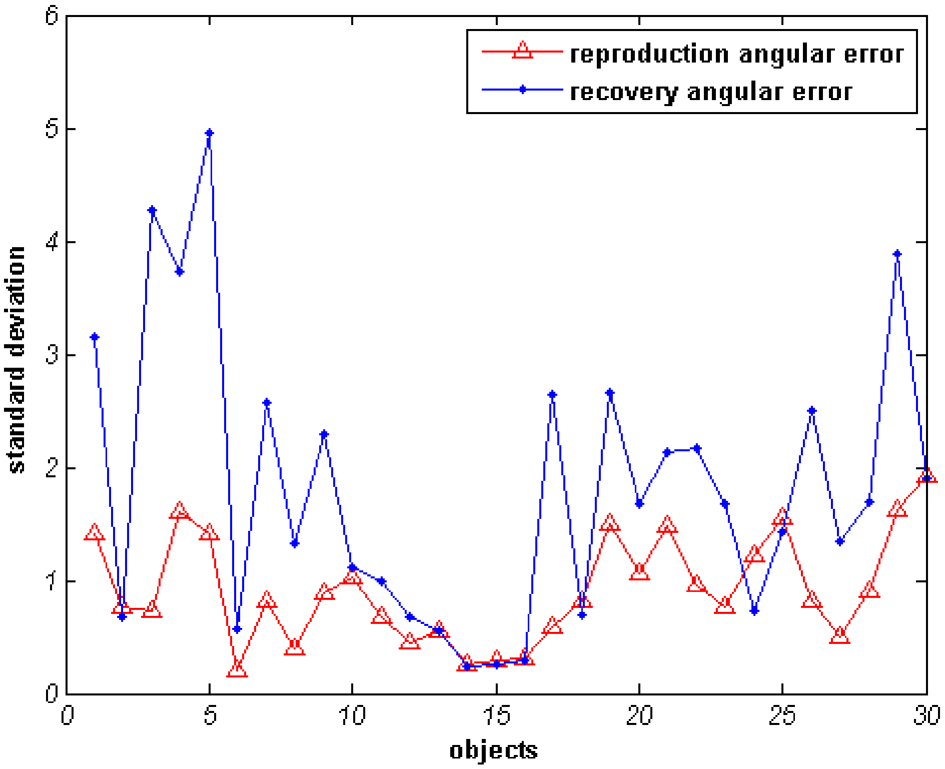


FIGURE 5.3: Standard deviation of recovery and reproduction angular errors for the 30 objects in the SFU dataset.

5.4.1 Similar Scenes with Different Illuminants

Figure 5.4 shows an example of the same object from SFU dataset being captured under different illuminations. This is a good example of the same scene being captured under different illuminants.

For each image in Figure 5.4, the illuminant is estimated using six algorithms (see the first column in Table 5.21). We assess the correlation of the algorithms using both the recovery and reproduction angular errors. In Table 5.21, we tabulate Pearson's r coefficient of correlation [Sprent and Smeeton, 2007]. A correlation of one means the errors would be proportional to one another, 0 no correlation and -1 maximum negative correlation. Interestingly, for the six algorithms tested there is a low correlation between the reproduction and recovery angular errors.

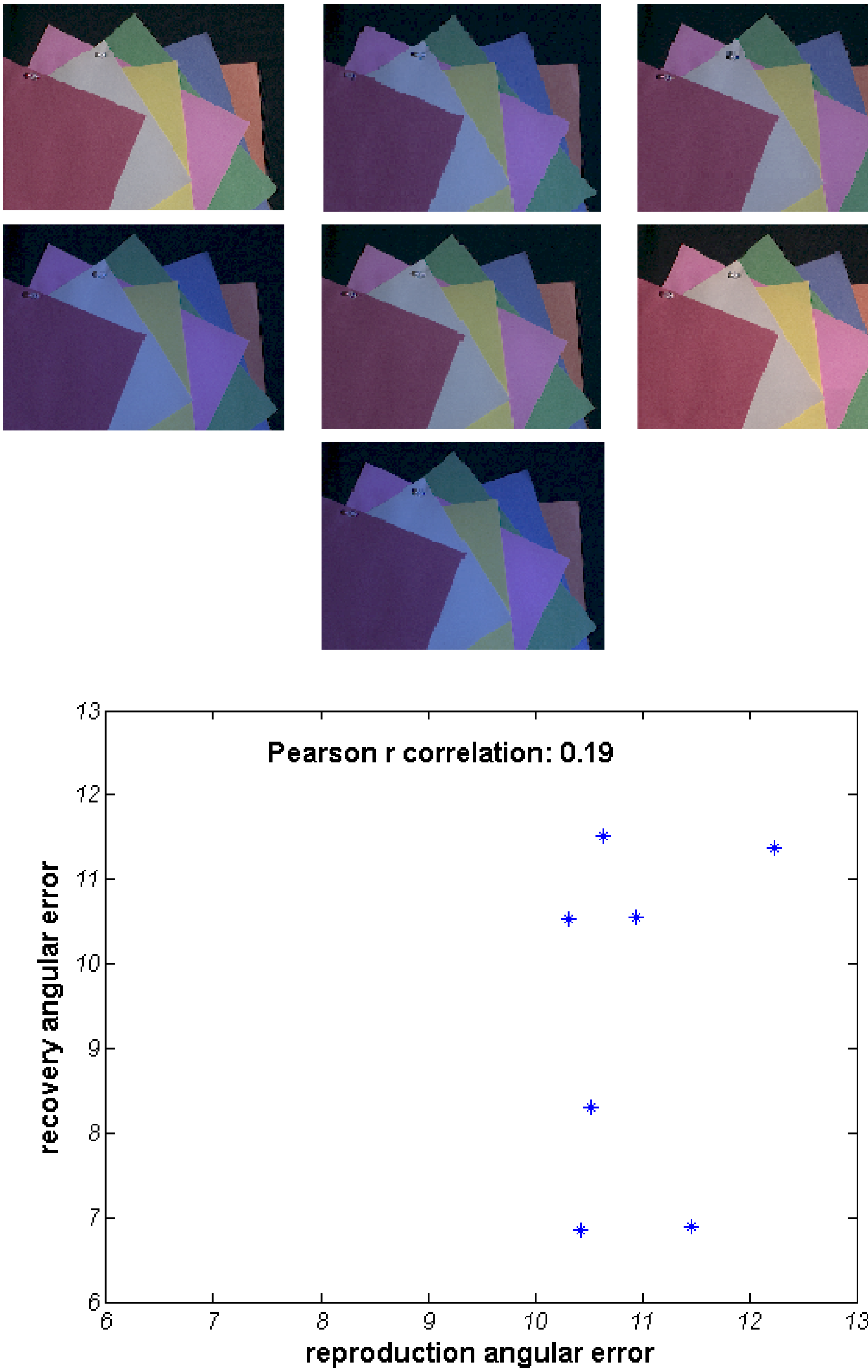


FIGURE 5.4: Correlation of reproduction and recovery angular errors for 1^{st} order grey-edge ($p - norm = 3$, $\sigma = 3$) algorithm applied on a set of images in the SFU dataset. The number on the plot shows the Pearson's r correlation value between the two errors. The images are not colour corrected.

In Figure 5.4, the plot of correlation between the two errors for the 1^{st} order grey-edge algorithm can be seen. As you can see the error values are highly

Table 5.21: Results of Pearson's r correlation test for the reproduction and recovery errors for several algorithms on a set images from SFU dataset

Algorithm	Pearson's r
1^{st} order grey-edge ($p = 3, \sigma = 3$)	0.19
2^{nd} order grey-edge ($p = 4, \sigma = 2$)	0.04
grey-world	0.55
Shades of grey ($p = 6$)	0.09
Edge gamut mapping ($\sigma = 7$)	0.29
Pixel gamut mapping ($\sigma = 8$)	0.21

uncorrelated. As expected the reproduction error is stable but for the given fairly constant reproduction error the recovery error varies widely.

5.4.2 Diverse Scenes

To study the correlation of recovery and reproduction angular errors for the diverse scenes Gehler-Shi dataset [Gehler et al., 2008; Shi and Funt, 2010] is used which contains different images of indoor and outdoor situations.

In Figure 5.5 we show a few different scenes. On the right side of Figure 5.5 the reproduction and recovery angular errors for the 1^{st} order grey-edge algorithm for the Gehler-Shi dataset is shown.

In Table 5.22, the Pearson's r values are reported for a group of algorithms on all the images of Gehler-Shi dataset. The correlation values are almost close to one for all the algorithms. This is a significant result as it shows that on average for typical viewing conditions the legacy recovery error can be used to rank algorithms. The flaw in its formulation, while important and worth remedying does not invalidate the historical development and ranking of algorithms using datasets such as Gehler-Shi and the recovery errors. That is, the best algorithms today are better than those of five years ago and these in turn are better than the venerable grey-world and MaxRGB algorithms.

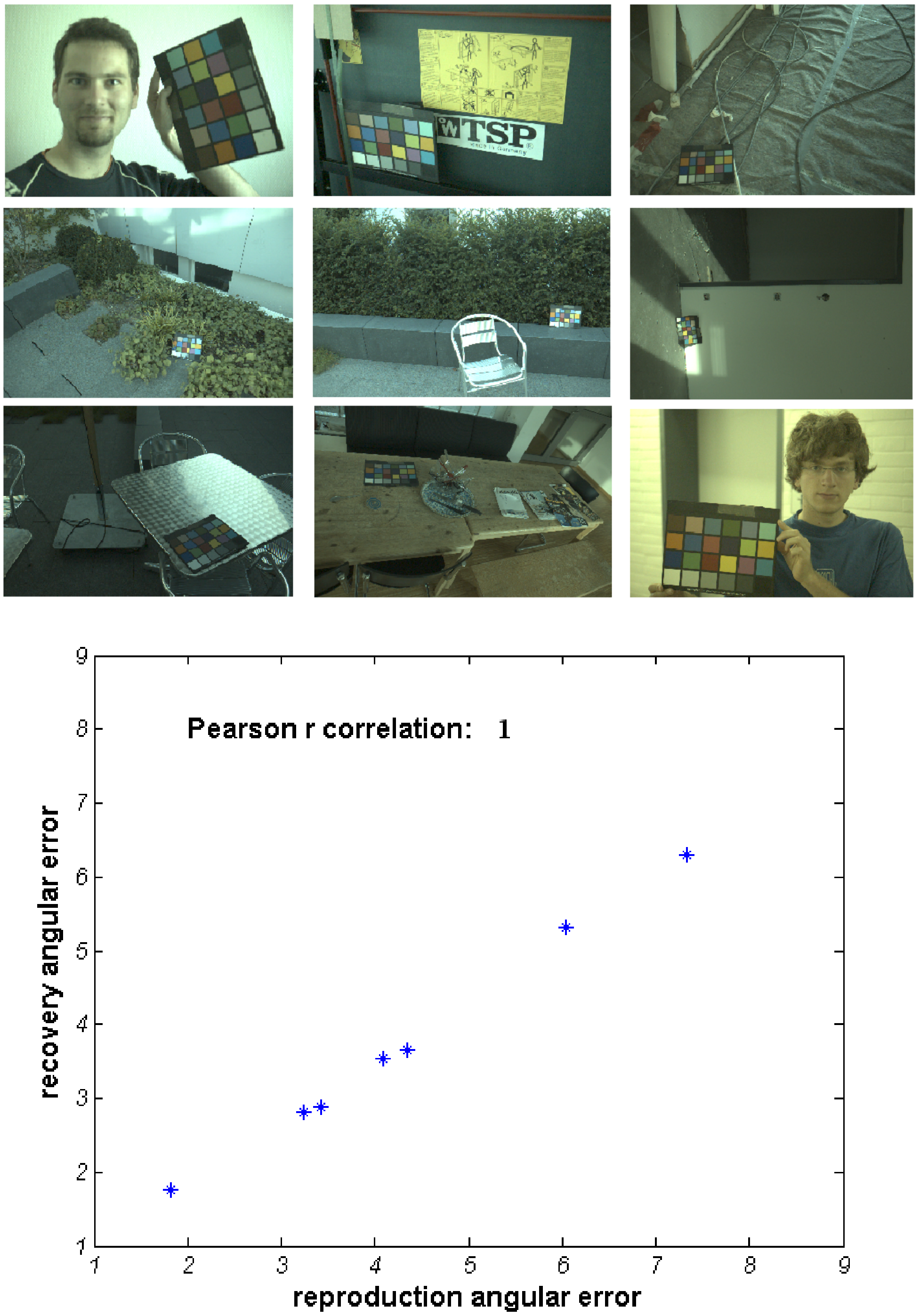


FIGURE 5.5: Correlation of reproduction and recovery angular errors for 1^{st} order grey-edge (3, 3) algorithm applied on a set of images in Gehler-Shi dataset. The number on the plot shows the Pearson's r correlation value between the two errors. The images are not colour corrected.

However, it is also important to note that the correlation statistic is a "broad brush". While the correlation analysis gives us confidence that the results in the literature (reporting the relative performance of algorithms) are in good order;

Table 5.22: Results of Pearson's r correlation test for the reproduction and recovery errors for several algorithms on all the images of Gehler-Shi dataset

Algorithm	Pearson's r
1^{st} grey-edge $(p = 3, \sigma = 3)$	0.95
2^{nd} grey-edge $(p = 5, \sigma = 6)$	0.95
grey-world	0.99
Shades of grey $(p = 5)$	0.98
MaxRGB	0.97
Pixel gamut mapping $(\sigma = 5)$	0.99
Edge gamut mapping $(\sigma = 3)$	0.96

previously, in this section we showed that there are some changes in the overall rankings when the two error metrics are used.

5.5 Conclusions

In this chapter, we observed that reproduction angular error might rank algorithms differently from recovery angular error. 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 favour of switching to the new reproduction error instead of continuing to use the legacy recovery error.

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.

Further, we studied the correlation between the reproduction and recovery angular errors for a given algorithm on images of similar and diverse scenes. We noticed the low correlation between the errors in the case of images of the same scene captured under different illuminations. Such a result was expected as the premise of reproduction angular error is that it is stable to changes in the illuminant compared to recovery angular error, which is more dependent on the illuminant. On the other hand, we observed that when the scenes are diverse the results of reproduction and recovery metrics for the same algorithm are very correlated. This observation is important as it establishes that the development of illuminant estimation algorithms is in good order. However, since we expect to capture images of the same scene as the illumination changes, we recommend the adoption of reproduction angular error.

Chapter 6

A Novel Framework for Evaluation of Illuminant Estimation Algorithms Based on a Palette of Colours

In Chapter 4, we introduced the reproduction angular error which is the angle between the true and the estimated reproduced white. A white surface is a good representative of the colour of light. However, the estimated illuminant is to be used to reproduce a range of colours in an image free of any cast caused by the illuminant colour. To this end, in this chapter a novel framework is proposed which measures how well a whole colour chart is reproduced [Finlayson and Zakizadeh, 2015].

6.1 Introduction

The pictorial description of reproduction angular error (introduced in Chapter 4) is shown in Figure 6.1. On the very left side of Figure 6.1 you see a white patch simulated as if it is captured under one of the relatively chromatic lights of SFU dataset [Barnard et al., 2002c]. The colour of the white patch is obviously affected by the blue colour of illuminantion and the result is the strong blue cast on the white surface. The reproduction angular error (as it is shown in Figure 6.1) can be simply defined as the angle between the RGB vector of a reproduced white patch by the ground-truth illuminant and the one reproduced using the estimated illuminant. In other words, the colour components of a white patch captured under the light $\underline{\rho}^E$ are divided (component-wise) by the R^E, G^E and B^E values of the ground-truth (measured) light $\underline{\rho}^E$. This results in the colour corrected white patch with $[R\ G\ B]^t = [1\ 1\ 1]^t$ (see the top row of Figure 6.1). Now, the white patch colour is reproduced using the estimated illuminant ($\underline{\rho}^{Est} = [R^{Est}\ G^{Est}\ B^{Est}]$) which results in a colour, different from [1 1 1] (see the bottom row of Figure 6.1) since $\underline{\rho}^{Est}$ is only an estimation of $\underline{\rho}^E$. The reproduction angular error measures the angle between the colour vectors of the two reproduced white patches $C1$ and $C2$.

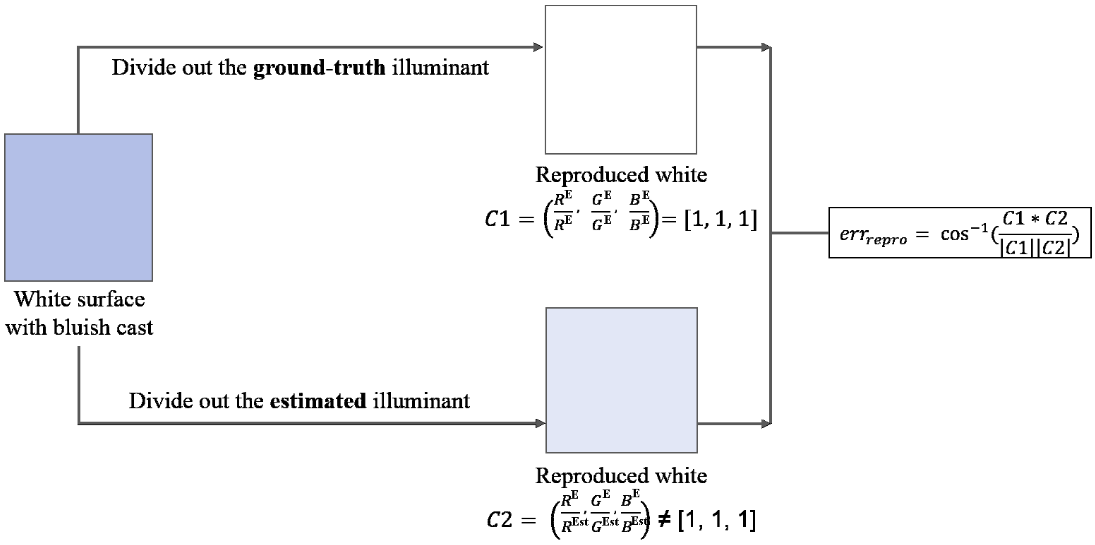

Divide out the **ground-truth** illuminant
White surface with bluish cast
Reproduced white
$C1 = \left(\frac{R^E}{R^E}, \frac{G^E}{G^E}, \frac{B^E}{B^E}\right) = [1, 1, 1]$
$err_{repro} = \cos^{-1}(\frac{C1 * C2}{|C1||C2|})$
Divide out the **estimated** illuminant
Reproduced white
$C2 = \left(\frac{R^E}{R^{Est}}, \frac{G^E}{G^{Est}}, \frac{B^E}{B^{Est}}\right) \neq [1, 1, 1]$

FIGURE 6.1: A pictorial description of reproduction angular error.

