Crowd-Sourced Data and its Applications for New Algorithms in Photographic Imaging

Michael David Charles Harris

A thesis submitted for the Degree of
Doctor of Philosophy

University of East Anglia
School of Computing Sciences

April 2015

Abstract

This thesis comprises two main themes. The first of these is concerned primarily with the validity and utility of data acquired from web-based psychophysical experiments. In recent years web-based experiments, and the crowd-sourced data they can deliver, have been rising in popularity among the research community for several key reasons – primarily ease of administration and easy access to a large population of diverse participants. However, the level of control with which traditional experiments are performed, and the severe lack of control we have over web-based alternatives may lead us to believe that these benefits come at the cost of reliable data. Indeed, the results reported early in this thesis support this assumption. However, we proceed to show that it is entirely possible to crowd-source data that is comparable with lab-based results.

The second theme of the thesis explores the possibilities presented by the use of crowd-sourced data, taking a popular colour naming experiment as an example. After using the crowd-sourced data to construct a model for computational colour naming, we consider the value of colour names as image descriptors, with particular relevance to illuminant estimation and object indexing. We discover that colour names represent a particularly useful quantisation of colour space, allowing us to construct compact image descriptors for object indexing. We show that these descriptors are somewhat tolerant to errors in illuminant estimation and that their perceptual relevance offers even further utility. We go on to develop a novel algorithm which delivers perceptually-relevant, illumination-invariant image descriptors based on colour names.

Acknowledgements

First thanks must undoubtedly go to my supervisory team: Graham Finlayson and Barry Theobald. The support when it was necessary, but freedom to tackle problems as I see fit has been greatly appreciated.

Thanks must also go to the EPSRC for funding my studies, and to the University of East Anglia for enabling them. Despite its foibles, the UEA is a rather lovable old Hector, and I shall miss it dearly.

For providing access to data for their web-based experiments, Yujie Mei and Guoping Qiu. Similarly Dave Connah, who as well as providing data and code, I think is safe to call a "top bloke".

James Tauber and Brian Rosner, for providing tremendous help and experience with all things web-related, and for the generous provision of hosting services for our web-based experiments. Furthermore, their contributions to open source software are invaluable to not only myself but enumerable people the world over.

This paragraph had begun "All those in the colour lab, especially…", but after attempting to enumerate each lab member who has given me particular help and support over the years, it quickly became apparent that the list contained everyone. Suffice it to say – thank you – to each and every member of the lab with whom I have had the pleasure of working.

To all in HWOps, for providing me with inspiration and incomparable tutelage, and a place I like to call home.

To all at Tridan IT, particularly Daniel Perry. To have a place, outside of academia, to share the real day-to-day implications of computer science and software engineering has been an excellent grounding.

To my family, for a lifetime of support and encouragement in my pursuit of science and engineering.

Finally, to my wife Alyson. For putting up with my stress and distraction for the last few years and for providing me with love, support and encouragement. For not only inspiring me, but teaching me, and sometimes downright forcing me to be a better scientist. She is the greatest scientist I know, and an even better person.

Contents

List of Figures

List of Tables

List of Algorithms

Publications

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

- M. D. Harris and G. D. Finlayson. Comparing a Pair of Paired Comparison Experiments: Examining the Validity of Web-Based Psychophysics. In *Proceedings of IS&T's Nineteenth Color and Imaging Conference*, pages 29–34, San Jose, California (USA), November 2011.

- M. D. Harris, G. D. Finlayson and J. Tauber. Web-based Image Preference. In *Journal of Imaging Science and Technology*, 57.2, 2013

- M. D. Harris and G. D. Finlayson. Temporal Stability of Ranks for Image Preference. In *Proceedings of the Twelfth International AIC (Association Internationale de la Couleur) Congress*, Newcastle Upon Tyne, England, 2013.

Glossary

C2G Colour to Greyscale

CIE Commission Internationale de l'Éclairage (International Commission on Illumination)

ECI European Colour Initiative

GMM Gaussian Mixture Model

HDR High Dynamic Range

ISO International Organization for Standardization

RGB Red, Green, Blue

RMS Root Mean Square

sRGB Standard Default Color Space for the Internet (Stokes et al., 1996)

TMO Tone Mapping Operator

Tone Mapping Operators

Drago *Adaptive Logarithmic Mapping For Displaying High Contrast Scenes* (Drago et al., 2003)

EMPJ *Photographic Tone Reproduction For Digital Images* (Reinhard et al., 2002)

Filter *Fast Bilateral Filtering For The Display Of High-Dynamic-Range Images* (Durand and Dorsey, 2002)

GD *Gradient Domain High Dynamic Range Compression* (Fattal et al., 2002)

Hier *Hierarchical Tone Mapping For High Dynamic Range Image Visualization* (Qiu and Duan, 2005)

LCIS *LCIS: A Boundary Hierarchy For Detail-Preserving Contrast Reduction* (Tumblin and Turk, 1999)

LocalHA *Tone-Mapping High Dynamic Range Images By Novel Histogram Adjustment* (Duan et al., 2010)

Mantiuk08 *Display Adaptive Tone Mapping* (Mantiuk et al., 2008)

Reinhard *Dynamic Range Reduction Inspired By Photoreceptor Physiology* (Reinhard and Devlin, 2005)

Ward *A Visibility Matching Tone Reproduction Operator For High Dynamic Range Scenes* (Larson et al., 1997)

Colour to Greyscale Operators

ALS *Grey Color Sharpening* (Alsam and Kolås, 2006)

BAL *Spatial Color-To-Grayscale Transform Preserving Chrominance Edge Information* (Bala and Eschbach, 2004)

GRU *The Decolorize Algorithm For Contrast Enhancing, Color To Grayscale Conversion* (Grundland and Dodgson, 2007)

LUM *Luminance* – Per-pixel luminance values.

RAS *Rasches method* (Rasche et al., 2005a,b)

SOC *Multispectral Image Visualization Through First-Order Fusion* (Socolinsky and Wolff, 2002)

Chapter 1

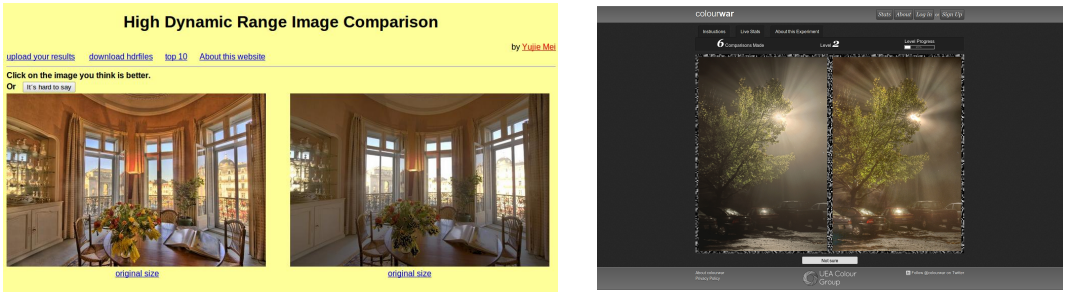
Introduction

This thesis introduces, and makes contributions to, several seemingly disparate sub-topics of colour science and photographic imaging. Unifying these disparate topics however, is the story of data, specifically data acquired from crowd-sourcing via the internet.

We begin by investigating the acquisition of this data. We explore some of the motivations behind web-based data collection, and note that, while other disciplines have been successfully exploiting this paradigm for quite some time, progress in the colour science field has been slow and, often, delivered unreliable data. We investigate the reasons behind this, and develop web-based experiments of our own which are capable of successfully delivering reliable results.

After exploring data acquisition, we use data gathered in this way to investigate the seemingly diverse topics of illuminant estimation, object recognition, and colour naming. After establishing the relationship between these topics and demonstrating the utility of computational colour naming (using a model constructed using crowd-sourced data), we develop a novel algorithm to assign illumination-invariant colour names to images, with the objective of enabling object indexing and human-led image search.

In considering how to procure reliable psychophysical data and use this in develop-



(a) *(b)*

Figure 1.1: Two differing approaches to web-based paired comparisons

ing robust imaging algorithms, the focus of this thesis is necessarily broad. This said, we have developed compelling, interlinked, practical systems that validate the implementation and practice of crowd-sourcing experimental data. We advocate a similar schema for others working on imaging problems where human judgement is an important criterion.

1.1 Outline of the Thesis

The thesis is organised as follows:

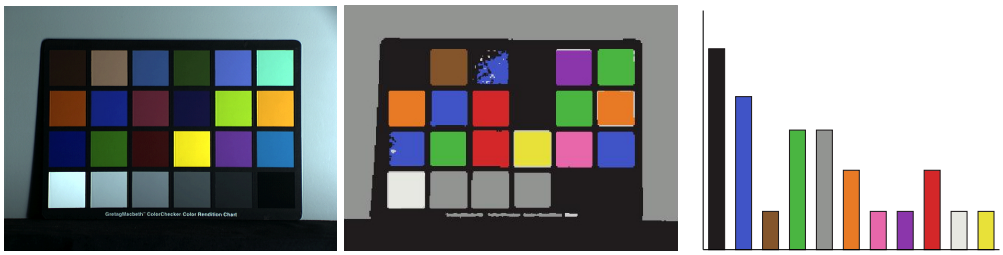
Chapter 2 first covers some general background. The topics covered in this chapter are those that recur several times in the later chapters and so require a foundational explanation before introducing the individual concerns of those later chapters.

Chapter 3 examines the concept of taking the paired comparison paradigm onto the web. We first evaluate an existing web-based experiment (Mei, 2010a) (as seen in fig. 1.1a) by replicating it under laboratory conditions and comparing the results from each variant. Disappointingly, we do not find strong correlation between the two sets of results. However, we then construct our own web-based experimental platform (fig. 1.1b), taking appropriate care over various aspects which are made apparent from the earlier comparison. These prove to be crucial to the success of web-based experiments, as the results acquired by our web-based platform are much more highly con-

cordant with the lab-based alternative. We corroborate these positive results by performing an additional experiment on the same web-based platform using an entirely different class of image processing algorithm, and compare the results to published lab-based findings (Connah et al., 2007).

Chapter 4 introduces a new statistical technique for the assessment of paired comparison experiments (web-based or otherwise). We observe that many researchers can struggle to recruit large numbers of observers to participate in experiments and so a measure of whether or not a sufficient number has yet contributed would be desirable. Similarly, for larger-scale experiments, there is utility in an indicator of when is an appropriate time to begin drawing conclusions from the data gathered so far. We build on commonly used (Thurstone, 1927) analyses of paired comparison data to construct a simple measure, although the concept is applicable to other analytical approaches.

Chapter 5 begins our exploration of applications for crowd-sourced data. We take freely-available data from an existing large-scale web-based colour naming experiment (Munroe, 2010) and use it to construct a model for computational colour naming. Inspired by Funt et al. (1998), who investigated whether a suite of commonly-used illuminant estimation techniques are sufficient to enable object recognition across multiple illumination conditions, we perform similar experiments to ascertain whether those same techniques are sufficient to enable consistent colour naming (using our new naming model). Upon discovering that colour names are somewhat resilient to inaccurate illuminant estimation, we postulate that they might better serve the object indexing problem than the histograms constructed from traditional colour space quantisations. We show that histograms derived from distributions of colour names in images (see fig. 1.2) can perform comparably to, and often better than, the traditional approach, all the while offering a more compact representation. Moreover, we show that this representation provides further utility: as it encodes perceptually-relevant image data, it can be used as a key by which to index images for searching by human-generated queries.



(a) Original image *(b) Colour names* *(c) Colour name histogram*

Figure 1.2: Colour-name-based histogram for example image

Chapter 6 builds upon the discoveries made in chapter 5. We note that, while colour names can provide useful colour descriptors, the approach in the previous chapter first requires an illuminant estimation step. Even though we have seen that colour names are, to an extent, resilient to inaccuracies in illuminant estimation, it would be desirable to omit this step altogether. As such, we seek to develop an algorithm which allows the designation of colour names to surfaces in images as they would appear under a canonical illuminant, regardless of the actual scene illuminant. In so doing we would be able to recover perceptually meaningful image descriptors, which would be useful for both machine object indexing and human image search, while being illumination-invariant. To deliver such an algorithm, we make use of some well-known properties of the diagonal model of image formation (described in section 2.1) and some constraints imposed by existing illuminant estimation techniques (section 2.2), and combine these within a boolean discrete relaxation framework. The described algorithm succeeds in meeting the stated objectives, albeit with some specific reservations.

Chapter 7 draws conclusions from this thesis.

Chapter 2

Background

This thesis introduces many subtopics in the fields of colour image processing, computer vision, and image understanding. The later chapters will separately introduce their concerns and include their own background sections to cover motivation, related work etc. This separate background chapter serves to provide a grounding in topics which require explanations that are too detailed for inclusion in other chapters, are pertinent to several other chapters, or are relevant but non-essential to the narrative of the later chapters.

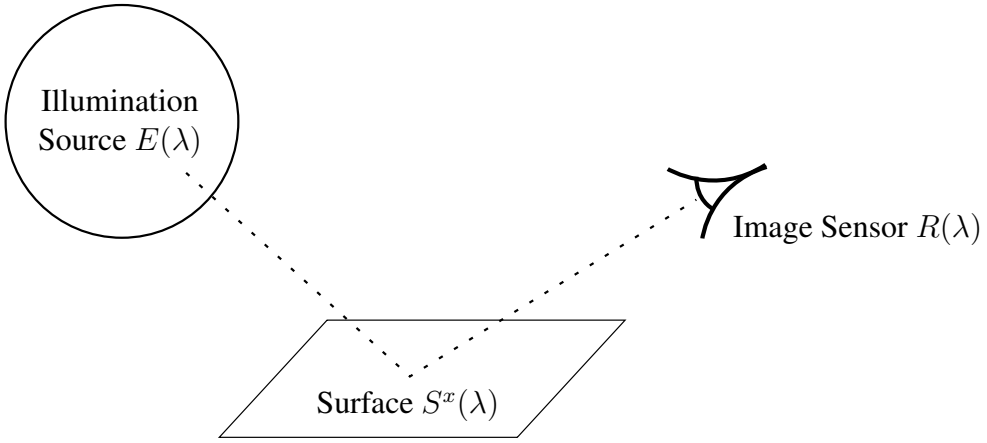


Figure 2.1: Model of image formation

2.1 Image Formation

As seen in fig. 2.1, we can model the process of image formation as the response to the product of the spectral power of the scene illuminant, the spectral reflectances of the surfaces in the scene, and the spectral sensitivity of the camera sensors. These physical variables can be brought together as a single equation:

$$\rho_k^x = \int_\omega E(\lambda)\, S^x(\lambda)\, R_k(\lambda)\, \mathrm{d}\lambda, \tag{2.1}$$

where $E(\lambda)$ is the spectral power distribution of the scene illuminant, which strikes a surface with reflectance $S^x(\lambda)$ at some spatial location x, and is collected by the camera sensor $R_k(\lambda)$ for each of k sensor classes (usually $k \in [R, G, B]$). We integrate over (usually) the visible spectrum ω to give a sensor response. This model, while not accounting for surface texture (Oren and Nayar, 1995), specular highlights (Lee, 1986; Shafer, 1985), or inter-reflections (Funt et al., 1991), provides a tolerable approximation of the actual camera response to a given scene (Wandell, 1987).

If the spectrum of the illuminant is equivalent to the output of eq. (2.1) with a pure

white surface (i.e. $S(\lambda) = 1$), then the camera response to the illuminant can be defined as

$$\rho_k^E = \int_\omega E(\lambda)\, R_k(\lambda)\, \mathrm{d}\lambda. \tag{2.2}$$

Equally, the response to a given surface under a pure white illuminant is written as:

$$\rho_k^{S,x} = \int_\omega S^x(\lambda)\, R_k(\lambda)\, \mathrm{d}\lambda. \tag{2.3}$$

It is useful to reformulate eq. (2.1) using eqs. (2.2) and (2.3) as:

$$\rho_k^x \approx \rho_k^E\, \rho_k^{S,x}. \tag{2.4}$$

This simplification (Borges, 1991; Worthey and Brill, 1986) is shown to hold for many typical sensors, so long as they are sufficiently narrowband (Finlayson et al., 1994), and in the case where the narrowband requirement is not met directly in the camera native space, it has been shown (Chong et al., 2007) to generally hold in some alternative basis (we can multiply the sensors, or equally the sensor responses, by a 3×3 matrix such that eq. (2.4) holds with regard to the new basis). It is then possible to model the process of illumination change in the language of matrix multiplication:

$$\begin{bmatrix} \rho_R^x \\ \rho_G^x \\ \rho_B^x \end{bmatrix} \approx \begin{bmatrix} \rho_R^E & 0 & 0 \\ 0 & \rho_G^E & 0 \\ 0 & 0 & \rho_B^E \end{bmatrix} \begin{bmatrix} \rho_R^{S,x} \\ \rho_G^{S,x} \\ \rho_B^{S,x} \end{bmatrix}. \tag{2.5}$$

This is known as the *diagonal model* of image formation (Finlayson et al., 1994), and can be represented more compactly as

$$\underline{\rho}^x \approx E\, \underline{S}^x, \tag{2.6}$$

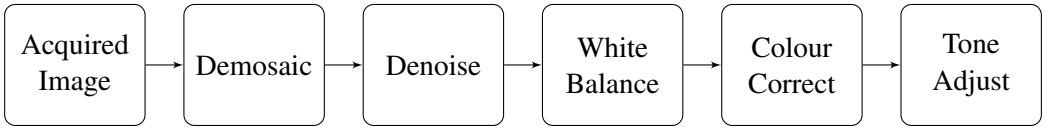


Figure 2.2: A typical image processing pipeline, adapted from Ramanath et al. (2005)

where E denotes a diagonal matrix in which the k^{th} diagonal element is ρ_k^E from eq. (2.2) and $\underline{S}^x$ is a vector in which the k^{th} element is $\rho_k^{S,x}$ from eq. (2.3). The RGB value at the pixel corresponding to physical location x becomes $\underline{\rho}^x$.

This model represents an approximation of the physical processes occurring during image formation, culminating in a sensor response from the imaging device. However, this raw form can be very different from the images used for display purposes. The process of rendering an image for display is complex (Ramanath et al., 2005), as alluded to in fig. 2.2, and has many vendor-specific variations. The topics discussed in this thesis will consider images and data at various stages throughout this process, but a complete end-to-end overview of every subprocess is outside of our current purview.

2.2 Illuminant Estimation

Illumination conditions have profound effects on the content of images, and a single surface can elicit very different pixel values from one image to another if the illumination conditions are altered. Illuminant estimation is the task of estimating the illumination conditions of a scene after an image has been taken. Often the definition of this task is extended by the desire to generate a new image which is free from the effects of the prevailing scene illuminant and re-rendered under a synthetic pure white light (white balance) or by generating descriptors for scene content which are invariant under changes in illumination (colour constancy).

This thesis does not seek directly to contribute any new methods of illuminant estimation but, as will be seen in later chapters, illumination impacts deeply on some

other topics under consideration. As such, we require a basic understanding of some of the more authoritative techniques in the illuminant estimation canon.

All of the illuminant estimation techniques discussed here adopt the diagonal model of image formation introduced in the previous section, and so have the objective of finding $\underline{E}$, given only the image values $\underline{\rho}$. Once this illuminant estimate is made (or indeed, if the illumination is known), removing it, and thus rendering the scene under a pure white illuminant, is straightforward (Drew and Funt, 1992; Finlayson and Morovic, 2000; Vrhel and Trussell, 1992):

$$\underline{\hat{\rho}}^x = E^{-1}\,\underline{\rho}^x. \tag{2.7}$$

As well as the restrictions imposed by the diagonal model, the methods discussed below all assume that every scene is illuminated by a single illuminant. This is an assumption which is often broken by real-world imagery – images taken outdoors are often illuminated by both skylight and direct sunlight, images taken indoors will often have an artificial light source as well as daylight from a window, and shadows and inter-reflections between objects can also be considered to be secondary illumination conditions. There are algorithms designed to consider multiple illuminants (Finlayson et al., 1995; Kawakami and Ikeuchi, 2009), but we will not discuss them here.

2.2.1 Max RGB

Attributed to the work of Land and McCann (1971), the Max RGB algorithm defines the illuminant estimate as the maximum pixel value in each image channel. Under the assumption that no surface can reflect more light than that which is incident upon it, the maximally reflected value must be the closest estimate to the illuminant (assuming there are no clipped pixels). If the image contains a white patch, which reflects all incident light, then this method works well. Equally, if there is a bright yellow and a bright blue

surface, then the per-channel maximum is the same as if there were a perfect white in the scene. This can be formalised thus:

$$\hat{\rho}_k^E = \max_x(\underline{\rho}_k^x). \tag{2.8}$$

2.2.2 Grey World

The Grey World algorithm (Buchsbaum, 1980) is founded on the assumption that the mean surface reflectance in a scene is achromatic. Under a pure white illuminant, a scene with high colour variation should have an average pixel value which equates to grey. By shifting the illuminant in a greenish direction, for example, the mean pixel value for the same scene should move in the same greenish direction. It follows from this observation that the illuminant can be estimated, subject to some unknown scaling factor, by simply taking the mean of all pixel responses in the image:

$$\hat{\rho}_k^E = \operatorname{mean}_x(\underline{\rho}_k^x). \tag{2.9}$$

2.2.3 Shades of Grey

Finlayson and Trezzi (2004) observed that the Max RGB and Grey World estimates can be posited as extremes of Minkowski family norms. Let $\underline{x} = [x_1, \ldots, x_N]$, then for any $p \geq 1$ a norm can be defined by:

$$\|\underline{x}\|_p = \left\{ \sum_{i=1}^{N} |x_i|^p \right\}^{1/p}. \tag{2.10}$$

An estimate for the scene illuminant can be made by taking a norm for each image channel $\underline{p}_k$:

$$\hat{\rho}_k^E = \mu(\underline{\rho}_k), \tag{2.11}$$

where the norm $\mu(\underline{x})$ is normalised by the number of pixels N present in the image channel

$$\mu(\underline{x}) = \frac{\|\underline{x}\|_p}{N^{1/p}}. \tag{2.12}$$

Max RGB and Grey World are equal to the L^1-norm and L^∞-norm respectively. Finlayson and Trezzi noted that these two algorithms will return a correct estimate if the maximum pixel is white, or if the average pixel is grey, respectively. They suggested that a middle ground, where the average pixel value is some shade of grey, might return better results. In effect, they proposed that brighter pixels are more important in illuminant estimation (Fredembach and Finlayson, 2008). After performing experiments with varying norms, they achieved favourable performance using an L^6 norm.

2.2.4 Grey Edge

Van De Weijer et al. (2007) proposed the Grey Edge hypothesis, an assumption that the average of the reflectance *differences* in a scene can be used as a more reliable cue for illuminant colour than the surfaces themselves. This edge-based algorithm can be seen as a pre-processing step for other colour constancy algorithms, i.e. first an image derivative is calculated, and then that derivative is used by Max RGB, Grey World etc. instead of the original image.

All of these simple statistical methods can be unified under one formalism as follows:

$$s\hat{\underline{\rho}}^E_{n,p,\sigma} = \left(\int \left| \frac{\delta^n \underline{\rho}^{x,\sigma}}{\delta x^n} \right|^p \mathrm{d}x \right)^{1/p}, \tag{2.13}$$

where the camera response at the spatial location x is given by $\underline{\rho}^x$. The image is first smoothed by a Gaussian filter with standard deviation σ to help compensate for image noise. The smoothed image is then differentiated with an order n differential operator

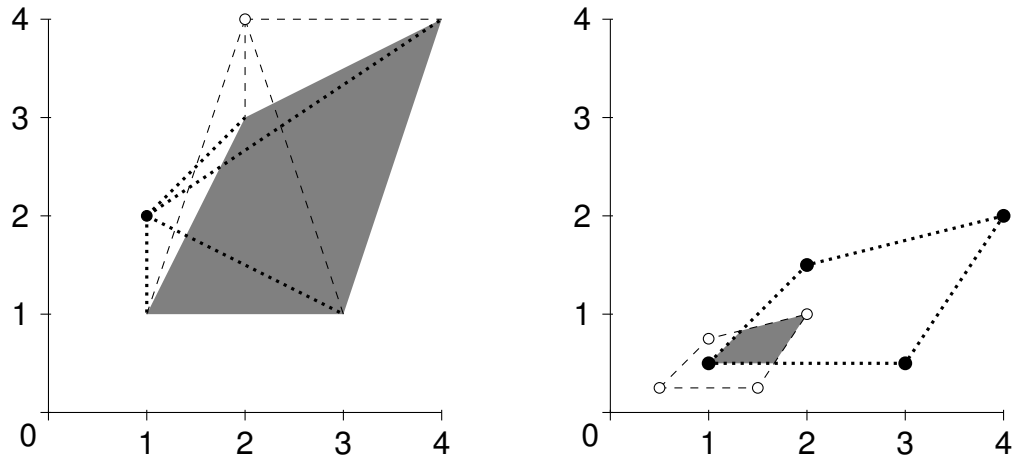
(where order 0 would indicate no differentiation – i.e. the Shades of Grey family of algorithms discussed above without utilising the Grey Edge observation). The Minkowski family p-norm is then calculated on the differentiated, smoothed image, giving the illuminant estimate $\hat{\underline{\rho}}^E$. The scalar s is an undetermined scaling factor, acknowledging that recovery of the magnitude of the illuminant is not possible using this method.

2.2.5 Gamut Mapping

The final method considered in this thesis takes a different approach to the above statistical methods. Gamut Mapping, as introduced by Forsyth (1992), relies on a priori knowledge of the surfaces which are likely to appear in images. First, a reference gamut is built of plausible pixel values under a known reference illuminant. The objective of the algorithm is then to find a plausible illuminant estimate which, after removal of the effects of the illuminant, shifts all the observed image pixel values so that they lay inside the reference gamut.

Consider the example in fig. 2.3. This example is presented in two dimensions to aid with visual understanding, but the principal holds in higher dimensions. Indeed it is quite feasible, and often advantageous, to carry out the algorithm in a two-dimensional chromaticity space (Finlayson, 1996). However, for work done later in this thesis, three-dimensional RGB space is used for consistency with the other methods described above.

In fig. 2.3a, an example reference (or *canonical*) gamut is shown by the shaded area. The gamut is constructed by sampling the pixel values of a large number of surfaces under a canonical illuminant, and is then represented by the convex closure of those pixel values. When an image is taken under a different illuminant, pixel values can be generated which lie outside the canonical gamut (note that this is not inevitable for every pixel – the gamuts of different illuminants will often have significant overlaps). If, for each of these pixel values, we generate a map from that value to each point on the convex hull of the canonical gamut, then the convex hull of those mappings represents



(a) Canonical gamut and two points mapped onto it

(b) Intersection of plausible mappings

Figure 2.3: Example of gamut mapping

the convex set of all possible mappings from the pixel value to any point within the canonical gamut. This is demonstrated in fig. 2.3a: the points outside the gamut are mapped onto it by the dotted and dashed lines – the gamut is, in this case, defined by four vertices and so each point has four mappings. In fig. 2.3b we see that those two sets of mappings have their own corresponding convex closures, the intersection of which defines the set of plausible mappings which will map every point into the canonical gamut.

As suggested by the fact that the intersection in fig. 2.3b is not a singular point, this method generally does not reduce to a single unique answer, and so a method is used to select one mapping from the plausible set. There are several approaches to this choice, such as choosing the mapping which maximises the volume of the image gamut (Forsyth, 1992) (i.e. results in the most colourful image), or using a statistical approach (Finlayson and Hordley, 1999).

While this thesis focuses on simple statistical methods which exploit properties of

the diagonal model of image formation, the wider field of illuminant estimation is large (Finlayson et al., 2002b; Finlayson and Schaefer, 2001; Funt et al., 1996; Gershon et al., 1987; Maloney and Wandell, 1986; Tan et al., 2003). Outside the scope of this thesis, some algorithms attempt to find and exploit physical properties in images, such as specular highlights, mutual illumination, and shadows. Others require detailed calibration or extensive training.

2.2.6 Evaluating Illuminant Estimation

Given these differing approaches to illuminant estimation, we need a way of comparing and evaluating the differing approaches. To do this, we compare the estimates generated by an illuminant estimation technique to ground-truth data which is known a priori. In so doing, we are again accepting the assumptions of the diagonal model, and that the scene we are investigating is frontally illuminated by a single illuminant and contains only Lambertian surfaces. Within these confines we can compare the vector representing the scene's true illuminant $\underline{\rho}^E$ to the estimate $\hat{\underline{\rho}}^E$. It is possible to use the RMS error for this, but as this error intrinsically encodes intensity, most authors opt for the intensity invariant angular error, which expresses the error as the angle between the two vectors:

$$\text{angular error} = \cos^{-1}\left(\hat{\underline{\rho}}^E \cdot \underline{\rho}^E\right), \tag{2.14}$$

where $\hat{\underline{\rho}}^E \cdot \underline{\rho}^E$ is the dot product of the normalised vectors containing the illuminant estimate and the ground-truth measurement.

2.3 Tone Mapping Operators

Later in this thesis, chapter 3 undertakes several comparisons of the output of a number of image processing algorithms. However, the purpose of the later work is not to directly compare and contrast those algorithms, but to understand more about the methods we use to evaluate observer preference among them. As the image processing algorithms introduced are not themselves under direct scrutiny, many differing collections of algorithms could suffice. The algorithms which were chosen fall in to two main categories: tone mapping operators (TMOs), and colour-to-greyscale (C2G) algorithms. We give a brief description of these two classes of algorithm below, and list the specific algorithms under comparison. However, since we are only concerned with how to compare their outputs, and not with the differences in the output themselves, we refrain from detailed descriptions of the differing approaches; there are many sources of such comparisons in the existing literature (Connah et al., 2007; Drago et al., 2002; Ledda et al., 2005; Čadík et al., 2008; Yoshida et al., 2005).

The first class of image manipulation algorithm are tone mapping operators. These are functions designed to map pixel values of high dynamic range images into a low dynamic range space such that those images can be viewed on low dynamic range monitors or printed using a conventional printer, all the while attempting to preserve the colour, contrast and brightness information present in the original image. Many approaches to this problem exist and have been evaluated in detail by several authors, such as Ledda et al. (2005). The operators used are:

Drago

Adaptive Logarithmic Mapping For Displaying High Contrast Scenes

Drago et al. (2003)

EMPJ

Photographic Tone Reproduction For Digital Images

Reinhard et al. (2002)

Filter

Fast Bilateral Filtering For The Display Of High-Dynamic-Range Images

Durand and Dorsey (2002)

GD

Gradient Domain High Dynamic Range Compression

Fattal et al. (2002)

Hier

Hierarchical Tone Mapping For High Dynamic Range Image Visualization

Qiu and Duan (2005)

LCIS

LCIS: A Boundary Hierarchy For Detail-Preserving Contrast Reduction

Tumblin and Turk (1999)

LocalHA

Tone-Mapping High Dynamic Range Images By Novel Histogram Adjustment

Duan et al. (2010)

Mantiuk08

Display Adaptive Tone Mapping

Mantiuk et al. (2008)

Reinhard

Dynamic Range Reduction Inspired By Photoreceptor Physiology

Reinhard and Devlin (2005)

Ward

A Visibility Matching Tone Reproduction Operator For High Dynamic Range Scenes

Larson et al. (1997)

The results of applying these operators to an example image (here we use the 'Belgium' scene from the dataset detailed in appendix A) are shown in fig. 2.4. For comparison, the raw linear image is shown in fig. 2.4a. This image appears dark because the range of brightnesses that can be captured (10000:1) is large compared to the range of brightnesses that can be reproduced on printed paper (100:1). This demonstrates the value of tone mapping algorithms – for making 10000:1 visible in 100:1.

2.4 Colour To Greyscale

The second class of image manipulation algorithm referred to later in the thesis are colour-to-greyscale algorithms. These are algorithms designed to reduce colour images, usually three-dimensional RGB, into one-dimensional greyscale images. There are many existing approaches to solving this problem, a collection of which are reviewed by Connah et al. (2007). We use the same collection of algorithms as Connah et al.:

ALS

Grey Color Sharpening

Alsam and Kolås (2006)

BAL

Spatial Color-To-Grayscale Transform Preserving Chrominance Edge Information

Bala and Eschbach (2004)

(a) Linear HDR image *(b) Drago* *(c) EMPJ*

(d) Filter *(e) GD* *(f) Hier*

(g) LCIS *(h) LocalHA* *(i) Mantiuk08*

(j) Reinhard *(k) Ward*

Figure 2.4: Tone mapping operators applied to 'Belgium' scene

GRU

The Decolorize Algorithm For Contrast Enhancing, Color To Grayscale Conversion

Grundland and Dodgson (2007)

LUM

Luminance

Per-pixel luminance values. Assuming an image colour space of sRGB, this is given by:

$$lum = 0.2172 \times R + 0.7152 \times G + 0.0722 \times B \tag{2.15}$$

RAS

Rasches method

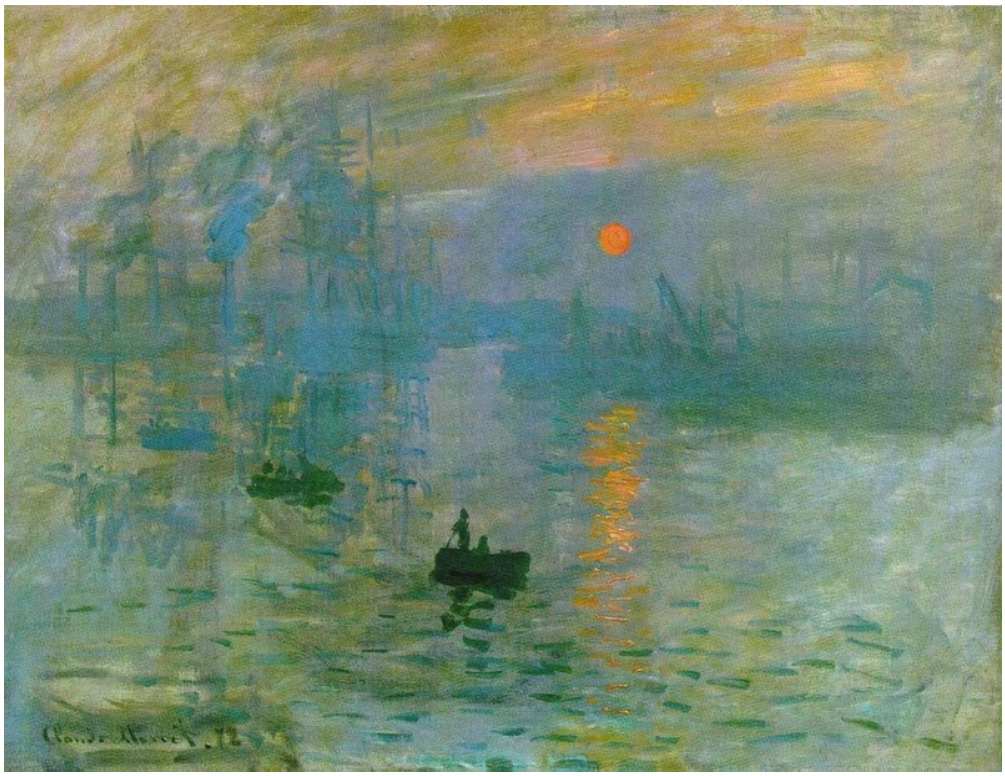
Rasche et al. (2005a,b)

SOC

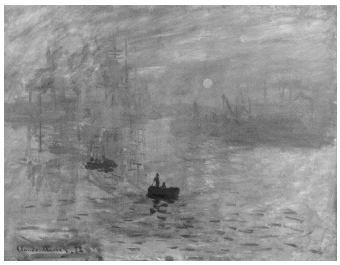
Multispectral Image Visualization Through First-Order Fusion

Socolinsky and Wolff (2002)

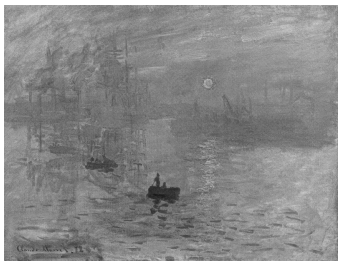
The results of applying these operators to an example image (here we use the 'Monet' scene from the dataset detailed in appendix B) are shown in fig. 2.5. This illustrates why the colour-to-greyscale problem has been the focus of so much research. In the 'LUM' image (fig. 2.5e), the 'sun' disappears and the semantic meaning of the reproduction is changed. In some sense a good grey scale reproduction should convey the same meaning as the colour original.