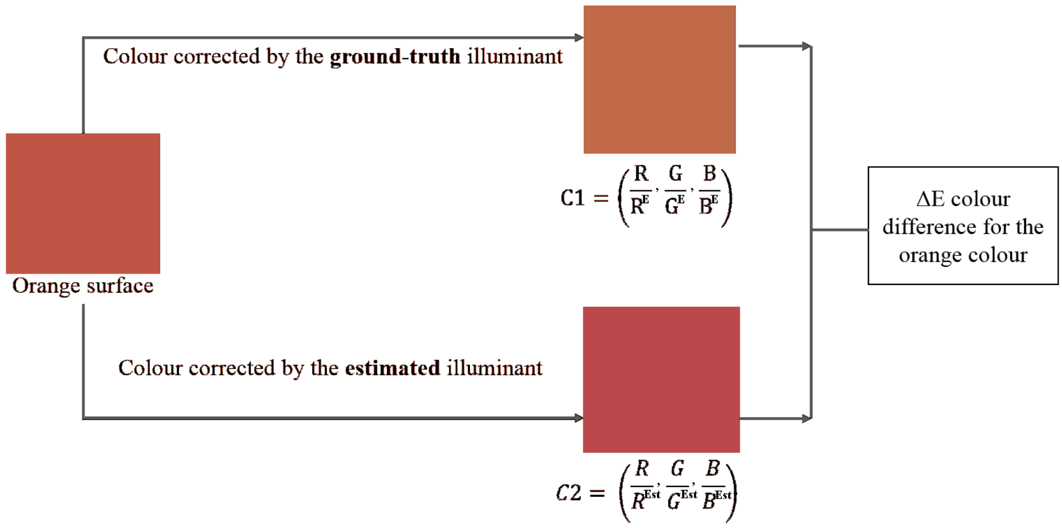

Colour corrected by the **ground-truth** illuminant
Orange surface
$C1 = \left(\frac{R}{R^E}, \frac{G}{G^E}, \frac{B}{B^E}\right)$
ΔE colour difference for the orange colour
Colour corrected by the **estimated** illuminant
$C2 = \left(\frac{R}{R^{Est}}, \frac{G}{G^{Est}}, \frac{B}{B^{Est}}\right)$

FIGURE 6.2: ΔE for reproduced colours.

In this chapter, we seek to measure the difference between a range of colours (not just a white patch) which are reproduced by the estimated and ground-truth illuminants. This idea is depicted in Figure 6.2. This figure shows an orange surface simulated as if being captured under a chromatic light. The colour of the surface is then corrected by dividing out the ground-truth light (top row) and the estimated light by grey-world algorithm [Buchsbaum, 1980] (bottom row). The quality of the reproduced orange colour by the estimated light is then evaluated using the colour difference formula CIE2000 ΔE [Sharma et al., 2005] (Appendix A) .

Figure 6.2 represents an idealised scenario. To calculate the ΔE colour difference (as it is explained in Appendix A) we require the XYZ values of colours and the illuminant. However, most colour constancy benchmark data sets don't provide such information, as these data sets are aimed to provide real photographic look images. Even if the XYZ values can be calculated (e.g. with spectral data sets such as [Foster et al., 2006]) the ground-truth and the estimated illuminants are only a simulation of those made by a digital camera. This is what makes the task of evaluating an algorithm's performance by comparing colours challenging.

In our approach we seek to generate a Macbeth colour checker as if it would appear under the actual and the estimated illuminant. We then calculate ΔE for the 24 patches.

The organisation of this chapter is as follows: In Section 6.2, the framework for evaluation of illuminant estimation algorithms based on a palette of colours is given. The results are discussed in Section 6.3. Section 6.4 concludes the chapter.

6.2 The Framework for Evaluation of Illuminant Estimation Algorithms Based on a Palette of Colours

We develop our model for the SFU dataset [Barnard et al., 2002c]. We use the set of spectra for 24 Macbeth colour checkers patches and the 23 lights from the SFU dataset [Barnard et al., 2002c]. For camera spectral sensitivity functions we use the Sony-DXC-930 CCD [Barnard et al., 2002c] which is used to make the SFU data set. SFU dataset is useful for our purpose since the spectral of the lights under which the images were captured using Sony-DXC-930 camera are also provided. Although, for the camera sensitivity functions any particular camera can be used in the problem formulation.

6.2.1 Generating Synthetic Colour-Checkers for a Target light

We adopt the standard model of image formation:

$$\rho_k = \int_{380}^{780} R_k(\lambda) E(\lambda) S(\lambda) d\lambda. \tag{6.1}$$

In Eq. 6.1, $R_k(\lambda)$ is the Sony-DXC-930 camera sensitivity functions for the three sensors $k = \{R,\ G,\ B\}$. $S(\lambda)$ is the spectra of the 24 patches of Macbeth colour checker and $E(\lambda)$ is the spectra of the 23 SFU lights. The data is provided from $\lambda = 380nm$ to $780nm$ (the visible spectrum) with 4nm intervals. We actually calculate Eq. 6.1 as a Reimann summation. Using Eq. 6.1 and for the 23 SFU lights we generate 24 RGBs (each for one patch of the colour checker). Figure 6.3 demonstrates different steps of the framework. The final product is shown as K (the matrix of 23 checkers) in Figure 6.3. These 23 synthetic checker images encapsulate our understanding of how checker appears under different lights.

In the second step, we wish to generalise this understanding so that we could, given the RGB of any target light, synthesise the appearance of the checker for any illuminant. Denoting the 24×3 RGBs for a Macbeth colour checker as M, we model M as a linear sum of three basis Macbeth colour checkers:

$$M \approx \sum_{i=1}^{3} M_i m_i, \tag{6.2}$$

where, m_i denotes a scalar weight and the optimal basis in a least-square sense are found using Principal (Characteristic) Vector Analysis [Maloney 1986] of the 23 synthetic Macbeth checker images. Crucially, we found the best basis models our data extremely well with the actual and 3-basis approximation being visually almost the same in appearance (the three basis capture 99% of the variance).

Now, we place the RGB for the white reflectance in the Macbeth checker (the 19^{th} patch) for each basis term M_i in the three columns of a calibration matrix Ω. Denoting an RGB of a light as $\underline{\rho}^E$, the linear combination of the columns of Ω defines the weights $\underline{m}$ used in Eq. 6.2:

$$\underline{m} = \Omega^{-1}\underline{\rho}^E. \tag{6.3}$$

In Eq. 6.3, the illuminant vector $\underline{\rho}^E$ could be the reference light ([1 1 1]), the ground truth light or the estimate made by an algorithm. Given $\underline{m}$, we can

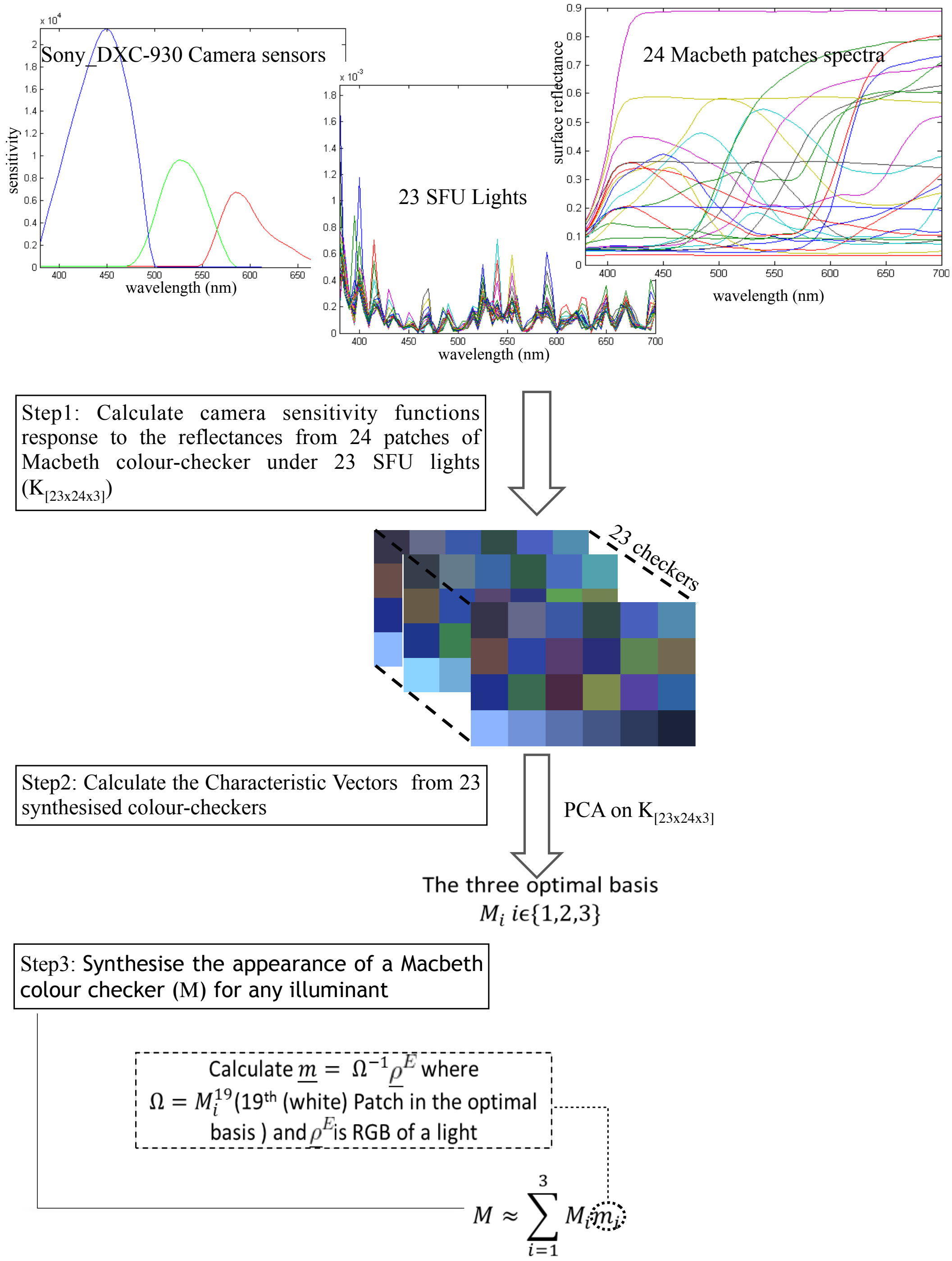

x 10^4
sensitivity
Sony_DXC-930 Camera sensors
wavelength (nm)
x 10^-3
23 SFU Lights
wavelength (nm)
surface reflectance
24 Macbeth patches spectra
wavelength (nm)
Step1: Calculate camera sensitivity functions response to the reflectances from 24 patches of Macbeth colour-checker under 23 SFU lights ($K_{[23x24x3]}$)
23 checkers
Step2: Calculate the Characteristic Vectors from 23 synthesised colour-checkers
PCA on $K_{[23x24x3]}$
The three optimal basis
M_i $i\epsilon\{1,2,3\}$
Step3: Synthesise the appearance of a Macbeth colour checker (M) for any illuminant
Calculate $\underline{m} = \Omega^{-1}\underline{\rho}^E$ where $\Omega = M_i^{19}$(19th (white) Patch in the optimal basis) and $\underline{\rho}^E$is RGB of a light
$M \approx \sum_{i=1}^{3} M_i m_i$

FIGURE 6.3: The general framework for generating a synthetic Macbeth colour-checker under a target light.

calculate the appearance of the checker using Eq. 6.2. Figure 6.4 (a) shows one input image from the SFU dataset [Barnard et al., 2002c] and Figure 6.4 (b) shows our synthesised Macbeth for this light.

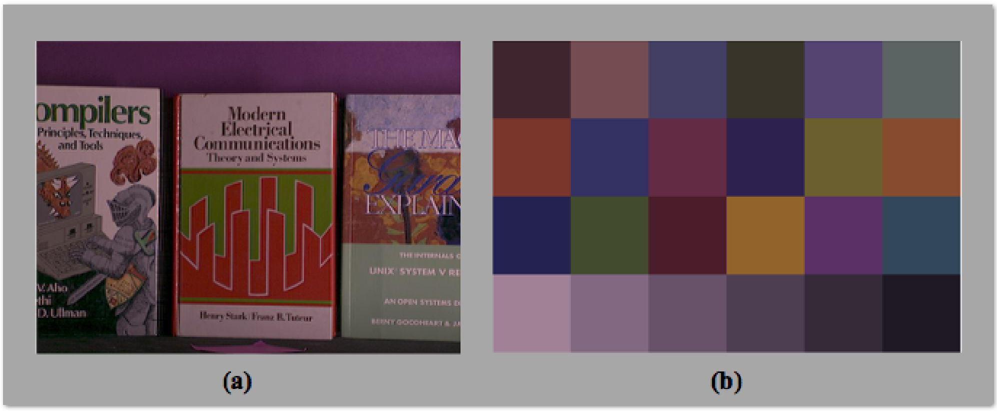


Figure 6.4: (a) An image from SFU dataset (captured under Sylvania warm white fluorescent light), (b) The synthetic colour checker under the ground truth light under which (a) was taken.

Now, we wish to consider the appearance of the checker when we make an image reproduction. That is we wish to reproduce the checker with the actual (ground-truth) light and compare it to the checker reproduced when an illuminant estimation algorithm is used to define the illuminant colour.

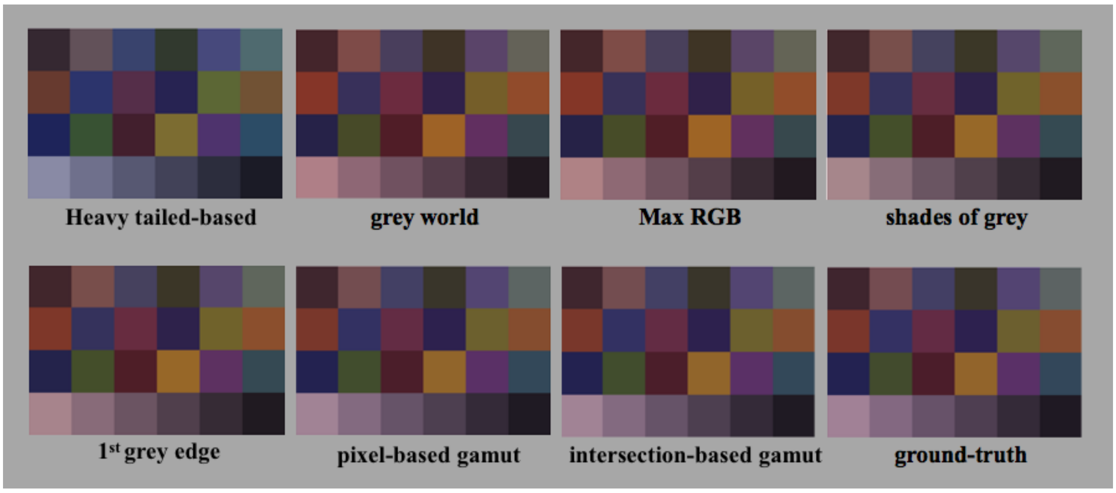


Figure 6.5: Synthetic colour-checkers under the estimates of the actual light by different illuminant estimation algorithms. The synthesised colour checker under the actual light (the last checker) is also included.

Figure 6.5 shows the synthesised colour checkers under the estimates of the same illuminant in Figure 6.4 (a) using different illuminant estimation algorithms (for the ease of comparison, we have also included the checker under the ground-truth illuminant in this figure). All images are scaled so that the brightest pixel value across the colour channels is one and a gamma of 0.5 is applied. It can be seen that depending on the performances by various algorithms the synthesised colour checkers look close to or very different from the one under the ground-truth illuminant (the last checker in Figure 6.5).

6.2.2 Modelling the Appearance of a Colour-Corrected Checker

So far, we have focussed on explaining how we synthesise the colours of the Macbeth colour checker for a target light. But, we ultimately seek to model the appearance of a checker under an actual light when it is corrected to the reference checker using the estimated illuminant (by an algorithm).

Denoting, respectively, the checkers under the reference white light ([1 1 1]), the actual coloured light and the estimated coloured light as M^{ref}, M^{act} and M^{est}, the estimated reproduction, $\tilde{M}^{ref}$, is calculated as:

$$\tilde{M}^{ref} = M^{act} T, \tag{6.4}$$

and T is calculated as:

$$T = [M^{est}]^{+} M^{ref}, \tag{6.5}$$

where, $[M^{est}]^{+}$ denotes the Moore-Penrose pseudoinverse [Ben-Israel and Greville, 2003]:

$$[M^{est}]^{+} = ([M^{est}]^{t}[M^{est}])^{-1}[M^{est}]^{t} \tag{6.6}$$

That is, T (in Eq. 6.5) is the least-squares fit from the checker viewed under the estimated light to the reference lighting conditions. This 3×3 matrix T

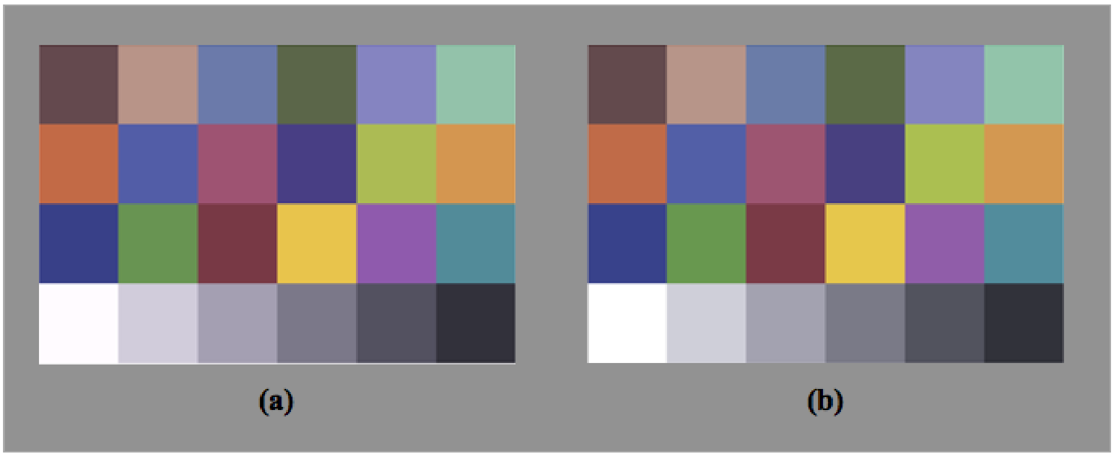


FIGURE 6.6: (a) The corrected checker by pixel-based gamut mapping algorithm. (b) The correct reproduction of the checker (i.e. synthesised checker under the reference white light [1 1 1]).

is then applied (as a correction matrix) to the checker under the actual light (Figure 6.4 (b)) to result in the corrected colour checker using the illuminant estimate. Figure 6.6 (a) shows the corrected checker by pixel-based gamut mapping algorithm [Gijsenij et al., 2010]. Figure 6.6 (b) is the correct reproduction of the checker, i.e. the checker under the reference white light ($\underline{\rho}^E = [1\ 1\ 1]$) where the colour of the white patch is equal to [1 1 1]. The checker in Figure 6.6 (a) is fairly similar to the checker under the reference white light (Figure 6.6 (b)). In this case, pixel-based gamut mapping algorithm does a good job colour-correcting the checker.