(a) Original Image

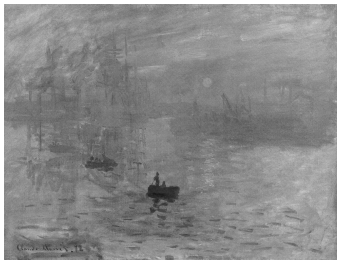


(b) ALS

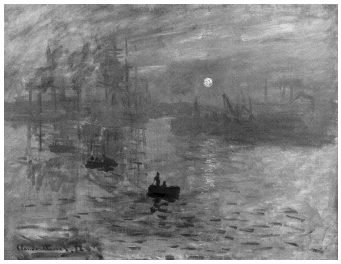


(c) BAL

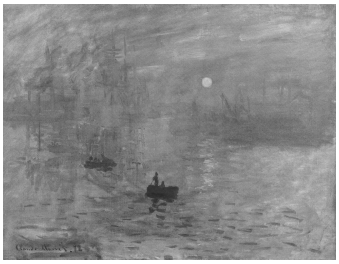
(d) GRU



(e) LUM



(f) RAS



(g) SOC

Figure 2.5: Colour-to-greyscale operators applied to ‘Monet’ scene

2.5 Paired Comparisons

In chapter 3, we shall be considering observer preference experiments between the algorithms described in sections 2.3 and 2.4. One of the most widely adopted standards for modern preference experiments is ISO 3664 (ISO, 2009), which makes assertions about many factors such as display calibration and ambient illumination. Like all standards, ISO 3664 was developed with input from many different organisations and academics. It represents a 'best practice' guide and is, in effect, a reasonable compromise to the methods employed in individual organisations. The factors that are frequently highlighted in the literature – which, it should be noted, are not necessarily specific to paired comparisons – are summarised below.

The display device used in the experiment should be calibrated to a standard such as sRGB (Stokes et al., 1996). This assures that the colour balance and intensity are regulated such that they can be accurately reproduced by other experimenters. The control of display characteristics is also tightly coupled with regulation of ambient illumination in the viewing environment. The room should be dimly lit, but not dark (to avoid eye strain), by a controlled illumination device at a specific colour temperature – D65 in most cases. The illumination source should be placed behind the display device, to avoid glare, and the display should have a 'hood' placed over it. The observer should also be given time before commencing the experiment to allow their eyes to adjust to the viewing conditions of the room. Observers are often pre-screened to determine any colour vision anomalies and only colour-normal observers are considered. Of course there are cases where non-colour-normal observers are desired, but for generic observer preference care should be taken to avoid skewing results with colour-anomalous observers.

The observer is positioned, and the interface designed, so that the observed images subtend an observable angle at the observer's retina which is below some threshold – such as 10°. This can be achieved simply by seating the observer at a fixed distance

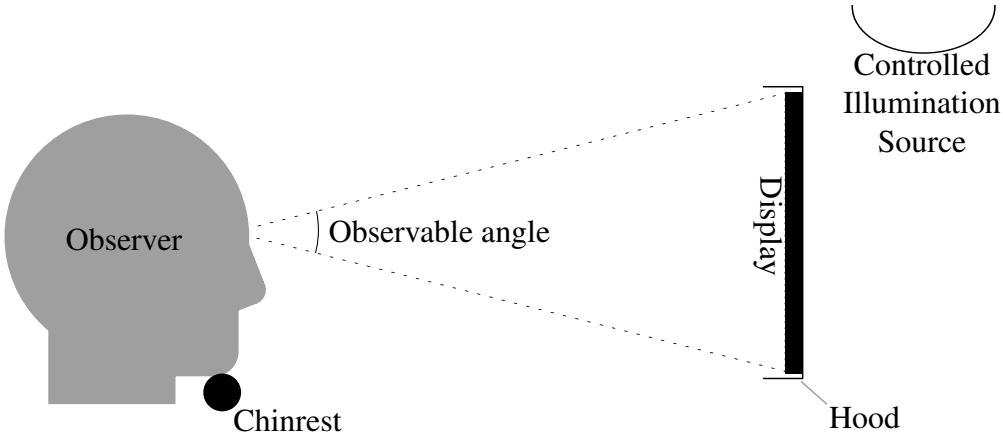


Figure 2.6: Experimental setup

from the display. If the exact viewpoint needs to be controlled, the observer might be asked to view the images with the aid of a chin rest, as in fig. 2.6. As well as taking care over the observable angle, it is also important to display the images against a neutral grey background – often a variegated or 'checkerboard' pattern as in fig. 2.7 – and to ensure that the images are visually separated.

There is some flexibility in how observers make a preference judgement. For example, an observer can be given some fixed amount of time to observe the images – ten seconds for example – after which point the images are removed and the observer makes their judgement. Alternatively, observers might be permitted to make their decision as soon as they see fit and be given as long as they wish to evaluate the images.

Typically, preference experiments are used to evaluate the outputs of several algorithms. For a single image processed by N algorithms there are 'N choose 2' pairs. Thus, clearly there are a large number of pairwise judgements to be made for even a small number of algorithms. Further, to remove bias in the selection procedure it is accepted practice to show each pair more than once with the order (which image is on the left) altered. Given that the number of comparisons which an observer must complete increases rapidly as the number of differing algorithms increases, the time taken to complete each comparison can be important in tackling eye strain and boredom among

Figure 2.7: Typical interface of a paired comparison experiment

observers. To ease observer fatigue, a preference experiment can be split into sessions typically lasting no more than thirty minutes. Often, experimenters limit the number of images and algorithms under review so that the whole experiment can be undertaken in thirty minutes (no more than one session).

2.6 Analysis of Paired Comparison Experiments

2.6.1 Thurstone's Law of Comparative Judgement

When seeking a preference metric of the perceived quality of several differing image treatments, an intuitive approach is to compare every treatment with every other in a pairwise fashion as described above, resulting in a 'tournament' of comparisons where the image that receives the greater preference 'wins' each comparison. The problem

then is aggregating the results from each comparison in the tournament into a definitive collection of preference scores. A common approach to this problem, which is still an active area of research (Keener et al., 1993), is the application of Thurstone's (Thurstone, 1927) law of comparative judgement.

Thurstone proposes that a discriminatory process between two stimuli, causing responses S_A and S_B, can be modelled as a normally distributed random variable, where the distribution represents the value of $S_A - S_B$ over many observations, under the assumption that S_A and S_B are themselves normally distributed. The mean of this distribution should give a good approximation of the true value of $S_A - S_B$. This approach allows us to make an estimate of the scale of $S_A - S_B$, even though observers do not make any explicit assertions of that scale, rather they are only ever asked to judge which of the two stimuli produces the 'greater' response. To accomplish this, Thurstone adopts some sets of assumptions, grouped by various cases which may apply to the experimental design. Here we shall only discuss case V, which is the most commonly applied case in the imaging science literature, and the case which we apply for the work described later in the thesis.

Given two stimuli, where the response S_A is judged as greater than S_B, it is not assumed that these responses will always be unanimous, owing to the variances in the scale values, σ_A^2 and σ_B^2. Rather the proportion of times that S_A is judged greater that S_B will give rise to a normal distribution such as that shown in fig. 2.8, where the shaded area represents the proportion of time that treatment A is preferred over treatment B, $P\left(S_A > S_B\right)$ or, equivalently, $P\left(S_A - S_B > 0\right)$. The case V solution imposes the assumption that $\sigma_A^2 = \sigma_B^2$. The value of σ_A^2 and σ_B^2 can be set at any arbitrary value, normally 1 is used such that the standard deviation of the distribution $S_A - S_B$ is $\sqrt{2}$. Given these assumptions, the measured value of $P\left(S_A > S_B\right)$ should follow the normal cumulative distribution function

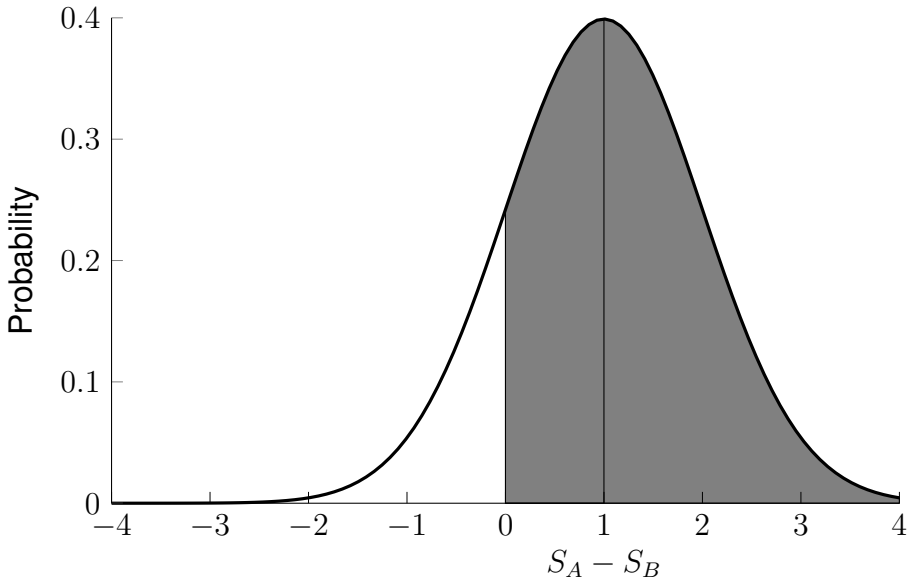


Figure 2.8: Distribution of $S_A - S_B$

$$H\left(S_A - S_B\right) = \frac{1}{2\sqrt{\pi}} \int_0^{+\infty} \exp\left(-\frac{1}{2}\left(\frac{t - \left(S_A - S_B\right)}{\sqrt{2}}\right)^2\right) dt, \tag{2.16}$$

with mean $S_A - S_B$ and standard deviation $\sqrt{2}$. There are other suitable alternative cumulative probability functions for H, which are discussed by Engeldrum (2000). From here, given the assumption $H\left(S_A - S_B\right) = P\left(S_A > S_B\right)$, it is possible to determine the scale value difference $S_A - S_B$ by inverting $H\left(\cdot\right)$. This gives the relation

$$S_A - S_B = H^{-1}\left(H\left(S_A - S_B\right)\right) = H^{-1}\left(P\left(S_A > S_B\right)\right). \tag{2.17}$$

The above description handles only the trivial case of comparing two differing treatments. To extend the model into the case of three or more stimuli, we can insert the probabilities of all the comparisons in the tournament into a *proportion matrix*. So, for the case of three treatments t:

$$P = \begin{bmatrix} P(S_1 > S_1) & P(S_1 > S_2) & P(S_1 > S_3) \\ P(S_2 > S_1) & P(S_2 > S_2) & P(S_2 > S_3) \\ P(S_3 > S_1) & P(S_3 > S_2) & P(S_3 > S_3) \end{bmatrix}. \tag{2.18}$$

To construct P, we first create a frequency matrix F, where each f_{ij} is a tally of the number of times that treatment i is preferred over treatment j. We then normalise F by the total number of observations n to give P. We can then construct our final *score matrix*, S, by calculating the score differences

$$S = H^{-1}(P) = \begin{bmatrix} S_1 - S_1 & S_1 - S_2 & S_1 - S_3 \\ S_2 - S_1 & S_2 - S_2 & S_2 - S_3 \\ S_3 - S_1 & S_3 - S_2 & S_3 - S_3 \end{bmatrix}. \tag{2.19}$$

From this score matrix we can then derive final score values for each treatment considered. We can observe from eq. (2.19) the sum of the first row of S

$$\frac{1}{t}\sum_{i=1}^{t}(S_1 - S_i) = S_1 - \bar{S}. \tag{2.20}$$

If we assume that the mean of the scale values $\bar{S} = 0$, then we can directly calculate the scale value S_1. Repeating this procedure for each row of S gives us the scale values for every treatment. Note that this summation method only suffices when the score matrix is *complete* – that is, every observer has completed every preference choice for every pair of treatments i and j. However, this is not always the case (as will become apparent in chapter 3). Under such circumstances it is said that the score matrix is *incomplete* or *unbalanced*, and to calculate the final score values the summation method is usually replaced with a least-squares approach (Morrissey, 1955).

It is important to note that this procedure will find the correct scale values assum-

ing that the underlying model assumptions hold. Testing the validity of these model assumptions is discussed in section 2.6.2.

For all the following statistical definitions, we continue to use the following notation: n = number of observers, t = number of algorithms (or *treatments*), F = frequency matrix, P = proportion matrix, S = score matrix. However, in some applications we modify n to be the number of *observations*, as it is common to structure experiments such that each observer views every image pair in both $[AB]$ and $[BA]$ orientations. Further, each of these orientations may be repeated, giving a total of four (or more) repetitions for each image pair.

2.6.2 Mosteller's Test

As described above, Thurstone's case V solution makes several assumptions about the data being analysed. Specifically that the variances for the underlying discriminal processes are equal and that the coefficient of the correlation between observer responses is zero. However, there are occasions when these assumptions do not hold and the case V solution is inadequate. To detect these situations, Mosteller (1951) put forth a chi-square test to evaluate the goodness-of-fit of the model to the data. This test is based on an arcsine transformation of a matrix of probabilities reconstructed from the final score matrix. This reconstruction converts the proportions into normal deviates with a mean value of zero and a constant variance $1/n$.

First we have to construct a matrix of probabilities assuming normal deviates (based on the assumptions of the case V solution). This matrix P' is constructed from the relation

$$p'_{ij} = P\left(S_i > S_j\right) = \frac{1}{\sqrt{2\pi}} \int_{-(S_i - S_j)}^{+\infty} \exp\left(-\frac{t^2}{2}\right) dt. \tag{2.21}$$

Once we have these probabilities p'_{ij}, as well as the original proportions p_{ij}, we

convert these into angles in radians, θ'_{ij} and θ_{ij}, by using the arcsine transformation

$$\theta'_{ij} = \sin^{-1}\left(2p'_{ij} - 1\right). \tag{2.22}$$

This transformation approximates H^{-1} from eq. (2.17) and converts the binomially distributed proportions into asymptotically normal random variables. With the values for θ'_{ij} and θ_{ij}, we can now calculate the χ^2 statistic, as formulated by David (1988), as

$$\chi^2 = n \sum_{i<j} \left(\theta_{ij} - \theta'_{ij}\right)^2, \tag{2.23}$$

where n is the number of observers. The degrees of freedom for the test are

$$(t-1)(t-2)/2. \tag{2.24}$$

When the χ^2 value obtained from this test is lower than the χ^2 value at some significance level p (for the given degrees of freedom), we cannot reject that, at that significance level, p_{ij} and p'_{ij} are from the same distribution, and so we accept that the case V solution is suitable for this data.

2.6.3 Score Difference Test

Upon compilation of a Thurstonian analysis, the outcome is a collection of assignments of scores to image treatments. From these scores it is possible to generate an ordinal ranking. However if the scores for two different treatments only differ by a small amount, we may be hesitant to assign a definitive ranking. To quantify this uncertainty, we can use the *score difference test*, described by Ledda et al. (2005).

This test groups a collection of scores such that two scores within the same group cannot be declared significantly different at a given significance level. Formally, we are grouping the scores (where *score*, in this case, refers to the row sum f_i of the frequency

matrix) so that the variance-normalised range of the scores within each group is less than or equal to some value R_α^+.

Calculating R_α^+ is equivalent to finding some R' such that $P\left(R \geq R'\right) \leq \alpha$. The distribution of the range R is asymptotically the same as the distribution of a variance-normalised range, W_t, of a set of normal random variables with variance $= 1$ and *t* samples (David, 1988). This gives us

$$P\left(W_{t,\alpha} \geq \frac{2R - \frac{1}{2}}{\sqrt{nt}}\right), \tag{2.25}$$

where $W_{t,\alpha}$ is the value of the upper percentage point of W_t at significance level α, which is tabulated in many statistics texts, e.g. Pearson and Hartley (1966). From here we can directly calculate the value of R_α^+ given the value of $W_{t,\alpha}$:

$$R_\alpha^+ = \left\lceil \frac{1}{2} W_{t,\alpha} \sqrt{nt} + \frac{1}{4} \right\rceil. \tag{2.26}$$

To this resultant integer value, R_α^+, we ascribe the following quality: if the score difference between two image treatments is less than R_α^+, those two treatments cannot be described as perceptually different at the chosen significance level, α.

2.6.4 Kendall's Coefficients of Consistency and Agreement

Kendall Coefficient of Consistency

As well as applying the widely used Thurstonian analysis to produce scores for our images, we can also use the frequency matrix F to derive some extra statistics which help to explain the behaviour of observers.

We would hope that, in general, observers are *consistent* when they make their preference choices. An inconsistency, in this case, refers to the situation where an observer prefers image A over B, and image B over C, but then prefers image C over A. Kendall and Smith (1940) define such an occurrence as a *circular triad*, and they can occur in

situations where the compared stimuli do not elicit very different responses, meaning that the observer has difficulty differentiating between them or, specifically to cases in image preference, in situations where different image treatments perform well in some image regions but not others, and the observer then chooses different image regions on which to base their preference for one comparison than they do for another.

When only a small collection of stimuli are being compared, it is simple to count these violations of consistency directly from the matrix F. However, for larger values of t, Kendall and Smith (1940) describe a process for calculating the frequency of the inconsistencies:

$$c = \frac{t}{24}\left(t^2 - 1\right) - \frac{1}{2}z, \tag{2.27}$$

where

$$z = \sum \left(f_i - \frac{(t-1)}{2} \right)^2, \tag{2.28}$$

and $f_i = \sum_{j=1}^{t} f_{ij}$, the row sum of the frequency matrix *for a single observer*.

Kendall and Smith (1940) compare the calculated count of inconsistencies to the maximum possible number of inconsistencies for that value of t. This normalised measure of consistency, Ω, has a maximum value of one in the case where there are no violations of consistency, and decreases to zero as the observed inconsistencies increase.

$$\Omega = \begin{cases} 1 - \dfrac{24c}{t^3 - 4t} & t \text{ even} \\ \\ 1 - \dfrac{24c}{t^3 - t} & t \text{ odd.} \end{cases} \tag{2.29}$$

Low values for Ω can be interpreted as an indicator that a particular observer was poor at making consistent preference choices. Alternatively, if Ω is low across many ob-

servers, it is an indicator that the stimuli being judged were too similar for the observers to make consistent choices.

It is important to note that, when giving summary statistics for an experiment, Ω is calculated separately for each observer and then averaged across all observers.

Kendall Coefficient of Agreement

Suppose we have a frequency matrix compiled for n observers. Each coefficient in this matrix can assume values in the range $0, \ldots, n$. The value of f_{ij} will be n iff $f_{ji} = 0$, which means that every observer agreed unanimously on their preference of the image pair $[i, j]$. If all observers are in complete agreement for every image pair then there will be $\binom{t}{2}$ coefficients equal to n, and $\binom{t}{2}$ equal to 0, with the remaining t elements lying on the diagonal. It would be entirely possible for this situation to arise even in the case of extremely low consistency. Conversely, a situation of complete disagreement would be evident if each coefficient had the value $n/2$ when n is even, or $(n \pm 1)/2$ when n is odd. If two observers make the same preference judgement on a pair of images $[i, j]$, we denote this as one agreement. It is possible to calculate the number of pairs of observers in agreement over each pair of images as in Kendall and Smith (1940):

$$\Sigma = \sum_{i \neq j} \binom{f_{ij}}{2}. \tag{2.30}$$

Σ is now a count of the total number of observed agreements. To convert this into a useful measure of agreement, we must normalise it by the maximum possible number of agreements given n and t

$$u = \frac{2\Sigma}{\binom{n}{2}\binom{t}{2}} - 1. \tag{2.31}$$

This gives the final measure of observer agreement, which can range from 1 in the case of perfect agreement, to $-1/(n-1)$ when *n* is even, and $-1/n$ when *n* is odd.

To gain some significance measure of the coefficient of agreement, we can use the χ^2 test described by Ledda et al. (2005) to test the null hypothesis that all observers made their preference judgements entirely at random.

$$\chi^2 = \binom{t}{2} (1 + u\,(n-1))\,. \tag{2.32}$$

This test has $\binom{t}{2}$ degrees of freedom.

2.6.5 Comparing Thurstonian Analyses

Kendall Rank Correlation Coefficient

To compare the results of two variations of a paired comparison preference experiment (as in chapter 3), we need a measure of computing the correlation between the two. Given the ordinal nature of the ranking derived from the scores output from a Thurstonian analysis, it follows to use a rank-correlation statistic such as Kendall's τ (Kendall, 1938).

To compute this statistic from two rank orders, those rankings must first be rearranged so that one is considered as a 'correct', or objective, order. For example, consider the two rankings A and B:

$$A = (2, 1, 5, 4, 3)$$
$$B = (1, 3, 4, 5, 2).$$

To rearrange these rankings, considering A objectively, the elements of A are rewritten such that they are in increasing order, while maintaining the corresponding elements of B:

$$A' = (1, 2, 3, 4, 5)$$
$$B' = (3, 1, 2, 5, 4).$$

Once this reordering is completed, a measure, k, of the ordered pairs within the ranking B' can be calculated:

$$k = \sum_{i<j} \begin{cases} 1 & \text{if } B'_j > B'_i \\ 0 & \text{otherwise} \end{cases} . \tag{2.33}$$

This can then be normalised to give the correlation coefficient τ.

$$\Sigma = 2k - \binom{t}{2}, \tag{2.34}$$

$$\tau = \frac{2\Sigma}{t(t-1)}. \tag{2.35}$$

Kendall (1938) gives a method for computing a significance measure, *p*, for τ. This measure is based on the likelihood of the observed correlation occurring given two independent variables. A low value for *p* indicates that a correlation to the extent of τ is unlikely to occur and so we reject the null hypothesis that the two variables are independent.

Sprow et al. Chi-Squared Goodness-of-Fit

From the Thurstonian analysis, we have access to more than just ordinal rank data. The scores give scale values as well as a rank ordering. In light of this, there may be some situations where a rank correlation statistic does not tell the whole story. Consider the scenario with three treatments *A*, *B*, *C* as shown in fig. 2.9. Shown are the scores

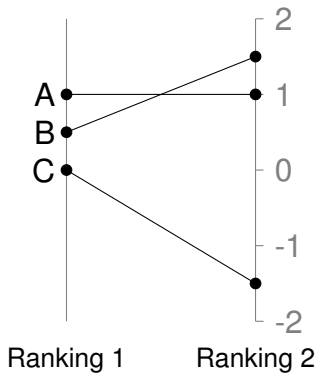


Figure 2.9: Rank position swaps do not reveal scale of score differences

across two experiments producing two different rankings. While the score for *A* remains constant across both rankings, *B* and *C* receive different scores. What is revealed is that while *B* exhibits a relatively small change in score (from 0.5 to 1.5), it creates a rank position swap. Meanwhile the difference in score for *C* is larger (from 0 to -1.5), but does not result in a rank position swap. Rank correlation statistics do not expose these situations and can penalise small changes in score while not penalising large changes.

To address this, Sprow et al. (2009) devised a χ^2 statistic, similar in construction to Mosteller's test (Mosteller, 1951). Instead of comparing the observed results of the experiment to an expected distribution based on normal deviates, this test treats one experiment as the 'observed' data, and the other as the 'expected' data. This statistic is defined as:

$$\chi^2 = \sum_{j<l} \left(\frac{n_{jl} \cdot n'_{jl}}{n_{jl} + n'_{jl}} \right) \cdot \left(\arcsin\left(2p_{jl} - 1\right) - \arcsin\left(2p'_{jl} - 1\right) \right)^2, \tag{2.36}$$

where P and P' are the proportion matrices of the 'expected' and 'observed' data respectively, and N and N' are matrices representing the total number of comparisons per pair in each of the experiments. This statistic accommodates for differing numbers of observers (and thus differing variance) between the two experiments and, due to its formulation, allows for unbalanced experiments, where each image pair is not necessarily

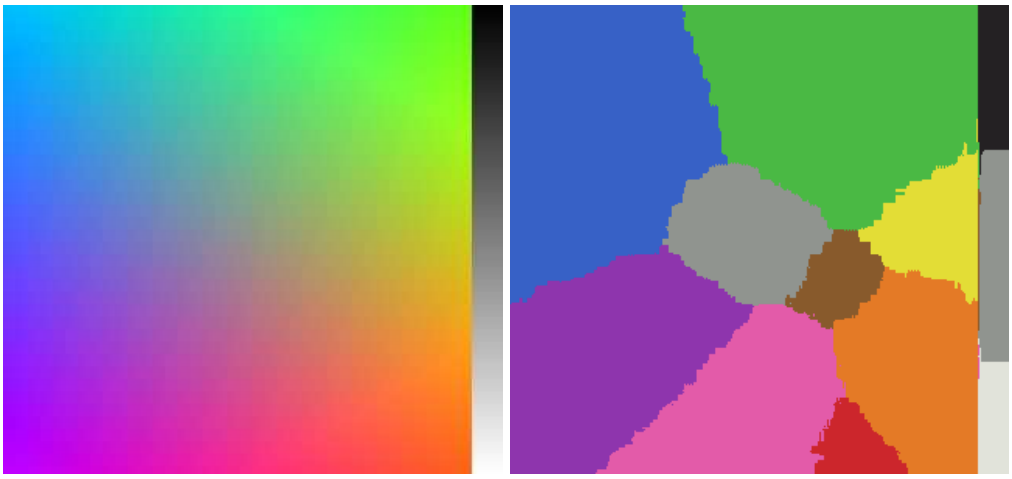
viewed an equal number of times to every other pair.

Much like Mosteller's, this test examines at what significance level can we assert that p_{ij} and p'_{ij} are from two different distributions. As such, and in juxtaposition with the significance measure for Kendall's τ, a low p–value from this statistic indicates a poor correlation.

2.7 Computational Colour Naming

In computer vision and image processing, computational colour naming may be defined as the task of assigning perceptual colour name labels to given numeric colour descriptors. For example, given the RGB triplet $(1, 0, 0)$, we would probably be seeking the colour name label "red". Colour names are important not only because they provide an efficient quantisation of a colour space, but also because they are of perceptual relevance to ourselves. Thus, we seek not only a mapping from numbers to names but a mapping that will make – at least broadly – the same colour designations that we do as humans. In this work, we do not seek to introduce new methods of computational colour naming – the colour naming model will be used as input to further processing in chapters 5 and 6 – and so here we provide a brief background and introduce our preferred method.

Clearly colour names are complex; we each have different perceptions of colour and differing vocabularies with which to describe them. Individuals also have differing levels of quantisation by which they categorise names: that which one person may label "pink", another may subdivide into "magenta", "fuchsia", "salmon", "bubblegum" etc. In spite of all this diversity, Berlin and Kay (1969) found that, across many languages and cultures, colour names develop hierarchically from a set of *basic colour terms* which slowly expands as a language evolves; English is considered to have eleven basic colour terms. However, this remains a contentious subject – authors who have studied cultures with differing 'visual diets' (Roberson et al., 2005, 2000) argue



(a) Input image *(b) Image reduced to eleven basic colours*

Figure 2.10: Colour name labelling

that cultural and linguistic differences can affect colour perception, while others maintain that colour names have a universal basis (Kay and Regier, 2003).

In light of this, it is unreasonable to conjecture a perfect computational solution to the problem – the objective is not yet well-defined. For our purposes, a reasonable expectation is that a colour naming model be able to correctly label pixel values with one of the basic English colour terms in such a way that is largely agreeable to a majority of human judgements, as in fig. 2.10, where each pixel in fig. 2.10b is coloured with the representative pixel value for that colour name. For the case of the $(1, 0, 0) \leftarrow red$ example this seems straight forward enough, but on the continuum between "red" and "orange", humans will place the boundary point at different locations from each other. To further exacerbate this, humans will make different choices under different circumstances – under different viewing conditions, if the colour is on a textured surface etc. For the purposes of this thesis however, we are not concerned with these details. All we desire is a model capable of reducing complex three-dimensional colour spaces into a discrete set of labels, in such a way that is largely harmonious with English speakers.

The literature introduces many computational approaches to colour naming, such as

simple multivariate probabilistic methods (Chuang et al., 2008; Heer and Stone, 2012) to more finely-tuned parametric approaches such as the *Triple Sigmoid with Elliptical center (TSE)* approach introduced by Benavente et al. (2008). We found that a model based on multivariate probability density functions suffices to arrive at a color naming model that meet the needs of the experiments in chapters 5 and 6.

Given some training data (discussed in chapter 5) of human labelled pixel values, we train a model comprising of a multivariate probability distribution for each colour name - we used the eleven English colour names defined by Berlin and Kay (1969). So, given a collection of RGB values which are labelled "pink" a model can be constructed from the mean and covariance of those data points, under the assumption that the distributions of those values are normal (which is likely an incorrect assumption in reality, but practically suffices). We experimented with a mixture of Gaussians approach (that is each individual colour name is represented by multiple Gaussian probability distributions, each with a weighting factor), and found that results were improved only marginally, and so for the sake of simplicity we continue with the use of a single multivariate probability distribution for each colour name.

Equation (2.37) recapitulates the general formulation of a multivariate normal distribution probability density function with mean μ and covariance matrix $\mathbf{\Sigma}$, where $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$.

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \tag{2.37}$$

For the case of our colour naming model in three-dimensional RGB space, $k = 3$ and $x_1, \ldots, x_k = R, G, B$. We assemble a full collection of these functions $f_1, \ldots, f_{11}$, one for each of the colour name categories.

After constructing this model, we can generate a likelihood vector $\underline{\lambda}_x$ of the prob-

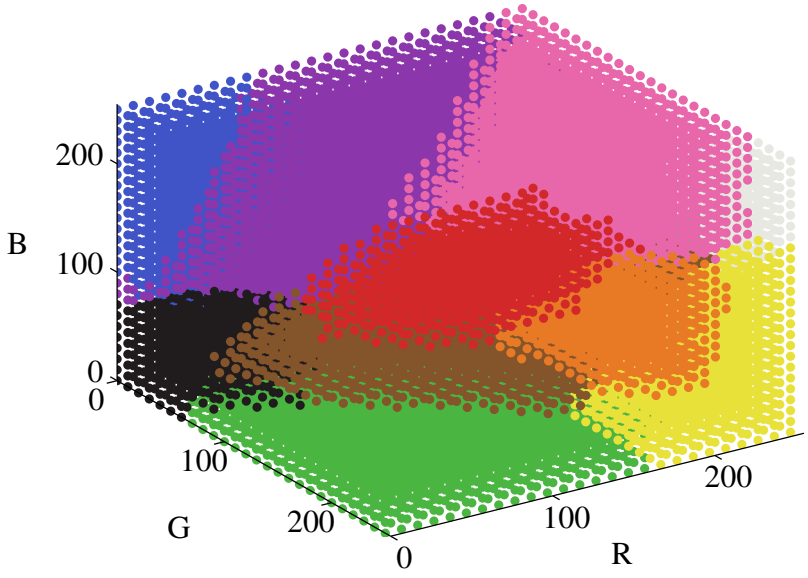


Figure 2.11: RGB cube populated with probability distributions for each colour name

abilities that x can be correctly labelled with each of the eleven colour names

$$\lambda_x = [f_1(x), \ldots, f_{11}(x)]. \tag{2.38}$$

From this likelihood vector we need to select one colour name to assign to x. The simplest approach to this is to choose the maximum likelihood

$$k_{max} = \arg \max_{k=1,\ldots,11} \{f_k(x)\}. \tag{2.39}$$

The name label for x is then the colour name associated with the index k_{max}

$$Label(x) = Label_{k_{max}}. \tag{2.40}$$

Figure 2.11 shows a depiction of the RGB cube sampled at uniform locations, the dot at each sample location is coloured with the representative pixel value for the colour name attributed to that sample location.

2.8 Object Indexing

The recognition and identification of objects in a scene is a foundational topic in the field of computer vision. At a very high level, we can decompose the task into two main subproblems – *a*) distinguishing individual objects in a scene and correctly isolating them from their surroundings, and *b*) identifying and understanding individual objects. Although it is of great interest, we shall not be considering the first of these subproblems in this work.