6.2.3 CIElab Colour Differences of the Reproduced Colours

Given the appearance of the reproduced Macbeth colour checker using the estimated illuminant ($\tilde{M}^{ref}$), we calculate the error for the i^{th} Macbeth colour checker patch as:

$$err_i = \|f(\tilde{M}_i^{ref}) - f(M_i^{ref})\| \tag{6.7}$$

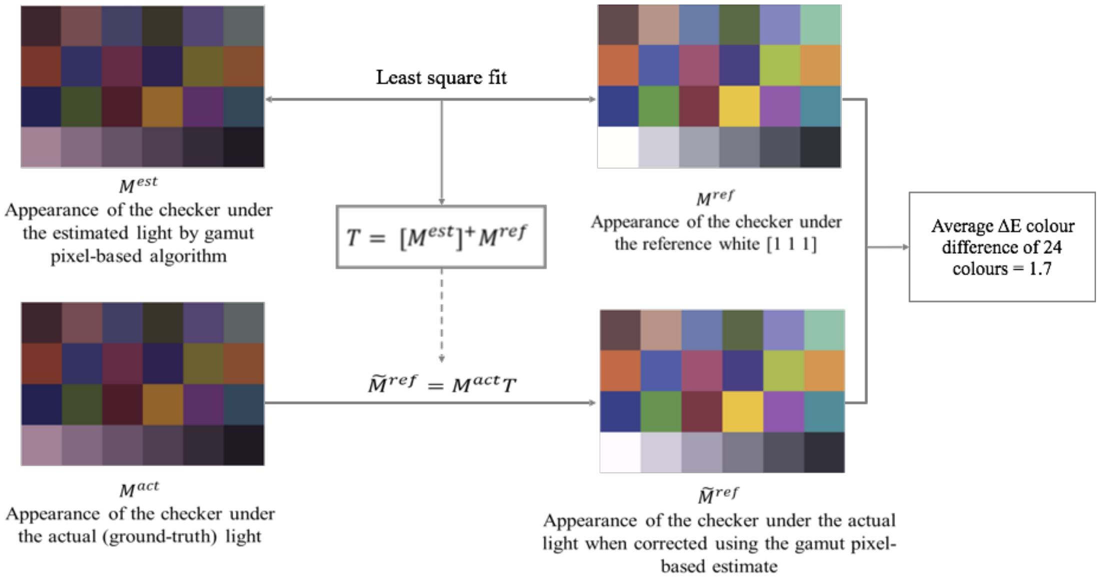

Least square fit
M^{est}
Appearance of the checker under the estimated light by gamut pixel-based algorithm
$T = [M^{est}]^{+} M^{ref}$
M^{ref}
Appearance of the checker under the reference white [1 1 1]
Average ΔE colour difference of 24 colours = 1.7
$\tilde{M}^{ref} = M^{act} T$
M^{act}
Appearance of the checker under the actual (ground-truth) light
$\tilde{M}^{ref}$
Appearance of the checker under the actual light when corrected using the gamut pixel-based estimate

FIGURE 6.7: Evaluating the quality of the colours in a reproduced colour-checker.

where f maps an RGB to CIE LAB (see Appendix A for mathematical formulation of RGB to CIE LAB conversion). Equation (6.7) can be any colour difference ΔE formulae but here we use the $\Delta E2000$ [Sharma et al., 2005].

Figure 6.7 is a pictorial discerption of Eq. 6.4 where a corrected colour checker ($\tilde{M}^{ref}$) is produced using the illuminant estimate by the pixel-based gamut mapping algorithm. $\tilde{M}^{ref}$ is then compared with the M^{ref} which would be the perfect Macbeth checker under a reference white light. The average ΔE difference in the reproductions is 1.7.

Figure 6.8 shows the colour checker in Figure 6.4 (b) which is white balanced using the estimates of different algorithms. The values on each checker are the the average of $\Delta E2000$ errors for the 24 patches of the checker.

6.3 Results

Here we use the 321 images from the SFU dataset [Barnard et al., 2002c]. This data set has linear images and a variety of objects are imaged under 11 lights

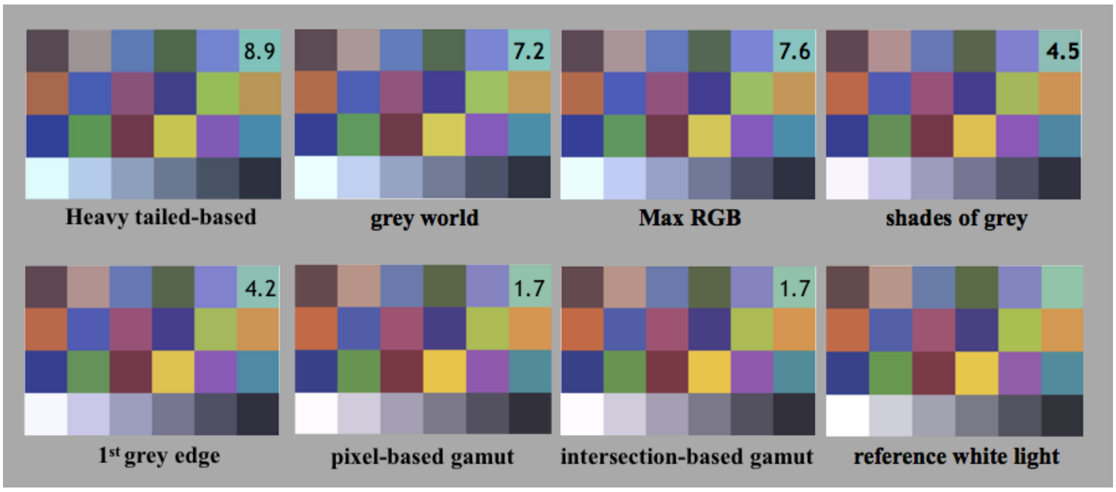


Figure 6.8: Synthetic white balanced colour-checkers by different illuminant estimation algorithms.

(ranging from quite yellowish to very blue). All images were captured with the SONY DXC-930. A variety of algorithms, including those listed in Table 6.1, were tested by [Gijsenij et al., 2011]. We can thus calculate for all Macbeth colour checker images and the overall median ΔE. Then according to this global median we can rank the algorithms. In Table 6.1, we list the algorithms and record the rank for the Recovery and Reproduction angular errors and the new calculated median ΔE colour differences.

While the rankings of all three metrics are almost similar it is clear that recovery angular error ranks algorithms a little differently from reproduction angular error. Further in Chapter 5 it was shown that the rankings are statistically different. And, this fact draws attention to the care the algorithm designer needs to take using the appropriate metric to assess their algorithm. The reproduction angular error assesses how well an algorithm reproduces white (i.e. when the estimated illuminant is divided out). The framework introduced in this chapter builds on this concept and accounts for the error for other surface colours. The ranks for the median ΔE errors are identical to the reproduction angular error.

Table 6.1: Comparison of ranking of algorithms based on reproduction angular errors and generalised reproduction errors.

	Recovery angular error		Reproduction angular error		ΔE	
Method	Median error	Rank	Median error	Rank	Median error	Rank
Gray-World	7.00	9	7.49°	9	7.02	9
MaxRGB	6.52	8	7.44°	8	6.13	8
Heavy tailed-based	3.45	**6**	4.11°	7	3.74	7
Shades-of-gray ($p = 7$)	3.72	**7**	3.94°	6	3.26	6
1^{st} grey edge ($p = 14$,$\sigma = 4$)	3.21	5	3.59°	5	3.12	5
2^{nd} grey edge ($p = 15$,$\sigma = 10$)	2.73	4	3.04°	4	2.88	4
pixel-based gamut ($\sigma = 4$)	2.27	**2**	2.83°	3	2.64	3
Edge-based gamut ($\sigma = 2$)	2.78	**3**	2.70°	2	2.59	2
Intersection-based gamut ($\sigma = 3$)	2.09	1	2.48°	1	2.46	1

6.4 Conclusion

In this chapter, we extended the idea of evaluating illuminant estimation algorithms based on the RGB values of a reproduced white patch (reproduction angular error), to study the quality of a range of colours (specifically the 24 patches in a Macbeth colour checker) reproduced by multiple algorithms.

We compared the reproduced colour patches by different illuminant estimation algorithms by looking at their CIE $\Delta E2000$ colour differences. In most cases, our evaluation based on reproduced colours matches the judgement we previously had using the reproduction angular error for a reproduced white patch. Indeed comparison of the reproduced images is a very efficient way of evaluation of illuminant estimation algorithms.

Chapter 7

Psychophysical Evaluation of Illuminant Estimation Algorithms

In Chapter 4, we, mathematically, demonstrated the stability of reproduction angular error when it evaluates the same algorithm's results for the same scene (only the illumination changes from one image to another).

Further, in Chapter 5, it was shown that the rank order of some algorithms for a benchmark dataset might switch if they are ranked using reproduction angular error instead of recovery angular error. In this chapter, we wish to divine whether observers judge image reproduction that correlates with reproduction angular error and/or the legacy recovery angular error.

7.1 Introduction

The contributions of this chapter are: first, we design a new experiment, based on image preference, which aims to evaluate whether reproduction or recovery error best correlates with judgements concerning the accuracy of image reproductions. Second, the experiment is implemented taking great care - and this is a first time, to our knowledge - to produce photographically plausible reproductions.

The chapter is organised as follows. In Section 7.2, the psychophysics experiment and the preparation of the data used in the study is explained. In Section 7.3, the results from the experiment are analysed. The chapter concludes in Section 7.4.

7.2 Psychophysics Experiment Set up and Data Preparation

7.2.1 Data Preparation

The images for our experiment are from the 200 images of the Canon EOS 600D camera from the recently created NUS dataset [Cheng et al., 2014].

In our experiment, illuminant estimates are 'divided out' from the raw images. Then, in a second step we apply a camera processing pipeline (in effect modelling colour correction, gamut mapping and tone correction [Ramanath et al., 2005]). The colour mapping process, i.e. mapping the raw sensor values to their corresponding RGB outputs, has been the subject of a number of studies (e.g. [Grossberg and Nayar, 2003, 2004; Kim and Pollefeys, 2008; Chakrabarti et al., 2009; Xiong et al., 2012; Kim et al., 2012]). However, most of those research are concerned with radiometric calibration, which is the process of recovering scene radiance from image intensities. For this experiment, we are interested in creating a camera output look for the white balanced images generated when we divide out

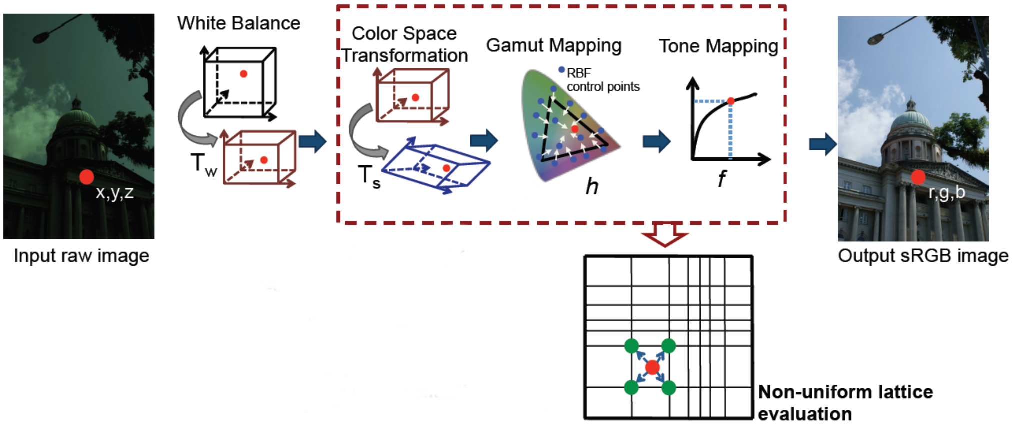


FIGURE 7.1: The simplified in-camera colour processing pipeline (from [Lin et al., 2012] with a little modification). A single sparse LUT has replaced several steps in the imaging pipeline.

the camera's estimate. Li *et al.* [Lin et al., 2012] suggest a calibrated (trained) sparse 3D lookup table (LUT), also known as lattice [Garcia and Gupta, 2009], suffices to map the raw sensor values from a particular camera to their corresponding RGB outputs. Crucially, to a good approximation, the same lattice can be used independent of the white point [Lin et al., 2012]. Using the calibrated lattice for a specific camera we can render white balanced images as if they have been passed through an in-camera colour processing pipeline like the simplified model in Figure 7.1.

Here for calibrating the lattice we have used a random selection of 50 raw images white balanced by the ground-truth illuminants and their corresponding output JPG images captured by Canon EOS 600D camera from the NUS dataset. We randomly select 50000 pixels from each image to generate the lattice. The above sampling results in 2500000 points for training the lattice. In our experiment, the dimension of the lattice is three (for R, G and B colour channels), the boundaries of the grid is set to be between zero and one and the size of the grid is 35 nodes in each dimension. So, we solve for a $35 \times 35 \times 35 = 42875$ lattice. Figure 7.2 gives a visual illustration of calibrating the lattice for the Canon EOS 600D camera based on a set of NUS dataset images captured by the same camera. Where for

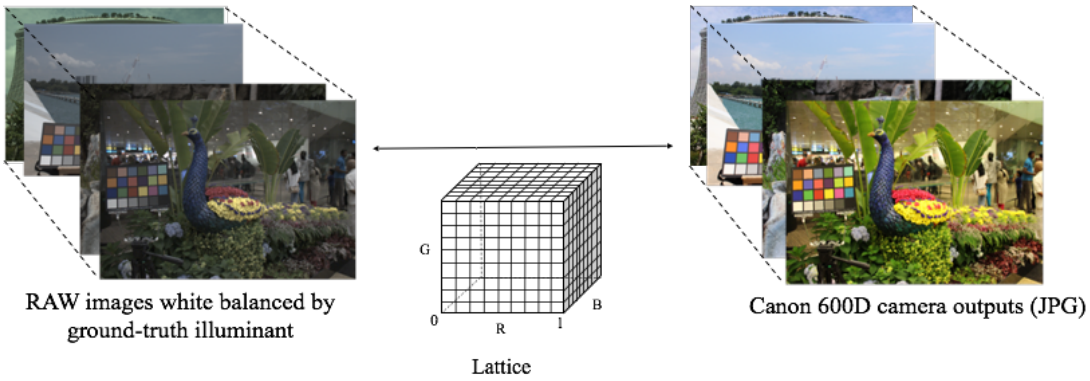

RAW images white balanced by ground-truth illuminant
G
0
R
1
B
Lattice
Canon 600D camera outputs (JPG)

FIGURE 7.2: Lattice calibration for the Canon EOS 600D camera from the NUS dataset.

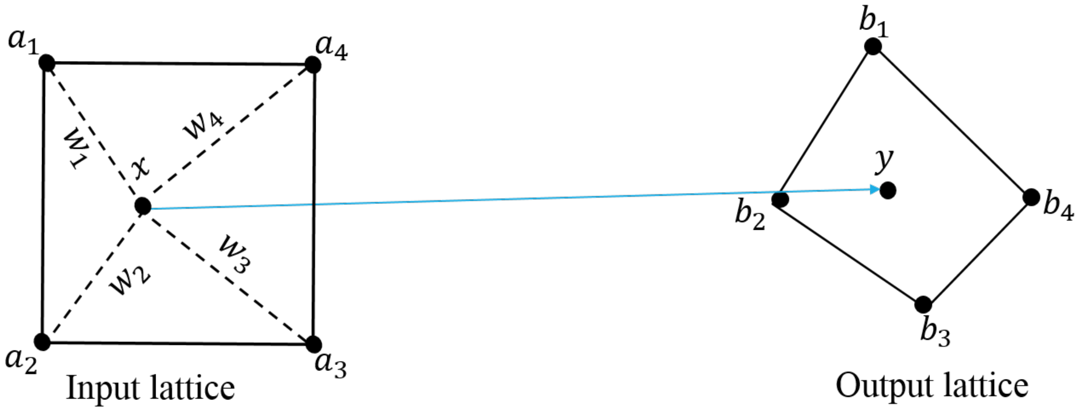


FIGURE 7.3: The application of the lattice: x is the linear combination of the bounded points a_i in the input lattice and its corresponding output point y is the same linear combination of the bounding points b_i in the output lattice.

simplicity a $9 \times 9 \times 9$ lattice is shown.

The optimisation to derive the lattice is presented in detail in [Garcia et al., 2012]. Once the lattice is calculated, given a raw shades of grey white balanced image the lattice is applied on each pixel of the raw images to generate its corresponding JPG equivalent. In Figure 7.3, we show the application of the lattice. Here, a point in an input coordinate system (e.g. raw) is presented as a linear combination of the rectangular region in which it falls (bounded by a_i). On the right of the figure we show the output lattice (b_i). The output value is the same linear combination of the output control points.

Figure 7.4 shows a few examples of images from Canon600D camera which are white balanced by the shades of grey algorithm and their equivalent JPG transformation by the calibrated lattice. The first row of this figure are the raw white balanced images, the images are raised to the 0.5 gamma value to make them visible. The second row are the results of applying the trained lookup table to the images in the first row. The last row are the actual JPG outputs from Canon EOS 600D. Note that the images in the second row look better that those in the first. This is important. Often illuminant estimation experiments (in computer vision or psychophysics) use raw images. Of course the actual camera outputs (third row) look best. This illustrates that for these images the shades of grey algorithm does not produce a reproduction as pleasing as that delivered by the camera's own proprietary algorithm.

Gamma correction or applying the lattice might effect the appearance of the images. However the same function is applied by the cameras after white balancing the image, regardless of white balanced method used and it improves the appearance of images.

7.2.2 Monitor

The images are presented on a high resolution professional LCD Backlit monitor (an HP DreamColor LP2480zx) with 1920 × 1200 pixels resolution. The monitor uses both a true 30-bit panel and an RGB LED backlight, providing over one billion possible colours and a wide colour gamut.

According to ISO 3664 standards, the calibration of the monitors for the psychophysics experiment is necessary. The monitor was calibrated using Spyder4Elite [spy, accessed Sept, 2016] prior to running the experiment. The calibration was carried out in the same environment in which the experiment would later take place. The experimental environment is a room illuminated with a dim light source (to avoid eye strain) provided behind the monitor to avoid glare.

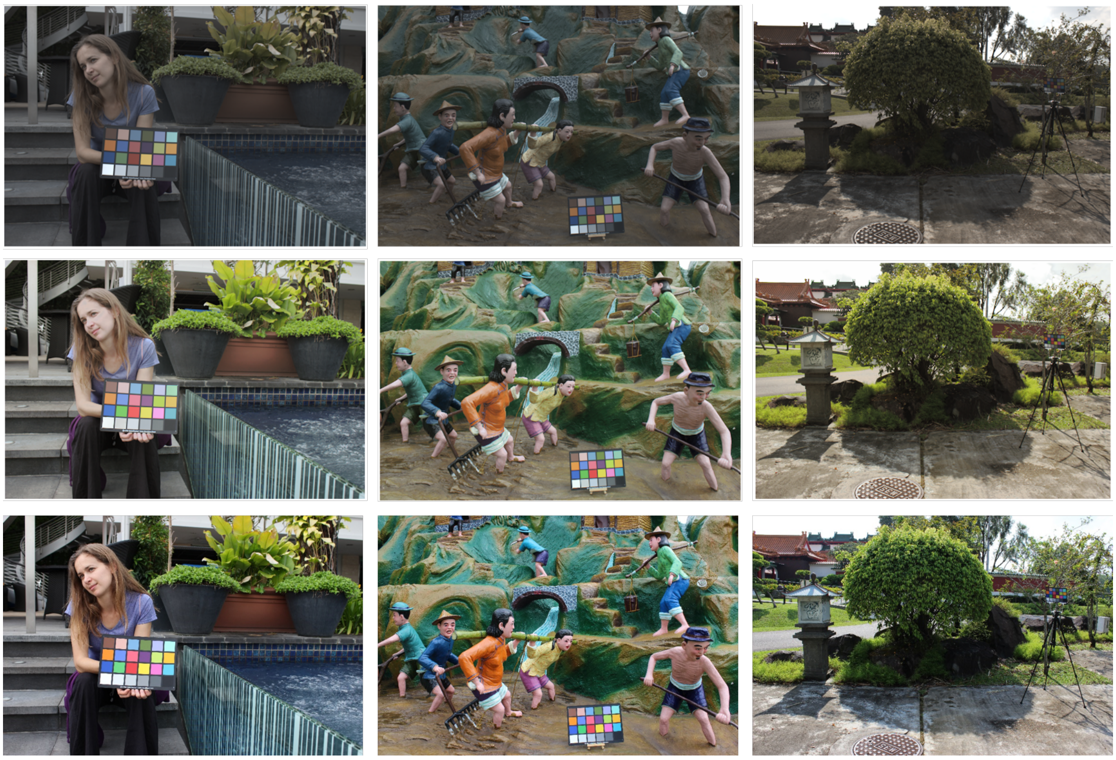


FIGURE 7.4: Examples of raw to JPG transformation using the calibrated lattice: The first row are the raw white balanced images by the shades of grey algorithm. The second row are the results of applied lattice. The last row are the actual JPG outputs from the camera (with the camera's properly white balanced algorithm).

7.2.3 Observers

All observers participated in the experiment have normal colour vision and normal to corrected-to-normal visual accuracy (all observers were asked to declare any visual deficiency including colour blindness). At the beginning of the experiment the observers are allowed to adapt their vision for 30 seconds by staring at a variegated grey screen. This adaptation period is necessary to allow the observer's vision to adjust to the viewing conditions. There were eight observers participating in this experiment with their age ranging from almost 25 to 65. The group of observers contained both male and female participants.

A diagram of experimental set up is shown in Figure 7.5.

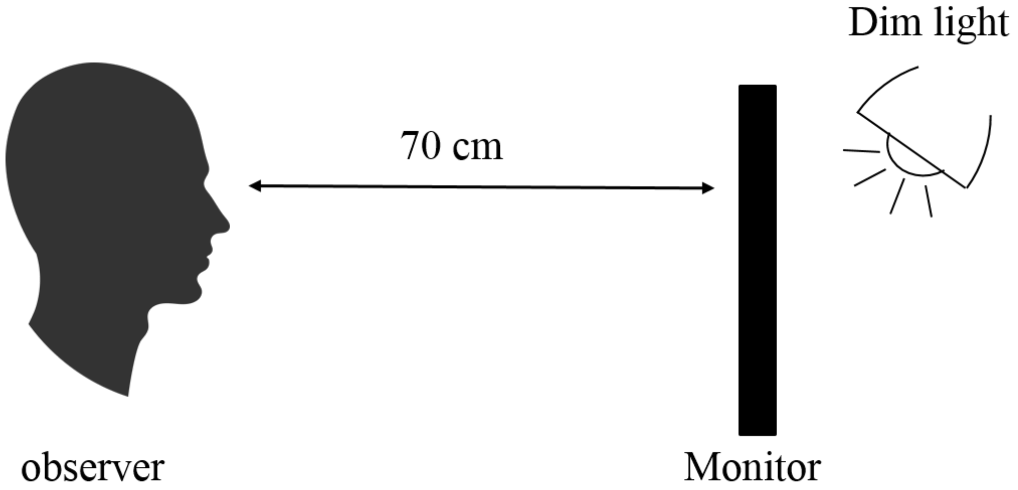


Figure 7.5: Set up for the psychophysics experiment.

7.2.4 Experimental Procedure

An observer is shown two pairs of images on a variegated grey background(like the one shown in Figure 7.6). The first pair contains a ground-truth (based on the physical white point) reproduction and that produced by algorithm a estimate. We, respectively, denote the two images in the first pair ${I_t}^1$ and ${I_a}^1$. A second image pair is calculated in the same way. A ground-truth image ${I_t}^2$ is produced and that for a second illuminant estimation algorithm ${I_b}^2$. Note the scene in the first image pair is different from the second and two different illuminant estimation algorithms are used.

Figure 7.6 shows an example from our experiment. Top left and right (${I_a}^1$ and ${I_b}^2$) respectively are the reproduction delivered using the edge-based and pixel-based gamut mapping algorithms. The other images (${I_t}^1$ and ${I_t}^2$) are the ground-truth reproduction. The images are selected carefully. Image ${I_a}^1$ is reproduced by the illuminant estimation algorithm a has a lower reproduction error than image ${I_b}^2$ reproduced by algorithm b (in this example the reproduction errors are, respectively: $3.76° < 8.73°$). Conversely, the recovery angular of algorithm a for image ${I_a}^1$ is higher than the recovery error of algorithm b for image ${I_b}^2$ (in this

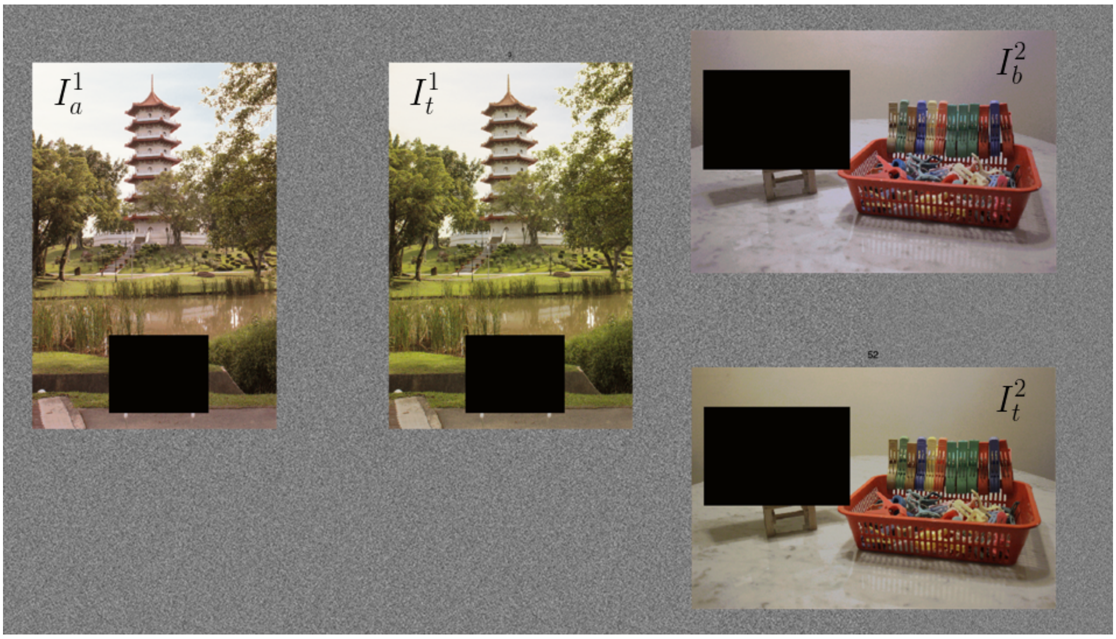


Figure 7.6: Screen setup for the experiment.

case 4.15° $>$ 3.46°). Seven pairs of images similar to the one in Figure 7.6 are chosen for each two algorithms.

In the experiment, the observer is then asked which image pair appears more similar. That is does $I_a{}^1$ look closer to $I_t{}^1$ compared with $I_b{}^2$ and $I_t{}^2$ (or vice versa). Note the observers do not know which image is corrected using the ground-truth illuminant and which by the estimate. We are interested in whether an observer judges $I_a{}^1$ to be closer to $I_t{}^1$ or $I_b{}^2$ to $I_t{}^2$. If the former, the reproduction error correctly predicts image reproduction. If the latter, it is recovery error that predicts observer's responses. The experiment is repeated for eight observers. Each image representation is repeated twice with the 'a' and 'b' pairs shown respectively left and right and the converse.

We compare four illuminant estimation algorithms in this experiment: 1^{st} grey edge, 2^{nd} grey-edge, shades of grey and pixel-based gamut mapping; here, denoted as GE1, GE2, SOG and GP respectively. Each algorithm is compared with the rest and in each pair of comparisons seven pairs of images are used. For instance, to compare 1^{st} grey edge and 2^{nd} grey-edge algorithm seven pairs of images are shown to the observers where in each pair one image is corrected by the 1^{st} grey edge and the second image is white balanced by 2^{nd} grey edge. Each pair of pairs has the 'swapping property' previously described. That is pair 'a' compared to 'b' can have lower recovery error and in reverse pair 'b' compared to 'a' has lower reproduction error.

7.3 Results and Discussion

7.3.1 Results of Chi-square Test

The Chi-square test [Conover, 1999] is a statistical test commonly used to compare observed data with the expected data. Here, the expected data is the number of pairs in which the image corrected by algorithm a is better than the image corrected by algorithm b according to the chosen metric. For instance if according to reproduction angular errors of the seven pairs of images, algorithm a is predicted to be better than algorithm b then the expected value would be seven. The observed value is the number of pairs where algorithm a is preferred over algorithm b by the observer. Chi-square is the suitable measure of the "goodness to fit" between the observed and expected values.

The chi-square test is used here to attempt to reject the null hypothesis that the observed and the expected data won't fit or in other words are independent.

With the expected (e) and the observed (o) values known, the Chi-square is calculated as the sum of the squared difference between:

$$\chi^2 = \frac{(o-e)^2}{e}. \tag{7.1}$$

It can be seen from the above calculation that it is intuitive to conclude that a large difference between the observed and expected values will result in accepting the null hypothesis which the independency of the two. If the observers agree with the results by reproduction or recovery angular error, then the difference should be small and the null hypothesis will be rejected. Clearly if χ^2 is zero then the expected and observed values are exactly the same and we can immediately reject the null hypothesis. In general, for small χ^2 we can reject the null hypothesis for some criterion amount we will not be able to reject the null hypothesis. This criterion amount is found by consulting the statistical tables.

Formally, to be able to accept or reject the null hypothesis, the calculated Chi-square value in Eq. 7.1 should be compared against the critical chi-square value in the corresponding table (e.g. [Conover, 1999]). The critical value is decided from the table of chi-square for a desired significance level (e.g. 5% or 0.05). If calculated chi-square value is greater than the critical chi-square value the null hypothesis is accepted and the observed and expected data will not fit. Otherwise, the null hypothesis is rejected and observers agree with the expected data. We have eight observers in our experiment, so the number of samples (observations) is eight. The critical chi-square value for seven degree of freedom with $p = 0.05$ is 14.07

In Table 7.1, the Chi-square values for the goodness of fitness between the observers' data and the expected values by reproduction angular error can be seen. Here, each cell of table contains two values $(x,\ y)$, where i represents the number of pairs of images for which algorithm a performs better than algorithm b according to reproduction angular error. The value j is the observers data, which shows the number of comparisons in which the observer has preferred algorithm a over algorithm b. For instance in the column indicated by GE1-GE2, the 1^{st} grey edge algorithm is compared against the 2^{nd} grey edge algorithm. Based on observer 1 for all seven pairs of images GE1 is better than GE2, or observer 3 has agreed with the GE1 superiority over GE2 only for four out of seven pairs. The expected number of pairs where algorithm a performs better than b according to the reproduction angular error is for all the seven pairs. However, we found that some observers were not consistent with their choices when they were shown the

Table 7.1: Chi-square values for comparing the results by the observers and the reproduction angular error. Each (x, y) represents (reproduction error, observer's data). (x, y) denotes there are x pairs for which the observer made a consistent judgement and for y ($y <= x$) of these pairs the observer agreed with the error metric.

	GE1-GE2	GE1-SOG	GE1-GP	GE2-SOG	GE2-GP	SOG-GP
observer 1	(7,7)	(6,5)	(7,4)	(7,7)	(7,6)	(6,4)
observer 2	(3,2)	(4,1)	(7,7)	(7,6)	(6,5)	(7,6)
observer 3	(7,4)	(7,5)	(7,6)	(7,6)	(7,7)	(7,7)
observer 4	(5,4)	(5,4)	(6,5)	(5,5)	(4,4)	(5,4)
observer 5	(6,4)	(7,7)	(7,6)	(7,5)	(6,6)	(7,6)
observer 6	(7,4)	(6,5)	(5,5)	(5,5)	(6,5)	(3,2)
observer 7	(6,4)	(6,6)	(6,4)	(7,7)	(7,6)	(7,6)
observer 8	(4,2)	(5,4)	(3,3)	(6,4)	(7,7)	(6,4)
Chi-square	**5.44**	**3.55**	**2.40**	**1.52**	**0.62**	**2.30**

same pair for the second time. If that was the case, we excluded that pair from the calculation of Chi-square for that specific observer. An example of such occurrence can be seen for observer 2, who has been consistent with his choices only for three pairs when comparing GE1 and GE2 algorithms. Since we are comparing four algorithms: GE1, GE2, SOG and GP, there are six columns of data which is the number of possible combinations of two out of four algorithms.

A comparison between the critical chi-square value for eight observers (which is 14.07 with the significance of $p = 0.05$) and the ones calculated in Table 7.1 shows there is no reason to reject the null hypothesis that the observed and expected values match. In other words, the observers agree with the prediction of the quality of the reproduced images by reproduction angular error.

Table 7.2 reports the same result but for comparison of the observers' data with the results by recovery angular error. Notice that the name of the algorithms in this table is switched, i.e. GE1-GE2 in Table 7.1 has changed to GE2-GE1 in Table 7.2. The high values of Chi-square in Table 7.2 for all six pairs of algorithms reject the null hypothesis that the observers data match recovery angular error's prediction.

Table 7.2: Chi-square values for comparing the results by the observers and the recovery angular error. Each (x, y) represents (recovery error, observer's data). (x, y) denotes there are x pairs for which the observer made a consistent judgement and for y $(y <= x)$ of these pairs the observer agreed with the error metric.

	GE2-GE1	SOG-GE1	GP-GE1	SOG-GE2	GP-GE2	GP-SOG
observer 1	(7,0)	(6,1)	(7,3)	(7,0)	(7,1)	(6,2)
observer 2	(3,1)	(4,3)	(7,0)	(7,1)	(6,1)	(7,1)
observer 3	(7,3)	(7,2)	(7,1)	(7,1)	(7,0)	(7,0)
observer 4	(5,1)	(5,1)	(6,1)	(5,0)	(4,0)	(5,1)
observer 5	(6,2)	(7,0)	(7,1)	(7,2)	(6,0)	(7,1)
observer 6	(7,3)	(6,1)	(5,0)	(5,0)	(6,1)	(3,1)
observer 7	(6,2)	(6,0)	(6,2)	(7,0)	(7,1)	(7,1)
observer 8	(4,2)	(5,1)	(3,0)	(6,2)	(7,0)	(6,2)
Chi-square	**22.44**	**31.55**	**34.40**	**40.52**	**42.62**	**32.30**

To analyse whether there is an agreement between the observers [Gijsenij et al., 2009a; Alfvin et al., 1997] the individual difference from the mean of observations have been calculated. For each observer, the correlation coefficient of x/y ratio by which the observer has agreed that algorithm a is better than algorithm b with the average of the same ratio for all the observers is computed. For all the pairs in Table 7.1 and all the eight observers the correlation coefficients calculated vary from 0.7 to 0.9 with an average of 0.8. Also, the correlation coefficient between the x/y ratios for the individual observers range from 0.6 to 0.9. The highest agreement between the observers was for the GE2-GP pair of algorithms and the lowest correlation was for the GE1-GE2 pair. This is expected as the 1^{st} and 2^{nd} order grey edge algorithms (GE1-GE2) are instances of the same algorithm and their performances are very close in many cases which makes the choice difficult for the observers.

7.4 Conclusion

Evaluation of illuminant estimation algorithms using the reproduction and recovery angular errors (in Chapter 4) shown there are sometimes disagreements between the two metrics regarding the ranking of a pair of algorithms. In this chapter, a psychophysical study was conducted to investigate with which of the two error metrics predicted the image preference judgements made by human observers.

The results of the experiments shows that in most cases the observers agree with the evaluation by reproduction angular error. In other words, where according to reproduction angular error algorithm a is performing better than b, in most cases the observers make the same choice. Although, there are cases where the observers disagree with the reproduction angular error's evaluation. However, the overall statistical analysis of the results using the Chi-square test shows the observers data highly agree with the results by reproduction angular error.

Perceptual analysis of images in terms of accuracy of reproduced colours is a difficult task since it could depend on many factors other than the accuracy of colours, such as content, etc. In digital photography the aim is not always reproducing the colours which are colourimetrically accurate but a reproduction of preference is sometimes more desired. To this end, in the experiment performed in this chapter, we also aimed to create a more photographic look for the raw images by passing them through an actual camera pipeline. This will provide the observers in the experiment with more natural photographic-look images and makes the task of comparison easier for them. To our knowledge, this is the first time in a psychophysics experiment concerning the quality of the colour corrected images that the images are rendered to a photographic look before the experiment.

Chapter 8

A Hybrid Strategy for Illuminant Estimation Targeting Hard Images

We notice that the largest switches between reproduction and recovery angular error were not for the mean and the median errors but rather for the max and 95% quantile errors. This is particularly important result. In general for computer vision and computational photography applications, the illuminant estimation algorithms in cameras work well. When they don't work, the failure cases that we notice, are for the images with the high recovery and reproduction errors. Specifically, the failure cases are for 95% and max errors. It is precisely these images that modern day illuminant estimation algorithms seek to solve the problem for and specifically for these images that we find that the ranking of algorithms change remarkably when reproduction angular error is compared with recovery angular error. This motivates us to study these images not by looking at the overall error but the individual errors to see if there exist 'hard' images that are challenging for multiple methods. Our findings indicate that there are certain images that are difficult for fast statistical-based methods, but that can be handled with more complex learning-based approaches at a significant cost in time-complexity. This has led us to design a hybrid method [Zakizadeh et al., 2015] that first classifies an image as 'hard' or 'easy' and then uses the slower method when needed, thus providing a balance between time-complexity and performance. In addition, we have identified dataset images that almost no method is able to process. We argue, however, that these images have problems with how the ground truth is established and recommend their removal from future performance evaluation.

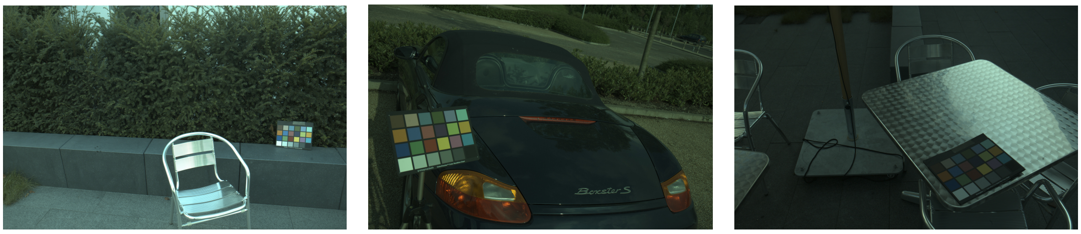
8.1 Introduction

In Chapter 2, we provided an overview of several illuminant estimation algorithms. We also classified these algorithms into different categories. However, concerning the complexity of the methods they can be roughly classified into two types: statistical-based methods and learning-based techniques. As mentioned in Chapter 2, statistical-based methods (e.g. [Land et al., 1977; Buchsbaum, 1980; Van De Weijer et al., 2007a; Finlayson and Trezzi, 2004; Gijsenij et al., 2012; Cheng et al., 2014]) directly estimate the illumination from statistics computed from the input image. These methods are fast and work irrespective of the type of camera used. Their performance, however, is generally not as good as learning-based methods. Learning-based methods (e.g. [Forsyth, 1990; Finlayson et al., 2001; Gijsenij and Gevers, 2007; Gehler et al., 2008; Gijsenij et al., 2010; Chakrabarti et al., 2012; Finlayson, 2013; Joze and Drew, 2012]) exploit the availability of training images that have labelled ground truth illumination. Learning-based methods generally give superior results over statistical methods, but at the cost of higher running-times and the need to be trained per camera. The selection of an illumination estimation method is generally guided by the need for performance vs. time-complexity, e.g. most onboard camera white-balance algorithms still use statistical-based methods.

The methods of performance evaluation of illuminant estimation algorithms and how the errors are reported for a benchmark dataset were covered in Chapter 3. We mentioned how different aggregate performance errors, such as mean, median, trimean and quantiles, are given over the whole dataset. The routine reporting of these statistics provides some insight to a method's performance across an entire dataset. Interestingly, however, none of the prior works have examined if there is any commonality in these statistics across the images in the dataset. For example, it is unclear if the bottom 25% results have shared images across different methods. This would be interesting finding as it would indicate the existence of images that multiple methods consistency perform poorly on. We term these images as 'hard images'. This lack of analysis serves as the impetus for the work in this chapter.