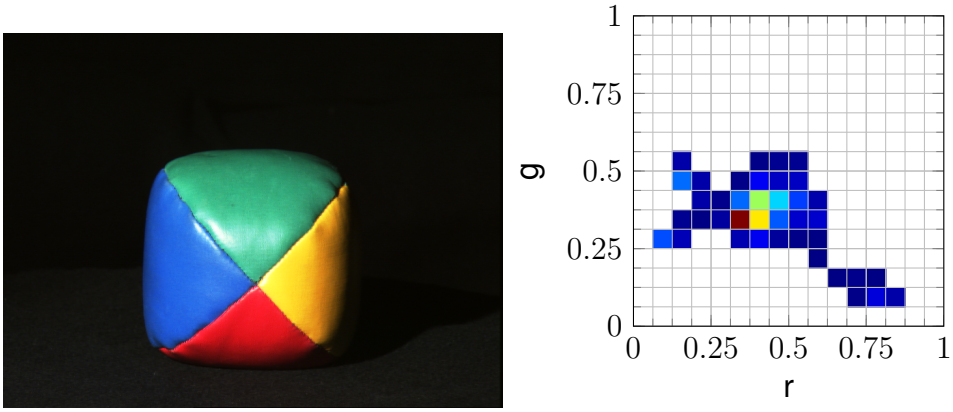
The second part of the problem, object identification, has many candidate approaches; objects carry many cues which can be useful for identification. Important among these cues are shape and colour (Beis and Lowe, 1994, 1997; Berens et al., 2000; Berretti et al., 2000; Chen et al., 1999; Hassan et al., 2009; Mahmoudi et al., 2003; Qiu, 2002; Schettini et al., 2002; Swain and Ballard, 1991; Yu, 2009). The geometric shape of an object can allow us to pattern-match and identify objects against a database. If we then introduce some machine learning techniques it is possible to identify objects by their subcomponents (Felzenszwalb and Huttenlocher, 2005), e.g. this object has four legs, therefore it might be a horse (of course with only this information to work with, we could also be looking at a chair – such techniques require more complex inputs in reality). If we stick with our four-legged example, it is easy to understand how matching the shape of a horse against a reference shape can lead to identification – many children's educational books will use examples such as this in teaching children to identify animals. However, such an approach is very fragile: a side-on view of a horse is very different from a front-on view. We could expand our reference dataset with many views of horses, but it might be more appropriate to consider some extra cues. Adding colour-based recognition provides some key benefits, such as rotational invariance and (limited) tolerance to changes in point-of-view; in our example case it will also give us the added benefit of determining horses from zebras. Clearly this is a complex problem, and approaches that utilise many cues will give the best results, but a comprehensive

review of the state-of-the-art is outside of the scope of this thesis. Instead, we shall be focussing on a solely colour-based object indexing approach by Swain and Ballard (1991).

The work of Swain and Ballard (1991), which is now approaching twenty-five years of age, has grown old gracefully and stood the test of time. Despite the fast pace of innovation in the field, Swain and Ballard's technique remains well-discussed and well-referenced, due in part to the fact that it is a very powerful, yet at the same time very simple approach to object indexing. Central to the technique is the concept of comparing the distribution of colours in a query object image to the distributions of colours in a database of reference object images. The query object is matched to the object in the database with a colour distribution most closely matching that of the query object image.

To perform this colour distribution matching, a database of *model images* must first be compiled. It should be emphasised here that this technique can only match against known objects – a query object of an apple can only be successfully identified if the database contains an image of an apple with the same (or very similar) colour distribution. Moreover we shall not discuss any safeguards against false positives – if a query image has no correct match in the database then the closest match will always be returned (this could be mitigated to a certain extent by requiring a certain threshold for a positive match, but this will not be considered here).

Once a corpus of model images has been assembled, each image is processed to acquire a colour histogram for each object in the database (every object image must first by masked to exclude any background pixels). To do this, the colour space (assume RGB for now) of the images is quantised into a predefined number of *bins*. For example, we could partition the RGB cube into $4 \times 4 \times 4$ bins, but the binning does not necessarily have to be symmetrical – we could use $8 \times 4 \times 2$ if we so wished. We will refrain from discussing optimal binning strategies at this stage, as this will be discussed later in the



(a) Example object *(b) Histogram*

Figure 2.12: Example object with corresponding histogram. For ease of visualisation, the shown histogram is in two dimensions, as described in section 2.8.1

thesis. After quantising the colour space, the number of pixels with colour values that fall into each bin are counted for each image; i.e. with a $2 \times 2 \times 2$ binning strategy, we would, for each image, calculate an 8-component vector where the first component contains the sum of the number of pixels which fall into the first bin etc. This vector represents the histogram for each image, and is the key for that object by which we will query the database. Figure 2.12 depicts an example image along with its corresponding histogram (for the two-dimensional (r, g) chromaticity representation – discussed more in section 2.8.1).

Figure 2.13 outlines how the technique is used to match a query image against the compiled database. Given a query image, we first calculate the query image histogram via the method described above, and then compare it to each histogram in the database. The model object with the most closely matching histogram is identified as the matching object.

To compare two histograms, Swain and Ballard define a measure of *histogram intersection* as:

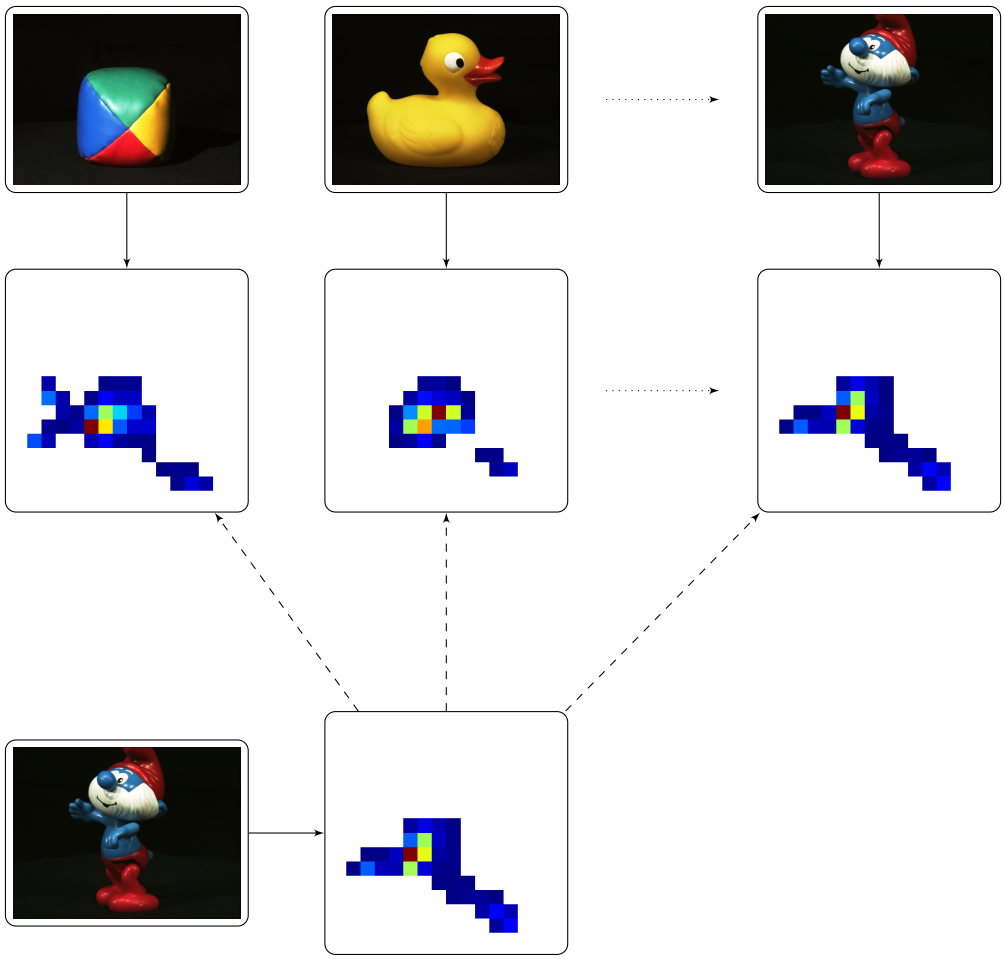


Figure 2.13: Summary of Swain and Ballard's (Swain and Ballard, 1991) histogram-based object indexing

$$H(T, M) = \frac{\sum_{j=1}^{n} min(T_j, M_j)}{\sum_{j=1}^{n} M_j}, \tag{2.41}$$

where T is the histogram of the test (or query) image, M is the histogram of the model object image, and n is the total number of histogram bins. This measure gives a score between 0 and 1, where 1 indicates that the histograms are identical. Selecting the closest match is as simple as selecting the model object image associated with the histogram with the highest score.

2.8.1 Object Recognition in Chromaticity Space

In their original work, Swain and Ballard construct their image histograms in three-dimensional RGB space as described above. However, for many applications, this introduces some problems. The RGB colour space encodes chromatic values as well as intensity information. The latter may be undesired, e.g. when the same object appears brighter or darker in different images it is the intensity information that causes the discrepancy. Once intensity is factored out the chromaticity values remain constant. To compound this issue, this object indexing approach is often used in concert with various illuminant estimation approaches and, as mentioned in section 2.2, many illuminant estimation techniques are unable to recover intensity information. We will further examine the implications of illuminant estimation for object recognition later in this thesis.

To circumvent these issues arising from constructing image histograms from a three-dimensional colour space which encodes intensity, some authors (Berens et al., 2000; Funt et al., 1998) choose to first convert their object images into a two-dimensional chromaticity space. In such colour spaces the intensity information of RGB is effectively discarded and we are left with just the chromatic component of the colour information. A commonly used standard chromaticity space is the (r, g) space:

$$r = \frac{R}{R+G+B}, \qquad g = \frac{G}{R+G+B}. \tag{2.42}$$

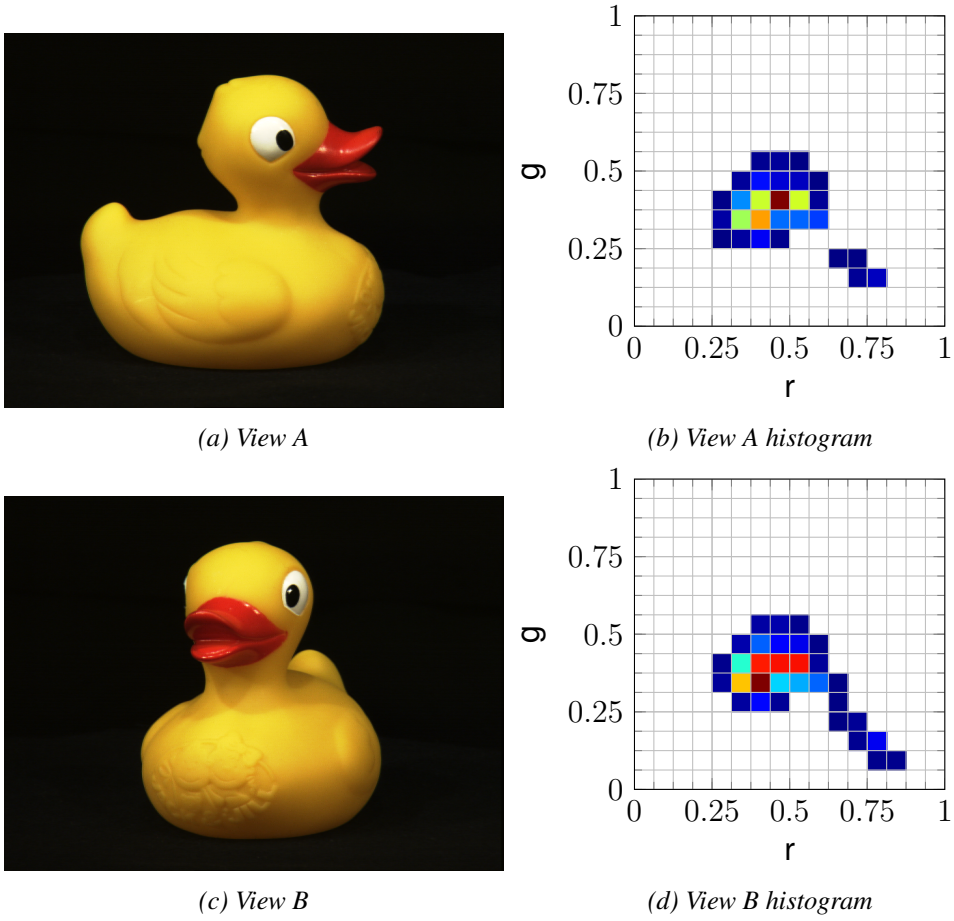
With object images converted into this colour space, we can now construct our histograms in two dimensions. In so doing, we may lose some discriminatory power when querying our database (this will be discussed later in the thesis), but we gain resilience to changes in intensity arising from varying exposures or illuminant estimation.

By constructing an object image database as described above, Swain and Ballard showed that they were able to distinguish between objects in a comprehensive object

database to a satisfactory degree (how to measure the precision of the matches returned by the approach is discussed below). The method is also resilient to changes in rotation of objects, moderate changes in point-of-view (as shown in fig. 2.14), partial occlusion and, if two-dimensional histograms are used, changes in intensity. The approach does have some distinct drawbacks however. Firstly, it is only capable of indexing images of single objects in isolation, when in reality objects are usually viewed as part of a broader scene with many other objects and varying backgrounds. Segmenting an image to extract the individual objects contained within is a nontrivial exercise in itself. Further, as alluded to when discussing the inclusion of illuminant estimation into the method, changes in illumination conditions can be extremely detrimental to the method. Although we can discard the intensity information and be resilient to the inability of many illuminant estimation techniques to recover intensity, slight errors in the recovery of the chromatic component can be very problematic (Finlayson et al., 2002a; Funt et al., 1998).

2.8.2 Evaluating Object Recognition Performance

To quantify the effectiveness of the method, we need a measure of correctness for object recognition. Swain and Ballard (1991) noted that the method does not necessarily return one object as the match candidate, rather it can be seen as a method of sorting the entire database by likelihood of matching the query object. From this observation they introduced the notion of a *rank* for the results output from the method. That is, if the resulting sorted list of database entries correctly places the desired object at the top of the list, then the rank is 1; if the method fails to deliver the correct result directly, but has at least placed the correct result in second place in the sorted list, then the rank is 2, and so on. From this rank concept, Swain and Ballard introduce a measure they refer to as the *match percentile*:



(a) View A *(b) View A histogram*

(c) View B *(d) View B histogram*

Figure 2.14: Two views of same object, with corresponding histograms

$$MP = \frac{N_{models} - rank}{N_{models} - 1}, \tag{2.43}$$

where N_{models} is the number of model objects in the database. This measure gives scores between 0 and 1. While a score of 1 suggests that the object was correctly recognised and placed at the top of the sorted list, 0 indicates that the correct match was in fact placed at the bottom of the sorted list. We can multiply this value by 100 to obtain a percentile score.

2.9 Discrete Relaxation

Many problems within the field of image processing (among many others), and the algorithms which are employed to solve them, are based on the concept of *constraint propagation* (Boykov et al., 2001; Hummel and Zucker, 1983; Pelillo, 1997; Waltz, 1975). Broadly speaking this is a computational approach which allows local constraints to be propagated across a global solution.

A trivial example would be to arrange the letters A, B and C using the constraints that:

1. A must come before B and
2. B must come before C.

If we consider only the first rule we could satisfy the constraint with any of the orderings ABC, ACB or CAB; if we only consider the second rule then any of ABC, BAC or BCA suffice. It is only by taking the two rules in unison, and *propagating* the constraints, that we come to the correct ordering ABC.

If we this reformulate this ordering challenge as a graph problem, we can picture a three-node graph as in fig. 2.15, where each node has to be *labelled* with one of A, B or C.

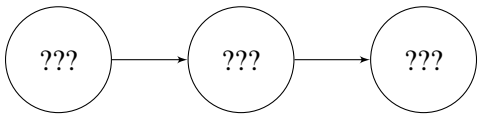


Figure 2.15: A to-be-labelled three node graph

Figure 2.16 shows the process of applying a discrete relaxation approach to our graph labelling problem. Not shown is a final pass through all nodes that would need to be made, wherein it is noted that no changes are made to each node. Once every node has been visited with no modifications taking place, we can return our final labelling.

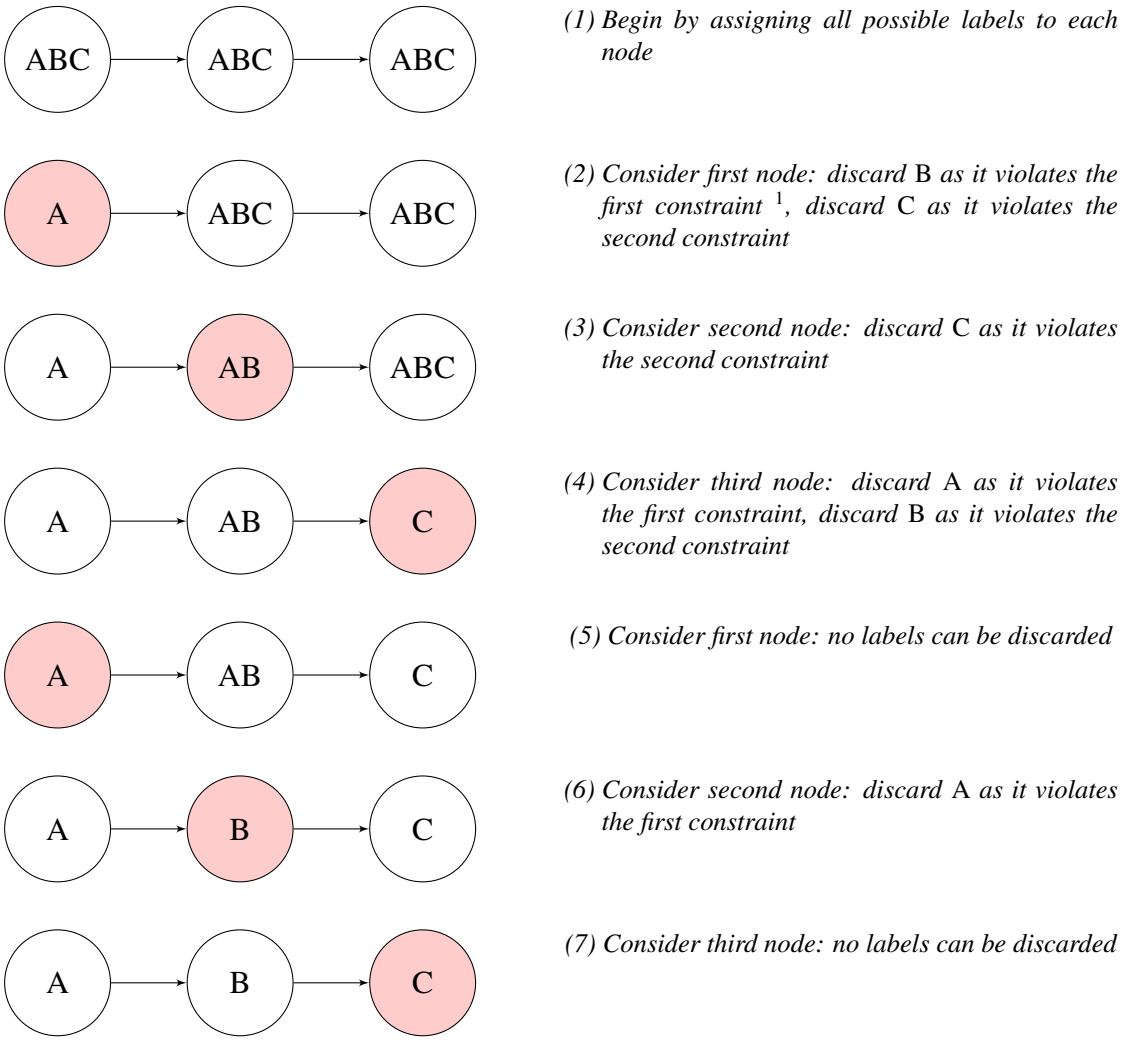


Figure 2.16: Solving a trivial labelling problem with discrete relaxation

We now introduce some vocabulary. In the language of relaxation labelling, the constraints described for our trivial example above are known as *binary constraints*, because they describe constraints that operate on relationships between two labels. We also have the concept of *unary constraints*, which constrain single labels in isolation; for example, we could have introduced the constraint "*A* can only be assigned to the first node" (indeed, if we had utilised that constraint in the walk through in fig. 2.16 we could

[1]The constraint is that "*A* must come before *B*" – a corollary of this is that *B* must follow *A*.

have arrived at the solution rather sooner). As we discarded labels, we did so because they were *inconsistent* with the constraints under consideration. All remaining labels are said to be *consistent*, and the final product is deemed to be a *consistent labelling*.

With our vocabulary in place we can introduce an informal description of the discrete relaxation algorithm as follows:

1. Assign labels to each object, adhering to unary constraints.
2. For each object in turn, consider the binary constraints and delete the inconsistent labels for that object.
3. If any object has no remaining consistent labels, stop – there is no consistent labelling available. Otherwise, repeat Step 2 until a consistent labelling is found.

This description outlines discrete relaxation at a conceptual level, but to begin to implement it we need to introduce some formalisms. Let

- $U = \{u_1, \ldots, u_n\}$ be a collection of n objects (the nodes in the graph in our example problem)
- $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ be a set of m labels (*A*, *B* and *C* in our example)
- E be the set of edges representing relationships between objects in U
- L be an $n \times m$ binary matrix where

$$L_{i,j} = \begin{cases} 1 & \text{if } u_i \text{ can be consistently labelled with } \lambda_j \\ 0 & \text{otherwise} \end{cases} . \tag{2.44}$$

- R be an $n \times n$ set of $m \times m$ binary *compatibility matrices* where

$$R_{i,j}(l,m) = \begin{cases} 1 & \text{if } u_j \leftarrow \lambda_m \text{ is consistent with } u_i \leftarrow \lambda_l \\ 0 & \text{otherwise} \end{cases} . \tag{2.45}$$

We can begin by constructing an initial labelling L under the conditions of the unary constraints. Each object to be labelled is represented by a row L_i in the matrix L, if λ_j satisfies the unary constraints for that object then the element in the j^{th} column of that row ($L_{i,j}$) will be 1, else 0.

We then need to construct a set of compatibility matrices R, representing the binary constraints. There will be one matrix for each relationship between any two objects ($n \times n$, although an optimised algorithm will not require a fully populated set of matrices), and each matrix will have the same number of rows and columns as the number of possible labels m. $R_{i,j}$ will contain a 1 in the l, m^{th} element if λ_m does not break the binary constraints for U_j, given that U_i has been labelled with λ_l. Each compatibility matrix only describes the relationship between one pair of objects in isolation, it does not encode any restrictions on the possible labels for those two objects given their context outside of that one relationship.

From this starting point we can run a consistency algorithm, such as that in algorithm 1, to propagate the constraints and arrive at a final labelling. Algorithm 1 is taken from Henderson (1990), which also gives a much more thorough introduction to discrete relaxation techniques in general.

This algorithm, and any other boolean discrete relaxation algorithm, prunes the possible labels that could be designated for each object. It does not guarantee that the final labelling will result in each object having one unique label assigned to it. In the extreme, consider the case where the initial labelling L contains all 1s, as does every compatibility matrix – this is a perfectly valid (empty) set of constraints and a labelling with no

Algorithm 1 A queue-based consistency algorithm, from Henderson (1990)

```
1:  function CONSISTENT
2:      Q ← {(i, j)|(i, j) ∈ E and i ≠ j}
3:      while Q ≠ ∅ do
4:          remove (k, m) from Q
5:          if ¬SUPPORT(k,m) then
6:              Q ← Q ∪ {(i, k)|(i, k) ∈ E and i ≠ k and i ≠ m}
7:          end if
8:      end while
9:  end function
10:
11: function SUPPORT(i,j)
12:     consistent ← true
13:     for all λ ∈ L_i do
14:         support ← false
15:         for all λ′ ∈ L_j do
16:             support ← (support or R_{i,j}(λ, λ′))
17:         end for
18:         if ¬support then
19:             consistent ← false
20:             L_i ← L_i − {λ}
21:         end if
22:     end for
23:     support ← consistent
24: end function
```

discarded labels would be a perfectly valid output of the algorithm. Conversely, it is perfectly plausible to construct a set of constraints which are impossible to consistently propagate, and it is feasible to attain a labelling with no consistent labels for one or even all objects. The more likely outcome is that the algorithm will discard some, but not necessarily all, labels for each object and pruning the remaining labels will require further processing dependent on the task at hand.

For a more real-world application of this method in image understanding, we turn to scene labelling. Consider the scene depicted in fig. 2.17a. We wish to label each object in that scene with the following possible labels:

B Background (the wall in the image, but we will use "Background" as we have another label using the letter W)

C Cupboard (or drawers, but we have another label using D)

D Desk

F Floor

L Lamp

M Monitor

W Window

In fig. 2.17b we can see a segmentation of the scene and all possible labels applied to each segment. Note that the segmentation shown is illustrative only and is not intended to be accurate – it serves to guide the eye to the correct area of the scene.

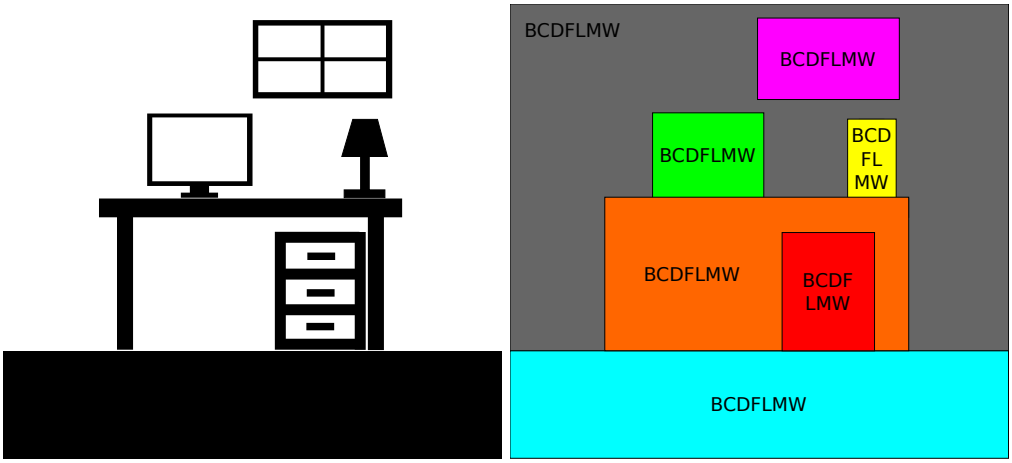
If we now introduce the following unary constraints, we can arrive at the labelling in fig. 2.17c:

- Background touches the top image border
- Cupboard is rectangular
- Floor touches the bottom image border
- Lamp is not rectangular
- Monitor is rectangular
- Window is rectangular

These unary constraints *can* be fairly powerful on their own – notice how the lamp (the yellow segment) has been reduced to two candidate labels by only its shape and position. However, if we now introduce the following binary constraints, we can use discrete relaxation to arrive at the final consistent, unique, labelling seen in fig. 2.17d:

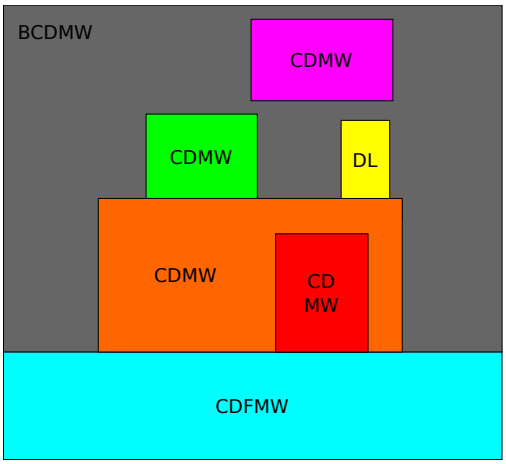
- Cupboard is under Desk
- Desk is below Window
- Floor is below all other labels
- Lamp is on top of Desk
- Monitor is on top of Desk
- Window is above Desk
- Window is surrounded by Background

Clearly this scene diagram and the constraints introduced above still only serve as an example. For a real-world scene labelling task we would need more advanced constraints.

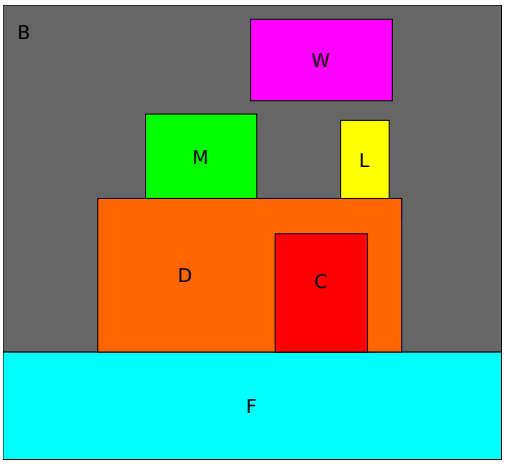


(a) Base Scene

(b) Initial labelling



(c) Labelling after enforcing unary constraints



(d) Final consistent labelling

Figure 2.17: Example scene labelling

Chapter 3

Web-Based Paired Comparisons

Paired comparison experiments are frequently used to gather observer preference data in many areas of image enhancement. However, due to the large quantity of comparisons each individual must complete, these experiments are typically carried out with few observers. Taking this method onto the web is a quick way of gaining a larger number of observers and preference judgements. This chapter examines the validity of web-based paired comparisons and whether the loss of control over viewing conditions causes significantly different results when compared to a lab-based alternative.

3.1 Introduction

The images we encounter every day are often the products of long chains of image processing algorithms. The display version of an image is often extremely different from the raw image data that was captured by the imaging device. The pipelines that take raw images and transform them into display-quality imagery are often exceedingly complex (the example in fig. 2.2 represents a stripped down process to highlight the key components), and each component has been tuned to meet the needs of that particular pipeline. Irrespective of the niche served by a specific pipeline (be it optimising for print, white balance, etc.), the parameters involved will have been tuned to optimise some measure of 'goodness', as determined by some human observer(s). 'Goodness', however, is a vague and subjective notion, and so when tasked with quantifying it many researchers will seek some more well-understood proxies such as brightness or contrast. But, it is often precisely this judgement of sheer observer *preference* that camera manufacturers and purveyors of image manipulation software must address.

This question of observer preference can be evaluated systematically in a psychophysical experiment, whereby two or more pipelines are evaluated in tandem and presented to observer(s). Raw image data for several scenes (which scenes to use will be dependent on the specific task at hand, but a suite of differing scenes should generally always be used to gather data points across as wide a sample of input data as possible) will be processed by the competing pipelines to deliver a collection of image reproductions. These reproductions are presented to an observer and their preference among them is recorded. In so doing it is important that care is taken with the preparation and presentation of the images, for we do not wish the results of our experiment to depend upon conscious or unconscious biases, or upon artefacts of how the images are viewed. The observer should be blinded to which pipeline produced which reproduction, and in many cases it may also be necessary that the observer is also naïve to the purpose of the evaluation. Further concerns are discussed in section 2.5.

For the two-alternative case, where one pipeline *A* is evaluated against some other *B*, a *paired comparison* paradigm is clearly sufficient – the observer can view a reproduction from *A* alongside one from *B* and make a clear indication of which they prefer. By repeating this across several differing scenes and averaging the preference judgements, it is possible to make an assertion as to which pipeline is, on average, preferred. But often we are interested in evaluating many more than two competing pipelines at once. How then, should we handle the addition of more competing pipelines? Perhaps it is still feasible to present reproductions from *A*, *B* and *C* to an observer and ask them for a direct preference, but this task quickly becomes far too hard for an observer as we add more and more options. To circumvent this, we adhere to the paired comparison paradigm and display reproductions to the observer in a pairwise fashion – to evaluate *A*, *B* and *C*, an observer would compare *[AB]*, *[AC]* and *[BC]*. Maintaining a simple binary decision for the observer for each iteration of the experiment keeps the task simple for the observer and so more reliable results should be collected.

However, the paired comparison paradigm is not without issue. As more competing pipelines are added to the experiment, the number of comparisons that an observer must make grows rapidly – for N pipelines there are $\binom{N}{2}$ ('N choose 2') pairs of reproductions. For even a modest number of observers and a small number of repetitions of each comparison, it takes a long experimental session to obtain complete image preference data. Such experimental sessions can be laborious and often boring for observers to complete, and so apathy may begin to affect experimental results. For this and many other reasons, it can be exceedingly hard for researchers to recruit sufficient numbers of observers. Further, the requisite preparation of viewing conditions may be challenging and time consuming for many researchers. So, the premise of conducting a preference experiment can be rather daunting and those that do often have to settle for observer numbers that are lower than may be desired – it is not uncommon to see experiments carried out with fewer than ten observers (Connah et al., 2007), which can be problem-

atic from the viewpoint of statistical significance.

So then, if researchers must continue to undertake paired comparison preference experiments (and the lack of a more efficacious alternative dictates that we do), is there anything that can be done to ease the practical burden? Some experimenters have begun to carry out pairwise comparison studies over the internet using an interface implemented in a regular web browser. Web-based paired comparison experiments can certainly provide a quick and easy method of gaining a potentially very large number of participants in exchange for a minimal amount of time and effort on the part of the researcher. But do these benefits come at the cost of reliable data?

Web-based experiments can lack the control over confounding variables that lab-based studies provide, and as such it is not obvious that they will deliver data that are useful. Of course controlled, lab-based, image viewing was adopted for a reason; we cannot, for example, calibrate a remote observer's monitor. However, it can be argued that having no control over these confounding variables gives a more 'real-world' representation of observers, and that the effects of the variance in these conditions will become minimised as the numbers of observers and differing viewing environments increase. Given a greater set of preference data we would like to be able to arrive at stronger conclusions (and so make stronger recommendations about which reproductions perform most favourably).

In this chapter we take an empirical approach to evaluating the validity of data acquired by web-based paired comparison experiments. First, we examine an existing web-based paired comparison experiment (Mei, 2010a), concerned with tone mapping operators for high dynamic range scenes, by carrying out a lab-based replicate and cross-examining the results from the web-based variant. In so doing we find that observer judgements made in the web-based experiment differ markedly from those made by observers in our lab (under controlled standard viewing conditions).