In this chapter, we describe an analysis on 12 leading illumination estimation algorithms belonging to both statistical- and learning-based methods. In particular, we enumerate over all combinations of five methods out of 12 to find the set of images where at least the majority (three or more) methods fail. We consider these images to be 'hard' for this subset of methods. Our findings indicate that there are, indeed, sets of hard images for different subsets (e.g. see Figure 8.1). More importantly, these subsets can be grouped depending if their methods belong to statistical-based or learning-based. To this end, we found that there are a number of 'hard' images for the fast statistical-based methods that can be handled by more complex learning-based approaches . This led us to develop a hybrid estimation approach that classifies the image as hard or easy depending on the results of the statistical-based methods. In the case an image is categorised as hard, it is likely that the results of the simple camera on-board white balancing algorithms are incorrect. Such hard images can be saved as raw on the camera for later off-line processing by slower, but more accurate, learning-based methods, such as the exemplar-based method [Joze and Drew, 2012]. This leads to better overall illumination estimation performance while reducing the overall time-complexity. Our analysis also has found that certain images in a well established benchmark dataset are hard for all methods. On closer examination we found that these images have issues that makes establishing the ground truth difficult and advocate for their removal for future evaluation.

The chapter is organised as follows: In Section 8.2, we analyse the estimates by statistical-based algorithms on Gehler-Shi dataset to see if there exist set of images where the algorithms perform poorly on ('Hard' images). In Section 8.3, we introduce the hybrid strategy for detecting hard images. Section 8.4 explains the experiments and results. The removal of certain images which we think should be



hard image hard image easy image

FIGURE 8.1: Examples of images from the Gehler-Shi dataset [Gehler et al., 2008; Shi and Funt, 2010] considered hard and easy based on our analysis of the performance of 12 different methods on the entire dataset.

removed from Gehler-Shi dataset is discussed in Section 8.5. Section 8.6 concludes the chapter.

8.2 Analysing Estimates by Algorithms on a Common Dataset

Gijsenij et al. [Gijsenij et al., 2011] performed a thorough evaluation of 15 methods on the Gehler-Shi dataset [Gehler et al., 2008; Shi and Funt, 2010]. Their work provided results for each of these 15 methods for each image in dataset. We use this comprehensive results for our analysis in this chapter.

From Gijsenij et al. [Gijsenij et al., 2011], we select 12 algorithms that have received the greatest attention in the published literature. We divide them into two groups. *Statistical-based methods* including: S1 = shades of grey [Finlayson and Trezzi, 2004], S2 = grey world [Buchsbaum, 1980], S3 = 1^{st} order grey edge [Van De Weijer et al., 2007a], S4 = 2^{nd} order grey edge and S5 = white-patch [Land et al., 1977]. *Learning-based methods* including: L1 = exemplar-based [Joze and Drew, 2012], L2 = color constancy using natural image statistics [Gijsenij and Gevers, 2007], L3 = edge-based gamut, L4 = pixel-based gamut, L5 = intersection-based gamut [Forsyth, 1990; Gijsenij et al., 2010], L6 = Bayesian method [Gehler et al., 2008] and L7 = spatial correlation [Chakrabarti et al., 2012].

As previously mentioned, Gijsenij et al. [Gijsenij et al., 2011] provides the complete results (estimated illumination) by each 12 method for each image in the Gehler-Shi dataset. This dataset contains a total of 568 images involving two cameras, a Canon 1D (86 images) and a Canon 5D (482 images). Because the learning-based methods are trained per-camera, we focus on the Canon 5D given that it has the most images. This gives us a total of 482 images with 12 results, corresponding to the associated methods S1-5 and L1-7.

Our analysis is intended to find images that are collectively hard for multiple methods. In this case, 'hard' images are those where multiple methods are unable to estimate the illumination within some error threshold. In this chapter, we use nine degrees error as this threshold, meaning that the estimated illumination has at least nine degrees (or more) angular difference from the ground truth illumination. Nine degrees is used as it represents a threshold that categorises typical error of the bottom 25% for most methods as reported by Gijsenij et al. [Gijsenij et al., 2011]. Thus, we are comparing the images that are reported to give the worse performances for the 12 methods.

When we examine which images in the dataset that have at least nine degrees of error for all 12 methods, we found there are only a few images (this finding is discussed in more detail in Section 8.5). This means that there is significant variation in the images that different methods perform poorly on. To provide a more manageable grouping, we consider all combinations of 5 methods from the 12 total (i.e. 12 choose 5). In particular, we enumerate all five combinations of the 12 methods which gives total of 792 combinations. Among these combinations, we are interested in those for which at least three out of five methods introduce errors higher than our threshold. This is illustrated in Figure B.1 which shows one out of the 792 combinations. The columns in Figure B.1 represent a unique image in the dataset. The rows represent the five different methods tested. A white-box means a method has failed for this particular image (i.e. produces a high error). A black-box means the method is successful. Three or more empty boxes for a

Combinations with **most** 'hard' images			Combinations with **least** 'hard' images		
Methods	failed images	Time (m)	Methods	failed images	Time (m)
S2 S3 S5 L3 L7	84	1.5	L1 L2 L4 L6 L7	31	12.6
S2 S3 S5 L3 L6	80	9.8	S4 L1 L2 L5 L7	27	3.7
S2 S3 S5 L3 L4	78	1.8	S1 S4 L1 L2 L6	24	11.2
S2 S3 S4 S5 L3	73	1	S2 L1 L2 L6 L7	22	11.7
S1 S2 S3 S4 S5	69	0.36	S2 S4 L1 L2 L7	19	2.87
S1 S2 S4 S5 L4	64	1.2	S2 S4 L1 L6 L7	18	11.1

Table 8.1: The five combinations out of 12 illuminant estimation algorithms in terms of number of images they fail for. We have highlighted the fastest (on the left) and slowest (on the right) combinations. Running time given are per image.

particular column represents an image where the majority of methods has failed. This is considered a 'hard' image for this particular combination of methods. For the example shown, the combination are methods (S1, S4, L1, L2, L6), and this set results in 24 hard images.

This procedure is performed for all combinations of 5 methods out of the 12. For each combination, we record the number of hard images per combination and sort the list of combinations based on the number of hard images. Table 8.1 includes the combinations with **most** and **least** 'hard' images. Almost all combinations with **most** 'hard' images include three or more simple statistical-based algorithms. The combinations with **least** 'hard' images are mostly dominated by learning-based methods.

Each method examined has a time complexity associated with it. The work by Gijsenij et al. [Gijsenij et al., 2011] did not report this time-complexity, however, more recent work has examined most of the same methods and reported the running-time [Cheng et al., 2014]. The only exception is that of the exemplar-based method (L1). For this method, we estimate its time to take approximately twice that of the gamut-based methods based on the running-time reported by the author [Joze, 2013]. The fastest and slowest combinations are highlighted in

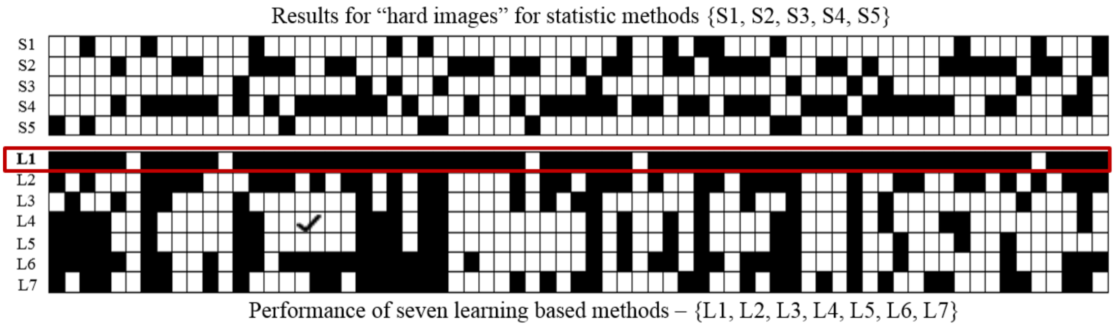


Figure 8.2: The hard images from the Gehler-Shi dataset [Gehler et al., 2008; Shi and Funt, 2010] for the five statistical-based methods. L1-L7 rows are the performance of the learning-based methods.

Table 8.1. The statistical-based methods in general have a much faster running-time than the learning-based methods. For example, the highest number of hard images is 84 that is achieved using the combination in the first row of section 'combinations with **most** hard images' in Table 8.1. This set of methods requires roughly 1.5 minute per image to run all 5 methods. The overall run-time is mainly attributed to the two learning-based techniques: (L3) edge-based gamut [Gijsenij et al., 2012] and (L7) spatio spectral [Chakrabarti et al., 2012].

The fifth largest number of failure images (out of 792 combinations) is for the set of the five statistical methods (S1-S5). This is highlighted on the left in Table 8.1. This only requires approximately 0.36 minutes per image and is the fastest of all the combinations. This is a very interesting finding. It shows that the statistical-based methods tend to collectively fail on the same images in the dataset. This means that we have a chance to examine these images to see if we can build a classifier that can predict if an image is 'hard' or 'easy' for this set of methods. The question now is can we find a method that performs well on the hard images for the statistical-based approaches.

Given the combination of five statistical-based methods and their associated hard images, we examine the performance of the learning-based methods. Figure 8.2 shows the results. The diagram shows all 69 of the hard images (where at least three or more of the learning-based methods fail). The rows below show the results

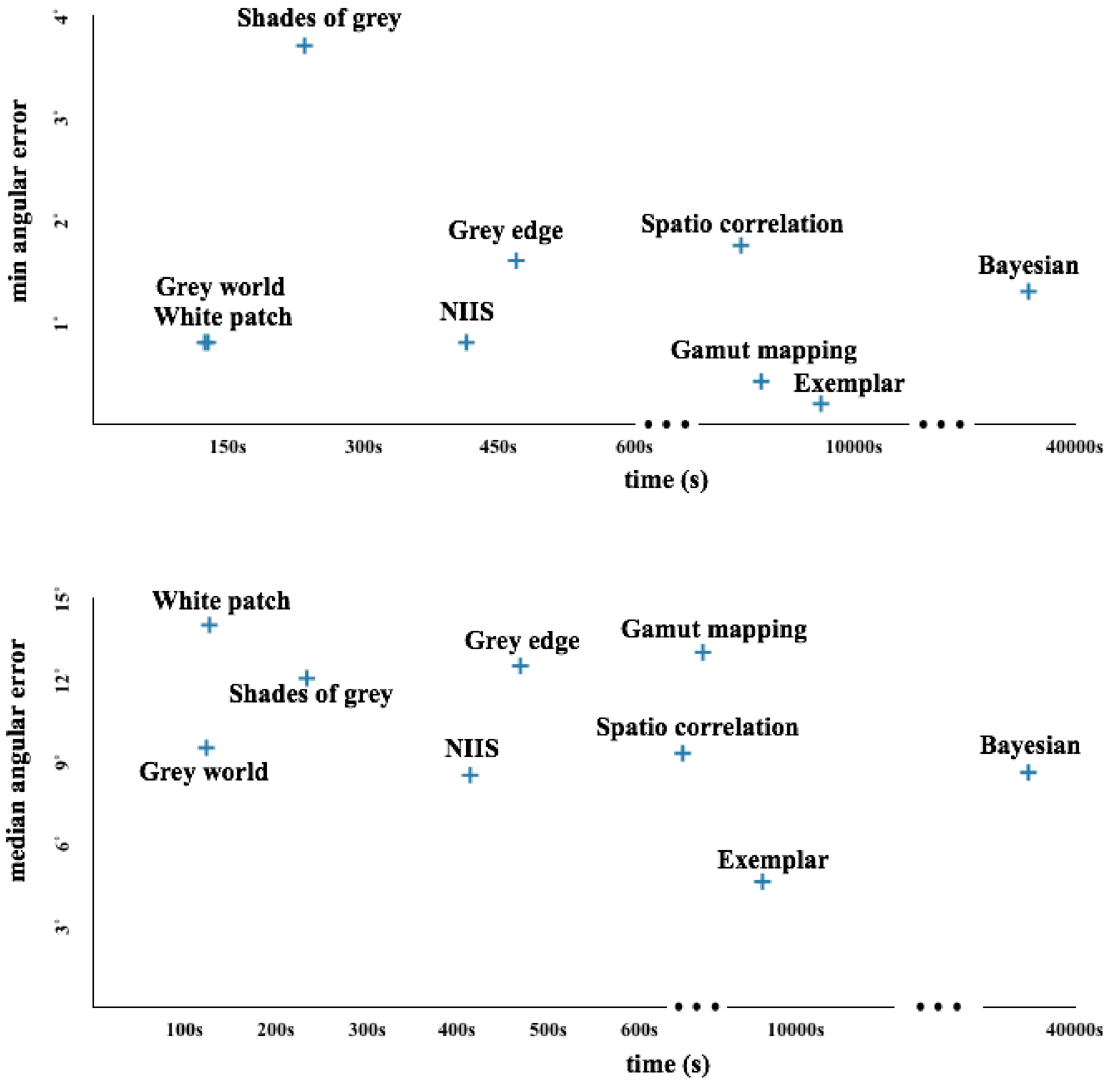


FIGURE 8.3: Computational time vs. performance of illumination estimation methods. The top plot shows the minimum angular error vs. computational time for the 69 hard images in Figure 8.2 and the bottom plot shows the median error vs. computational time for the same images. Although some fast algorithms such as white patch or grey world have low minimum angular errors but their medium error is very high. Among learning-based methods which are slower, exemplar-based has the lowest minimum and median angular error.

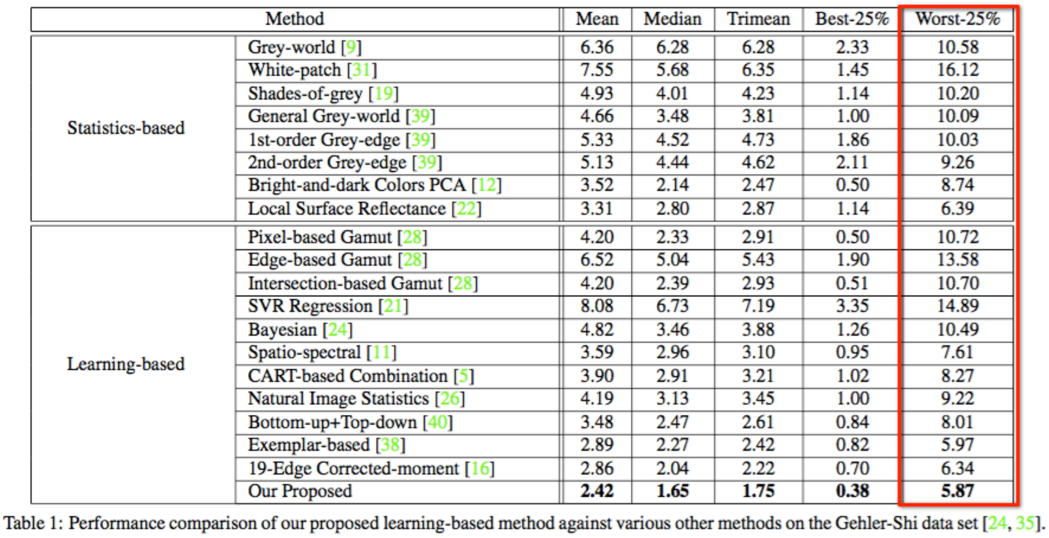
of the L1-L7 learning-based methods. It is interesting to note that there are some images considered hard for the statistical-based method that all learning-based method are successful on. Overall, however, the L1 (exemplar-based [Joze and Drew, 2012]) method does particularly well for the hard images, able to produce a better result on all except a few of the images.

Figure 8.3 shows the error vs. computational time for the 69 hard images in Figure 8.2. In the top plot the minimum angular error for the algorithms versus the algorithms' computational time for all the images is shown. The bottom plot shows the median errors for the same algorithms and the same images. Although some statistical based methods (which are fast) such as white patch algorithm have low minimum angular errors but their median error is very high. Exemplar-based algorithm has a very low minimum angular error (0.2°) and its median error is still lower than all the methods. Of course learning-based methods are time consuming and their usage is limited to a more powerful computational system.

Based on the analysis in this section, we have developed a hybrid method that first applies the statistical based approaches. As discussed in the next section, from this we can classify if the image is hard or easy. For images that are classified as hard, we propose that they are saved as raw (on the camera) for later to be processed off-line by learning based methods such as the exemplar-based (L1).

8.3 Hybrid Method for Targeting Hard Images

In this section, we describe our framework to classify images as hard or easy and then process them accordingly. As discussed in Section 8.2, an image is labelled as hard if at least three out of five simple statistical-based algorithms have an error beyond nine degrees. Nine degrees of angular error can be a reasonable threshold for an image to be considered hard. This can be derived by looking at most of the 25% worst performance errors reported for several illuminant estimation



Method		Mean	Median	Trimean	Best-25%	Worst-25%
Statistics-based	Grey-world [9]	6.36	6.28	6.28	2.33	10.58
	White-patch [31]	7.55	5.68	6.35	1.45	16.12
	Shades-of-grey [19]	4.93	4.01	4.23	1.14	10.20
	General Grey-world [39]	4.66	3.48	3.81	1.00	10.09
	1st-order Grey-edge [39]	5.33	4.52	4.73	1.86	10.03
	2nd-order Grey-edge [39]	5.13	4.44	4.62	2.11	9.26
	Bright-and-dark Colors PCA [12]	3.52	2.14	2.47	0.50	8.74
	Local Surface Reflectance [22]	3.31	2.80	2.87	1.14	6.39
Learning-based	Pixel-based Gamut [28]	4.20	2.33	2.91	0.50	10.72
	Edge-based Gamut [28]	6.52	5.04	5.43	1.90	13.58
	Intersection-based Gamut [28]	4.20	2.39	2.93	0.51	10.70
	SVR Regression [21]	8.08	6.73	7.19	3.35	14.89
	Bayesian [24]	4.82	3.46	3.88	1.26	10.49
	Spatio-spectral [11]	3.59	2.96	3.10	0.95	7.61
	CART-based Combination [5]	3.90	2.91	3.21	1.02	8.27
	Natural Image Statistics [26]	4.19	3.13	3.45	1.00	9.22
	Bottom-up+Top-down [40]	3.48	2.47	2.61	0.84	8.01
	Exemplar-based [38]	2.89	2.27	2.42	0.82	5.97
	19-Edge Corrected-moment [16]	2.86	2.04	2.22	0.70	6.34
	Our Proposed	**2.42**	**1.65**	**1.75**	**0.38**	**5.87**

Table 1: Performance comparison of our proposed learning-based method against various other methods on the Gehler-Shi data set [24, 35].

Figure 8.4: An example of 25% worst errors reported in the literature (table from [Gehler et al., 2008; Shi and Funt, 2010])

algorithms on a set of real images like Gehler-Shi dataset [Gehler et al., 2008; Shi and Funt, 2010]. An example of such an evaluation table is shown in Figure 8.4. In the table take from [Cheng et al., 2015a], we can see that the average of the 25% worst errors (the highlighted column) of all the illuminant estimation algorithms used in this evaluation is around nine. We label an image as easy if all five methods succeed, i.e have an error below the threshold. We set the threshold for easy images as eight degrees which is slightly lower than the hard images threshold. We use these labelled images as training data to build a classifier.

8.3.1 Features and Classifier

We have experimented with several image features to be used in designing a classifier to label a new input image as either hard or easy. One feature commonly used in learning-based colour constancy methods is the rg chromaticity values ($[r, g] = [R, G]/(R + G + B)$). These are typically used to compute a histogram over the r and g values as features. We found, however, that the distribution of the rg values had little correlation to image being labelled hard or easy. We also examined the chromaticity values with respect to the rg chromaticity curve of the ground-truth illuminants (i.e. the locus of ground-truth illuminants in chromaticity space). Again, we found that these had little correlation to whether an image was labelled as hard or easy.