Learning lessons from the shortcomings of the existing web-based experiment, we

then construct our own web-based experimental platform and re-examine the results from the lab-based counterpart in contrast with the new data. Adopting somewhat minimal advances in the control over image presentation proves to be crucial in making the web application work.

These two pieces of evidence suggest that image preference studies can be successfully transplanted to the web so long as sufficient care is taken over image presentation. Significantly, in a third contribution of the chapter we track the similarity or otherwise of the preference results as a function of the number of observations. We do obtain convergence between lab- and web-based preference data but only after *sufficient* preference judgements are made.

3.2 Background

Over the past quarter of a century since its invention, the world wide web has changed many aspects of modern life. Thanks to its unbridled proliferation, the web has given billions access to a vast corpus of information. But in addition to being a fantastic engine for the dissemination of information, the web has also become an excellent tool for harvesting new information. Web-based experiments are widely used in many fields outside of colour and imaging science (Kawrykow et al., 2012; Lakhani et al., 2013; Saunders et al., 2014), and so many attempts have been made to gather data from participants over the web within the field. Several of these attempts are introduced and examined by Birnbaum (2004). However, a large amount of the successful among these studies have followed survey-based formats, suggesting that the presentation and viewing conditions of the experiment have little or no impact on the results gathered. In colour science however, the environment around the participant, the screen upon which they are observing any displayed images, and ambient lighting conditions, along with numerous other factors, can all play a significant role in the participant's responses.

For studies that do concern these factors, we can begin by noting the work of Rasmussen (2008), who used a web-based experiment to investigate defect detection. Observers were presented with two duplicates of the same image, one of which had been modified to exhibit some 'defect', or noise, and the time taken for observers to identify which of the two images was defective was recorded. The results of this experiment were not compared to any lab-based alternative, but as every comparison had a correct answer, the authors could quantify the level of correctness of the observers, which was generally positive. The reported level of engagement was affected by the manipulation of the data: some data points were discarded according to some filtering steps, such as removing user sessions below a certain accuracy level, or excluding observers who did not complete a minimum of one hundred observations. This was done in an attempt to remove the effects of spurious participants. This particular study also required a calibration stage to be completed by observers, and so represents a more restrictive kind of experiment to what we envisage in this chapter. By the time observers actually begin contributing meaningful data, they have already spent some significant time completing the calibration stage. While this can be seen as a good thing – an observer who has invested their time into the experiment may feel more of a sense of ownership and so be more likely to contribute high quality data – it is more often seen as a high barrier to entry, and so discourages potential observers from participating. In an attempt to combat this, observer engagement was encouraged by presenting the experiment in a game-like format: observers were challenged to identify the defects within the quickest time possible. Engagement was further incentivised by the inclusion of a monetary reward for top performers.

In other web-based work where presentation and viewing environment may affect results, Zuffi et al. carried out a web-based readability test (Zuffi et al., 2007) as part of a larger study examining a suite of differing experimental paradigms, with varying degrees of control over viewing conditions (Zuffi et al., 2008). The web-based portion

of this research attempted to isolate the thresholds for lightness differences between text and background colour on web pages, and was compared to a lab-based control experiment. Similar results between the two replicates were indeed found. Unfortunately the authors do not describe the details of their recruitment process for either the lab-based or, more crucially, the web-based experiments. Interestingly however, they do reveal that in this experiment there were actually fewer web-based participants than those in the lab. This is quantified in terms of *observations* as opposed to direct observer numbers, with 664 observations for the lab-based experiment and 546 on the web. From this we can suggest that the web experiment was not well advertised and that perhaps some number of those participating in it were also participants in the lab-based experiment or may have been 'expert observers' (colleagues and friends of the authors themselves). Despite these potential biases, that the two experimental formats produced similar results is encouraging.

Moving on to the research specifically involving the paired comparison paradigm, there has unfortunately been a comparatively small number of paired comparison experiments carried out on the web. Those that have been attempted have shown varying degrees of success, but there has been little effort in empirically comparing the results gathered to any 'ground truth' lab-based data. Some notable attempts to date are studies by Jiang et al. (2011) and Sprow et al. (2009).

As part of a larger study, Jiang et al. (2011) performed a web-based paired comparison experiment and contrasted it with a lab-based counterpart using the same dataset. The wider study involved three experiments concerning reproductions of fine art pieces. Two lab-based paired comparison experiments were carried out, one with and one without a hardcopy original of the art piece present. A variant of the no-hardcopy experiment was then transplanted onto the web – for obvious reasons no web-based version of the experiment with the hardcopy present could be carried out. Interestingly, but perhaps not surprisingly, it was found that the two lab-based variants showed little

correlation, suggesting that the presence of the hardcopy leads observers to make different preference choices. However, and more profoundly for our current objectives, strong correlation at the 95% level was found between the web-based experiment and the no-hardcopy lab-based experiment (i.e. two variants which both lack the hardcopy reference, but which only differ in being lab- or web-based). Eighty-eight observers were reported for the web-based experiment.

A study by Sprow et al. (2009) focussed on web-based and lab-based variants of a paired comparison preference experiment concerning a gamut mapping task. The experiment presented an sRGB reference image as well as two images mapped to various device gamuts by competing gamut mapping algorithms. This study attracted a larger number of participants to the web-based experiment – around 700, and 70 observers for the lab-based experiment, highlighting one of the key motivations for our pursuit of web-based experiments. Generally, very strong correlation ($>$90%) was shown between the two sets of results. These results are important, and promising, for our objective but they do come with some notable caveats: many observers were friends, relatives and coworkers of the authors and many were also recruited by solicitation via the ECI (European Colour Initiative) mailing list, which may have caused a strong bias toward expert observers. Some observers also participated in both variants, with 43 of the 70 observers in the lab variant contributing to the approximately 700 total for the web variant. This particular study utilised a questionnaire and adjustment/characterisation images to gather extra data about observers' display devices. This extra intrusion was kept as minimal as possible, but would likely still drive away a substantial amount of possible observers had they not been recruited directly from the colour community.

In recent years, there have been several web-based experiments in the wider field of colour science, e.g. the colour naming experiments of Moroney (2003) and Mylonas et al. (2013). These examples have been extremely successful in exploiting the power of the internet to collect data at a large scale, and arguably, due to the requirement of such

a large range of participants, could not have been done otherwise. Similarly Darrodi (2012) utilised a large-scale web-based experiment to gather data from a wide range of participants about colour semiotics. Again, this research would arguably not have been feasible without exploiting a web-based paradigm – indeed, the data gathered by the web-based experiment allowed for greater insight and further conclusions to be made that were not delivered by a similar lab-based experiment.

Non-academic projects, such as Munroe's colour naming experiment (Munroe, 2010) (this will be discussed in more detail section 5.2.1), which attracted over 220,000 participants[1], and the 'typewar' platform (Tauber, 2009) show the huge potential for mass data collection and the public interest in scientific research performed in this way. This concept of 'crowd-sourcing' data is not new to the internet, but it has recently undergone a rise in popularity due in part to the surge in adoption of social networking sites and their integration with third-party services. This leads us to a significant concern which researchers should be cognizant of when undertaking web-based research – attracting participants and maintaining engagement. Recruitment through mailing lists and pre-existing contacts is effective, but it carries the problem of introducing a sampling error in that the participants already have a vested interest in the results and/or are expert observers. Casual web users have little or no commitment to the study in which they are voluntarily participating, and the task of keeping them engaged and entertained without introducing bias into the results can be problematic. The offer of a material reward for participation, or for top contributors, has been used in the past but it introduces the problem of participants manipulating the system for their own reward, without taking any care over their responses.

The contributions in this chapter begin by investigating the web-based paired comparison experiment launched in 2010 by Mei (2010a). This experiment collected user preferences of differing reproductions of high dynamic range scenes processed by a

[1]This estimate is based on user sessions, it does not account for participants taking part more than once.

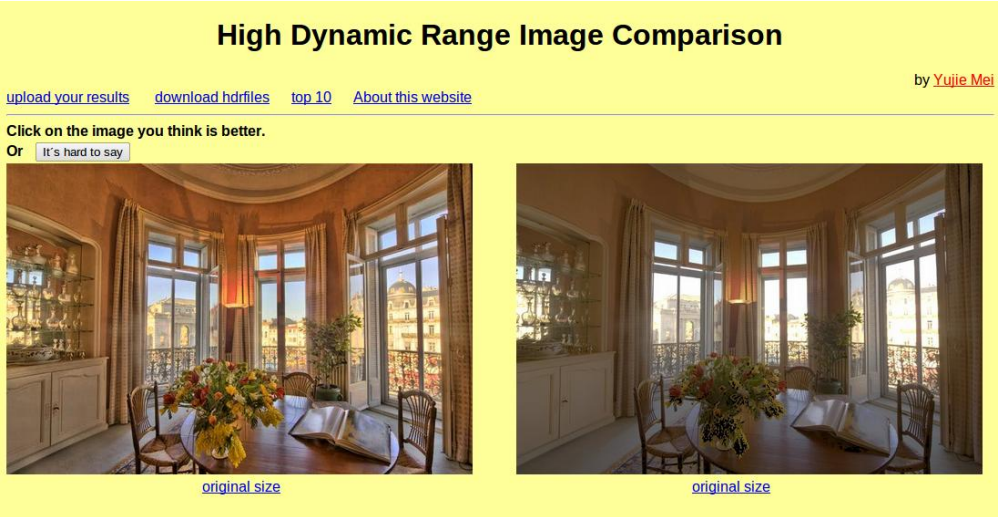


Figure 3.1: Interface of the web-based paired comparison experiment of Mei (2010a)

suite of tone mapping operators, as viewed through a visitor's web browser on their own computer. The full results of this work are reported in Qiu et al. (2011) – in summary, thirteen different scenes (listed in appendix A) were used to evaluate the TMOs listed in section 2.3[2]. Upon arrival at the site the visitor was presented with two images of the same scene treated by two different TMOs, and could click on either one to submit a preference. Alternatively the visitor had the option to click a button to indicate a lack of preference, or a 'tie' situation. The interface was as shown in fig. 3.1. While the experiment was running, the results of the preference choices were collated, ranked and made available online (Mei, 2010b).

We wish to evaluate the validity of the results gathered by Mei's web experiment (hereafter referred to as the *Nottingham-Web* experiment), as contrasted to a lab-based alternative. The next section describes our approach to this evaluation.

[2] Abbreviated TMO names have been kept consistent with those used in Mei's web experiment.

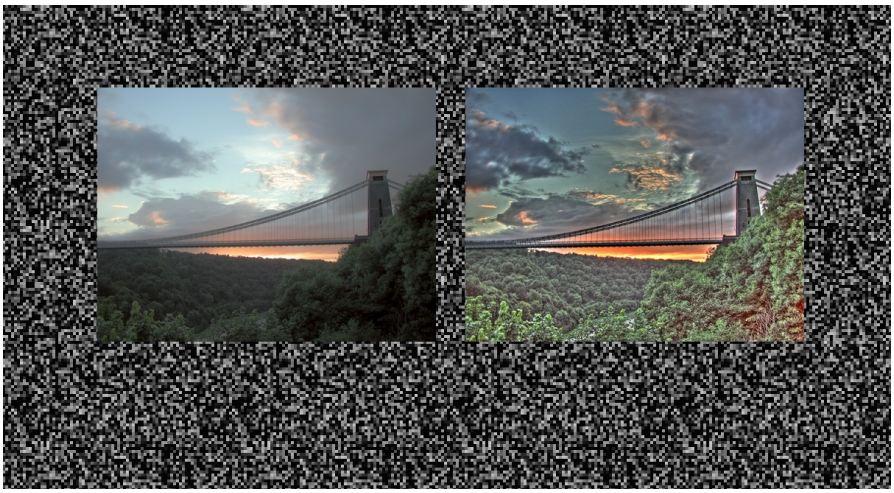


Figure 3.2: Interface of our lab-based TMO experiment

3.3 Experimental Design – Evaluating the Validity of an Existing Web-Based Experiment

To compare results with the web-based research, a controlled paired comparison experiment (hereafter referred to as the *Lab-TMO* experiment) was carried out with fourteen unpaid participants who were naïve to the objective of the experiment. Viewing conditions were prepared in accordance with ISO standard 3664:2009 (described in section 2.5), and images were displayed on a HP LP2480ZX monitor calibrated to sRGB standard (Stokes et al., 1996). The interface of the *Lab-TMO* experiment was as depicted in fig. 3.2. Observers were not positioned with a chinrest, but the chair on which they sat was set at a fixed distance from the monitor and the height was adjusted according to the observer. With this arrangement the average image size subtended at the retina was approximately 6° visual angle, with approximately 1° of padding between the two images. Viewing time was not limited but was monitored – the average viewing time was 5.5 seconds per image pair.

The *Lab-TMO* pairwise comparison was run using the same collection of scenes and TMOs as used in the *Nottingham-Web* experiment. As in the web experiment,

different subsets of the algorithms were used for each of the different scenes. The original reason for the absence of some scene-operator combinations in the *Nottingham-Web* experiment is unknown but, lacking the ability to retroactively acquire any missing data, the *Lab-TMO* experiment used the same subsets for consistency in results. There were 2 scenes for which 6 algorithms were evaluated (giving $(\frac{6\times5}{2})\times$ 2 = 30 pairs), 5 scenes where 7 algorithms were tested (105 pairs), another 4 scenes where 8 algorithms were tested (112 pairs), 1 scene where with 9 algorithms (36 pairs), and 1 final scene with 10 algorithms (45 pairs). In grand total there were 328 pairs of images. Each pair was viewed in $[AB]$ and $[BA]$ orientations, where *A* and *B* are images for the same scene processed by two different tone mapping algorithms, making a total of $328 \times 2 = 656$ comparisons per observer. Due to this large amount of comparisons undertaken, the average observer completed the experiment in one hour, however this was split into sessions lasting no more than thirty minutes each in order to minimise eye strain and loss of concentration among observers.

The images used in the lab-based experimental variant were taken directly from Mei (2010a), and resized with bicubic resampling to fit within the intended observable angle at a standardised viewing distance of approximately one metre. Note that the images displayed to participants in the *Lab-TMO* experiment were exactly the same as in the *Nottingham-Web* experiment (save for displayed size); it is solely the change in environment and presentation which is of interest.

The instructions given to the user in the *Nottingham-Web* experiment were "Click on the image you think is better", with a tie option given as "Or *It's hard to say*" (emphasis indicates clickable button text). The instructions given in the *Lab-TMO* conditions were modified slightly, as the user did not click on images to indicate preference, but had separate physical buttons to select either image or the tie option, as such the instructions given were "Choose the image you think is better, or press [the tie button] if it is hard to say", while the physical buttons were demonstrated.

3.4 Results – Validity of an Existing Web-Based Experiment

It has become commonplace to analyse paired comparison data of this kind by using Thurstone's law (Thurstone, 1927) of comparative judgement, as described in section 2.6.1. However, a Thurstonian analysis of the *Nottingham-Web* study was not compiled, nor is the original raw data available to create one. Instead, the authors used what they called the 'Image Quality Ranking Index' (or IQRI), detailed in Qiu et al. (2011). This index for a particular reproduction t is defined as:

$$IQRI_t = \frac{v}{w_t + \frac{d_t}{2}}, \tag{3.1}$$

where w is the number of wins for reproduction t, d is the number of tie situations involving t, and v is the total number of votes cast across all comparisons involving t; a lower IQRI score indicates a more favourable ranking.

Clearly, we wish to compare our experimental results from *Lab-TMO* with the web-based *Nottingham-Web* rankings (available at Mei (2010b)). In the absence of Thurstone data, we do this by comparing the IQRI rankings of both experiments using the Kendall rank correlation coefficient (Kendall, 1938), as described in section 2.6.5. This is a measure of the level of correlation between two sets of ranked data, giving a score ranging from 1, indicating perfect correlation, to -1, indicating that one ranking is correlated with the inverse of the other. A score of 0 indicates that the two rankings are uncorrelated.

This correlation coefficient was computed for the IQRI rankings for all scenes. Table 3.1 shows, for each scene, the value for the Kendall rank correlation coefficient, τ, and where there is a significant similarity, the corresponding p-value.

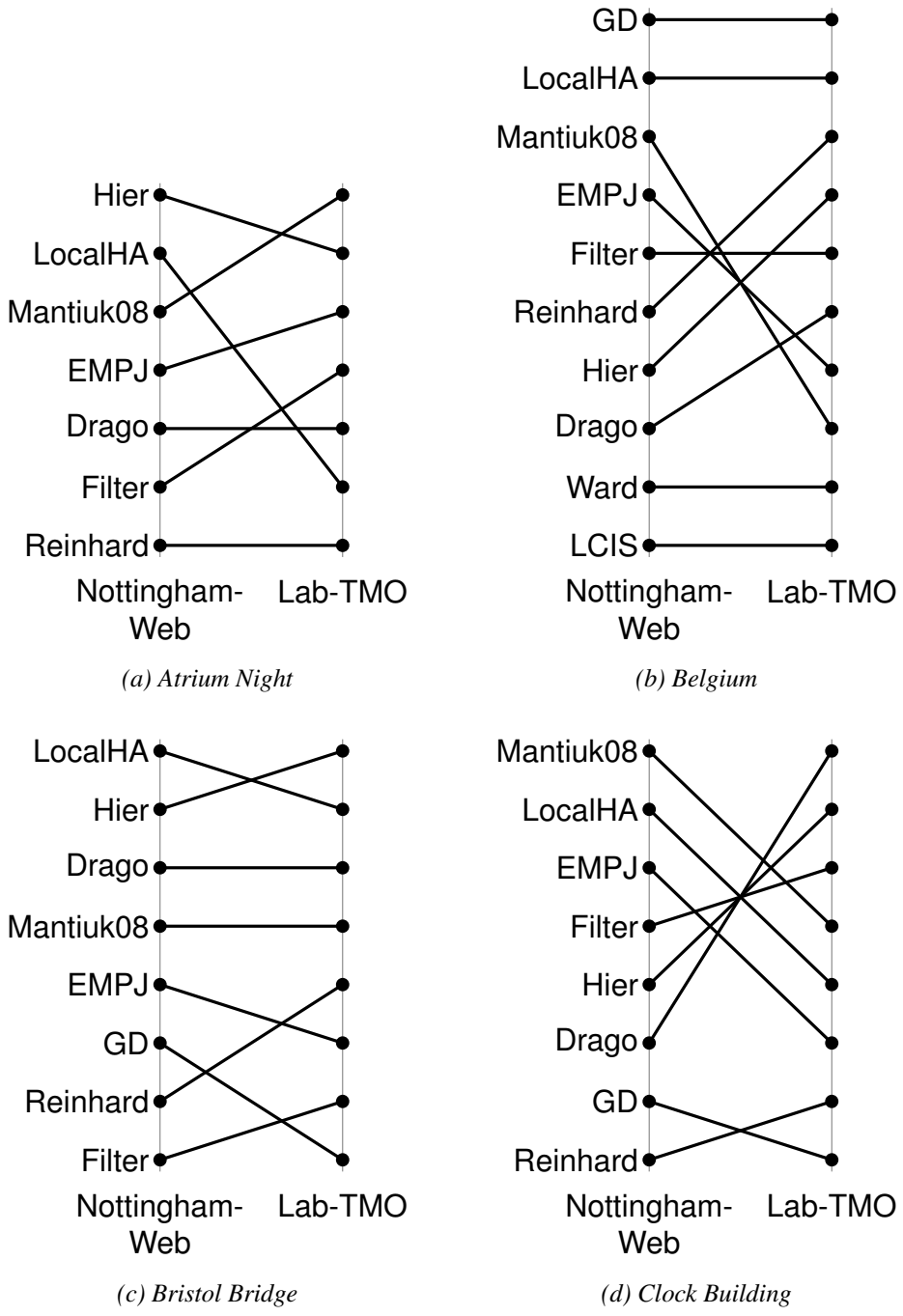
As shown in table 3.1, the 'Synagogue' and 'Tinterna' scenes both have very high rank correlation ($p < 0.01$ and $p < 0.05$ respectively); however, the rank correlations for

Table 3.1: Rank correlations for all scenes in the *Nottingham-Web* and *Lab-TMO* experiments

Scene	τ	Significance
Atrium Night	0.4286	
Belgium	0.5111	$p < 0.05$
Bristol Bridge	0.7143	$p < 0.05$
Clock Building	0.0714	
Fog	0.4444	
Foyer	0.3333	
Indoor	0.5238	
Memorial	0.5000	
Synagogue	0.7857	$p < 0.01$
Tahoe	0.4667	
Tinterna	0.8667	$p < 0.05$
Tree	0.2381	
Venice	0.1429	

the 'Clock Building' and 'Venice' scenes produce drastically different results. Overall, only four of the thirteen scenes produced rankings which were correlated across the two experiments at the 95% level. Figure 3.3 provides a visual representation of the rank correlations for all scenes. These parallel coordinate graphs display the *Nottingham-Web* ranking on the left axis and the *Lab-TMO* ranking on the right. Crossing lines provide a visual cue to the level of correlation between the two rankings.

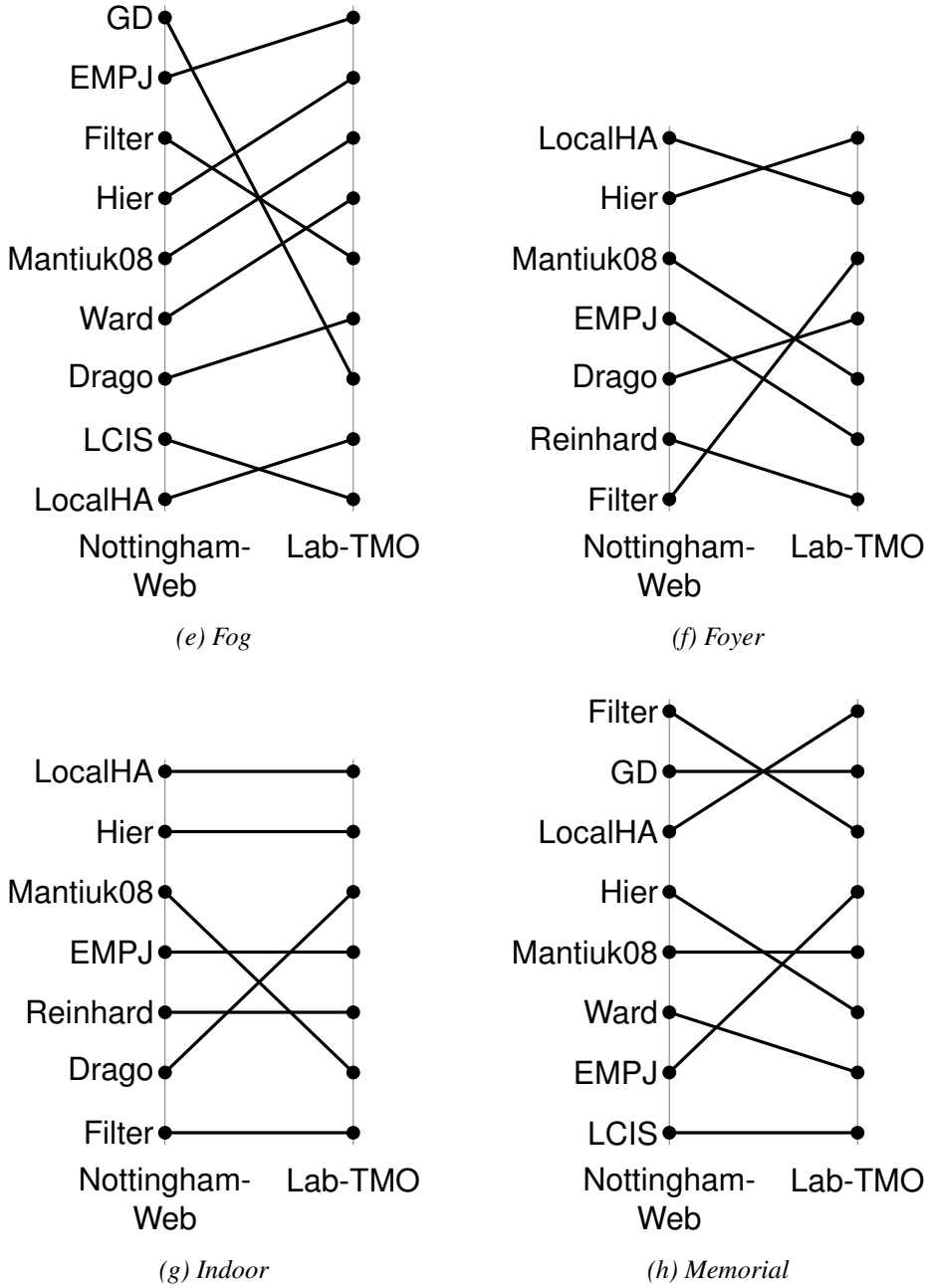
In light of the discrepancy between the two sets of rankings, it is important to evaluate the quality metrics of the *Lab-TMO* data in isolation, so that we can uncover any statistical artefacts which may impact our lab-to-web comparison. The Kendall coefficients of agreement among observers, and of intra-observer consistency were calculated for the data from the lab-based experiment; for future reference, we also calculated the Mosteller score for each scene (all these statistics are described in section 2.6).



(a) Atrium Night *(b) Belgium*

(c) Bristol Bridge *(d) Clock Building*

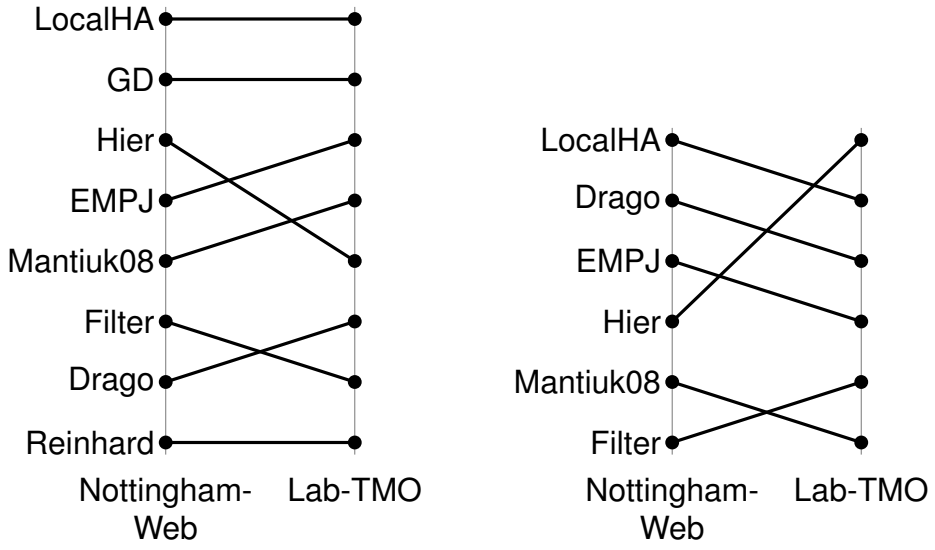
Figure 3.3: Rank correlations between *Nottingham-Web* and *Lab-TMO* variants, for all scenes, based on the IQRI metric



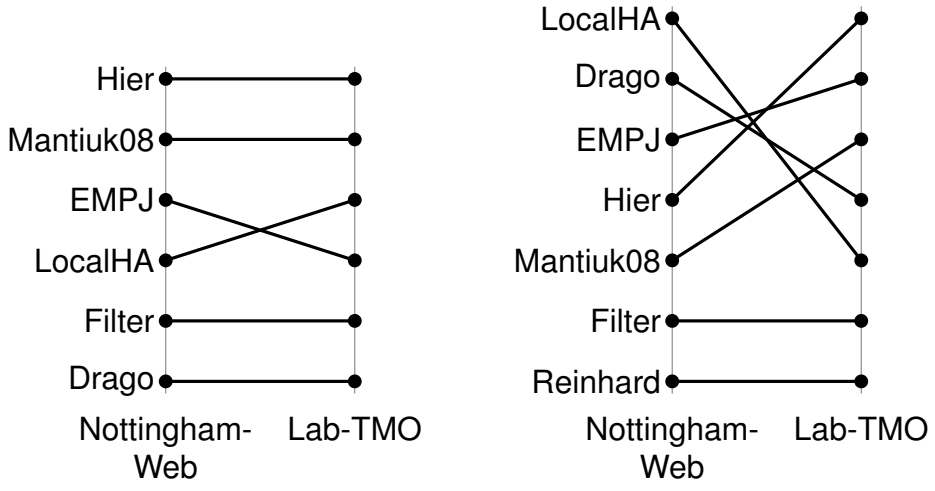
(e) Fog *(f) Foyer*

(g) Indoor *(h) Memorial*

Figure 3.3: Rank correlations between *Nottingham-Web* and *Lab-TMO* variants, for all scenes, based on the IQRI metric *(cont.)*

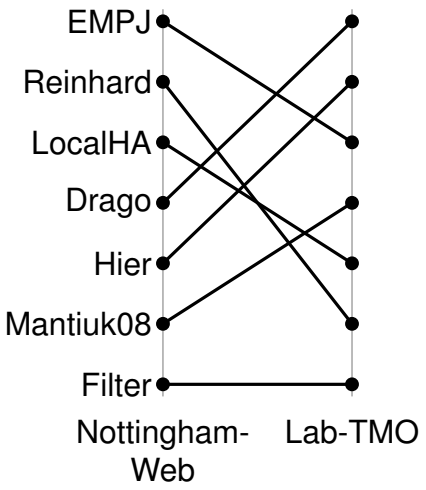


(i) Synagogue *(j) Tahoe*



(k) Tinterna *(l) Tree*

Figure 3.3: Rank correlations between *Nottingham-Web* and *Lab-TMO* variants, for all scenes, based on the IQRI metric *(cont.)*



(m) Venice

Figure 3.3: Rank correlations between *Nottingham-Web* and *Lab-TMO* variants, for all scenes, based on the IQRI metric *(cont.)*

Table 3.2 shows the summary statistics for all scenes; the columns under the 'Agreement' and 'Consistency' headings show that, remarkably, all scenes showed significantly high inter-observer agreement ($p < 0.001$ for all scenes) and also high levels of intra-observer consistency. The lower consistency score for the 'Belgium' and 'Foyer' scenes may suggest that observers were basing their decisions on different image features depending on the image pair presented. Upon inspection of the different reproductions of those scenes, it is evident that some operators perform well in the highlights but fail in the shadows, while some others perform conversely. Observers may have chosen to favour highlight performance for some image pairs, and shadow performance for others.

The columns under 'Mosteller' show the χ^2 score and corresponding significance level (p-values greater than 0.05 are omitted for clarity) for the Mosteller test, which shows that, for the majority of the tone-mapped scenes, the Thurstone Case V solution adequately describes the preference data. However, the significantly high scores for the 'Synagogue' and 'Tahoe' scenes should be noted – these suggest that, for these scenes,

Table 3.2: Summary statistics for all scenes in the *Lab-TMO* experiment

Scene	Mosteller		Agreement			Consistency
	χ^2	Significance	u	χ^2	Significance	Ω
Atrium Night	16.438		0.280	179.643	$p < 0.001$	0.691
Belgium	42.425		0.239	335.571	$p < 0.001$	0.592
Bristol Bridge	15.052		0.222	195.821	$p < 0.001$	0.719
Clock Building	24.126		0.433	355.357	$p < 0.001$	0.800
Fog	13.727		0.229	258.393	$p < 0.001$	0.694
Foyer	6.313		0.155	108.679	$p < 0.001$	0.577
Indoor	13.883		0.194	130.857	$p < 0.001$	0.707
Memorial	18.268		0.252	218.286	$p < 0.001$	0.646
Synagogue	36.019	$p < 0.05$	0.252	218.536	$p < 0.001$	0.815
Tahoe	19.535	$p < 0.05$	0.225	105.929	$p < 0.001$	0.633
Tinterna	18.058		0.274	126.000	$p < 0.001$	0.718
Tree	18.532		0.287	183.679	$p < 0.001$	0.719
Venice	4.668		0.227	149.429	$p < 0.001$	0.694

the assumptions of the Case V solution may not hold and that these scenes should be treated with some caution when we later compare the web-based results to these lab-based results.

The data in this table convey that the observers in our lab made consistent preference judgements, and, finally, that the observers agreed with each other on image preference choices to a significantly high degree.

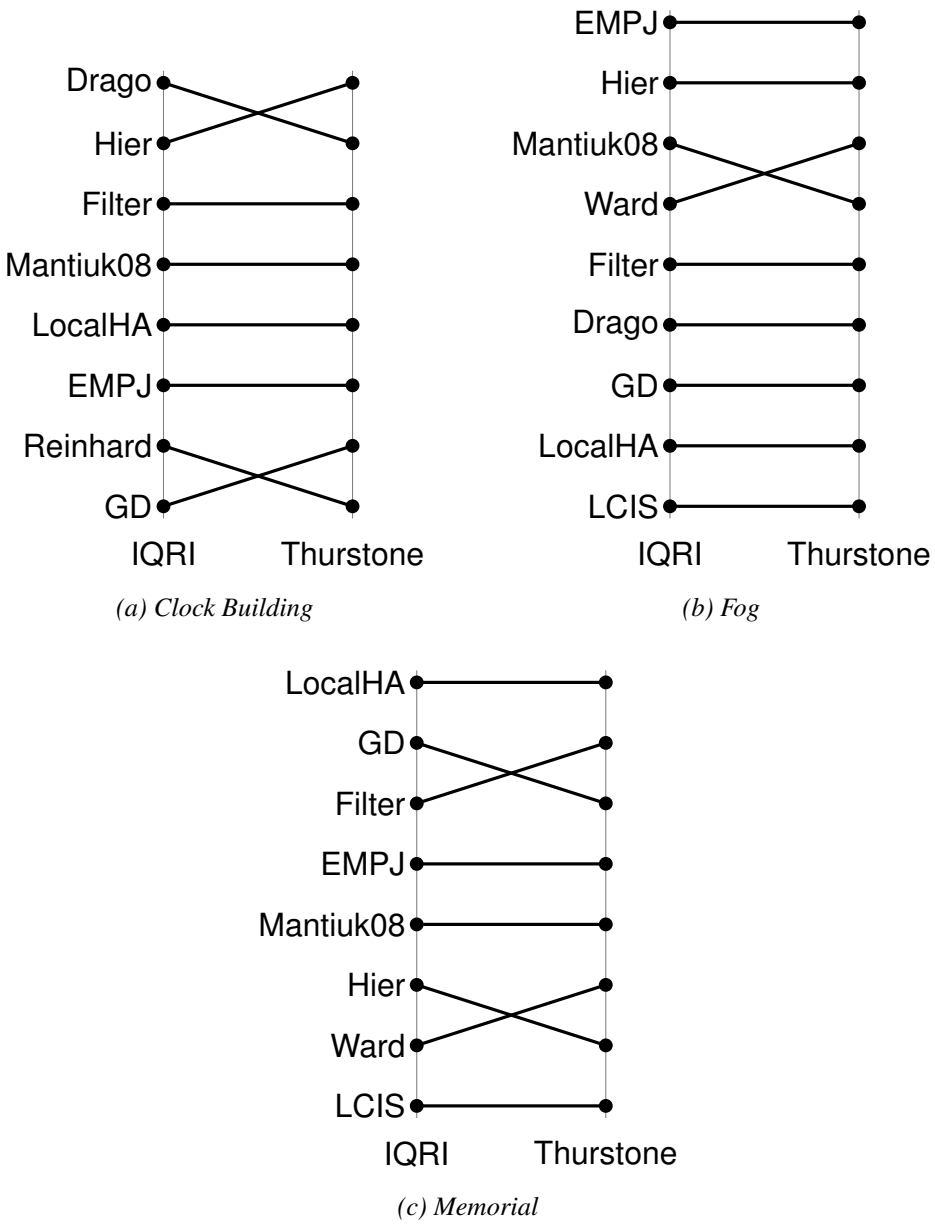
These results may initially seem disappointing in the light of our objective of performing preference experiments on the web in order to make their administration easier for researchers. However, we can begin to suggest some plausible reasons for the seemingly poor performance of the *Nottingham-Web* experiment, which will be discussed in section 3.8. To gain some further insight however, we developed our own web-based experimental platform to perform more in-depth analysis of the contrasting res-

ults between lab-based and web-based experimental variants. In so doing we will be able to generate full Thurstonian analyses for the experiments, and so we will be dispensing with the IQRI metric. We do not assert that the IQRI metric is in any way erroneous, but it is not the commonly accepted standard for the analysis of paired comparison experiments. Table 3.2 shows that the Thurstone Case V solution is, in most cases, sufficient for the task of analysing preference data for these tone mapping operators (although notably not in all cases). If we compare the rankings generated for the *Lab-TMO* experiment by both the IQRI and traditional Thurstone approaches, we find that, broadly, the same rankings are produced. However, the results are not identical for all scenes – fig. 3.4 shows the scenes for which differing rankings are produced.

3.5 Experimental Design – A New Web-Based Platform

Extending from the work in the previous section, we observe that the *Nottingham-Web* experiment suffered from low numbers of participants and did not control for some factors which could still plausibly be controlled and/or monitored even in a web-based scenario (further discussed in section 3.8). In light of this we opted to implement our own web-based research platform (Harris, 2011) so that we could gain greater control over the web-based data collection. We will now compare observer preferences from this new web-based platform and from controlled experiments carried out in our own lab, as introduced in section 3.3.

One of the limitations of the *Nottingham-Web* web-based experiment was that, due to the design of the page and the size of the images compared, only an estimated 20% (Google Inc, 2009) of visitors to the site would be able to observe the entirety of both images in a pair on their screen without scrolling. Worse still, for an estimated 50% of observers the resolution of their display device would cause the page layout to display one image stacked atop the other, meaning that the observer would have to



(a) Clock Building *(b) Fog*

(c) Memorial

Figure 3.4: Correlations of rankings based on IQRI and Thurstone metrics for *Lab-TMO* experiment. Only scenes which do not have perfect correlation are shown

scroll vertically between the two images, and would never be able to make a direct comparison of both images on the screen at the same time. In our system, the layout is fixed so that images will always be shown side-by-side, and the data gathered show that 87% of observers were able to see the entirety of both images at the same time without any scrolling. To facilitate this, we resized the images to a smaller scale than was used in the *Nottingham-Web* experiment. We used the same size images for both our lab- and web-based experiments. All images were resized using bicubic resampling and, for the web experiment, we ensured that there would be no client-side rescaling of the images.

In previous similar experiments, authors have often recruited observers through friends and colleagues, or at conferences or through mailing lists etc. as noted in section 3.2. Obviously this can lead to an unrealistic sample of observer populations, as those recruited from within the community are likely to be expert observers, and anyone who is personally recruited is likely to feel an obligation to complete a large number of preference choices, or to spend more time scrutinising their decisions in order to 'get it right'. We therefore opted against personal recruitment and targeted the wider online audience for our experiments. The project was publicised through social media and advertised through various other websites unrelated to colour science. For example, the project was built using some popular open source tools including *Django* and *Pinax*, and so the project was promoted among those open source communities. This attracted many observers who were enthusiastic to participate, but were not 'expert observers' from a colour science perspective. Thanks to this recruitment policy, participants were attracted much more organically and represent a much better sample of internet users 'in the wild'. We actively avoided directly recruiting from colour communities.

Observers were free to complete as many or as few preference choices as they wished. If an observer submitted only a handful of preference choices these were added to the pool of data with equal weighting to those submitted by an observer who submitted hundreds.

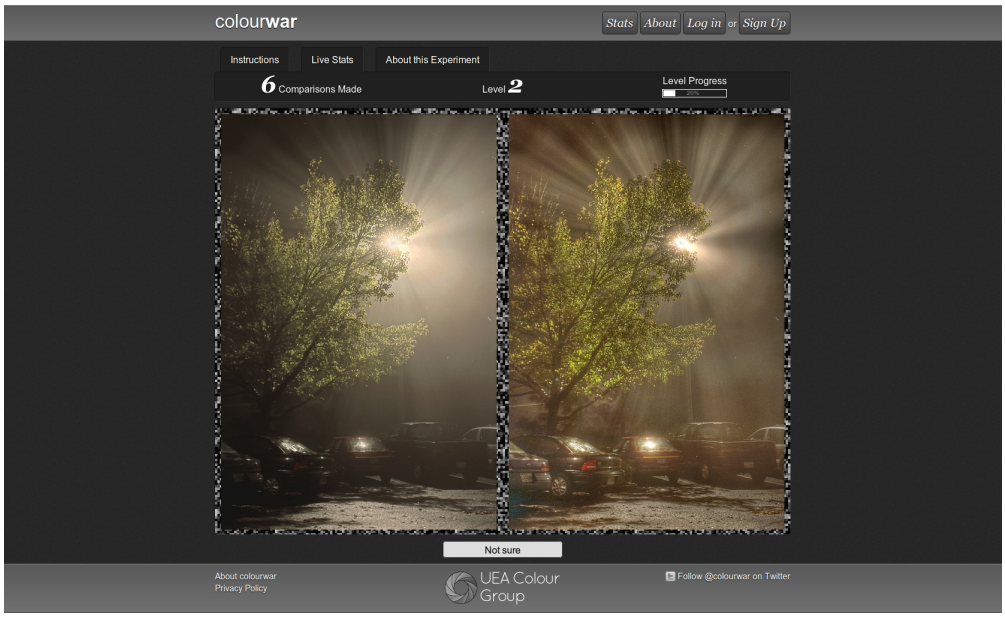


Figure 3.5: Interface of the our web experiment

Also in opposition to some previous approaches, we opted to have no calibration process, questionnaires or adjustment images. Observers visiting the site were immediately presented with their first preference choice, as depicted in fig. 3.5. Primarily it was thought that immediate presentation of the task at hand would be more likely to engage observers and encourage them to partake; presentation of welcome pages, splash screens, or anything of the sort are well known to increase the 'bounce rate' on websites. It is also noted that even if a calibration process were implemented, it would likely be of little value: observers' viewing conditions are likely to change with time, especially on mobile devices. Furthermore, observers could be employing multiple displays, returning to the site on multiple devices, or they could be using a device with an auto-dimming or otherwise automatically adjusting display.

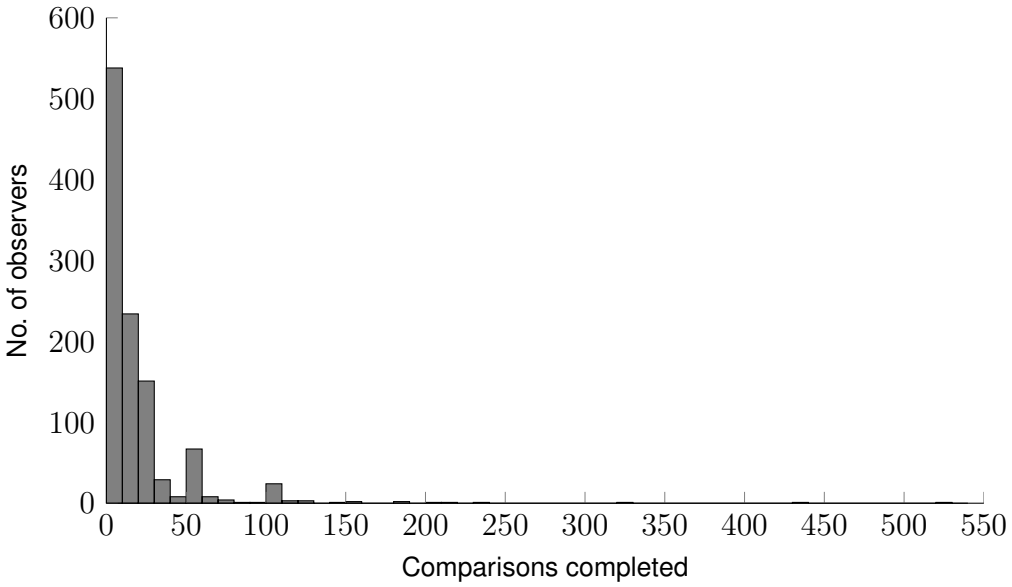


Figure 3.6: Comparisons completed per observer

3.6 Results – A New Web-Based Platform

Here we present data collected by our web-based platform over a span of one year of operation, during which time over twenty-six thousand preference judgements were submitted by more than one thousand observers. The mean number of comparisons per observer is 18.9, with a standard deviation of 35.3 – the distribution of completed comparisons per observer is shown in fig. 3.6. Unfortunately, due to its unbalanced nature, we cannot complete the same summary statistics as above for the web-based data. Expecting web observers to complete every possible combination of images, in order to facilitate the balanced paradigm, is simply unreasonable. Indeed if we omit all unbalanced sessions from our data we would be left with only two complete, balanced, sessions.

Our web-based variant of the *Lab-TMO* experiment is hereafter referred to as the *Web-TMO* experiment. Table 3.3 shows how the *Web-TMO* data compare to the *Lab-TMO* data – we are now considering how the Thurstonian analysis of one variant cor-

Table 3.3: Correlations for all scenes in the *Lab-TMO* and *Web-TMO* experiments

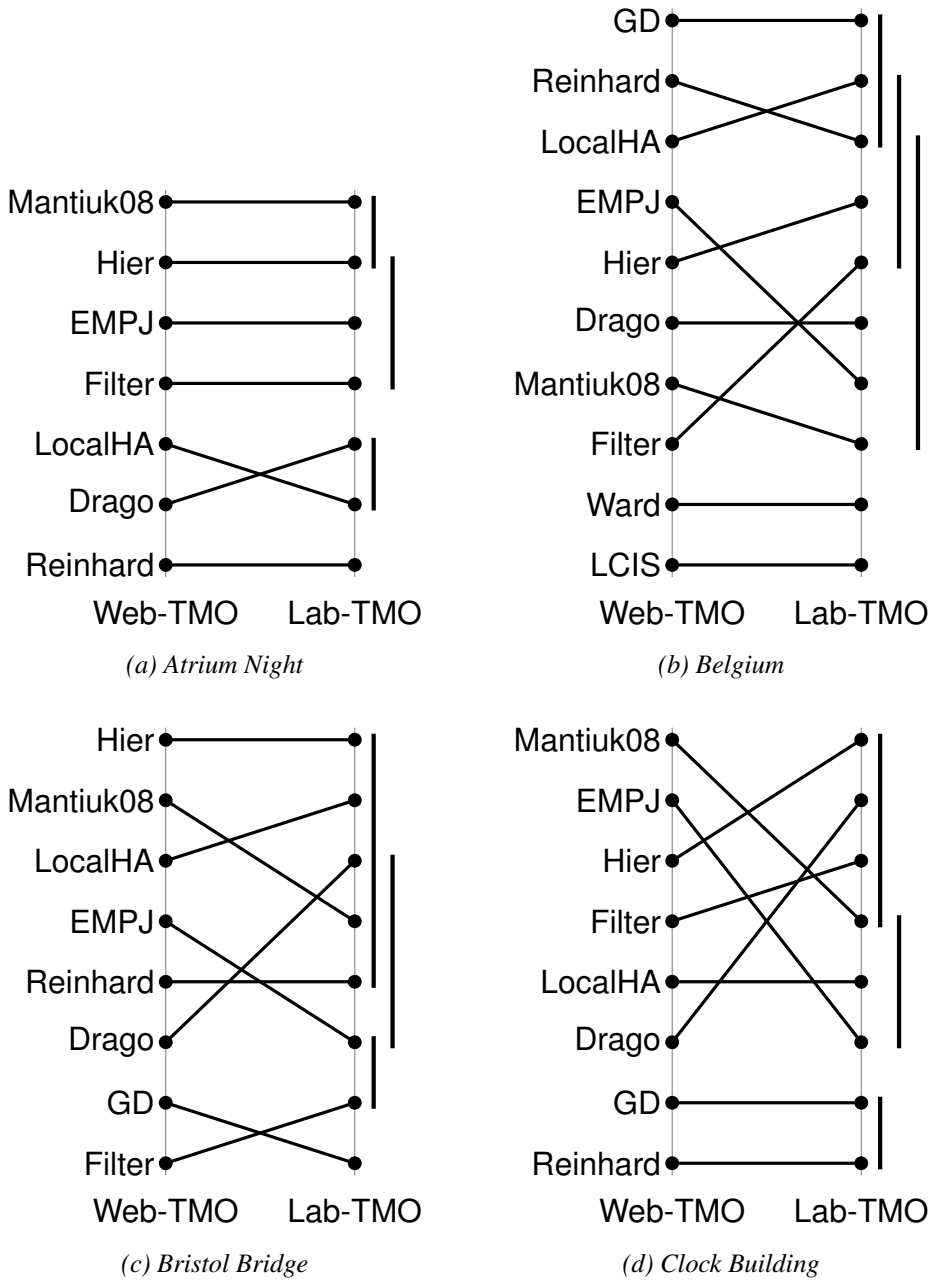
Scene	Kendall Rank Correlation		Sprow Goodness-of-Fit	
	τ	Significance	χ^2	Significance
Atrium Night	0.905	$p < 0.01$	23.123	
Belgium	0.733	$p < 0.01$	48.589	
Bristol Bridge	0.571	$p < 0.05$	72.106	$p < 0.001$
Clock Building	0.357		150.678	$p < 0.001$
Fog	0.333		98.427	$p < 0.001$
Foyer	0.333		83.700	$p < 0.001$
Indoor	0.714	$p < 0.05$	17.081	
Memorial	0.643	$p < 0.05$	34.599	
Synagogue	0.857	$p < 0.01$	26.182	
Tahoe	0.467		48.377	$p < 0.001$
Tinterna	0.867	$p < 0.05$	20.508	
Tree	0.810	$p < 0.05$	62.485	$p < 0.001$
Venice	0.619		41.896	$p < 0.01$

relates with the other. The results of both the Kendall rank correlation coefficient and the Sprow goodness-of-fit test (as described in section 2.6.5) are shown. Recall the disparity in the significance measures for the Kendall and Sprow statistics – a low p-value for the Kendall rank correlation coefficient suggests a strong correlation, while a low p-value for the Sprow goodness-of-fit test suggests a poor correlation. We can see that eight of the thirteen scenes give significantly correlated rank orderings. However, for the 'Clock Building', 'Fog', 'Foyer', 'Tahoe' and 'Venice' scenes, both of the Kendall and Sprow measures agree that those scenes showed poor correlation, although we should bear in mind the results given above of the Mosteller test as applied to the *Lab-TMO* data, which suggest that the 'Synagogue' and 'Tahoe' scenes may be ill-suited for the case V solution.

Interestingly, for 'Bristol Bridge' and 'Tree', significant rank correlation is achieved

but the Sprow test indicates a poor goodness-of-fit. This is examined in further detail in section 3.6.1, with the aid of data from a second suite of image processing algorithms.

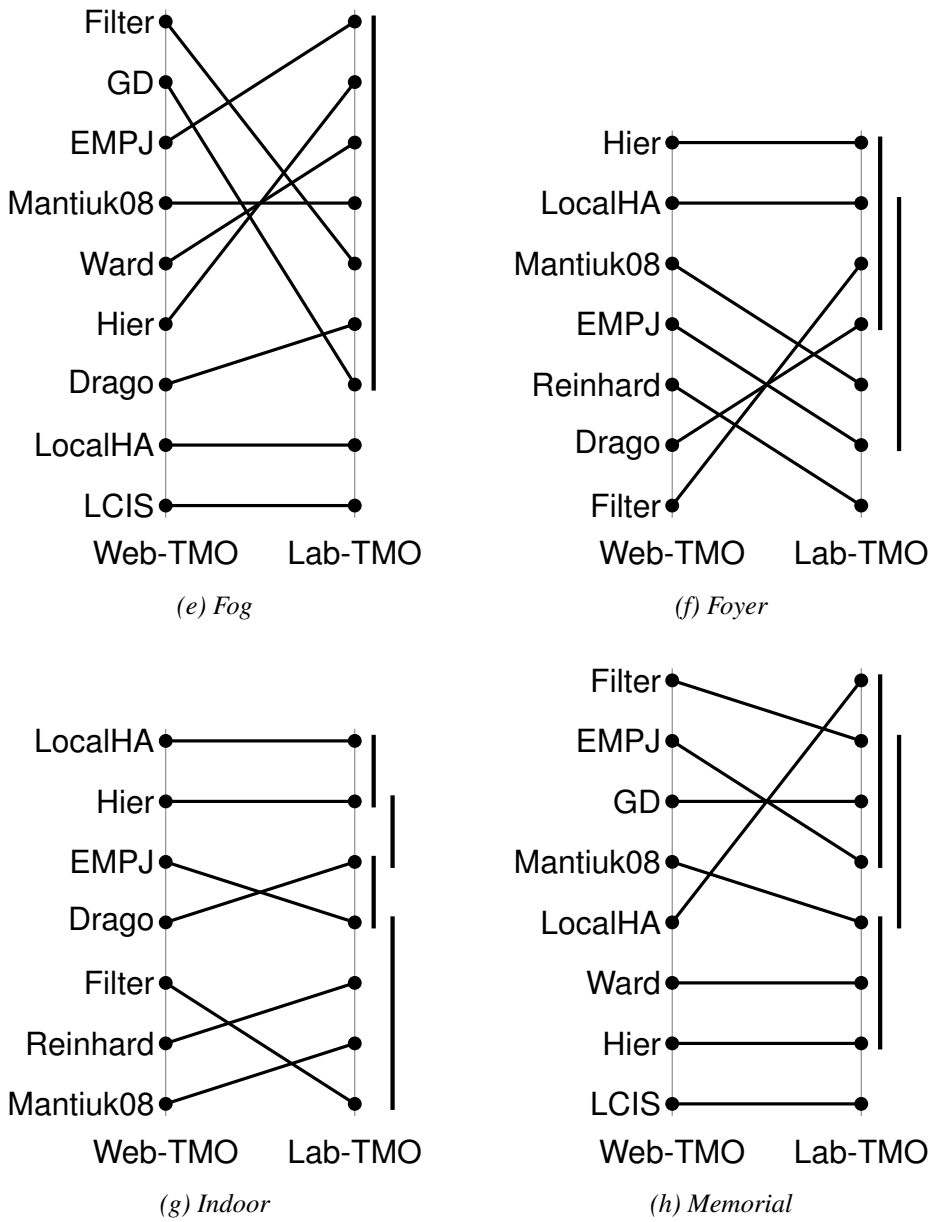
Figures 3.7 and 3.8 show the rank position swaps and the results of the Thurstonian analyses for both web-based and lab-based variants of the TMO experiments. Figure 3.7 is presented similarly to fig. 3.3, this time with *Web-TMO* rank on the left axis, and the *Lab-TMO* rank on the right. However, these graphs have been augmented by the addition of the vertical lines to the right of the right-hand axis. These lines represent the groupings made by the score difference test described in section 2.6.3; for any collection of two or more algorithms grouped by one of these lines it is proclaimed by the score difference test that these algorithms cannot be asserted to be perceptually dissimilar at the chosen significance level, which in this case is $\alpha = 0.05$. This statistic was calculated for the lab-based data, as we are treating our *Lab-TMO* experiment as a ground truth, but also as the score difference test requires data from a balanced experiment. The addition of this statistic adds an interesting perspective to our data; consider for example the 'Atrium Night' scene described in fig. 3.7a. We can see that there is only one rank position swap between the *Web-TMO* and *Lab-TMO* rankings (which is supported by table 3.3), but the addition of the score difference test reveals that this swap is between two algorithms which are not perceptually dissimilar. In light of this, we may be more willing to accept this rank position discrepancy between the two rankings and suggest that the results produced by each variant are similar enough to assert that they are the same. This insight becomes even more interesting when considering scenes such as 'Fog' – this scene has many rank position swaps, and indeed the Kendall τ correlation in table 3.3 is very low. However, we can see from fig. 3.7e that all but two of the algorithms are declared by the score difference test to be not perceptually dissimilar – when those algorithms are applied to this scene, at least. So, such discordance between the two rankings is perhaps not surprising – indeed the algorithms which are not declared dissimilar by the test are not affected by any rank position swaps. Similar trends



(a) Atrium Night *(b) Belgium*

(c) Bristol Bridge *(d) Clock Building*

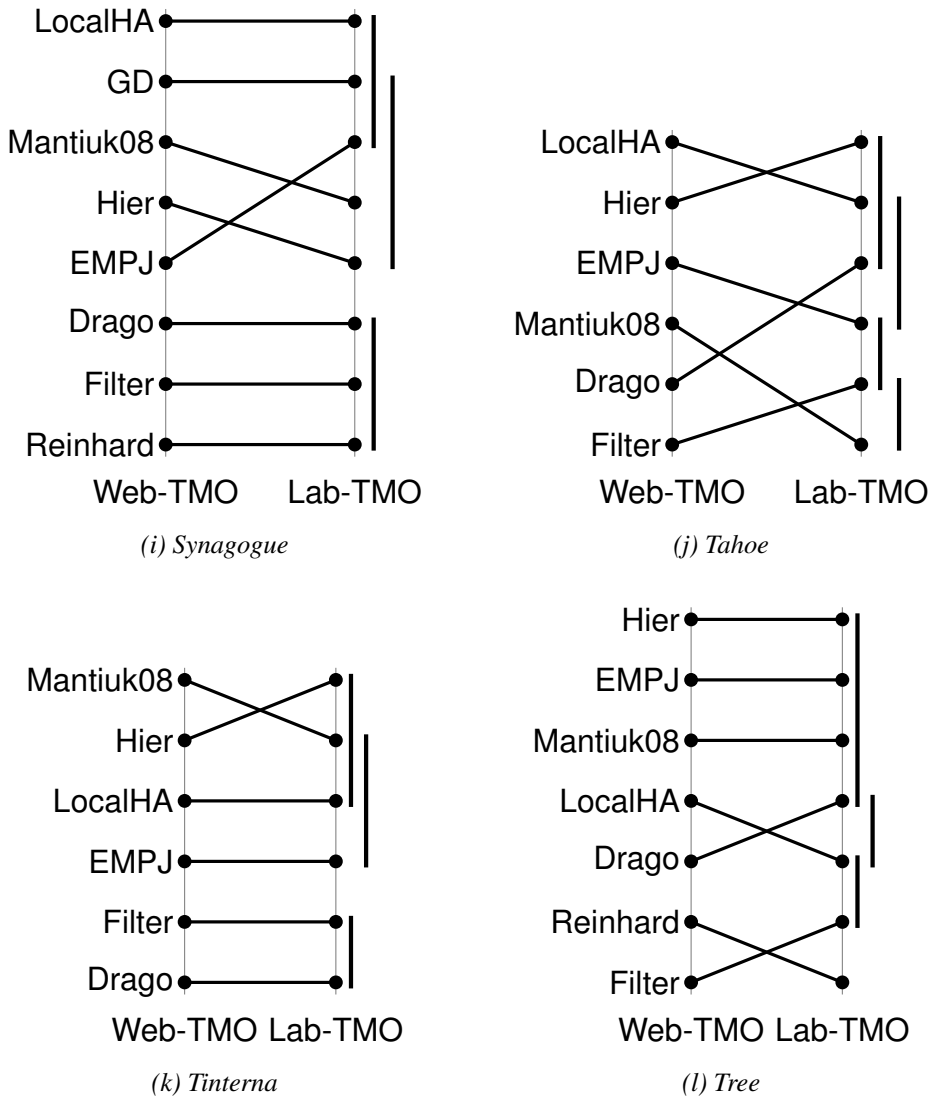
Figure 3.7: Rank correlations between *Web-TMO* and *Lab-TMO* variants, for all scenes, based on Thurstone Case V scores



(e) Fog *(f) Foyer*

(g) Indoor *(h) Memorial*

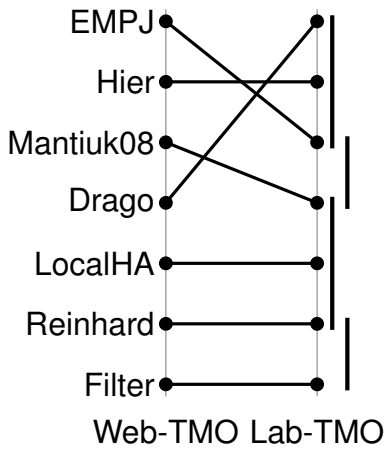
Figure 3.7: Rank correlations between *Web-TMO* and *Lab-TMO* variants, for all scenes, based on Thurstone Case V scores *(cont.)*



(i) Synagogue *(j) Tahoe*

(k) Tinterna *(l) Tree*

Figure 3.7: Rank correlations between *Web-TMO* and *Lab-TMO* variants, for all scenes, based on Thurstone Case V scores *(cont.)*

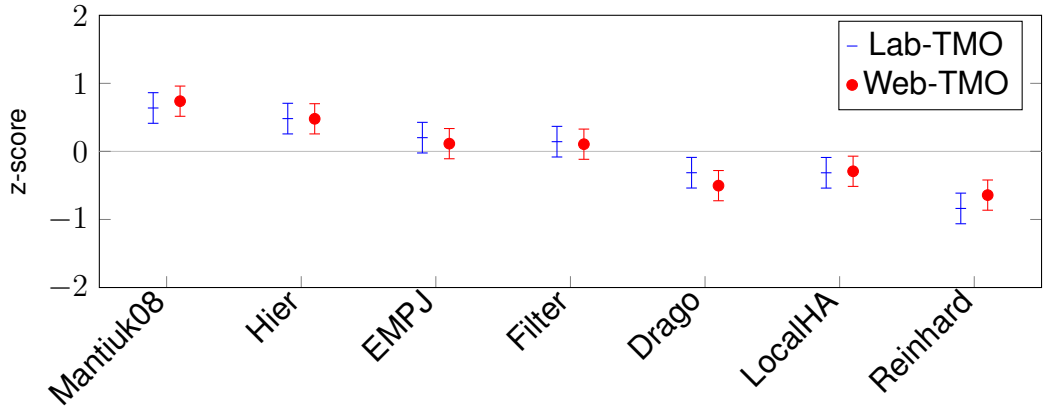


(m) Venice

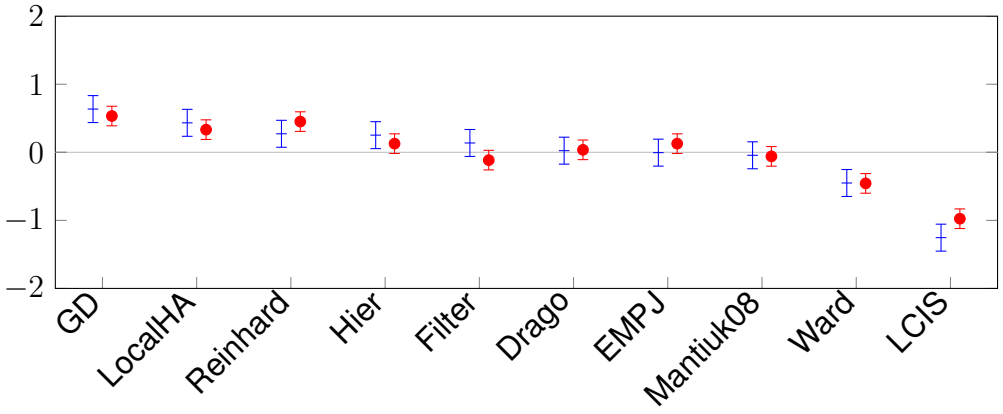
Figure 3.7: Rank correlations between *Web-TMO* and *Lab-TMO* variants, for all scenes, based on Thurstone Case V scores *(cont.)*

can also be seen in the 'Belgium', 'Bristol Bridge', 'Indoor', 'Synagogue' and 'Tinterna' scenes, where even though the rankings for these scenes may have many swaps, they are always between algorithms which, according to this test, are not dissimilar at the 95% level.

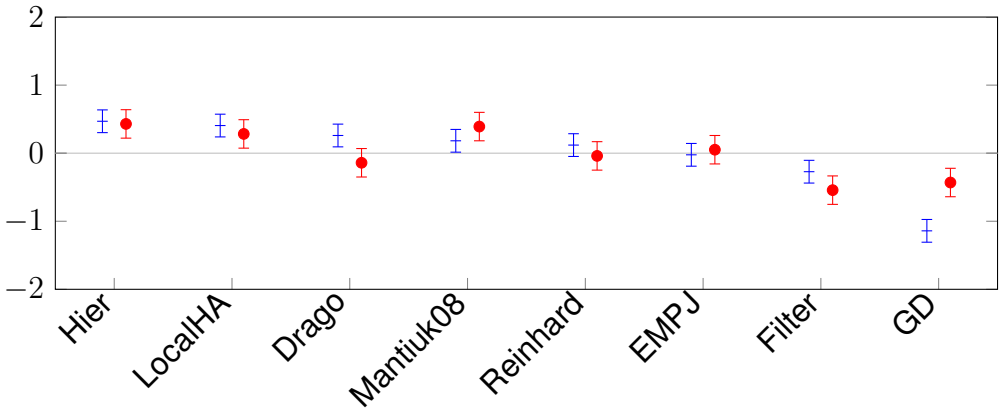
Figure 3.8 reveals some interesting features from the scores for each algorithm derived by the Thurstone analysis of each variant. While rank position swaps are not so easy to visualise in these graphs, they help to reveal differences in absolute scores. For example fig. 3.8l can help to explain why table 3.3 shows strong Kendall rank correlation (supported by fig. 3.7l) for the 'Tree' scene while also reporting significant poor correlation according to the Sprow test. We can see in fig. 3.8l that, while the *Web-TMO* scores reveal the same ordinal trend as the *Lab-TMO* scores (save for the two swaps), they are shifted negatively for all algorithms except *LocalHA* and *Reinhard*. Since these z-scores from a Thurstone analysis must always sum to zero (under the case V assumptions), the more positive scores in the *Web-TMO* experiment for these two algorithms have had the effect of dragging the scores for all the other algorithms in a negative direction.