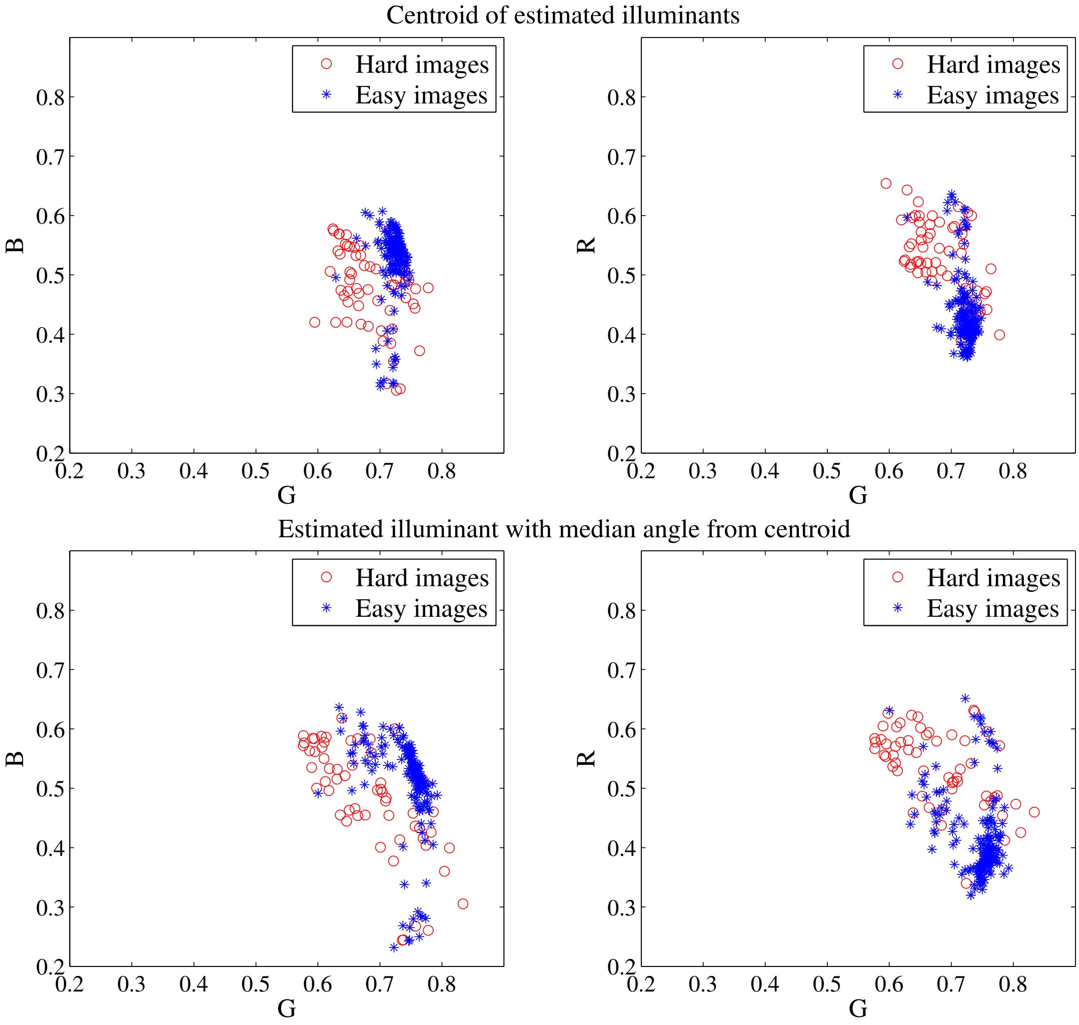


FIGURE 8.5: The top two diagrams show the centroids of five estimated illuminants for hard and easy images. The bottom two diagrams are the selected illuminants out of five estimations with median angle from the centroid. The features for easy images form a cluster in both cases.

The lack of success with chromaticity values led us to examine features defined in the full 3D RGB space. In particular, we looked at the mean (centroid) location of the five estimated illuminants provided by the statistical methods (S1-S5). Figure 8.5 (top) shows the distribution of these centroid of the estimated illuminants for a set of hard (red) and easy (blue) images from Gehler-Shi dataset. We can see that these form two distinct clusters of points. We also calculated the angle between each of five estimated illuminants and the centroid of the estimates. Among five estimates we selected the one with the median angle from the centroid. The points in the last two plots of Figure 8.5 (bottom) belong to the selected estimated illuminants. While there is a discernable pattern in the data, it is not as distinguishable as that with the the clusters of centroid.

Based on the observation in Figure 8.5, we experimented with classifiers using five different features: 1) The centroid of the five estimated illuminants; 2) the estimated illuminant selected out of the five estimates with the median angle from the centroid; 3) feature 1 and the standard deviation of the five estimated illuminants; 4) feature 2 and the standard deviation of the five estimated illuminants; and 5) the standard deviation of the five estimated illuminants. The features were used to train a support vector machine (SVM) [Cortes and Vapnik, 1995] classifier based on the implementation of Chang and Lin [Chang and Lin, 2011].

Table 8.2 shows the overall accuracy of the SVM classifier with all the features as well as how accurate the model classifies hard and easy images. We found that the simple centroid feature produced the best results over all the five features and use it in our overall framework.

Feature	Overall accuracy	Hard image accuracy	Easy image accuracy
1. Centroid	93.6%	85%	96.6%
2. Median from centroid	86.7%	68.3%	94.3%
3. Standard deviation (std)	82%	42.3%	95.9%
4. Centroid + std	89.7%	68.1%	95.9%
5. Median + std	85.4%	59.2%	94.7%

Table 8.2: Performance of the SVM classifier with different features.

Figure 8.6 shows the receiver operating characteristic curve which plots the true positive rate ($TPR = \frac{TP}{TP+FN}$, where TP is true positive and FN is false negative) against the false positive rate ($FPR = \frac{FP}{FP+TN}$, where FP is false positive and TN is true nagative) at various threshold settings. In this case, the curve shows the trade-off between the accuracy of the classifier in classifying hard images (rate of images correctly classified as hard) versus the classifier's error (probability of an easy image being returned as a hard). The value one on the axis represents 100% accuracy, 0.5 means 50% and so on.

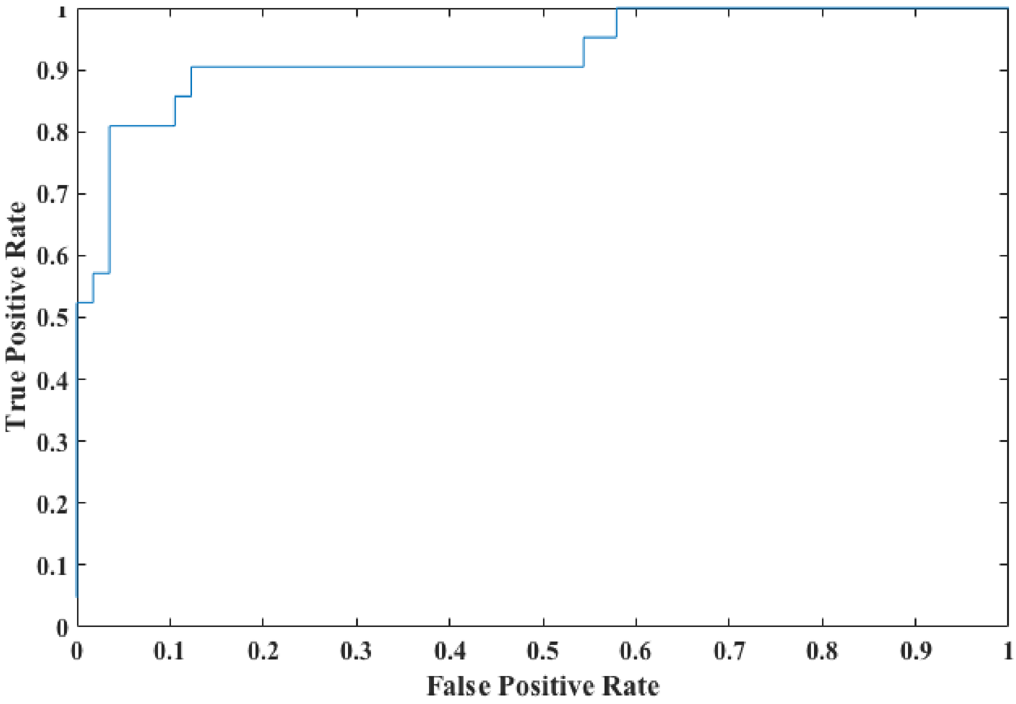


Figure 8.6: The receiver operating characteristic curve which shows the trade-off between the rate of hard images being truly classified as hard and the rate of images being falsely classified as hard.

8.3.2 Overall Procedure

The overall framework of our hybrid strategy can be seen in Figure 8.9. For a given input image, its illumination is estimated by the five statistical-based methods (S1-S5). The centroid (mean) of the five estimates is calculated and used with the SVM to predict if the image is hard or easy. If the image is classified as hard, we use a learning-based method such as the exemplar-based method [Joze and Drew, 2012] to process the image to obtain the final illumination estimate.

If the image is classified as easy, we have five estimates to choose from. A straightforward option would be to use the average of these estimations. This is reported in our experiments in the next section. However, another option is to use this information to get a better prediction of the illuminant. In particular, the recent 'corrected moments' work by Finlayson [Finlayson, 2013] showed that a correction matrix can be pre-computed using the ground-truth illuminants from the training-data to correct the estimates of the existing simple derivative-based statistical methods. In this case, we can use the result of the two derivative-based methods S3 and S4 (1^{st} grey edge and 2^{nd} grey edge) to build the correction matrix. We found this approach gives notably better results over using the average of the S1-S5 scores. This is also reported in the experiments in the following section.

8.4 Experiments and Results

We have tested our hybrid strategy on the Gehler-shi [Gehler et al., 2008; Shi and Funt, 2010] dataset using different features mentioned in Section 8.3. To generate a set of labelled data we categorise hard and easy image based on their thresholds (here we set eight degrees for easy images and nine degrees for hard images as explained in Section 8.2). Out of 482 images of *Canon 5D* from *Gehler-shi* [Gehler et al., 2008; Shi and Funt, 2010] dataset, this results in 233 labelled images. The sets of training and test images are made by 3-fold cross validation, i.e. each fold has 155 training and 78 test images. The SVM classifier based on the 'centroid' feature is built on the training set and the accuracy of it is examined on the test images. The model's performance on this set of 78 test images showed an accuracy of 93.6% with 85% for classifying hard images and 96.6% for classifying easy images. Table 8.3 shows the result of the model applied on a set of unlabelled images.

The performance of the five statistical methods (S1-S5) for all images are shown in the first row of the Table 8.3. The L1 column shows the error of exemplar-based method for all images. The exemplar-based has an overall good performance for all images but is significantly slower than the S1-S5 methods combined.

In our hybrid algorithm, we use our SVM to classify the input images. In the first column of the proposed section of Table 8.3 the average of statistical-based methods is used as our estimate. By excluding hard images we have avoided the high error of S1-S5 that is obtained when applied to all images. It is interesting to note that median of the average of S1-S5 is less than the median of the individual methods. As previously mentioned, we also use the corrected-moment illuminant estimation method [Finlayson, 2013] to further improve the results. This method uses a cross validation procedure to build a correction matrix that takes the results from the S3 and S4 estimates and refine the result based on the ground-truth illuminants of training data. Table 8.3 shows the (corrected) algorithm performance. This allows us to get an additional gain on the performance of the statistical based methods. Note that the approach in [Finlayson, 2013] still has trouble on the hard images and the use of the exemplar-based method is significantly better and therefore necessary for the hard images.

Our results show that this strategy of using fast statistical-based methods can give us good performance on the easy images, while identifying the difficult images and passing them to a slower, but more accurate learning-based approach. While the overall running-time is slow due to the use of the learning-based method, our approach can reduce this by almost half while giving similar performance. Moreover, the results for easy images can be obtained in a matter of seconds.

The list of images from Gehler-Shi dataset classified as hard are given in Appendix C.

	S1	S2	S3	S4	S5	L1	Proposed	
							average	corrected
All	4.37°	7.04°	4.81°	4.73°	6.46°	2.4°		
Easy	3.5°	6.9°	4.26°	4.7°	4.7°	2.1°	3.42°	**2.4°**
Hard	6°	7.04°	6.1°	4.8°	12.9°	2.91°	2.91°	**2.91°**
Time (per image)	$3.4s$	$1.8s$	$6.8s$	$8s$	$1.85s$	$1.96m$	$21.9s$ + ($1.96m$ per hard image)	
Time (total)	$18.5m$	$9.8m$	$36.9m$	$43.5m$	$10m$	$10.7h$	$4.5h$	

Table 8.3: The median errors of the proposed hybrid framework treating hard and easy images differently. In comparison we show the errors of fast statistical algorithms (S1 to S5), as well as time complexity of exemplar-based method [Joze and Drew, 2012] (L1).

8.5 Removal of False Hard Images

As mentioned in Section 8.2 we found nine images that almost all methods failed on from the Gehler- Shi [Gehler et al., 2008; Shi and Funt, 2010] dataset. We were keen to see if there were some characteristics to the hard images that no method could resolve, however, on careful inspection of these images we realise it was due to the position of the colour chart in the scene. Figure 8.7 shows the removed images.

In all of these images, the colour checker board that is used to provide the ground truth illumination is placed under a different illumination than the rest of the scene. This means the scene is lit by two different illuminations, but in the cases of these nine images, the dominant illumination arguable does not fall on the colour checker board. These images do not represent fair test cases and should be removed as they introduce negative results for evaluation and are erroneously used by learning-based methods for training. We have provided an updated version of the Gehler-Shi [Gehler et al., 2008; Shi and Funt, 2010] which excludes these nine images and their measured ground truth illuminations[1].

[1] http://colour.cmp.uea.ac.uk/datasets/GehlerFalse.html

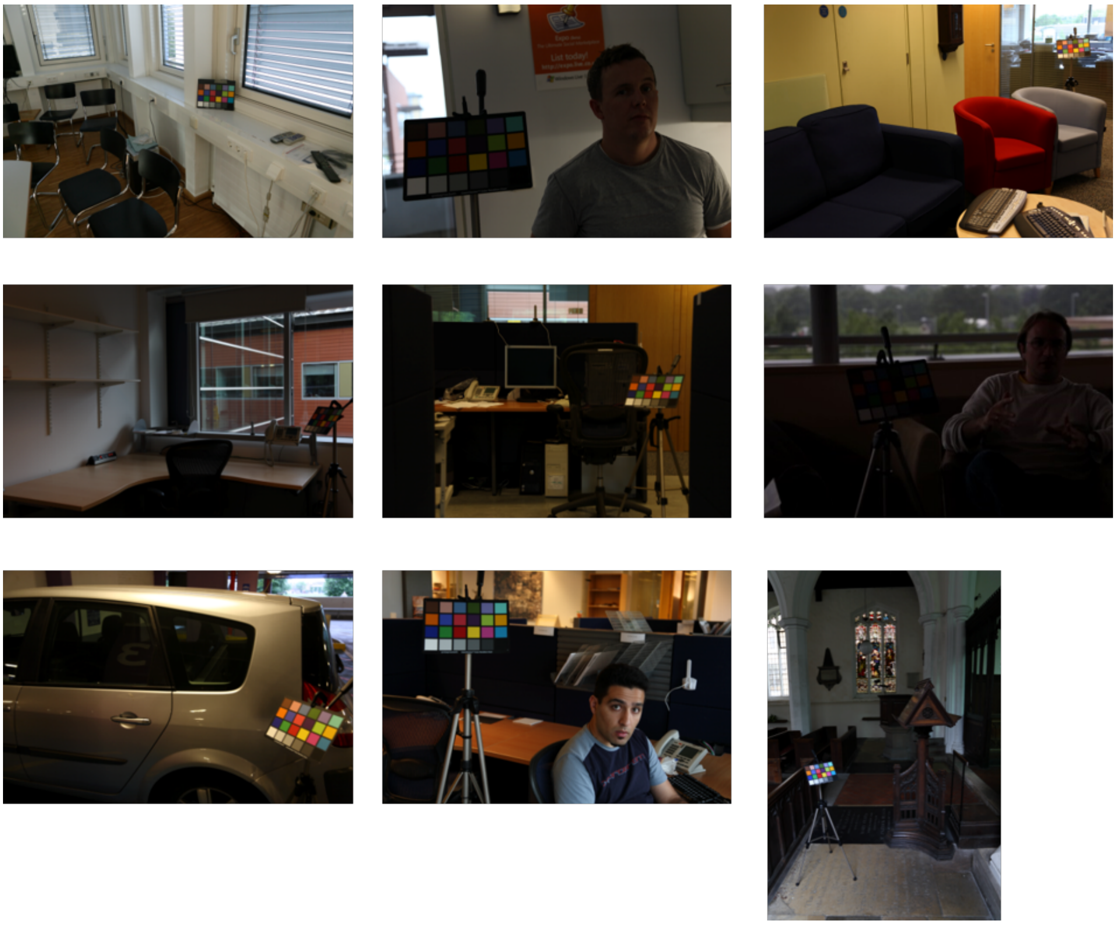


FIGURE 8.7: Images that all methods incorrectly estimate the illumination on from the Gehler-Shi dataset [Gehler et al., 2008; Shi and Funt, 2010].

8.6 Conclusions

This chapter has analysed the performance of multiple colour constancy methods to examine if methods fail on the same images. As far as we are aware, this is the first work to examine the relations of the hard images across different colour constancy methods.

Our analysis revealed that there are common 'hard' images for subsets of methods. One of these subsets with a large number of hard images is composed of all fast statistical-based colour constancy methods. We also observed that there exist some learning-based methods that give excellent performance on this set of hard images, but at a significant cost in running-time. Based on these observations, we proposed a hybrid method that classifies an image as hard or easy and then processes it accordingly. This allows easy images to be processed quickly. Easy images white-balancing could even be performed onboard the camera itself. For the images classified as hard, learning-based methods such as the exemplar-based method [Joze and Drew, 2012] are applied to give good results. We note that learning-based methods will continue to improve in terms of performance and speed. Recent work by Bianco et al. [Bianco et al., 2015] and Cheng et al. [Cheng et al., 2015c] provided similar estimation performance to the exemplar-based method (L1) used in our work, but at a significantly faster running-times. These methods can be easily incorporated into our overall framework's running time, however, we note that learning-based methods will still need to be performed off-line and therefore require the determination of which images are 'hard' and require such off-line processing.

Our analysis has also identified nine images in the widely used Gehler-Shi [Gehler et al., 2008; Shi and Funt, 2010] dataset that were problematic for all 12 methods we examined. We have found that these images have problems with how the ground-truth is established and we recommend their removal from the dataset for the future studies.