(a) Atrium Night

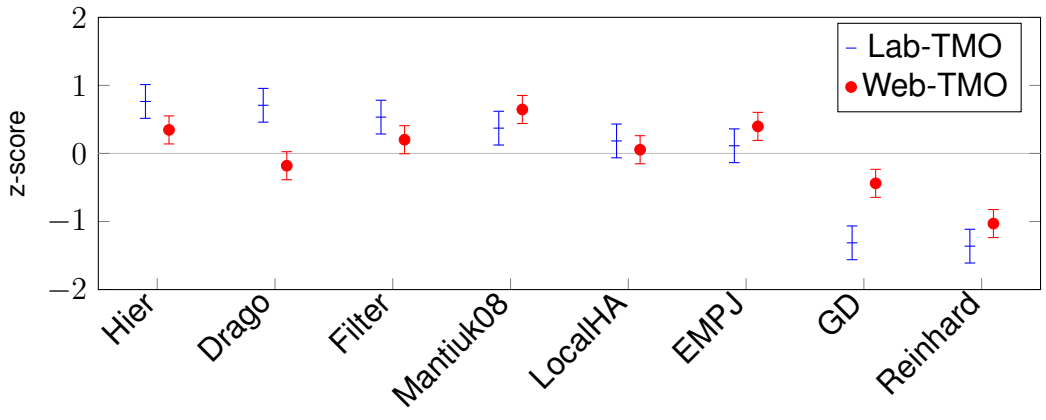


(b) Belgium

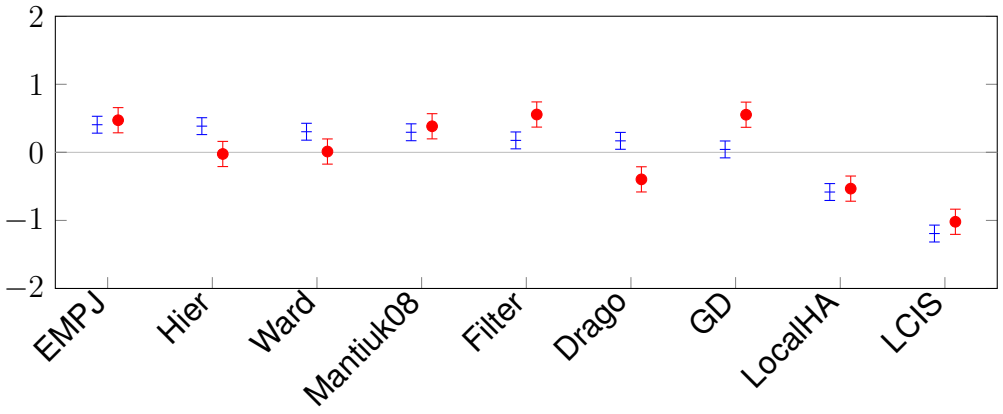


(c) Bristol Bridge

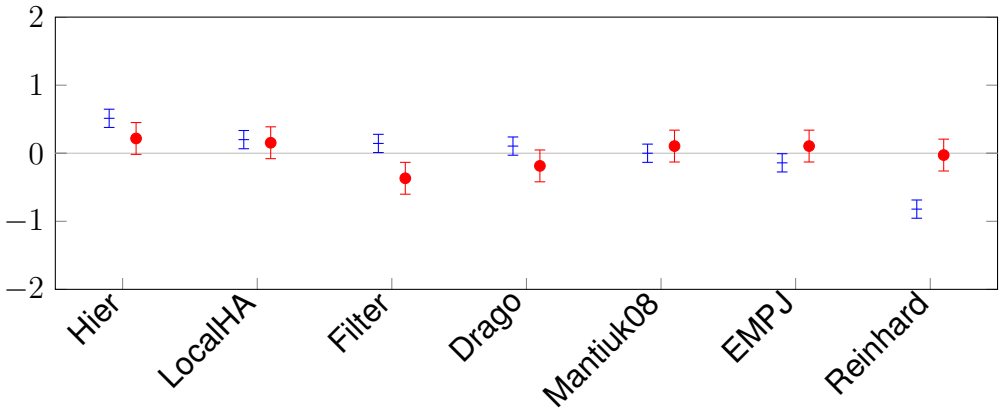
Figure 3.8: Thurstone Case V scores for *Lab-TMO* and *Web-TMO* variants, for all scenes



(d) Clock Building

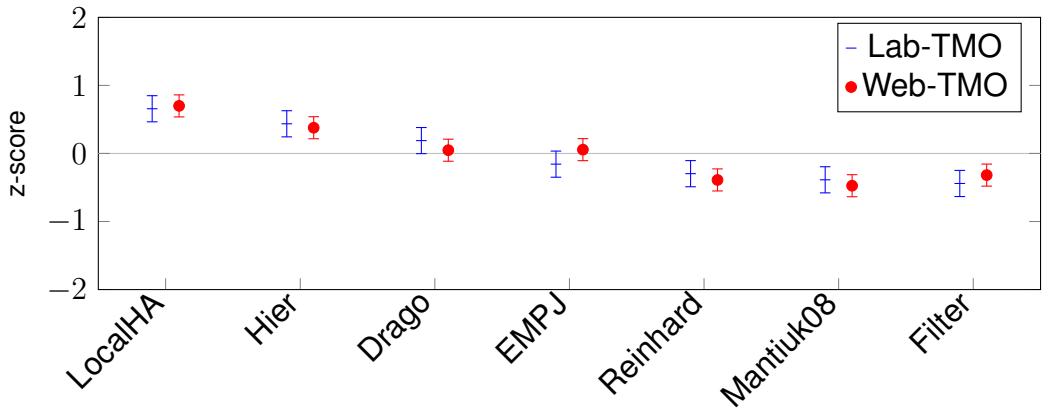


(e) Fog

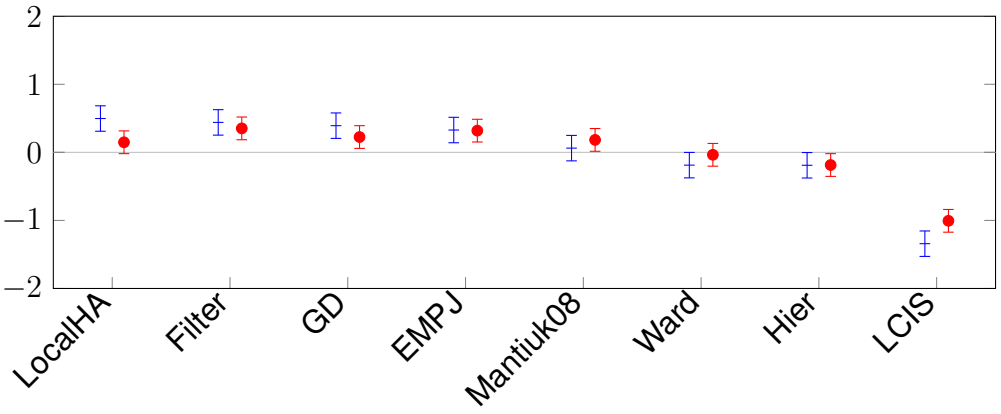


(f) Foyer

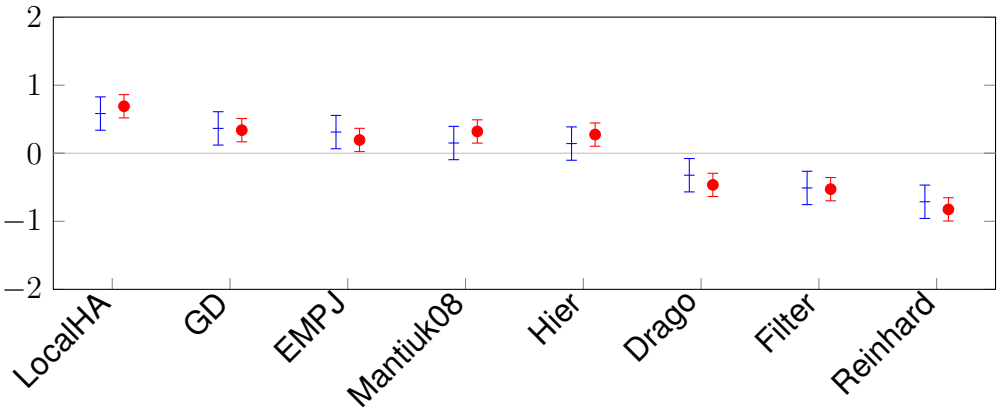
Figure 3.8: Thurstone Case V scores for *Lab-TMO* and *Web-TMO* variants, for all scenes *(cont.)*



(g) Indoor

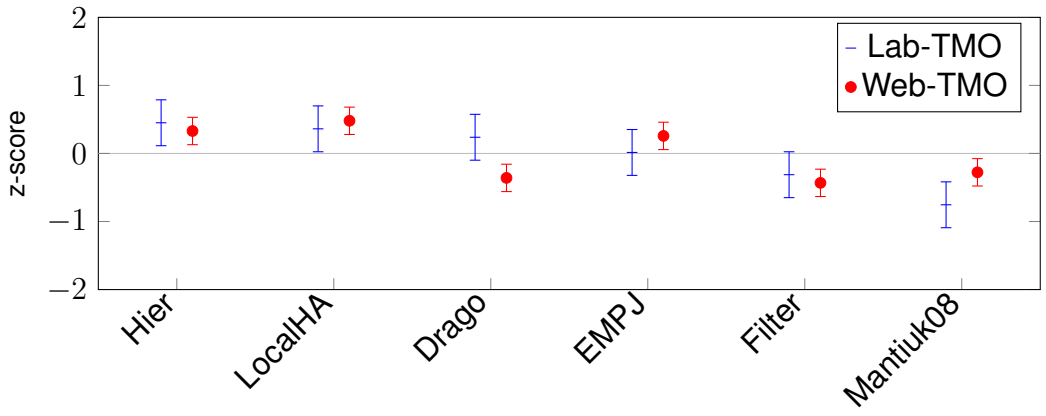


(h) Memorial

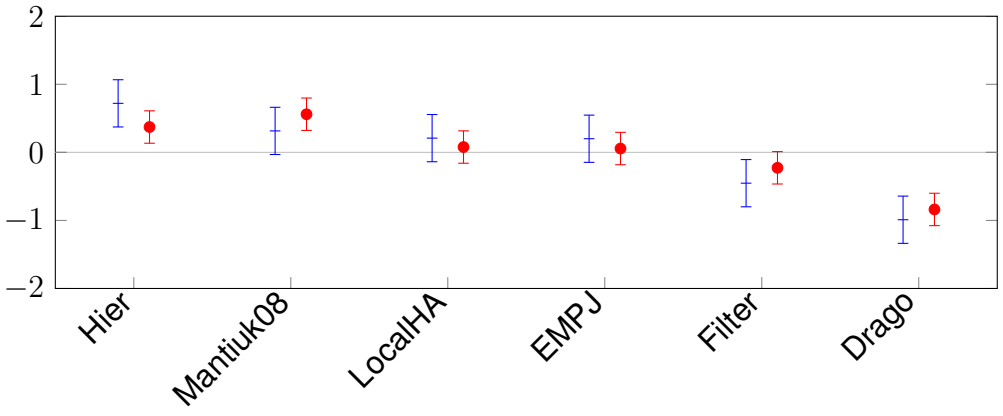


(i) Synagogue

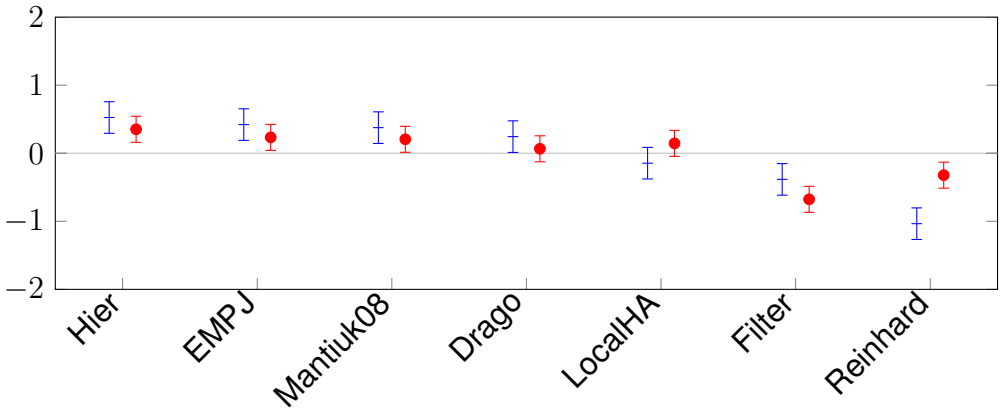
Figure 3.8: Thurstone Case V scores for *Lab-TMO* and *Web-TMO* variants, for all scenes *(cont.)*



(j) Tahoe

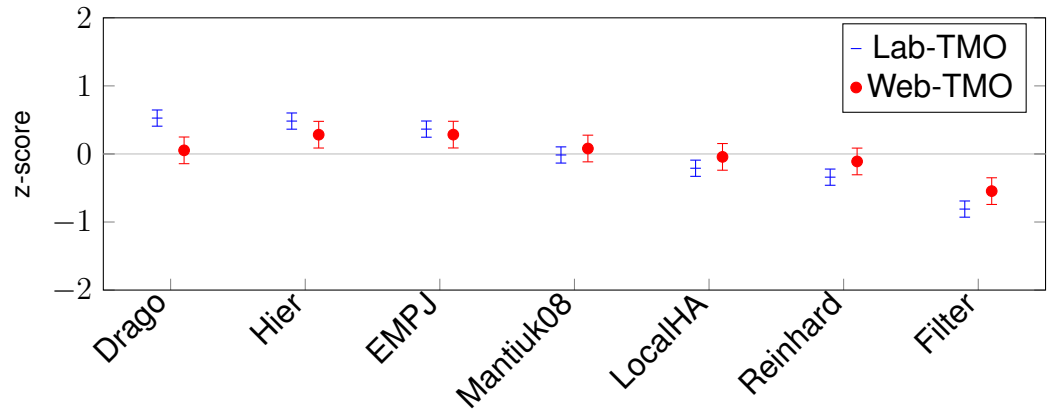


(k) Tinterna



(l) Tree

Figure 3.8: Thurstone Case V scores for *Lab-TMO* and *Web-TMO* variants, for all scenes *(cont.)*



(m) Venice

Figure 3.8: Thurstone Case V scores for *Lab-TMO* and *Web-TMO* variants, for all scenes *(cont.)*

3.6.1 Adding a Second Dataset

To corroborate the results from our TMO experiments, we also ran a further experiment examining observer preference for colour-to-greyscale algorithms. For the lab-based variant of this experiment (hereafter referred to as *Lab-C2G*), we used existing data published by Connah et al. (2007). The C2G operators compared by Connah et al. are listed in section 2.4[3], and the images used to test these operators are listed in appendix B. The control conditions for the *Lab-C2G* experiment are summarised in Connah et al. (2007), but they were largely similar to those we used for the *Lab-TMO* experiment and met the same standards requirements. We are not concerned with a full analysis of the results generated from the *Lab-C2G* experiment from the perspective of evaluating C2G operators, again we are only concerned with how the lab-based data compare to a web-based replicate. To establish a reference for what level of confidence we can hold the lab-based data to, we recapitulate the summary statistics for the *Lab-C2G* data in table 3.4, with the addition of the results of the Mosteller test (which was not calculated in Connah et al. (2007)).

[3] Abbreviated C2G operator names have been kept consistent with those used in Connah et al. (2007).

Table 3.4: C2G experiment: summary statistics for lab data

Scene	Mosteller		Agreement			Consistency
	χ^2	Significance	u	χ^2	Significance	Ω
Girl	4.362		0.040	28.833	$p < 0.05$	0.714
Hats	3.026		0.061	36.000	$p < 0.01$	0.604
Heron	11.569		0.521	194.833	$p < 0.001$	0.885
Monet	24.199	$p < 0.01$	0.435	165.167	$p < 0.001$	0.807
Parrot	13.172		0.386	148.000	$p < 0.001$	0.818
Poppies	9.070		0.226	92.833	$p < 0.001$	0.755

We can see that, as in the *Lab-TMO* experiment, there were high levels of intra-observer consistency for all scenes. However, for the 'Girl' and 'Hats' scenes, the inter-observer agreement was slightly lower – it is still significantly high ($p < 0.05$ and 0.01 respectively) but it is not at the $p < 0.001$ level as in the other scenes. The reasons for the poorer performance for these scenes are discussed by Connah et al. (2007); it is suggested that, for these scenes in particular, the compared algorithms all perform similarly and different observers may be selecting different criteria to judge the minor differences in these images.

The Mosteller test shows positive results for five of the six scenes but, as with 'Synagogue' and 'Tahoe' from the TMO experiment, we should perhaps be wary when considering the 'Monet' scene due to its significantly high χ^2 score.

Our web-based variant of the C2G experiment, hereafter referred to as *Web-C2G*, ran parallel to the *Web-TMO* experiment on our web-based research platform discussed in the previous section. Observers were randomly assigned to one of the two experiments on their first visit to the site, but could opt-in to a different experiment if they so wished. Similarly, if an observer completed all the comparisons for a particular experiment (a feat managed by only two observers out of over one thousand), they would be assigned to the other upon their next visit. Table 3.5 shows how the web data compare

Table 3.5: C2G experiment: correlations between lab and web results

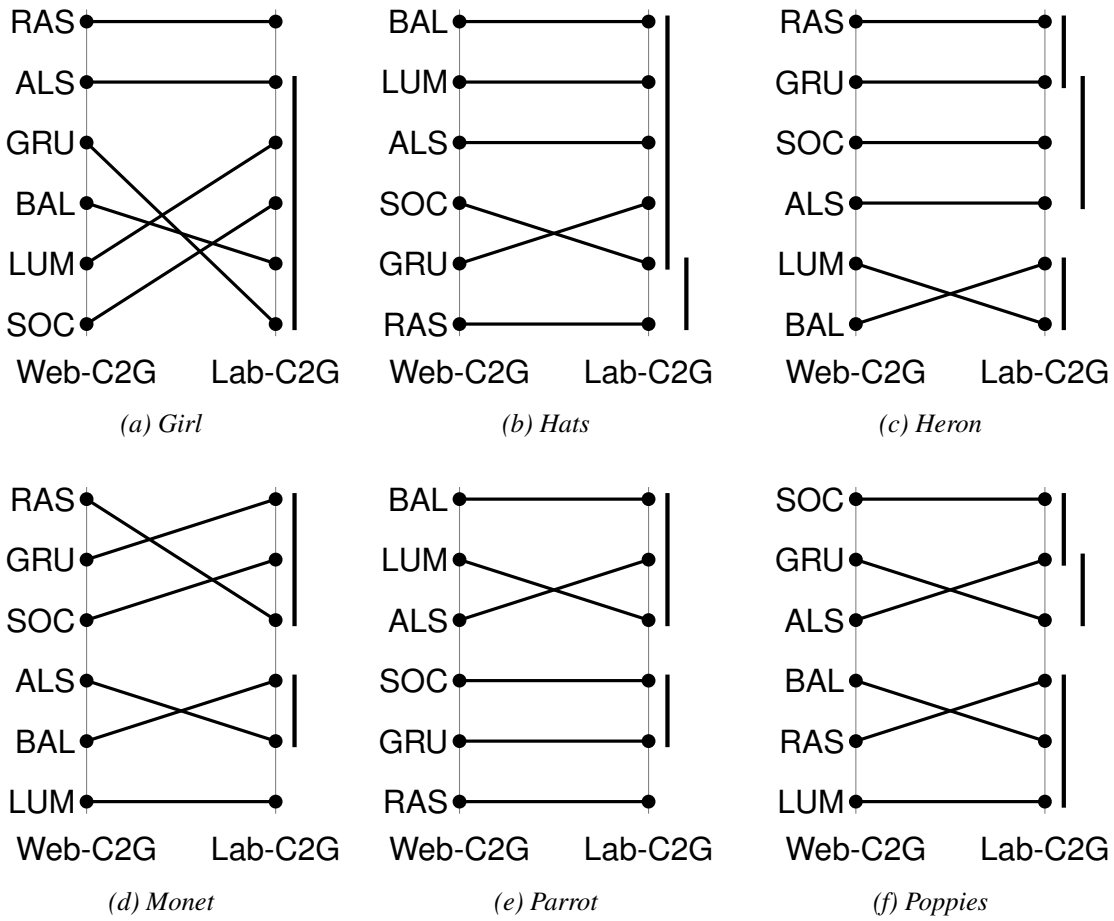
Scene	Kendall Rank Correlation		Sprow Goodness-of-Fit	
	τ	Significance	χ^2	Significance
Girl	0.333		17.970	
Hats	0.867	$p < 0.05$	15.422	
Heron	0.867	$p < 0.05$	99.281	$p < 0.001$
Monet	0.600		48.534	$p < 0.001$
Parrot	0.867	$p < 0.05$	29.811	
Poppies	0.733	$p < 0.05$	27.162	

to the lab data for the C2G experiment – much like table 3.3, these data represent the first year of data collection.

Four out of six scenes give significantly correlated rank orderings, while 'Monet' exhibits weak correlation according to both the Kendall and Sprow measures – although we should once again bear in mind the results of the Mosteller test which suggest that the 'Monet' scene is ill-suited for the case V solution.

Figures 3.9 and 3.10 show the rank position swaps and the results of the Thurstonian analyses for both web-based and lab-based variants of the C2G experiments. As with fig. 3.7, the graphs in fig. 3.9 have been augmented by the groupings discerned by the score difference test at the $\alpha = 0.05$ level.

The 'Girl' scene presents an interesting situation: it exhibits weak rank correlation according to the Kendall measure, but favourable goodness-of-fit according to the Sprow measure. Figure 3.10a shows the results of the Thurstonian analysis of the 'Girl' scene for both the lab and web variants plotted on the same axes. It is evident that the scores are very similar in both experiments, but the minor fluctuations happen to cause significant rank differences. The ordinal rankings, seen in fig. 3.9a, produce many rank position swaps between the two variants, but they are all within the bounds of the perceptibly similar. This highlights the danger of relying solely on a rank correlation



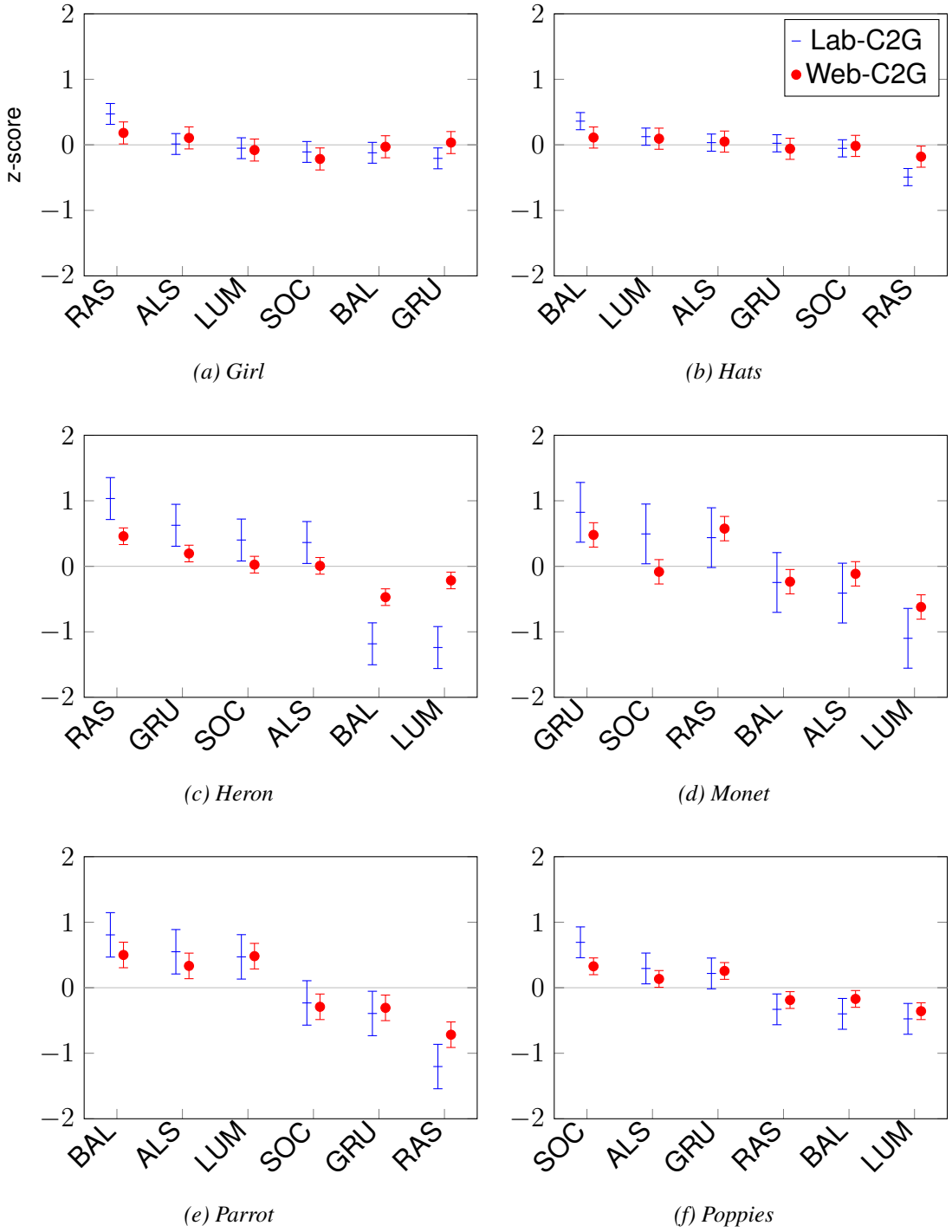
(a) Girl *(b) Hats* *(c) Heron*

(d) Monet *(e) Parrot* *(f) Poppies*

Figure 3.9: Rank correlations between *Web-C2G* and *Lab-C2G* variants, for all scenes, based on Thurstone Case V scores

measure to quantify the similarities or otherwise of our lab- and web-based variants.

Notably, if we carry out the score difference test for all scenes in the C2G experiments, this same explanation holds true for every scene that does not exhibit significantly high rank correlation – the rank position swaps are always among those algorithms which are, according to the score difference test applied to the lab-based data, not significantly dissimilar. This is an important point to underline – for every scene that does not exhibit strong rank correlation, the rank position swaps causing that weak correlation are all among algorithms which are not perceptibly dissimilar. The same is not always true for the TMO experiments, but does hold in many cases.



(a) Girl *(b) Hats*

(c) Heron *(d) Monet*

(e) Parrot *(f) Poppies*

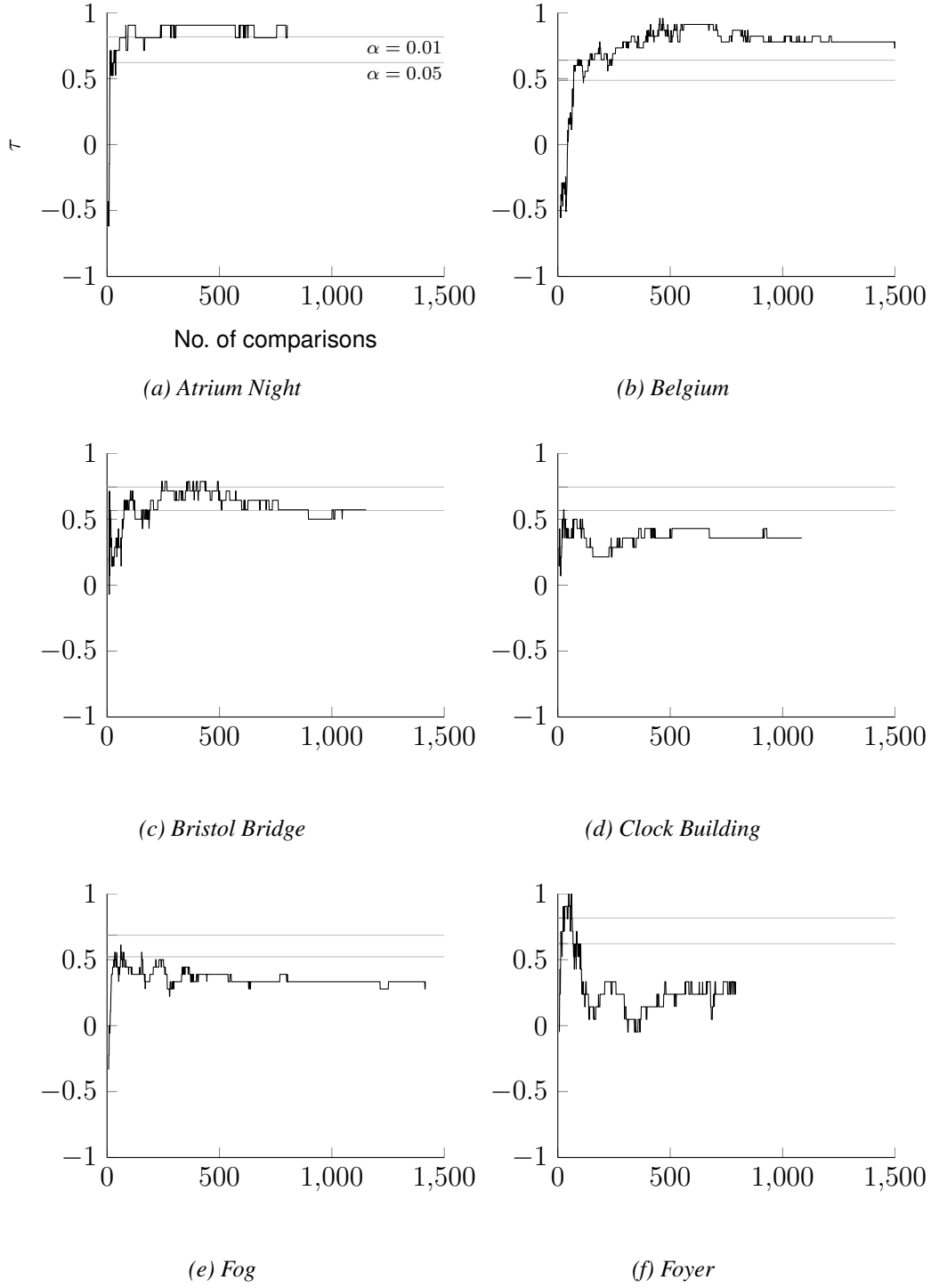
Figure 3.10: Thurstone Case V scores for *Lab-C2G* and *Web-C2G* variants, for all scenes

Another interesting situation arises for the 'Heron' scene, as well as 'Bristol Bridge' and 'Tree' from the TMO experiment: significantly strong rank correlation is achieved but the Sprow test indicates a poor goodness-of-fit. Figure 3.10c shows how this can be the case for the 'Heron' scene – the rank orderings are very similar, with only one position swap between the 'BAL' and 'LUM' algorithms, however the web results are somewhat muted in comparison to the lab results. This could be due to the larger number of observers for the web experiment. The results for 'Bristol Bridge' and 'Tree' show similar properties.

3.7 Correlation Over Time

A feature of our web-based platform is the ability to compute all the statistics used above in real time. This means that we can examine the correlation between the lab-based and web-based variants as a function of time or, equivalently, the number of comparisons completed. In so doing, we will consider the TMO and C2G experiments in unison. Figures 3.11 and 3.13 show, for the TMO experiments and the C2G experiments respectively, Kendall rank correlation between the lab- and web-based variants as a function of the number of comparisons made in the web variants, while figs. 3.12 and 3.14 show the correlation based on the Sprow measure, again as a function of the number of comparisons. The grey horizontal lines show the value of τ for the Kendall graphs, and χ^2 for Sprow, required to be significant at the 95% and 99% levels.

If we first consider those scenes where both statistical measures are in agreement that high correlation was achieved, namely 'Atrium Night', 'Belgium', 'Indoor', 'Memorial', 'Synagogue' and 'Tinterna' for the TMO experiments, and 'Hats', 'Parrot' and 'Poppies' for C2G, we can see that significant correlation is achieved for all but one of those scenes after approximately three hundred comparisons, or about sixteen individual observers (based on the mean number of comparisons per observer of 18.9).



(a) Atrium Night

(b) Belgium

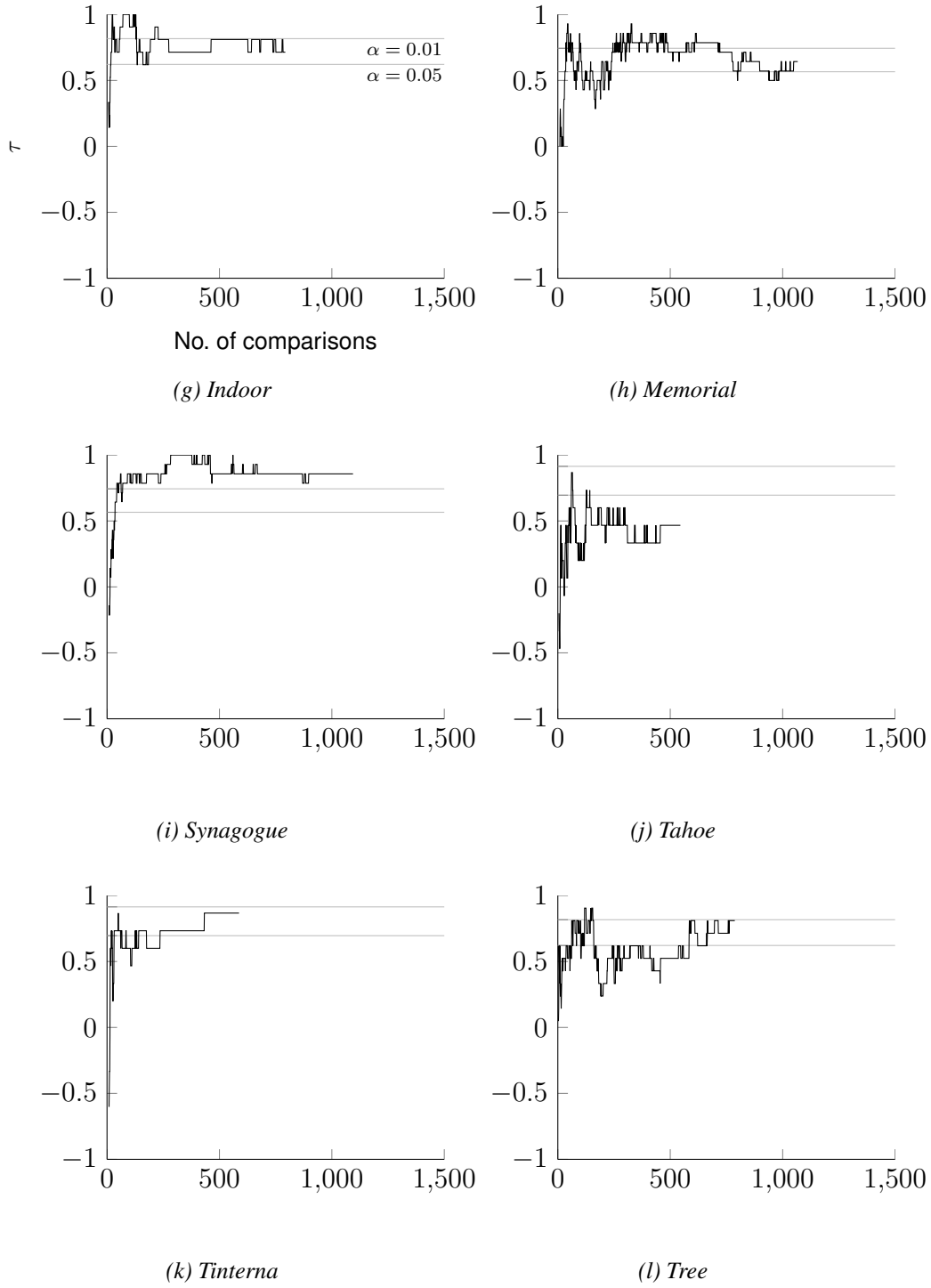
(c) Bristol Bridge

(d) Clock Building

(e) Fog

(f) Foyer

Figure 3.11: Correlation over time for all scenes in the TMO experiments, based on the Kendall rank correlation coefficient

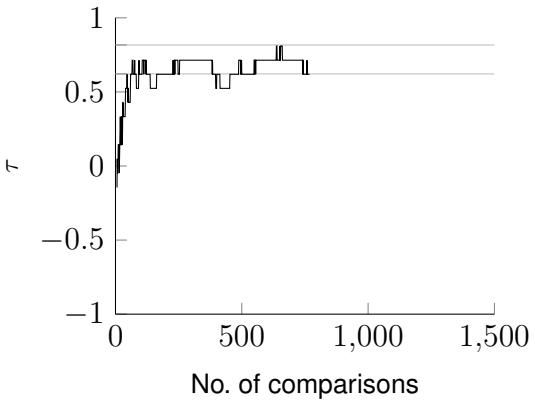


(g) Indoor *(h) Memorial*

(i) Synagogue *(j) Tahoe*

(k) Tinterna *(l) Tree*

Figure 3.11: Correlation over time for all scenes in the TMO experiments, based on the Kendall rank correlation coefficient *(cont.)*



(m) Venice

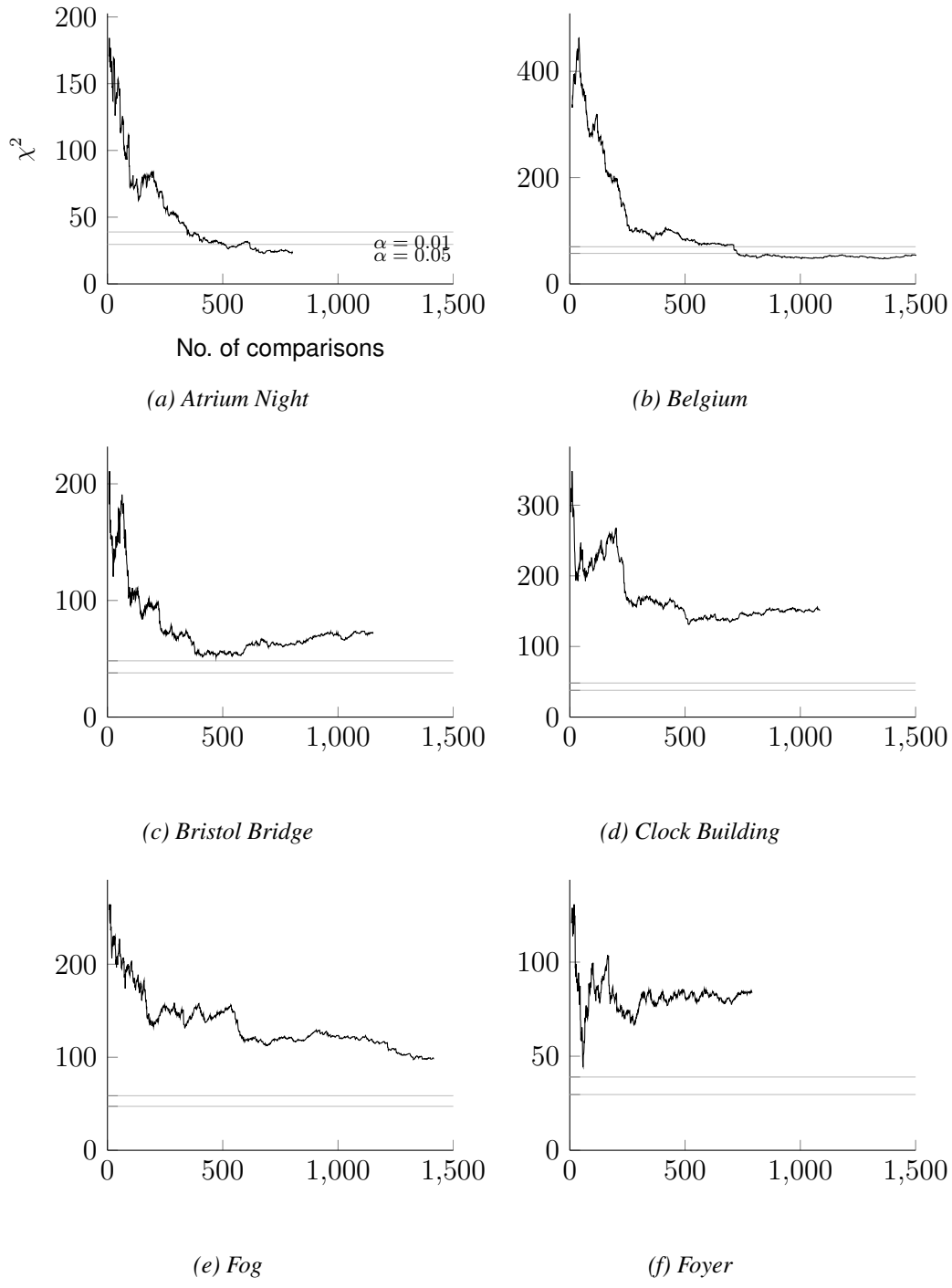
Figure 3.11: Correlation over time for all scenes in the TMO experiments, based on the Kendall rank correlation coefficient *(cont.)*

While the 'Hats' scene required a little longer for the Kendall rank correlation statistic at approximately five hundred comparisons (or ≈ 27 observers).