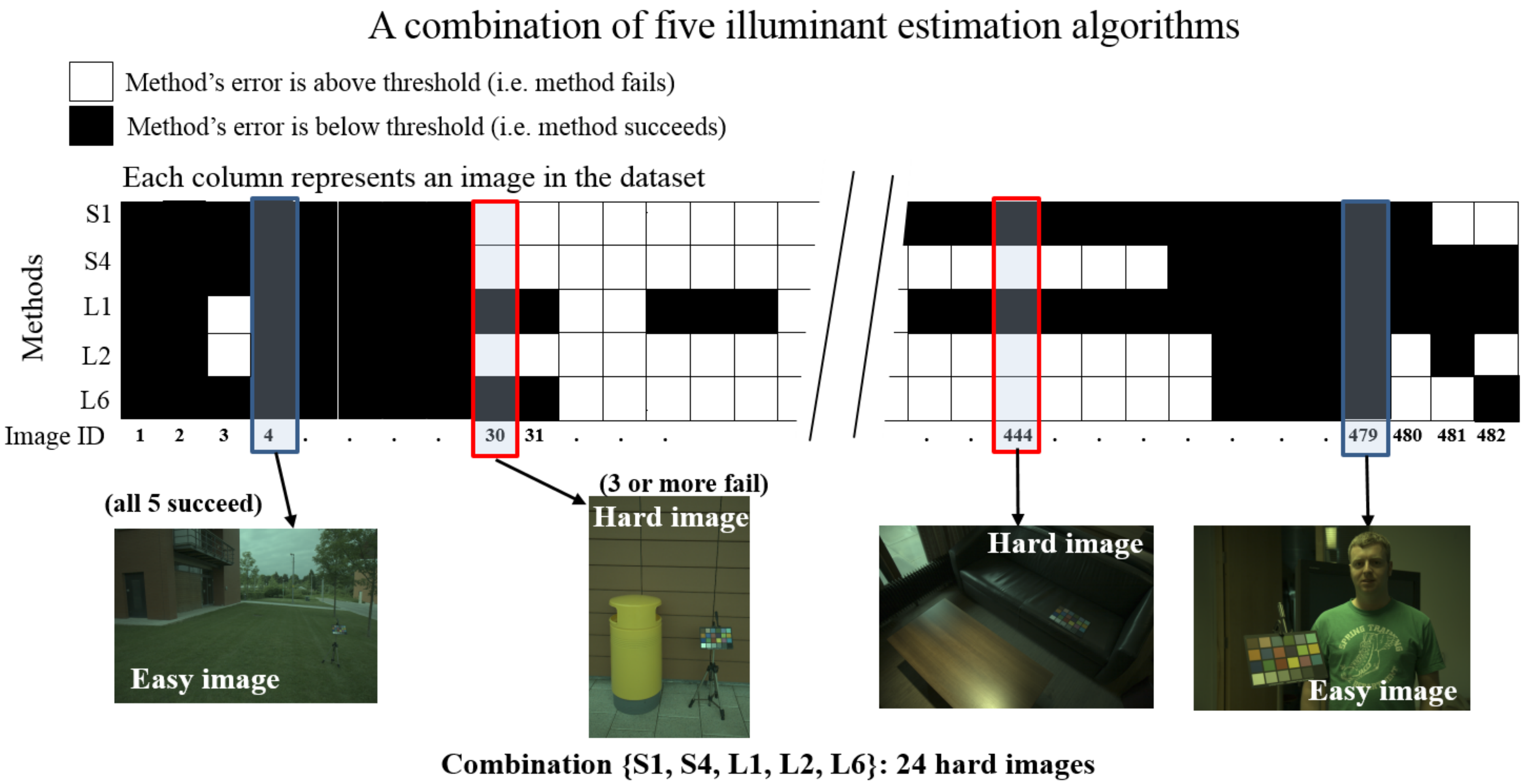

A combination of five illuminant estimation algorithms
Method's error is above threshold (i.e. method fails)
Method's error is below threshold (i.e. method succeeds)
Each column represents an image in the dataset
Methods
S1
S4
L1
L2
L6
Image ID
1
2
3
4
30
31
444
479
480
481
482
(all 5 succeed)
Easy image
(3 or more fail)
Hard image
Hard image
Easy image
Combination {S1, S4, L1, L2, L6}: 24 hard images

FIGURE 8.8: A combination of five illuminant estimation algorithms. This combination results in 24 hard images out of 482 images of Canon5D from Gehler-Shi dataset [Gehler et al., 2008; Shi and Funt, 2010].

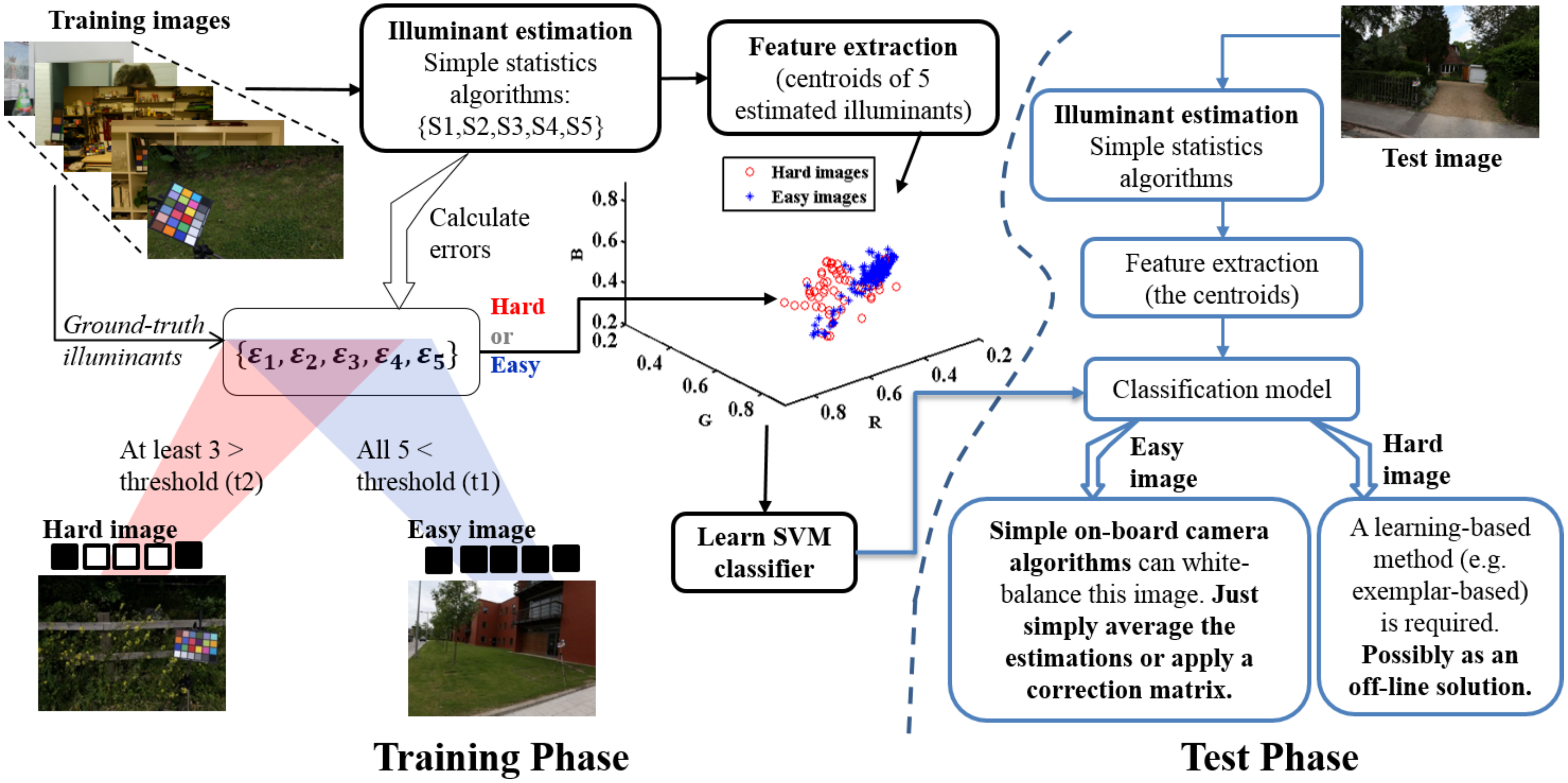

Training images
Illuminant estimation
Simple statistics algorithms: {S1,S2,S3,S4,S5}
Feature extraction
(centroids of 5 estimated illuminants)
Hard images
Easy images
Calculate errors
Ground-truth illuminants
$\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5\}$
Hard
or
Easy
B
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
G
0.8
0.6
0.4
0.2
R
At least 3 > threshold (t2)
All 5 < threshold (t1)
Hard image
Easy image
Learn SVM classifier
Training Phase
Test image
Illuminant estimation
Simple statistics algorithms
Feature extraction
(the centroids)
Classification model
Easy image
Hard image
Simple on-board camera algorithms can white-balance this image. Just simply average the estimations or apply a correction matrix.
A learning-based method (e.g. exemplar-based) is required. Possibly as an off-line solution.
Test Phase

FIGURE 8.9: The proposed framework which focuses on illuminant estimation for hard images. The classifier categorises images into hard and easy. The easy images can be treated using fast statistical-based techniques. Images classified as hard are processed using a slower, but more accurate learning-based method.

Chapter 9

Conclusion and Future Research

This chapter summaries the contributions of this thesis including the analysis of flaws by recovery angular error metric for illuminant estimation. Further a new metric, reproduction angular error, for illuminant estimation is proposed which is more inline with the reproduction of white balanced images. Other contributions include a psychophysical study for relating performance metric and human observer preference and finally a proposed hybrid method of illuminant estimation for detection of images which are hard for most illuminant estimation algorithms. Finally, the possible future research directions are discussed.

9.1 Summary

This thesis has contributed to the performance evaluation of illuminant estimation algorithms. Considering the large body of literature in illuminant estimation and proposal of new methods every year, the evaluation and comparison of illuminant estimation algorithms is of great importance.

Chapter 1 provided a brief background on the problem of illuminant estimation. In Chapter 2 the background of image formation which is often used when discussing illuminant estimation is presented and a illuminant estimation algorithms are surveyed. Existing benchmark datasets used for illuminant estimation was presented in Chapter 3.

The number of illuminant estimation algorithms proposed over years and the fact that the topic is still of interest to the computer vision researchers today predicates the need to evaluate the performance of the algorithms. The reliability of the most widely used metric in illuminant estimation, the recovery angular error, was re-examined in this thesis. It was demonstrated that this metric is flawed and so its adoption could lead to misjudgement about the performance of an algorithm depending on the lighting condition. The same scene viewed under different lights with the illuminant estimated with the same algorithm delivers the same reproduction (when the light colour is divided out) but the recovery angular error can vary. It was shown that the range of recovery angular error is very large. For instance certain lights like cyan, magenta and yellow can induce large recovery error, but red, green and blue lights often have smaller errors. **Reproduction angular error**, which is an improvement of recovery angular error was proposed as a solution to the flaw in Chapter 4. Reproduction angular error measures the angle between the reproduced white surface by the ground-truth illuminant and the one by the estimate of an algorithm. The performance of a wide range of illuminant estimation algorithms were re-evaluated for different colour constancy benchmark datasets. The results and their analysis by different statistical tests are provided in Chapter 5. Further, the correlation between the two metrics, reproduction and recovery angular errors, was studied for different scenes and the same scenes with different illuminants. The analysis showed where the scenes are the same and the illuminant differs, there is hardly any correlation between the two metrics. The correlation between the two metrics is stronger for different scenes.

Studying the state of the art on performance evaluation of illuminant estimation algorithms, it is felt that there is a gap for evaluation methods based on reproduced colours rather than measuring the difference between the reproduced white by the estimated illuminant and the ground truth or the difference between the estimated and ground truth illuminants. In Chapter 6, a novel framework is introduced to fill this gap which evaluates illuminant estimation algorithms based on a palette of colours. The colour differences between the actual and reproduced colours are calculated using the CIE lab colour difference formula. We found a strong correlation between the errors of CIE lab and reproduction angular errors which mean reproduction angular error can be used as a proxy.

The psychophysics study conducted in Chapter 7 demonstrated a relatively strong correlation between the reproduction angular error and human perception. The study was mainly set up to investigate whether the observers agree with the switches in the ranking of an algorithms pair estimating the illuminant of a pair of images. The experiment showed that the observers in most cases agree with the rank order given to the pair of algorithms by reproduction angular error.

Most of research in illuminant estimation, has been focused on a summary of statistics for performance of the algorithms over a benchmark dataset. None of the prior works have examined if there is any commonality in these statistics across the images in the dataset. A hybrid framework is proposed in Chapter 8 which recognise images for which most of simple and widely-used algorithms fail. Many recent algorithms are learning-based techniques that due to their complexity might be suitable as an offline solution rather than an on-board colour constancy method. Using the proposed hybrid strategy, the images which require further processing by complicated techniques can be labeled in a camera.

9.2 Future Research

We propose the reproduction angular error is a great improvement over the mostly used existing method (recovery angular error) and should be adopted by the community for the evaluation of illuminant estimation algorithms in the future work. Moreover, if a new metric is used then there is potential for algorithm development whose performance optimises that metric.

Regarding the hybrid strategy for detecting hard images, we are keen to extend our idea to additional colour constancy datasets. Currently, we were only able to apply this approach to the Gehler-Shi dataset as it has sufficient number of images. More recent datasets (e.g. NUS 9-camera) have more overall images, but fewer images per camera (only around 200 images per camera). We did attempt to apply this approach to the older Greyball dataset [Ciurea and Funt, 2003] but found the dataset is inappropriate given that it is low-resolution video footage (320×240) and is not properly linearised. We also found that this dataset had a large number of hard images due to improper position of the Grey-ball used for the ground truth. This points to the need of additional datasets in the colour constancy community and is an area which can be focused on for future work.

Appendix A

$\Delta E2000$ Colour Difference Formula

CIEXYZ tristimulus values of a colour is converted to its corresponding CIELab values as:

$$\begin{aligned} L &= 116f(Y/Yn) - 16 \\ a &= 500(f(X/Xn) - f(Y/Yn)) \\ b &= 200(f(Y/Yn) - f(Z/Zn)) \\ &\begin{cases} f(x) = x^{1/3} & \text{if } x > .008856 \\ f(x) = 7.787x + 16/116 & \text{if } x \leq .008856 \end{cases} \end{aligned} \tag{A.1}$$

Here, X_n, Y_n and Z_n are the CIE XYZ tristimulus values of the reference white point.

The $\Delta E2000$ colour difference [Sharma et al., 2005] between the two colours L_1 a_1 b_1 and L_2 a_2 b_2 is calculated as:

$$\Delta E_{00} = \sqrt{(\frac{\Delta L'}{k_L S_L})^2 + (\frac{\Delta C'}{k_C S_C})^2 + (\frac{\Delta H'}{k_H S_H})^2 + R_T \frac{\Delta C'}{k_C S_C}\frac{\Delta H'}{k_H S_H}} \tag{A.2}$$

where k_L, k_C and k_H are usually unity, and

$$\Delta L' = L_1 - L_2 \tag{A.3}$$

$$\Delta C' = C_1' - C_2' \tag{A.4}$$

where C_1' and C_2' are defined as:

$$C_1' = \sqrt{{a_1'}^2 + b_1^2} \quad \text{and} \quad C_2' = \sqrt{{a_2'}^2 + b_2^2} \tag{A.5}$$

and a_i' is defined as:

$$a'_i = a_i + \frac{a_i}{2}\left(1 - \sqrt{\frac{\overline{C}^7}{\overline{C}^7 + 25^7}}\right) \quad \text{where } \overline{C} = \frac{C_1 + C_2}{2} \tag{A.6}$$

We define ΔH as:

$$\Delta H' = 2\sqrt{C_1' C_2'}\sin(\Delta h'/2) \tag{A.7}$$

where:

$$h_1' = \begin{cases} \tan^{-1}(b_1/a_1') & \tan^{-1}(b_1/a_1') \geq 0 \\ \tan^{-1}(b_1/a_1') + 360^\circ & \tan^{-1}(b_1/a_1') < 0 \end{cases}$$

$$h_2' = \begin{cases} \tan^{-1}(b_2/a_2') & \tan^{-1}(b_2/a_2') \geq 0 \\ \tan^{-1}(b_2/a_2') + 360^\circ & \tan^{-1}(b_2/a_2') < 0 \end{cases}$$

$$H' = \begin{cases} (h'_1 + h'_2 + 360°)/2 & |h'_1 - h'_2| > 180° \\ (h'_1 + h'_2)/2 & |h'_1 - h'_2| \leq 180° \end{cases}$$

Further S_L, S_C and S_H are defined as:

$$S_L = 1 + \frac{0.015(\overline{L}' - 50)^2}{\sqrt{20 + (\overline{L}' - 50)^2}} \quad \overline{L}' = (L_1 + L_2)/2 \tag{A.8}$$

$$S_C = 1 + 0.045\overline{C}' \quad \overline{C}' = (C'_1 + C'_2)/2 \tag{A.9}$$

$$S_H = 1 + 0.015\overline{C}'T \tag{A.10}$$

where

$$T = 1 - 0.17\cos(\overline{H}' - 30°) + 0.24\cos(2\overline{H}') + 0.32\cos(3\overline{H}' + 6°) - 0.20\cos(4\overline{H}' - 63°)$$

Finally:

$$R_C = 2\sqrt{\frac{\overline{C}'^7}{\overline{C}'^7 + 25^7}} \tag{A.11}$$

$$R_T = -R_C \sin(2\Delta\theta) \tag{A.12}$$

where

$$\Delta\theta = 30 \exp\left\{-\left(\frac{\overline{H}' - 270^\circ}{25}\right)^2\right\}$$

If the colour values are initially provided in RGB they need to be converted to XYZ. To convert the values from RGB to XYZ (and vice versa) the colour profile of the device captured (e.g. a Sony camera) or displaying (an HP monitor) the RGB colours should be know. For instance, for an experiment involving displaying images on a monitor, the monitor can be set to display colours with an sRGB ICC profile [Consortium et al., 2004]. In this case, the conversion will be a mapping between sRGB and XYZ.

Or where the calculations require mapping the RGB values captured by a camera (eg. Sony-DXC-930) to the XYZ values, the mapping between the RGB values and the corresponding XYZs can be solved for so the camera sensitivity functions (if known) be fit to their XYZ corresponding values of the CIE 1931 Colour Matching Functions (CMFs)[Wyszecki and Stiles, 1982a].

Appendix B

CIE 1931 Chromaticity Diagram

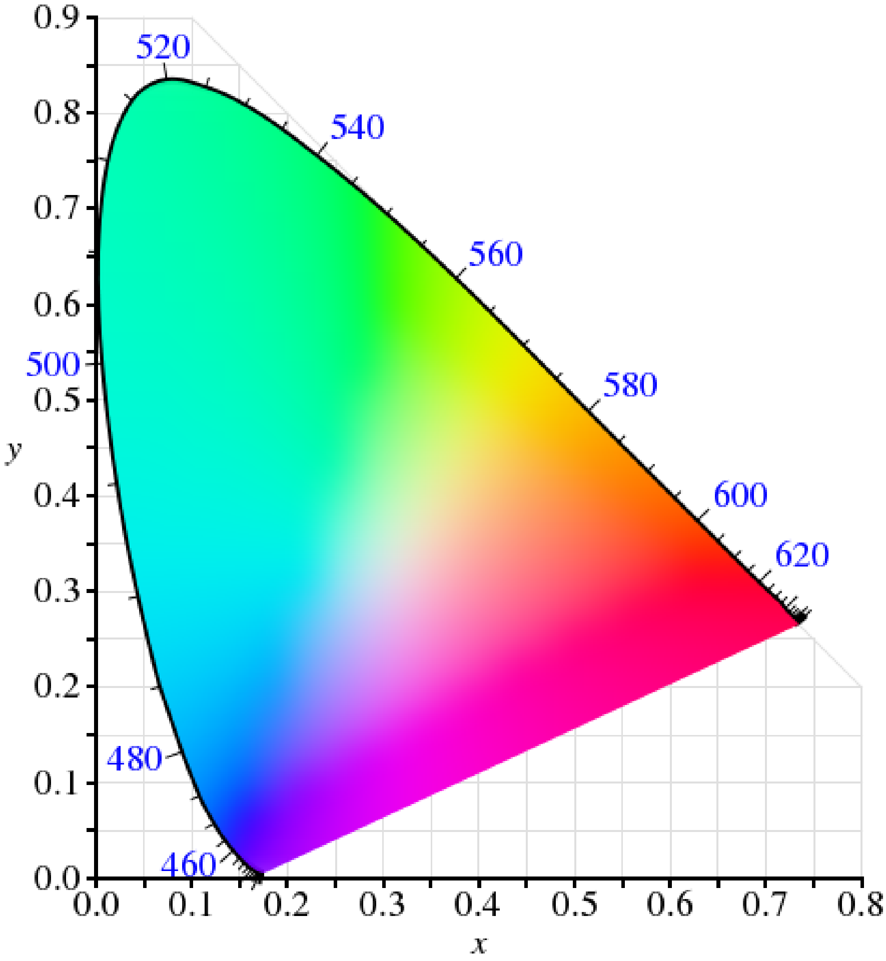


Figure B.1: CIE 1931 chromaticity diagram.