Conversely, for the remaining scenes we can observe that the levels of correlation between the lab- and web-based rankings, while not achieving significantly high levels, do still stabilise after approximately five hundred comparisons. This suggests that five hundred comparisons is sufficient to obtain a stable result from a cohort of generic web users, but that this result cannot necessarily be relied upon to correlate with a lab-based experiment. This stability despite lack of correlation may also suggest a deeper underlying difference in preference metric for observers on the web.

3.8 Discussion

These results compare the outcomes of two very different experimental paradigms. Although both are paired comparison experiments and both are comparing the same collections of images, the levels of control in the lab-based experiments contrast greatly with the almost total lack of control in the web-based counterparts. It is not the intention of this work to examine why one algorithm (TMO or C2G) is preferred over another

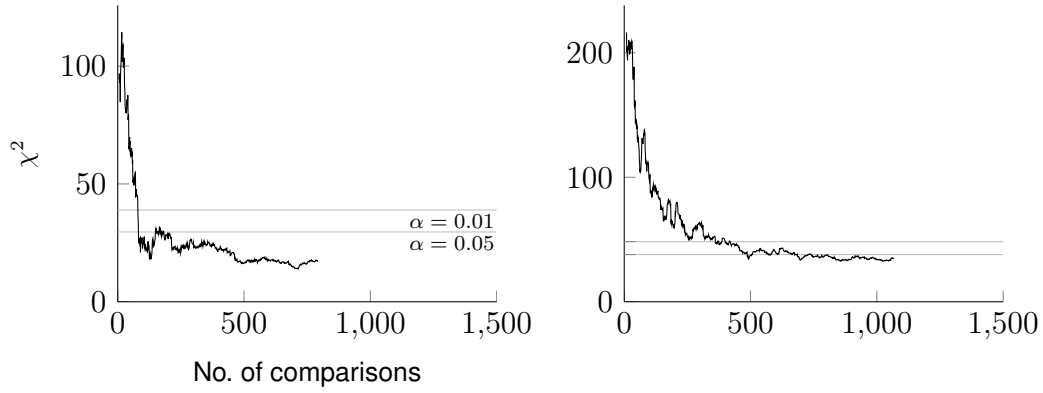


(a) Atrium Night *(b) Belgium*

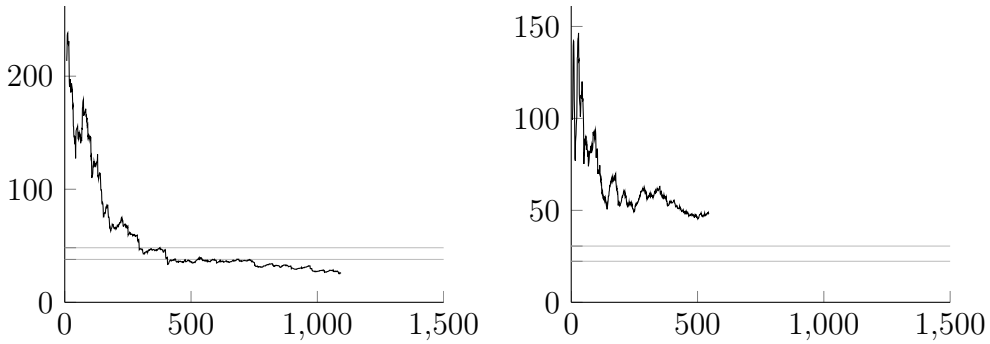
(c) Bristol Bridge *(d) Clock Building*

(e) Fog *(f) Foyer*

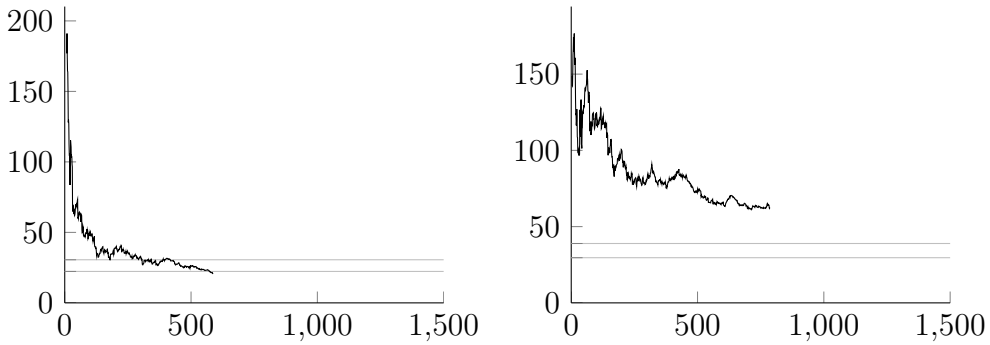
Figure 3.12: Correlation over time for all scenes in the TMO experiments, based on the Sprow et al. measure of correlation



(g) Indoor *(h) Memorial*

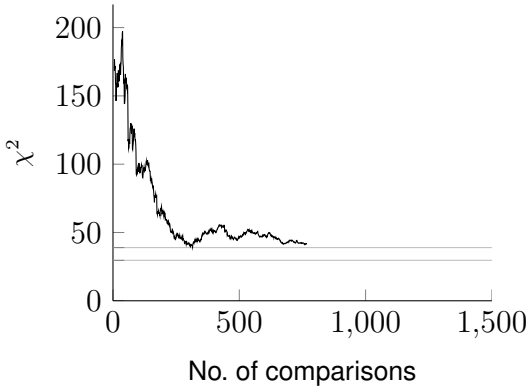


(i) Synagogue *(j) Tahoe*



(k) Tinterna *(l) Tree*

Figure 3.12: Correlation over time for all scenes in the TMO experiments, based on the Sprow et al. measure of correlation *(cont.)*



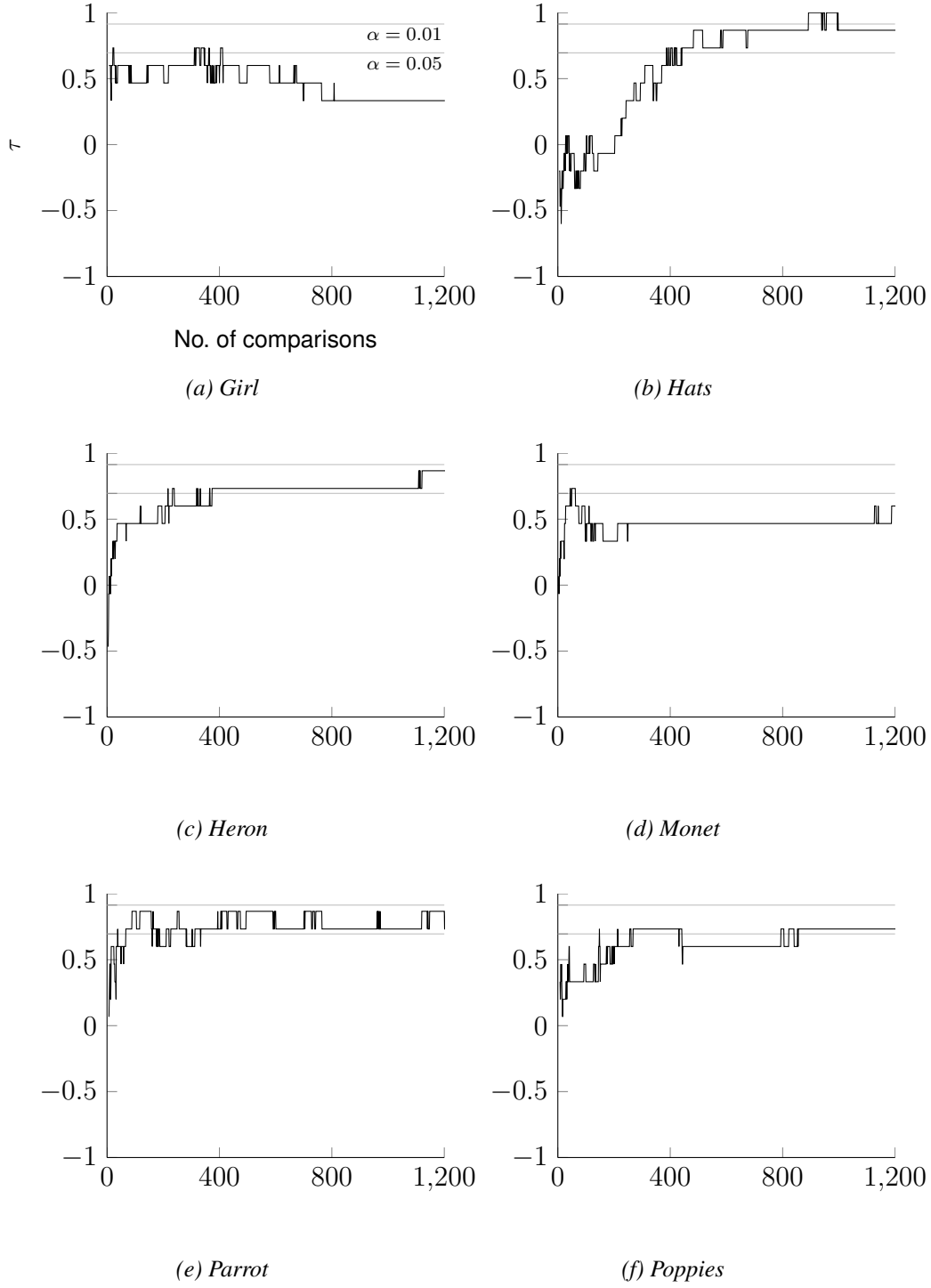
(m) Venice

Figure 3.12: Correlation over time for all scenes in the TMO experiments, based on the Sprow et al. measure of correlation *(cont.)*

in each particular experiment – rather, the data of interest pertains to the extent of the similarity between the juxtaposed sets of rankings, and what factors can account for any differences.

It is noted that our *Web-TMO* experiment achieved higher levels of correlation with the *Lab-TMO* counterpart than did the *Nottingham-Web* experiment. If we examine the differences in how the *Web-TMO* experiment was conducted in contrast with the *Nottingham-Web* experiment, we can uncover some perhaps important differences that could go some way to explaining the conflicting results.

Many conventions of displaying images to a participant were not sufficiently addressed by the interface of the *Nottingham-Web* experiment. Aside from the many aspects of the environment which are beyond the feasible control of any web-based interface (such as ambient lighting, viewing angle, viewing distance, and screen resolution), the presentation of the *Nottingham-Web* experiment introduced some complications of its own. For example, the images were displayed against a bright yellow background, bordered by other colourful interface elements. Conversely, the *Web-TMO* (and *Web-C2G*) experiment employed a neutral background, with a variegated surround around the displayed images. The web-based platform used for the *Web-TMO* and *Web-C2G*



(a) Girl

(b) Hats

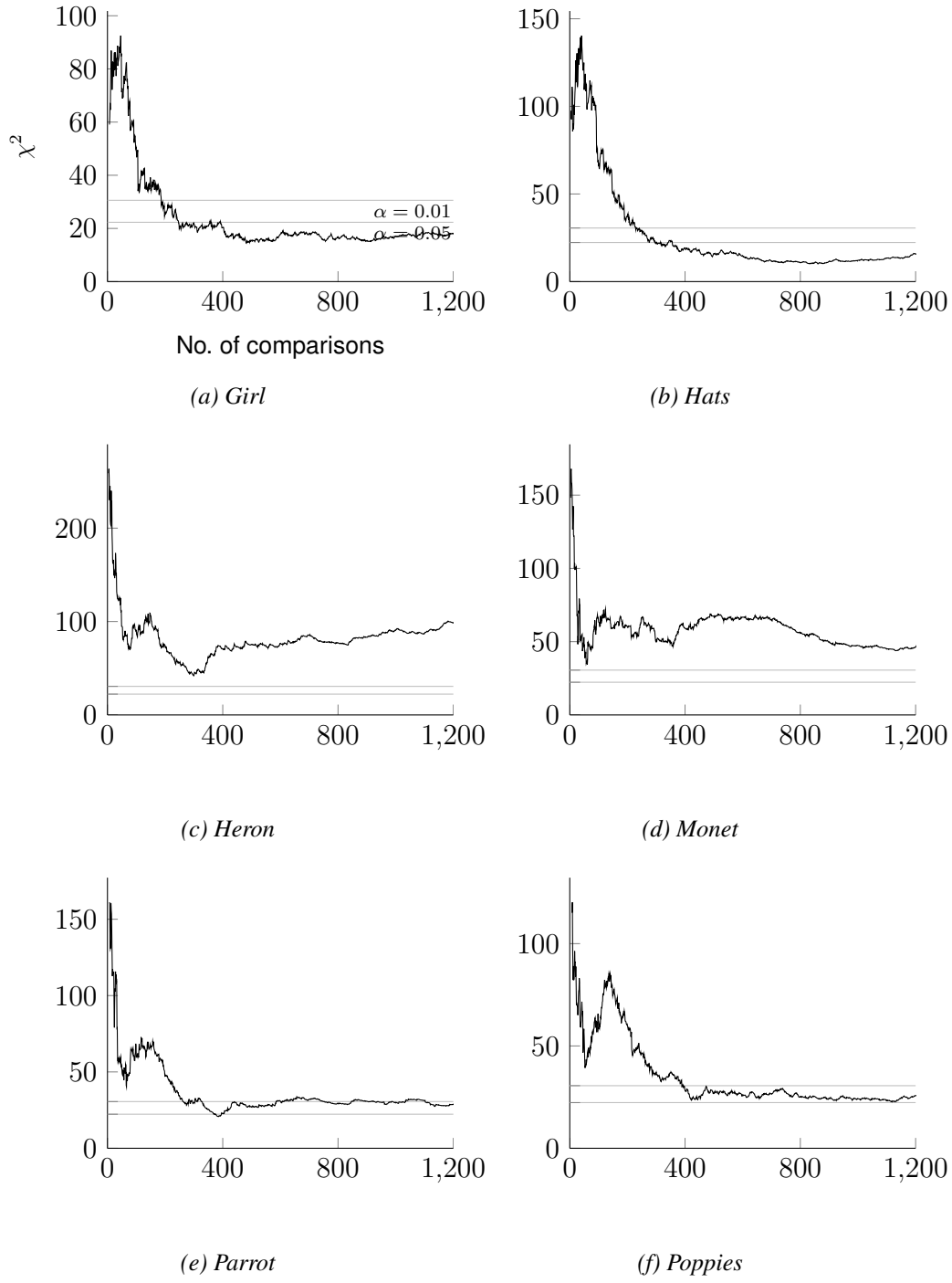
(c) Heron

(d) Monet

(e) Parrot

(f) Poppies

Figure 3.13: Correlation over time for all scenes in the C2G experiments, based on the Kendall rank correlation coefficient



(a) Girl *(b) Hats*

(c) Heron *(d) Monet*

(e) Parrot *(f) Poppies*

Figure 3.14: Correlation over time for all scenes in the C2G experiments, based on the Sprow et al. measure of correlation

experiments had several supporting webpages in addition to the main experimental interface. For example the Thurstone scores and several other statistics were computed in real time and displayed for review by any interested party, similarly descriptions of the algorithms under scrutiny and the images used were made available. All of these pages were also displayed with a consistent neutral colour scheme in an attempt to reduce any ordering effects if a participant viewed any of this information and then went on to contribute data.

With the *Nottingham-Web* experiment, the images to be compared were not resized on the server but were sent to the participant's browser at full resolution and resized in-browser. Due to the different implementations of image resizing across browsers, this means that some observers will have been presented with images resized using bicubic resampling, some with bilinear, and some with nearest neighbour. This will undoubtedly have resulted in the creation of image artefacts for some observers, but less so for others. Meanwhile our web-based experiments employed consistent server-side resizing of the images.

Even though the images were resized in the *Nottingham-Web* experiment, the layout in which they were displayed was not consistent. For most scenes many participants had to scroll to see the entirety of each image in the displayed pair. Worse still, the lack of control over layout means that those with a low screen resolution will have had to scroll to see one image stacked vertically atop the other, meaning that they would not be viewing both images on the screen at the same time and so could not make a direct comparison. As discussed in section 3.5, our platform accounted for this.

After making a preference choice in the *Nottingham-Web* experiment, a 'thank you' page was displayed to the participant. This page was redirected back to the main screen after a delay of one second via the use of a 'meta refresh'. A subset of web users are likely to have this functionality disabled and may have attempted to navigate back to the main screen either by using their browser's back button, which would have presented

them with the same image pair once more, or possibly by refreshing the 'thank you' page, which would have resubmitted their preference choice to the server. This is part of a wider possible source of sampling error: the *Nottingham-Web* experiment did not disallow multiple completions of the same comparison by the same observer. Not only does this permit data distortion through malicious intent, it allows more enthusiastic participants to contribute repeated data and so, intentionally or not, skew the results. The average web user has little incentive to complete all comparisons, as they are not under monitored conditions, and so may become bored with a web-based experiment fairly quickly and only submit a small number of preference choices. The visitors who are more likely to donate a larger number of comparisons are those who are already interested in such studies, such as other researchers and photographers. These expert observers will likely have inherently different preference choices to the general population. Our web-based platform employed more sophisticated randomisation of image pairs, data anonymisation and tamper-evident hashing to combat this. Participants were assigned image pairs randomly and could only submit their preference choice for a given pair once, unless they completed an entire experiment and restarted. Participants were not aware of which pair of algorithms they were comparing (unless they were such an expert observer that they could identify the algorithm from the image appearance).

Although it is feasible that these implementation details may have introduced biases into the *Nottingham-Web* experiment, these effects would likely be negligible. The more influential differences between the two experimental paradigms are more likely to be psychological effects elicited by the different contexts in which observers participated.

For consistency with the *Nottingham-Web* experiment, the *Lab-TMO* and *Web-TMO* experiments had a 'tie' option available to participants, allowing them to opt out of submitting a preference judgement for a particular image pair. This 'skip' option was rarely used: only 2.7% of comparisons were skipped in the *Lab-TMO* variant, and 4.5% in the *Web-TMO* counterpart. However, this slight increase in opt-outs for the web

experiment may go some way toward explaining the muted results we see in scenes such as 'Heron'. The ability of observers to opt out of a preference choice may lead to loss of data in situations where two image versions are very similar. If observers were forced to make a choice, they may take more time and consideration in choosing an image version which, albeit very slightly, outperforms the other; however, if they are given the ability to opt out they may quickly decide that the two versions are too similar to make a preference judgement, and those detailed discrimination data become lost. In the general case this should not incur too much penalty. If one image obviously outperforms another then the observer is unlikely to choose the 'tie' option. However, if two algorithms perform very similarly, as is the case with several of the algorithms compared in these experiments, then these detailed preference choices could lead to rank position swaps.

After completing the *Lab-TMO* experiment, observers were consulted about the factors which influenced their preference decisions. Many revealed that they used different image features to inform their decision about different scenes; rather than taking the image as a whole they used specific regions or features of each scene to influence their decision. Further to this, observers noted that certain images had certain recurring artefacts generated by some TMOs but not others, and would intentionally seek these artefacts out upon being presented with an image pair of a certain scene. These cues to decision making are learned as the observer completes more comparisons. An observer beginning the experiment may take more time considering the image as a whole before making their decision, but as they continue they learn which salient image features to look for. This could be an important factor separating the lab and web variants. It is known that the observers in the web variants did not all complete large numbers of comparisons before ceasing their participation. This implies that the rankings of the web variants are likely to be made up of a greater number of observers each undertaking a smaller number of comparisons, which in turn means that each comparison in the

web variants is more likely to have been made by a participant who is still unaware of these image features. This is another interesting point in favour of carrying out these kinds of experiments on the web – from this perspective at least, the balanced paradigm can actually be detrimental. As observers contribute larger amounts of comparisons, they learn to 'cheat', and so the data they continue to contribute may be biased. Meanwhile a web-based observer would likely cease participation before this inclination had manifested.

During consultation, the majority of observers in the *Lab-TMO* experiment mentioned the ambiguity in the instructions given. These were chosen to be as similar as possible to those in the *Nottingham-Web* experiment, and it is easy to see how differences of interpretation could arise. The prompt 'choose the image you think is better' could be interpreted as 'choose the image you think most represents a natural scene' or 'choose the image you think has more artistic merit' or even 'choose the image you would prefer to hang on your wall', all of which could produce vastly different results. Observers noted that, because they were partaking in the experiment under laboratory conditions, they felt that they should choose images which looked more natural. It is plausible that observers of the *Web-TMO* variant may have interpreted the prompt as in the latter interpretations above, considering that the sort of images traditionally associated with 'HDR photography' and 'tone mapping', especially among online photo sharing websites such as Flickr, are those over-saturated, extremely crisp images such as those shown in fig. 3.15, which are seen to be more artistic. If we suggest that the lab-based observers were choosing images which appeared more natural, while the web observers were choosing images which were more artistic (usually distinctly unnatural), then the two sets of observers were deriving completely different judgement metrics from similar instructions, due to the context in which the instructions were given (a formal, laboratory environment, or the informal environment of the internet). This may go some way to explaining the stability in the web results despite lack of correlation

(a) `https://www.flickr.com/photos/dcab2/3419699844`

(b) `https://www.flickr.com/photos/whosdadog/3106530236`

Figure 3.15: Examples of images found on image sharing websites such as Flickr when searching for 'HDR photography'

with the lab results, which was noted at the end of the previous section.

It is clear the question being asked of the observer is important. Prompts can easily be interpreted in many different ways depending on their environment. However, often in these kinds of experiment, we are seeking general observer preference. In both the TMO and C2G cases (and in many more like them), we are not looking for observer opinion on a specific metric such as 'which image appears more saturated?', but we are seeking to quantify a quality as broad and expansive as general observer *preference*.

All of these points share a common theme: transplanting paired comparison experiments onto the web does not, necessarily, mean the complete surrender of all control over the experiment. With consideration over presentation, and large numbers of observers, it is entirely possible to achieve reliable results.

3.9 Conclusions

The results in this chapter compare the outcomes of two differing experimental techniques. At the start of the chapter, we compared an existing web-based preference experiment to a lab-based replicate, and we go on to carry out the same task using the same lab-based data except with our own web-based counterpart. Given the similarity of the experiments, it is surprising that we do not find similar results. Comparing our *Lab-TMO* results to those of the *Nottingham-Web* experiment, we find only four of thirteen scenes show significantly high rank correlation, but comparing those same lab-based results to the results gathered from our *Web-TMO* experiment we achieve significant rank correlation for eight of the same thirteen scenes.

In light of this preferable performance obtained by the *Web-TMO* experiment (and indeed by *Web-C2G*), we can begin to suggest some key features of web-based experiments that future researchers should consider. Arguably the most important feature is to replicate standardised viewing conditions as closely as is feasible. Our web-based ex-

periments displayed images in a consistent side-by-side layout against a neutral, variegated grey background, with neutrally coloured interface elements surrounding the main interface – this is perhaps the starkest contrast to the bright yellow background and colourful interface elements seen in the *Nottingham-Web* experiment. However, these considerations have to be made in tandem with other objectives. For example, it could be argued that the optimum interface would be a full-screen neutral background displayed with only the images under comparison – i.e. with no other interface elements – but this interface would be hard to navigate for the user and takes no consideration into providing instruction. In taking experiments onto the web we have to be cognizant of the general expectations of web-based interfaces and design our experimental platform within those confines. Building interfaces in a web browser also introduces other concerns. Web-based observers will be completing experiments using a wide variety of different devices, which means we have to be aware of issues arising from differing display resolutions and layouts (portrait or landscape – particularly for handheld/mobile devices). Steps can be taken to ensure consistent display within reasonable limits – it is possible, for example, using modern web technologies, to ensure that images will always be presented side-by-side and that no scrolling will be necessary to view images in their entirety – a precaution not taken in the *Nottingham-Web* experiment. Due to inconsistencies in browser-based rescaling implementations, we recommend all rescaling be done server-side.

It was noted in section 3.8 that the *Nottingham-Web* experiment displayed a 'thank you' page after each comparison was submitted. In that section the possible detrimental implications of that particular implementation were discussed, but we suggest that such a step be omitted altogether, as it interrupts the observer and can disrupt their visual adaptation to the main experimental interface. In further technical implementation considerations, we suggest that future observers be aware of possible malicious intent and biasing of results by expert observers. Our web-based experiments made appropriate

use of randomisation, hashing, and data anonymisation to ensure that it was not apparent which image was generated by which algorithm (unless the observer was experienced enough to tell from the visual appearance of the image). These steps also ensured users could not repeat the same preference choice many times (comparisons were intentionally repeated as discussed in section 3.3, but observers could not bias results by voting beyond this designed repetition).

Also noted in section 3.8 was that the *Nottingham-Web* experiment employed a 'tie' option, which was replicated in the *Lab-TMO* and *Web-TMO* experiments, but not in *Lab-C2G* or *Web-C2G*. Such an option should be carefully considered, as it allows observers to opt out of preference choices at the expense of lost data – we advocate for a forced choice paradigm with no such option. We also recommend judicious consideration over the instructions given to observers – these should be as precise as possible to protect against misinterpretation.

Our experiments attracted sufficiently large numbers of participants, indeed the *Nottingham-Web* experiment did not reach the five hundred comparisons level which, according to our results, seems to be the point at which stable results are achieved. Future researchers should be mindful that they do not attempt to draw conclusions from experiments before sufficient judgements have been made – chapter 4 introduces a possible consideration.

Finally, we recommend a statistically robust Thurstonian analysis of results (as opposed to IQRI for example). Such an analysis, in concert with the supplemental analyses discussed in section 2.6, permits a thorough understanding of the results attained through paired comparison experiments, and can help to reveal any deficiencies (for example if competing algorithms do not produce results which are perceptually different to a significant degree).

Lab-based paired comparisons, with all the control and standardisation under which they are typically carried out, are seen by many as the 'correct' way of performing

visual psychophysics, while web-based techniques are criticised and often disregarded for their lack of traditional control. However, we have shown that when sufficient care is taken over presentation and the other practicalities of web-based experiments, often the results from web-based paired comparisons can closely correlate with those carried out under laboratory conditions. It is also shown that, when the results do not correlate, this can be attributed to lack of discriminatory power among the images being compared, or is indicative of an underlying problem in the images under comparison that may suggest they are ill suited for this type of experiment in general.

We observe that convergence in results can be met, so long as careful consideration is given to image presentation, the phrasing of the prompt given to the observer and whether or not general web users may have a predisposition to favour certain images that a lab observer may not. We also note that many previous studies in this area have exhibited poor results that may be attributable to small numbers of observers, or to samples of web users that are not generally representative of the observers on the web at large.

Chapter 4

Temporal Stability of Ranks for Image Preference

When evaluating observer preference among differing image processing algorithms, we are often interested in assigning a rank order to a collection of competing algorithms. In so doing, we will often structure experiments as described in section 2.5. However, these types of experiments can present practical and logistical obstacles such as those discussed in chapter 3, namely the setup of the experimental interface and viewing conditions, and time and expense of recruiting sufficient numbers of observers. Given these challenges of performing preference experiments, it would be desirable to have a measure of the stability of the ranking obtained from those observers that have completed the experiment to date. With such a measure, it would be possible to assess whether the current number of observers represents a sufficient sample size. If so, then it may no longer be necessary to continue the laborious process of recruiting more observers. For web-based experiments, the measure could indicate when to cease the study.

In this chapter, we use the data from some existing published preference experiments, and from some new experiments discussed in chapter 3, to show that a measure of the stability of a ranking can be determined solely from its current state. To de-

rive this measure we use a novel perturbation analysis of the score matrix (described in section 2.6.1) constructed during the analysis of paired comparison experiments. We determine the minimum number of anomalous observers (i.e. those who are, for each comparison, voting contrary to the current consensus) that would be required to change the current ranking to a significant degree.

4.1 Introduction

Psychophysical experiments, in particular paired comparison experiments, are a key part of the evaluation process of many image processing developments. Yet, despite their pervasiveness, there are still many differing approaches to their execution emerging in the literature. A key contributor to this variance, as mentioned in chapter 3, is the difficulty associated with recruiting observers. This difficulty, and the differing resources available to researchers to address it, can cause observer numbers to vary greatly between experiments, from the handful in Connah et al. (2007), to hundreds as in Sprow et al. (2009), and into the thousands as seen in chapter 3. However it does not necessarily follow that we should reject conclusions from experiments that have lower numbers of observers. It is quite feasible that lower sample sizes could be sufficient to yield valid results. What would be useful then, is some measure of the robustness of the current results of an experiment, at which point the quantity of observers attained so far can be deemed sufficient to draw reliable conclusions.

Over time, as data is gathered from more observers, the analyses of psychophysical experiments generally stabilise (although not always – if two or more images perform similarly then the perceived preference among them, especially from a rank ordering perspective, could remain perpetually unstable), and at such point continuing to recruit more observers is no longer strictly necessary. In order to test the stability of a set of results at a given point in time, we seek to test the resilience of that result set to change by re-posing the question of "is the current quantity of observers sufficient to draw reliable conclusions?" as "assuming all new observers are in ubiquitous disagreement with the current results, how many new observers are the current data resilient to?".

To answer this question, and in so doing meet our objective of quantifying sufficient observer numbers, we introduce a method built upon Thurstonian (Thurstone, 1927) analyses of paired comparison experiments. The method could feasibly be modified for use with other methods of analysis, or with different experimental paradigms.

4.2 Anomalous Observers

The proposed method centers on the notion of simulated *anomalous observers*. We define an anomalous observer to be an observer whose preference judgements are always contrary to the current consensus. At a given point in time (after some number of real observers have completed the experiment), we seek to determine how many anomalous observers the current ranking is resilient to, or equivalently, how many anomalous observers would be required to affect significant change in the current results.

Our simple approach is to compile a preference matrix after each real-world observer completes the experiment, and then simulate the addition of anomalous observers to the experimental results. The effect of the simulated observers on the preference matrix is calculated after each simulation and, given some significance measure, the output of our stability measure is the number of anomalous observers required to cause significant change. The implication of this concept is that the anomalous observers represent a worst-case scenario – if a result set is resilient to the addition of n anomalous observers, then it would require *at least* n additional real observers to cause a change in the results. By comparing n with the current number of real observers, we can estimate the likelihood that the current sample size is sufficient.