Appendix C

List of Images Classified as Hard by Hybrid method

Table C.1: The list of images from Gehler-Shi dataset classified as hard by the hybrid algorithm presented in Chapter 8 (all the images belong to Canon 5D camera, although the image numbers are as they appear in the dataset).

112	249	387	498
123	254	390	499
137	296	394	508
146	297	399	519
175	315	401	521
192	321	408	522
194	324	409	523
200	327	412	534
202	333	413	550
213	336	448	551
214	338	452	553
215	339	460	556
241	352	464	557
243	363	467	559
244	380	480	566
248	385	483	568

Bibliography

Colorchecker classic. http://xritephoto.com/ph_product_overview.aspx?ID=1192. Accessed: 2016-06-5.

Datacolor spyder4elite. http://spyder.datacolor.com/portfolio-view/spyder5elite/, accessed Sept, 2016.

Alaa E Abdel-Hakim and Aly A Farag. Csift: A sift descriptor with color invariant characteristics. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1978–1983, 2006.

Vivek Agarwal, Besma R Abidi, Andreas Koschan, and Mongi A Abidi. An overview of color constancy algorithms. *Journal of Pattern Recognition Research*, 1(1):42–54, 2006.

Vivek Agarwal, Andrei V Gribok, and Mongi A Abidi. Machine learning approach to color constancy. *Neural Networks*, 20(5):559–563, 2007.

Richard L Alfvin, Mark D Fairchild, et al. Observer variability in metameric color matches using color reproduction media. *Color Research & Application*, 22(3): 174–188, 1997.

Lawrence E Arend, Adam Reeves, James Schirillo, and Robert Goldstein. Simultaneous color constancy: papers with diverse munsell values. *JOSA A*, 8(4): 661–672, 1991.

Nikola Banic and Sven Loncaric. A perceptual measure of illumination estimation error, 2015.

Kobus Barnard and Brian Funt. Camera characterization for color research. *Color Research & Application*, 27(3):152–163, 2002.

Kobus Barnard, Vlad Cardei, and Brian Funt. A comparison of computational color constancy algorithms.I: Methodology and experiments with synthesized data. *IEEE Transactions on Image Processing*, 11(9):972–984, 2002a.

Kobus Barnard, Lindsay Martin, Adam Coath, and Brian Funt. A comparison of computational color constancy algorithms. ii. experiments with image data. *IEEE transactions on Image Processing*, 11(9):985–996, 2002b.

Kobus Barnard, Lindsay Martin, Brian Funt, and Adam Coath. A data set for color research. *Color Research & Application*, 27(3):147–151, 2002c.

Jonathan T Barron. Convolutional color constancy. In *IEEE International Conference on Computer Vision*, 2015.

Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, 2003.

Daniel Berwick and Sang Wook Lee. A chromaticity space for specularity, illumination color-and illumination pose-invariant 3-d object recognition. In *Computer Vision, 1998. Sixth International Conference on*, pages 165–170. IEEE, 1998.

Simone Bianco and Raimondo Schettini. Adaptive color constancy using faces. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1505–1518, 2014.

Simone Bianco, Gianluigi Ciocca, Claudio Cusano, and Raimondo Schettini. Improving color constancy using indoor–outdoor image classification. *IEEE Transactions on Image Processing*, 17(12):2381–2392, 2008.

Simone Bianco, Gianluigi Ciocca, Claudio Cusano, and Raimondo Schettini. Automatic color constancy algorithm selection and combination. *Pattern recognition*, 43(3):695–705, 2010.

Simone Bianco, Claudio Cusano, and Raimondo Schettini. Color constancy using cnns. In *IEEE CVPR Workshops, Deep Vision: Deep Learning in Computer Vision*, 2015.

David H Brainard and William T Freeman. Bayesian color constancy. *Journal of the Optical Society of America A*, 14(7):1393–1411, 1997.

Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.

Vlad C Cardei and Brian Funt. Committee-based color constancy. In *Color and Imaging Conference*, volume 1999, pages 311–313. Society for Imaging Science and Technology, 1999.

Vlad C Cardei, Brian Funt, and Kobus Barnard. Estimating the scene illumination chromaticity by using a neural network. *JOSA A*, 19(12):2374–2386, 2002.

Ayan Chakrabarti. Color constancy by learning to predict chromaticity from luminance. In *Advances in Neural Information Processing Systems*, pages 163–171, 2015.

Ayan Chakrabarti, Keigo Hirakawa, and Todd Zickler. Color constancy beyond bags of pixels. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–6. IEEE, 2008.

Ayan Chakrabarti, Daniel Scharstein, and Todd Zickler. An empirical camera model for internet color vision. In *BMVC*, volume 1, page 4. Citeseer, 2009.

Ayan Chakrabarti, Keigo Hirakawa, and Todd Zickler. Color constancy with spatio-spectral statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1509–1519, 2012.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3): 27, 2011.

Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*, 31(5):1049–1058, 2014.

Dongliang Cheng, Brian Price, Scott Cohen, and Michael S. Brown. Effective learning-based illuminant estimation using simple features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015a.

Dongliang Cheng, Brian Price, Scott Cohen, and Michael S Brown. Beyond white: Ground truth colors for color constancy correction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 298–306, 2015b.

Dongliang Cheng, Brian Price, Scott Cohen, and Michael S Brown. Effective learning-based illuminant estimation using simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1000–1008, 2015c.

Hamilton Y Chong, Steven J Gortler, and Todd Zickler. The von Kries hypothesis and a basis for color constancy. *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

Florian Ciurea and Brian Funt. A large image database for color constancy research. *11th IS & T/SID Color and Imaging Conference*, 2003(1):160–164, 2003.

D Coffin. Dcraw: Decoding raw digital photos in linux. https://www.cybercom.net/~dcoffin/dcraw/, 2008.

International Electrotechnical Commission et al. Colour measurement and management in multimedia systems and equipment-part 2-1: Default rgb colour space-srgb. Technical report, IEC 61966-2-1, 1999.

William Jay Conover. *Practical nonparametric statistics, Third Edition.* John Wiley & Sons, New York, 1999. ISBN 0-471-16068-7.

International Color Consortium et al. Image technology colour managementarchitecture, profile format, and data structure. *Specification*, 201(1):2004–10, 2004.

TN Cornsweet. Vision perception. *Vision perception*, 1970.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Peter B Delahunt and David H Brainard. Does human color constancy incorporate the statistical regularity of natural daylight? *Journal of Vision*, 4(2):1–1, 2004.

Michael DZmura, Geoffrey Iverson, and Benjamin Singer. Probabilistic color constancy. *Geometric representations of perceptual phenomena*, pages 187–202, 1995.

Marc Ebner. Color constancy based on local space average color. *Machine Vision and Applications*, 20(5):283–301, 2009.

Ronald Fagin, Ravi Kumar, and D Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

Graham Finlayson and Steven Hordley. Selection for gamut mapping colour constancy. *Image and Vision computing*, 17(8):597–604, 1999.

Graham Finlayson and Steven Hordley. Improving gamut mapping color constancy. *IEEE Transactions on Image Processing*, 9(10):1774–1783, 2000.

Graham Finlayson and Roshanak 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.

Graham D. Finlayson. Color in perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):1034–1038, 1996.

Graham D Finlayson. Corrected-moment illuminant estimation. *IEEE International Conference on Computer Vision (ICCV)*, pages 1904–1911, 2013.

Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. *12th IS & T/SID Color and Imaging Conference*, 2004(1):37–41, 2004.

Graham D. Finlayson and Roshanak Zakizadeh. Reproduction angular error: An improved performance metric for illuminant estimation. *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5*, 2014.

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, 1994a.

Graham D Finlayson, Mark S Drew, and Brian V Funt. Spectral sharpening: sensor transformations for improved color constancy. *Journal of the Optical Society of America A*, 11(5):1553–1563, 1994b.

Graham D Finlayson, Brian V Funt, and Kobus Barnard. Color constancy under varying illumination. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 720–725. IEEE, 1995.

Graham D Finlayson, Paul M Hubel, and Steven Hordley. Color by correlation. In *Color and Imaging Conference*, volume 1997, pages 6–11. Society for Imaging Science and Technology, 1997.

Graham D. Finlayson, Steven D. Hordley, and Paul 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.

Graham D Finlayson, Steven D Hordley, and P Morovic. Colour constancy using the chromagenic constraint. In *2005 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 1079–1086. IEEE, 2005.

Graham D Finlayson, Roshanak Zakizadeh, and Arjan Gijsenij. The reproduction angular error for evaluating the performance of illuminant estimation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP (99):1–8, 2016.

David A Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision*, 5(1):5–36, 1990.

David H Foster, Kinjiro Amano, Sérgio MC Nascimento, and Michael J Foster. Frequency of metamerism in natural scenes. *Journal of the Optical Society of America A*, 23(10):2359–2372, 2006.

Clément Fredembach and Graham Finlayson. Bright chromagenic algorithm for illuminant estimation. *Journal of Imaging Science and Technology*, 52(4):40906–1, 2008.

Brian Funt and Lilong Shi. The rehabilitation of maxrgb. In *Color and Imaging Conference*, volume 2010, pages 256–259. Society for Imaging Science and Technology, 2010.

Brian Funt and Weihua Xiong. Estimating illumination chromaticity via support vector regression. In *Color and Imaging Conference*, volume 2004, pages 47–52. Society for Imaging Science and Technology, 2004.

Brian Funt, Vlad Cardei, and Kobus Barnard. Learning color constancy. In *Color and Imaging Conference*, volume 1996, pages 58–60. Society for Imaging Science and Technology, 1996.

Brian Funt, Vlad Cardei, and Kobus Barnard. Neural network color constancy and specularly reflecting surfaces. In *Proc. of AIC Color*, volume 97, 1997.

Brian Funt, Kobus Barnard, and Lindsay Martin. Is machine colour constancy good enough? In *European Conference on Computer Vision*, pages 445–459. Springer, 1998.

Eric Garcia and Maya Gupta. Lattice regression. In *Advances in Neural Information Processing Systems*, pages 594–602, 2009.

Eric Garcia, Raman Arora, and Maya R Gupta. Optimized regression for efficient function evaluation. *IEEE Transactions on Image Processing*, 21(9):4128–4140, 2012.

Peter V Gehler, Carsten Rother, Andrew Blake, Tom Minka, and Toby Sharp. Bayesian color constancy revisited. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

Ron Gershon, Allan D Jepson, and John K Tsotsos. From [r, g, b] to surface reflectance: Computing color constant descriptors in images. In *IJCAI*, pages 755–758, 1987.

Theo Gevers and Arnold WM Smeulders. Color-based object recognition. *Pattern recognition*, 32(3):453–464, 1999.

Arjan Gijsenij and Theo Gevers. Color constancy using image regions. *International Conference on Image Processing (ICIP) (3)*, pages 501–504, 2007.

Arjan Gijsenij and Theo Gevers. Color constancy using natural image statistics and scene semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):687–698, 2011.

Arjan Gijsenij, Theo Gevers, and Marcel P Lucassen. A perceptual comparison of distance measures for color constancy algorithms. In *European Conference on Computer Vision*, pages 208–221. Springer, 2008.

Arjan Gijsenij, Theo Gevers, and Marcel P Lucassen. Perceptual analysis of distance measures for color constancy algorithms. *Journal of the Optical Society of America A*, 26(10):2243–2256, 2009a.

Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Physics-based edge evaluation for improved color constancy. In *CVPR*, pages 581–588, 2009b.

Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Generalized gamut mapping using image derivative structures for color constancy. *International Journal of Computer Vision*, 86(2-3):127–139, 2010.

Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475–2489, 2011. URL http://colorconstancy.com/.

Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Improving color constancy by photometric edge weighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):918–929, 2012.

Michael D Grossberg and Shree K Nayar. Determining the camera response from images: What is knowable? *IEEE Transactions on pattern analysis and machine intelligence*, 25(11):1455–1467, 2003.

Michael D Grossberg and Shree K Nayar. Modeling the space of camera response functions. *IEEE transactions on pattern analysis and machine intelligence*, 26 (10):1272–1282, 2004.

Steven D Hordley. Scene illuminant estimation: past, present, and future. *Color Research & Application*, 31(4):303–314, 2006.

Steven D Hordley and Graham D Finlayson. Re-evaluating colour constancy algorithms. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 1:76–79, 2004.

Steven D Hordley and Graham D Finlayson. Reevaluation of color constancy algorithm performance. *Journal of the Optical Society of America A*, 23(5): 1008–1020, 2006.

Paul M Hubel. Foveon technology and the changing landscape of digital cameras. In *IS&T/SID Color and Imaging Conference*, pages 314–317, 2005.

Paul M Hubel, Graham D Finlayson, and Steven D Hordley. White point estimation using color by convolution, April 3 2007. US Patent 7,200,264.

Jun Jiang, Dengyu Liu, Jinwei Gu, and Sabine Susstrunk. What is the space of spectral sensitivity functions for digital color cameras? In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 168–179. IEEE, 2013.

Hamid Reza Vaezi Joze. *Estimating the colour of the illuminant using specular reflection and exemplar-based method*. PhD thesis, Applied Sciences: School of Computing Science, 2013.

Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based colour constancy. *British Machine Vision Conference (BMVC)*, pages 1–12, 2012.

Hamid Reza Vaezi Joze, Mark S Drew, Graham D Finlayson, and Perla Aurora Troncoso Rey. The role of bright pixels in illumination estimation. In *Color and Imaging Conference*, volume 2012, pages 41–46. Society for Imaging Science and Technology, 2012.

Deane B Judd. Hue saturation and lightness of surface colors with chromatic illuminatio. *JOSA*, 30(1):2–32, 1940.

Seon Joo Kim and Marc Pollefeys. Robust radiometric calibration and vignetting correction. *IEEE transactions on pattern analysis and machine intelligence*, 30 (4):562–576, 2008.

Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S Brown. A new in-camera imaging model for color computer vision and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2289–2302, 2012.

Helga Kolb. Simple anatomy of the retina. 2012.

Edwin H Land and John McCann. Lightness and retinex theory. *JOSA*, 61(1): 1–11, 1971.

Edwin Herbert Land et al. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977.

Hsien-Che Lee. Method for computing the scene-illuminant chromaticity from specular highlights. *JOSA A*, 3(10):1694–1699, 1986.

H Leibowitz. Relation between the brunswick and thouless ratios and functional relations in experimental investigations of perceived shape, size, and brightness. *Perceptual and Motor Skills*, 1956.

Bing Li, Weihua Xiong, Weiming Hu, and Ou Wu. Evaluating combinational color constancy methods on real-world images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1929–1936. IEEE, 2011.

Hai Ting Lin, Zheng Lu, Seon Joo Kim, and Michael S Brown. Nonuniform lattice regression for modeling the camera imaging pipeline. In *European Conference on Computer Vision*, pages 556–568. Springer, 2012.

Rui Lu, Arjan Gijsenij, Theo Gevers, Vladimir Nedovi, De Xu, and Jan-Mark Geusebroek. Color constancy using 3d scene geometry. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1749–1756. IEEE, 2009.

Stuart Lynch, Mark Drew, and Graham Finlayson. Colour constancy from both sides of the shadow edge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 899–906, 2013.

Laurence T Maloney. Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *JOSA A*, 3(10):1673–1683, 1986.

Laurence T Maloney and Brian A Wandell. Color constancy: a method for recovering surface spectral reflectance. *JOSA A*, 3(1):29–33, 1986.

Andrew Moore, John Allman, and Rodney M Goodman. A real-time neural system for color constancy. *IEEE Transactions on Neural networks*, 2(2):237–247, 1991.

Albert Henry Munsell. Munsell book of color. 1950.

Sérgio MC Nascimento, Flávio P Ferreira, and David H Foster. Statistics of spatial cone-excitation ratios in natural scenes. *JOSA A*, 19(8):1484–1490, 2002.

Seoung Wug Oh and Seon Joo Kim. Approaching the computational color constancy as a classification problem through deep learning. *Pattern Recognition*, 61:405–416, 2017.

Konstantinos Plataniotis and Anastasios N Venetsanopoulos. *Color image processing and applications*. Springer Science & Business Media, 2013.

Rajeev Ramanath, Wesley E Snyder, Youngjun Yoo, and Mark S Drew. Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1):34–43, 2005.

Charles Rosenberg, Martial Hebert, and Sebastian Thrun. Color constancy using kl-divergence. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 239–246. IEEE, 2001.

Charles Rosenberg, Alok Ladsariya, and Tom Minka. Bayesian color constancy with non-gaussian models. In *Advances in neural information processing systems*, page None, 2003.

Guillermo Sapiro. Color and illuminant voting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(11):1210–1215, 1999.

Gerald Schaefer, Steven Hordley, and Graham Finlayson. A combined physical and statistical approach to colour constancy. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 148–153. IEEE, 2005.

Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985.

Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

Lilong Shi and Brian Funt. Re-processed version of the gehler color constancy dataset of 568 images. *Simon Fraser University*, 2010. URL http://www.cs.sfu.ca/~colour/data/.

David Slater and Glenn Healey. The illumination-invariant recognition of 3d objects using local color invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(2):206–210, 1996.

P. Sprent and N.C. Smeeton. *Applied Nonparametric Statistical Methods, Fourth Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2007. ISBN 9781584887010. URL http://books.google.co.uk/books?id=arn4eBAWodwC.

Andrew Stockman and Lindsay T Sharpe. The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision research*, 40(13):1711–1737, 2000.

Michael Stokes, Mattew Anderson, Srinivasan Chandrasekar, and Ricardo Motta. A standard default color space for the internet?srgb, 1996. *URL http://www.w3. org/Graphics/Color/sRGB*, 2012.

Robby T Tan, Ko Nishino, and Katsushi Ikeuchi. Color constancy through inverse-intensity chromaticity space. *Journal of the Optical Society of America A*, 21 (3):321–334, 2004.

Louis L Thurstone. Psychophysical analysis. *The American journal of psychology*, 38(3):368–389, 1927.

Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing*, 16(9):2207–2214, 2007a.

Joost Van De Weijer, Cordelia Schmid, and Jakob Verbeek. Using high-level visual information for color constancy. *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8, 2007b.

Javier Vazquez-Corral, C Párraga, Ramon Baldrich, and Maria Vanrell. Color constancy algorithms: Psychophysical evaluation on a new dataset. *Journal of Imaging Science and Technology*, 53(3):31105–1, 2009.

Johannes von Kries. Beitrag zur physiologie der gesichtsempfindung. *Arch. Anat. Physiol*, 2:505–524, 1878.

Johannes von Kries. Chromatic adaptation. *Festschrift der Albrecht-Ludwigs-Universität*, pages 145–158, 1902.

Brian A Wandell. The synthesis and analysis of color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):2–13, 1987.

Gerhard West and Michael H Brill. Necessary and sufficient conditions for von kries chromatic adaptation to give color constancy. *Journal of Mathematical Biology*, 15(2):249–258, 1982.

James A Worthey and Michael H Brill. Heuristic analysis of von kries color constancy. *JOSA A*, 3(10):1708–1712, 1986.

Meng Wu, Jun Sun, Jun Zhou, and Gengjian Xue. Color constancy based on texture pyramid matching and regularized local regression. *JOSA A*, 27(10): 2097–2105, 2010.

G Wyszecki and WS Stiles. Color science: Concepts and methods, quantitative data and formulae. *John Wiley&Sons, New York*, 1982a.

Gunter Wyszecki and Walter Stanley Stiles. *Color science*, volume 8. Wiley New York, 1982b.

Ying Xiong, Kate Saenko, Trevor Darrell, and Todd Zickler. From pixels to physics: Probabilistic color de-rendering. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 358–365. IEEE, 2012.

Roshanak Zakizadeh and Graham D Finlayson. The correlation of reproduction and recovery angular errors for similar and diverse scenes. In *IS&T/SID Color and Imaging Conference*, pages 196–200, 2015.

Roshanak Zakizadeh, Michael S Brown, and Graham D Finlayson. A hybrid strategy for illuminant estimation targeting hard images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–23, 2015.