4.2.1 Choice of Appropriate Significance Measure

The choice of significance measure, to determine the significance or otherwise of the effects of the anomalous observers, is important. However, due to differing objectives across experiments, it is infeasible to suggest a measure that works in all circumstances. For example, in many pieces of research the objective is simply to discover a rank ordering of some collection of competing image processing algorithms – the individual scores representing the differences in scale between each algorithm is of no concern. In such cases, the simple approach of declaring any change in rank ordering to be signific-

ant may well be sufficient.

In other cases, there may be certain treatments with which the experimenter is more concerned. If the psychophysical experiment is being carried out in order to evaluate the effectiveness of a new image processing algorithm, it may be that the experimenter only considers rank position changes concerning that particular algorithm important, and effectively ignores any changes amongst the 'also-ran' treatments.

Alternatively, a more comprehensive tool such as Kendall's rank correlation coefficient (Kendall, 1938) (described in section 2.6.5) could be used, with some prescriptive significance level as a threshold. However, a problem with only considering ordinal rank correlation is that rank position swaps between treatments with only small intervals between them would be considered as equivalent to rank position swaps between treatments separated by large intervals (as demonstrated in fig. 2.9), which in many cases would be undesirable. To address this situation a measure such as that defined by Sprow et al. (2009) (described in section 2.6.5), which is based on a chi-square test, could be employed. For these purposes, the real results can be used as the 'expected' distribution, while the results after the addition of the anomalous observers are treated as the 'observed' distribution.

Any of these approaches, and many others, could be viable, depending on the task at hand. In section 4.3 it is shown that the measures discussed above reveal similar trends when applied to our resilience test, although they may produce differing absolute values.

4.2.2 Creating Anomalous Observers

Once an appropriate significance measure has been chosen, the implementation of our method is simple. First, a 'ground truth' preference matrix F is compiled from the data provided by the real observers to date (these preference data may then be transformed into some other representation as required by the chosen significance measure or, if the experimenter is only concerned with a subset of the treatments, a submatrix of F may

Algorithm 2 Simple algorithm to increase perturbation of the frequency matrix

```
1: function RESILIENT
2:     c ← 1
3:     while ¬SIGNIFICANT(F, F + cP) do
4:         c ← c + 1
5:     end while
6:     return c
7: end function
```

be used). If n treatments are being compared, then F is an $n \times n$ matrix, and F_{ij} denotes the number of times algorithm i is preferred over algorithm j.

To simulate the data for one anomalous observer a new matrix P is created of the same size as F where

$$P_{ij} = \begin{cases} 1 & \text{iff } F_{ij} < F_{ji} \\ 0.5 & \text{iff } F_{ij} = F_{ji} \\ 0 & \text{otherwise} \end{cases} . \tag{4.1}$$

P is then multiplied by the number of repetitions in the particular experiment - a common paradigm is to display every image pair twice: once in $[AB]$ format, then again as $[BA]$, often this is then repeated once more, to give a total of four repetitions for each image pair. At this point P represents one anomalous observer voting contrary to consensus for the entire experiment.

To arrive at our final quantity of anomalous observers required to affect significant change, we follow algorithm 2. First, a counter $c = 1$ is initialised, then P is multiplied by c. The chosen significance measure is then applied to F and $F+cP$ - if the difference is significant then exit and return c, otherwise increment c and loop until the change is significant.

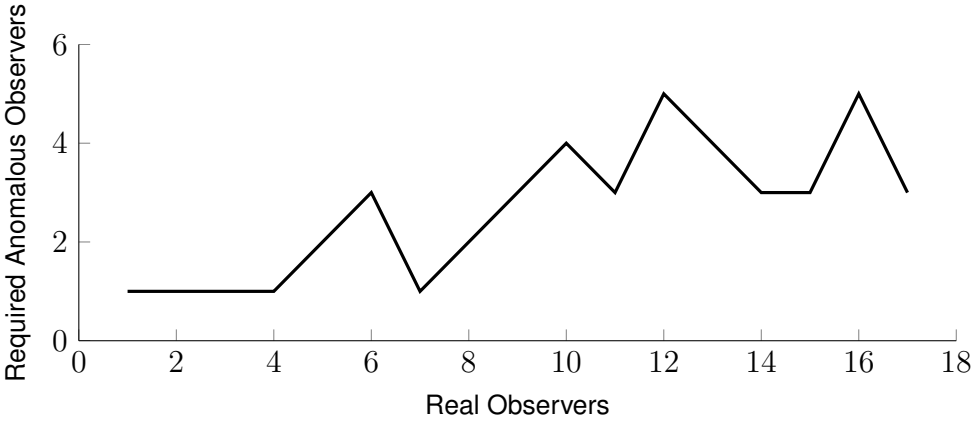
4.3 Results

To demonstrate what can be revealed by the use of this technique, we use existing data from some paired comparison experiments. Figure 4.1 shows how the number of anomalous observers required to affect significant change in results increases (generally) with the number of real observers. This is perhaps intuitive: given more observations, the ranking will become more resilient to change, but with this new statistical tool it is possible to quantify this observation.

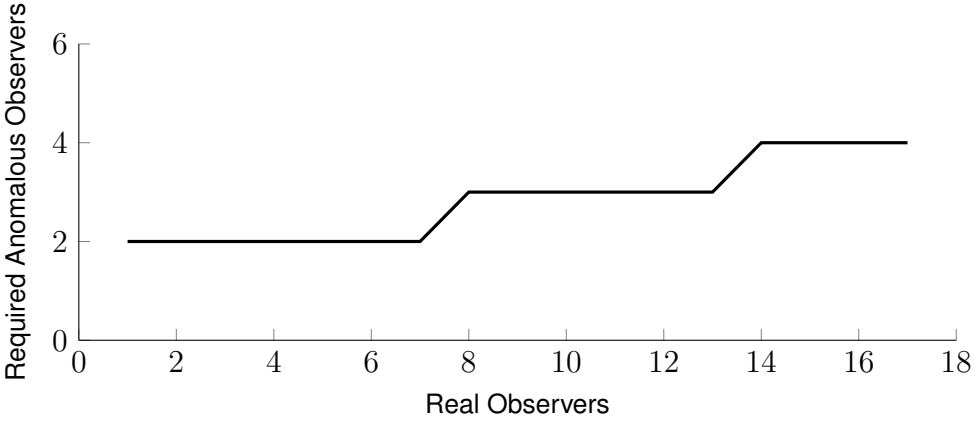
Figures 4.1a and 4.1b show the resilience of the rankings produced by the *Lab-C2G* experiment (Connah et al., 2007) introduced in chapter 3, using rank order change and the Sprow et al. chi-squared test respectively as the significance measures. In these examples, any rank ordering change was deemed significant, and an alpha level of 0.05 was used as the significance criterion for the Sprow et al. test.

The spikes in fig. 4.1a demonstrate the sensitivity of using rank order change as the significance measure. These fluctuations likely arise from the fact that, while real observer numbers are low (as is the case for this particular experiment), each single observer (real or anomalous) can have a significant impact on the current ranking – the larger differences in images are identified early on, but the more nuanced differences may take many more observers to identify. While these smaller differences are surfacing, small changes in score can result in big changes in rank ordering – exactly the problem described in fig. 2.9 and the motivation behind the introduction of the Sprow et al. test. The smoothness benefit of the more incisive Sprow et al. measure is clear to see in fig. 4.1b.

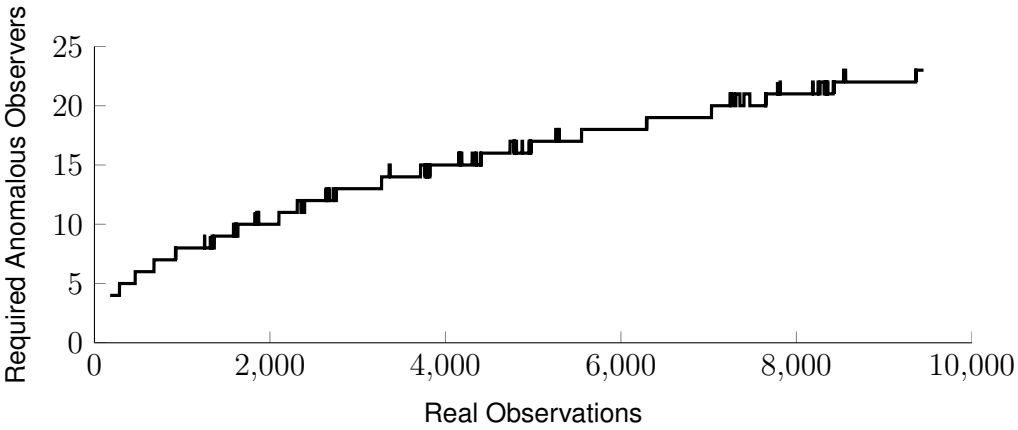
Figure 4.1c shows data taken from the *Web-C2G* experiment discussed in chapter 3. As the web-based data is unbalanced (not every observer necessarily completes every preference judgement), the x-axis in this plot shows the number of *observations* made – it is still assumed, however, that one simulated anomalous observer completes every preference choice, and so the results represent a worst-case scenario. In this case, for



(a) Lab-C2G *experiment (Connah et al., 2007), any rank order change is significant*

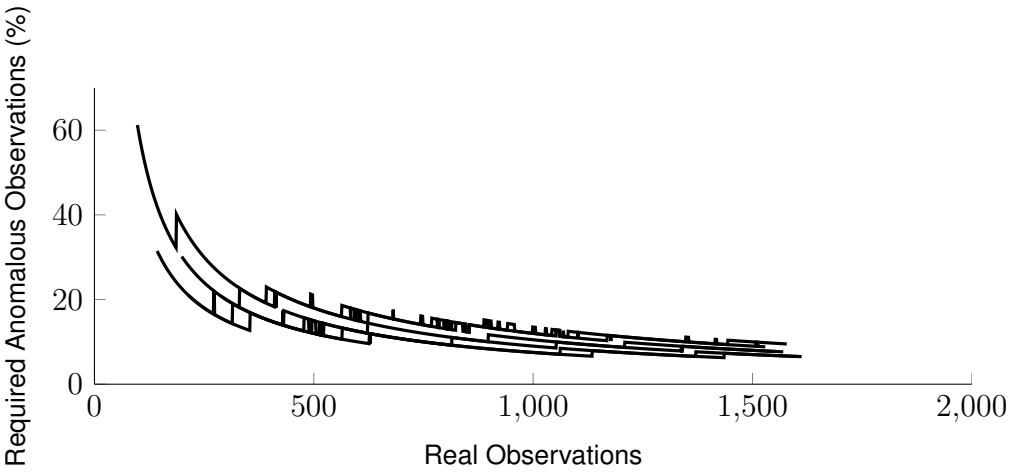


(b) Lab-C2G *experiment (Connah et al., 2007), using Sprow et al. measure of significance*

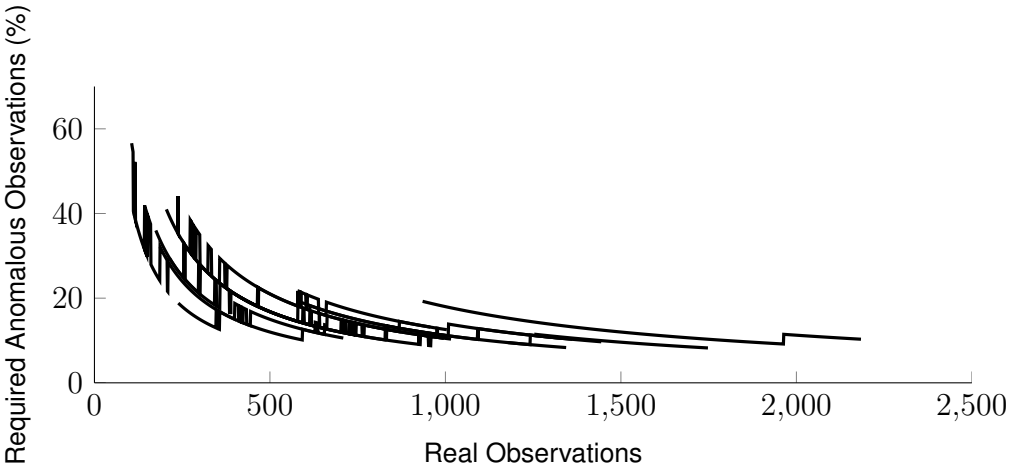


(c) Web-C2G *experiment, any rank order change is significant*

Figure 4.1: Resilience of rankings to anomalous observers



(a) Web-C2G



(b) Web-TMO

Figure 4.2: Resilience of ranks generated by *Web-C2G* and *Web-TMO* experiments, expressed as a percentage of real observations made. Any rank order change is considered to be significant. The plots contain one line for each scene in both experiments

simplicity, we deem any rank order change to be significant. These data show that our approach can be similarly applied to large-scale data as well as to experiments with smaller observer numbers, and in so doing reveals the same general trends. However, as shown in fig. 4.2, the larger scale data also reveals some interesting further insights.

Figure 4.2a depicts the required anomalous observations for the *Web-C2G* experi-

ment as percentages of the current numbers of real observations. The data are broken into multiple plots, one for each different scene in the experiment, but all using the same collection of image treatments. The reason for the downward trend may not at first be apparent, but this is due to the initial small numbers of real observers being similar in scale to the required anomalous observers – the base case being results after one real observer require only one anomalous observer to affect change, which in percentage terms is 100%. As the number of real observers increases, so then does the required number of anomalous observers, but it is not a linear relationship – hence the shape of the plots. Interestingly, stability is achieved at approximately 10%, and so we may prescribe for future iterations of this experiment that 10% is a target amount for reliable results. In fig. 4.2b, this analysis is repeated for the *Web-TMO* experiment (also in chapter 3), and similar convergence is found at approximately 10%. The trends in fig. 4.2 suggest that all the scenes in both our web-based experiments achieved stable rankings and, in support of the assertions in chapter 3, did so at around the 500 comparisons level.

4.4 Conclusions

This chapter presents a new technique for quantifying the resilience of a ranking from a psychophysical experiment to anomalous data, and demonstrates how it can be used to estimate whether sufficient observers have completed a given experiment to provide reliable results and conclusions. Given an appropriate significance measure, this technique can be used to provide a worst-case estimate of the quantity of new observers required to change the results to a significant degree.

Unfortunately it is not possible to prescribe some target value for our metric that experimenters can use as a generally applicable objective. The construction of this method is entirely dependent on the experiment at hand, the data gathered therefrom, and what the experimenter deems to be significant change in their results. For example,

from an experiment with two or more very similarly performing image treatments may arise a situation where the rank ordering of those two treatments is in constant flux with each new observer - in this situation our metric, if using rank order change as the significance measure, would report that the results are only resilient to one anomalous observer no matter what the quantity of real observers. Due to this limitation, this metric should not be used in isolation - the context of the particular experiment should always be considered, and the value delivered by our metric may be subject to further inspection.

We observed that our own web-based experiments depicted in fig. 4.2 achieved stability after approximately 500 comparisons had been completed, but this observation was made with the benefit of hindsight - our experiments continued beyond this point and so we have sufficient data to make the observation. However, the objective of the development of the metric, and indeed the reason why future experimenters may wish to apply it, is so that experiments can be ceased at the point when sufficient comparisons for stable results have been completed. So then, how can such a decision be made without the benefit of hindsight? Recall that the quantity delivered by the metric is a worst-case estimate - i.e. a value of 10 indicates that *at least 10* new observers, voting contrary to consensus, are required to affect change. In reality, new observers are unlikely to be completely anomalous. A plausible method for estimating the likelihood of recruiting anomalous observers would be a Monte Carlo simulation based upon the current observer population. For example, if our metric indicated that the current ranking is resilient to a number of anomalous observers equal to 10% of the current number of real observers, a simulation that extracted random samples of 10% of the current population and generated rankings from only their input would be able to suggest the likelihood of recruiting that amount of anomalous observers given their current distribution in the sampled population.

It is hoped that future work can build upon this metric to deliver a measure that is

more statistically robust - i.e. which can, without further inspection, and without being data dependent, deliver a measure of the statistical power of the current number of observers recruited into a particular experiment.

Chapter 5

Illuminant Estimation for Colour Naming

With the validity of data sourced from large-scale web-based experiments demonstrated in chapter 3, this chapter continues with a demonstration of the utility of such data. We use existing data delivered by a very large-scale web-based colour naming experiment to train a computational colour naming model, and with that model seek to answer the question of whether existing illuminant estimation techniques can be used to provide the colour naming model with data of sufficient quality such that stable colour names can be delivered across changes in illumination. In so doing, we observe that colour names provide a perceptually important representation of colour, and so we test whether colour names can be successfully utilised as a meaningful representation of image content, by means of an object indexing task.

5.1 Introduction

There is an ever-growing body of research in the field of illuminant estimation, and while this is an important (and popular) endeavour, progress has been slow. Over recent decades there have been many papers published which have inched the state-of-the-art closer toward a solution, but there remains much room for improvement. The literature in this field includes some work which poses the question of whether the state-of-the-art is good enough to produce satisfactory results for some concrete objective. Specifically in the cases of Funt et al. (1998) and Finlayson et al. (2002a) that objective is object indexing, as introduced in section 2.8. In this chapter we revisit the approaches used by Funt et al. (1998) and Finlayson et al. (2002a), but also introduce colour naming as an objective – i.e. using the same suite of algorithms used to test the object indexing objective, can we successfully colour-correct images taken under varying illumination conditions such that they appear the same as those taken under a canonical white illuminant, to such a degree that the colour names designated to the colours in the image by a computational colour naming model are the same?

After demonstrating the stability or otherwise of colour names across illuminant estimates of varying accuracy, we then, in section 5.4, reintroduce the object recognition problem using the knowledge gained from the colour naming experiment. Specifically, we investigate the feasibility of using the distribution of colour names present in an image as an image descriptor with which we can perform object recognition. From there, in section 5.5, we suggest that colour names are an inherently human-like way of describing image content and demonstrate a simple method by which the same scheme can be used to facilitate both machine object recognition and image search based on human queries. This chapter concludes in section 5.6.

5.2 Background

In 1998, Funt et al. (1998) sought to examine whether the field of illuminant estimation had matured to a point where results from the contemporary methods were good enough to be useful, in particular for the case of colour-based object recognition. To this end, colourful objects are represented by their histograms and object identity is defined as the closest histogram found in an object histogram database. Unfortunately, the conclusion at that time was that the best methods of the time were not reliably accurate enough to deliver results which were sufficient for successful object recognition. Later, Finlayson et al. (2002a) re-examined the same question with the addition of a newer illuminant estimation method (Finlayson et al., 2002b) and, contrary to the previous work, concluded that the performance of this newer method was sufficient for successful object recognition; though the success of that method lay in using an algorithm that required extensive, detailed, calibration. Unfortunately the use of such an algorithm is not always possible, for instance when dealing with images from the web, as little information besides the image data itself is known. These negative results are unfortunate as colour has been shown to be a useful cue for object recognition and image indexing when the illuminant colour is known (Flickner et al., 1995; Swain and Ballard, 1991).

In complementary research, colour naming has also proven to be useful for object recognition, image indexing and image search (Flickner et al., 1995). In image search, the query "red car" requires some model of colour naming in order to be served satisfactorily. The task of separating red cars from all others is, however, dependent on being able to disambiguate the surface colour of the car from the illumination conditions. For example the lighting conditions of a car showroom with red spotlights may cause a white car to appear reddish, and so presenting an image of that car would be inappropriate for our "red car" query. Retrieving pictures of a red car clearly also requires a semantic understanding of what a "car" looks like, which is of great interest (Everingham et al., 2010), but is not within the scope of this thesis.

In this chapter, we evaluate a collection of simple illuminant estimation algorithms (described in section 2.2) which, crucially, require little or no calibration. In section 5.3, we take a similar approach to that of Funt et al. (1998), but instead of using object recognition performance as a metric, we analyse the ability of these algorithms to produce images which can be correctly labelled with colour names.

5.2.1 Munroe Dataset

To build our colour naming model (as described in section 2.7), we used a freely-available colour naming dataset compiled via web-based data collection by Munroe (2010), which was also used by Heer and Stone (2012) and Beretta and Moroney (2012). As part of Munroe's data collection exercise, participants were first asked to complete basic demographic information before continuing to label samples from the sRGB (Stokes et al., 1996) cube which were displayed against a white background. Participants could complete as few or as many responses as they wished; there were no limits on participation levels and no limits on time to label each sample. There were no constraints on the labels that participants could use. Due to this, we took some pre-processing steps when using this data set to remove spurious and "spammy" data. Munroe calculated a spam score for each participant based upon the rarity of their labels in comparison to other participants, and using the same labels for many highly differing colour samples. We only used data from participants with a spam score lower than the median. The dataset also contains data from many languages - we extract only those labels contributed by participants who self-reported as native English speakers. Finally we extracted only data corresponding to the basic colour terms of Berlin and Kay (1969); this extends to discarding data with names such as "dark green" and "light green" and using only labels of the basic form "green".

For our experiments we attempted multiple approaches at the above pre-processing steps, such as conflating responses for "dark green" and "light green" into the stemmed

"green" label, and using different thresholds for spam scores. However, we found that, within reasonable limits, these differing configurations had little impact on the outcome of our experiments. We also note the work of Moroney and Beretta (2011) and Beretta and Moroney (2012), who experimentally validated these data by means of a controlled, lab-based, validation experiment. They showed that excessive filtering of the data is not necessary.

It should also be noted that, following the lead of other authors (Heer and Stone, 2012), we built our colour naming model in the CIE L*a*b* colour space. That is, the Gaussian mixture model described in section 2.7 is constructed in CIE L*a*b*, and the process of assigning a colour name label to a pixel value includes an implicit conversion into that colour space. However, for the sake of simplicity we shall ignore this detail going forward and simply refer to the process of assigning colour name labels given RGB pixel values.

5.3 Resilience of Colour Names to Illuminant Estimation Errors

Funt et al. (1998), and later Finlayson et al. (2002a), evaluated how performance in an object recognition task degraded with illuminant estimation accuracy. To do this, both pieces of work used the colour indexing approach of Swain and Ballard (1991), which represents images as colour histograms and then uses *histogram intersection* (see section 2.8) as a metric of image similarity. The colour histogram forms a query which is matched against a database of histograms for an object data set – the closest matching histogram identifies the query object. The ability of this technique to function after inaccurate illuminant estimation reduces to whether pixels are assigned to the same histogram bins as they would be given perfect illuminant estimation. Herein lies the problem: with anything other than perfect illuminant estimation, some colour values shift

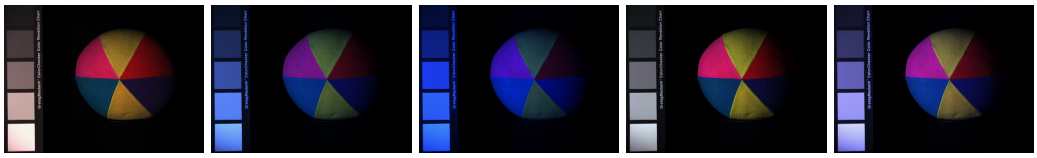


Figure 5.1: Differing illumination conditions in SFU Object Recognition dataset (Funt et al., 1998). The same object is shown under five different lighting conditions

across the boundaries of histogram bins. This becomes more of a problem as the level of quantisation of the colour space becomes finer (in order to support better discrimination between similar object histograms), or as the illuminant estimation becomes less accurate.

Figure 5.1 demonstrates the importance of illuminant estimation accuracy; under these radically varying lighting conditions (for the purposes of illustration these images have not been corrected for the illumination conditions), it easy to see how the colour values can drift from one histogram bin to another as the illumination conditions change. However, even with no correction for illumination, a human-made labelling of the colour names present in the object would likely be consistent across the different images.

In this chapter, we are interested in how computational colour naming degrades with illuminant estimation accuracy. Our hypothesis is that colour names, because they represent a coarse, perceptually important, quantisation of colour, will be sufficiently stable to support object recognition even when illuminant estimation is far from accurate. For context, we also test the coarseness of traditional colour histograms – although Funt et al. (1998) found colour constancy not to suffice for object recognition, they used a fine quantisation of colour space. Inaccurate illuminant estimation implies that a colour value (RGB triplet) corresponding to the same surface is mapped to different histogram bins under different illumination conditions. The more inaccurate the estimation, the poorer the colour-based recognition. Before presenting object recognition results based purely on colour names, we wish to investigate the quantity of colour values which are

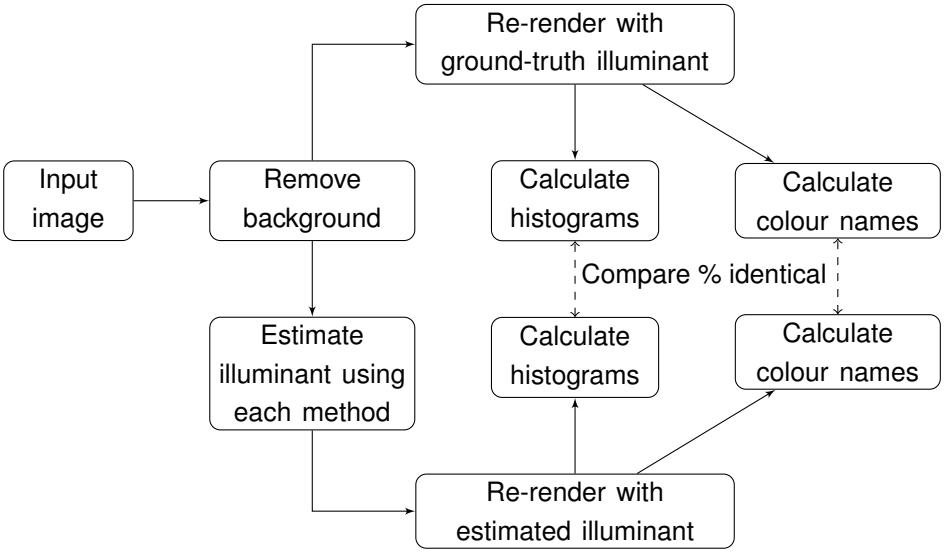


Figure 5.2: Process for determining correctness of bin/name assignments

mapped to the same histogram bin using varying quantisations of the colour space, and the quantity which are identified with the same colour name, under different illumination conditions normalised using a range of colour constancy algorithms.

The mean results for the collection of illuminant estimation algorithms described in section 2.2 are presented in table 5.1. The test images used were from the published datasets of Barnard et al. (2002) – we used a hybrid set of all the images in the *mondrian*, *specular*, *metallic* and *fluorescent* collections (71 objects under nine common illuminants). A simple threshold technique was used to isolate the objects in the images from the background (the objects are presented against a black background). To quantify bin/name correctness, each image is first treated with each illuminant estimation method (the illuminant is estimated, and then the scene is re-rendered under a white illuminant, based on this estimate) and then categorised by histograms with several levels of bin quantisation, and also labelled with the colour naming model described above. These categorisations are compared to the categorisations with perfect illuminant estimation (the data sets include ground-truth data), and the score given to each image is the proportion of pixels categorised identically – see fig. 5.2. These res-

Table 5.1: Correctness of bin/name assignments (expressed as percentage of correctly assigned pixels) with varying illuminant estimation errors. Values shown are the means across all images

Method	Error	Histograms (%)				Names (%)
		16×16	8×8	4×4	2×2	
Do Nothing	17.12°	25	41	67	83	67
Actual	0.00°	100	100	100	100	100
Grey World	10.97°	33	51	72	80	73
Max RGB	10.03°	37	58	79	87	79
Shades of Grey	8.06°	47	64	81	88	83
General Grey World	7.82°	49	65	81	87	84
1st Order Grey Edge	7.22°	49	65	83	89	84
2nd Order Grey Edge	6.75°	51	67	84	90	85
Pixel-Based Gamut Mapping	6.37°	53	69	84	90	87
Edge-Based Gamut Mapping	7.11°	39	57	82	89	81

ults are averaged across all images and aggregated by the illuminant estimation method used. The angular error of the illuminant estimates made by the different algorithms is also detailed in table 5.1.

When the actual illuminant is used, all normalised images are perfectly colour constant and 100% of the colours are mapped to the same names and bins. Also, as a general trend, the better the illuminant estimate, the more colours are stably mapped to the same bin/name. Encouragingly the colour name designations are better than the bin assignments of all but the 2×2 histograms.

As noted above, illuminant estimation is subject to an unknown scaling factor. To counter this, previous researchers have built their image histograms in an intensity-invariant chromaticity space such as that described in section 2.8.1. These two-dimensional histograms are those evaluated in table 5.1. However, for colour naming we require full three-dimensional colour data, so that "black" can be distinguished from

Figure 5.3: Effect of exposure correction step

"grey" and "white", for example. To facilitate this we implemented a naïve "exposure correction" step after illuminant estimation. The images were scaled so that the pixel value at the 95th percentile became equal to 255. This simple approach, as demonstrated in fig. 5.3, sufficed for correct colour naming on our test images (with varying test data the scaling factor would likely need to be adjusted – we used slightly differing scaling factors for the other datasets described below). The name correctness measure in table 5.1 is taken after this scaling step. This step is of particular importance for the SFU datasets (Barnard et al., 2002; Funt et al., 1998), as they are deliberately underexposed to avoid any clipped pixels.

The results for the chromaticity histograms shown in table 5.1 show that stability is greater for more coarsely binned histograms. This is intuitive, since pixel values have to "move" further from their correct value before they would be miscategorised. Indeed the name-based histograms (with 11 bins – one for each colour name) seem to fit in between those histograms with the most similar coarseness ($4\times4 = 16$ bins, and $2\times2 = 4$ bins). The response to this observation could be to opt for coarser histograms for the task of object recognition; however, coarser histograms are defined by fewer parameters, and so offer less discriminatory power among large databases of such histograms.

Table 5.1 reveals that the stability of colour name labelling of pixel values after inaccurate illuminant estimation is comparable to the stability of histogram binning for similarly quantised histograms, with a minimum of 73% of pixel values labelled consistently for the poorest performing algorithm (ignoring the 67% 'Do Nothing' result).

5.4 Object Indexing

Funt et al. (1998), and Finlayson et al. (2002a) already evaluated the performance of various illuminant estimation techniques in the framework of the colour indexing method of Swain and Ballard (1991) described in section 2.8. However, they used pixel chromaticity values as described in section 5.3 to build their histograms. We shall do the same for comparison, but also build a new set of histograms based on the distribution of colour names in each image. For our first experiment we used the same collection of images as Funt et al. (1998), referred to as the *SFU Object Recognition* dataset.

Figure 5.5a describes the results comparing mean illuminant estimation error against object recognition performance for the various chromaticity histograms, and for the name-based histograms. Each data point corresponds to one of the algorithms tested in table 5.1. We used the same object recognition score as Finlayson et al. (2002a) – the *match percentile*, described in section 2.8.2. The *average match percentile* indicated in figs. 5.5 and 5.6 is simply the mean of the aggregated match percentiles.

We see in fig. 5.5a that the name-based histograms outperform the chromaticity histograms quite significantly, and so it is worth revisiting the paradigm of employing chromaticity histograms for this purpose. The rationale for the use of chromaticity histograms is to account for the unknown scaling factor of the illuminant estimation methods. However, we are applying our "exposure correction" step for this same purpose in order to facilitate colour naming based on three-dimensional data. In light of this, we could construct three-dimensional RGB histograms using the same "corrected"

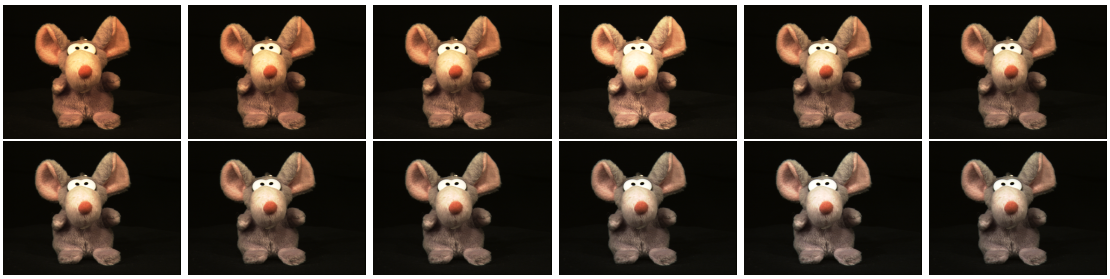
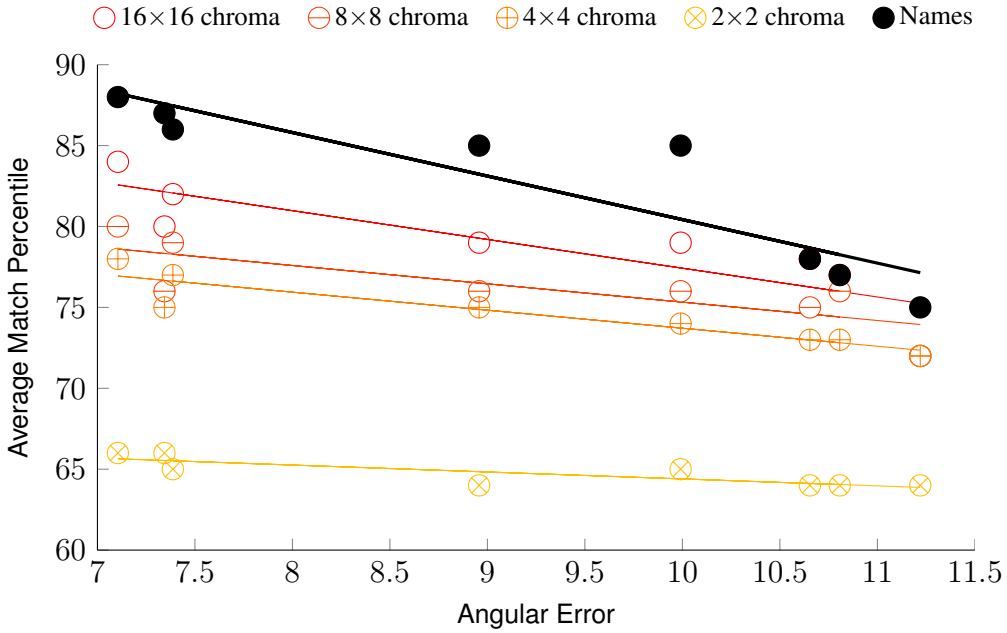


Figure 5.4: Differing illumination conditions in ALOI dataset (Geusebroek et al., 2005). Lighting conditions vary from 2175K in the top left image to 3075K in the bottom right

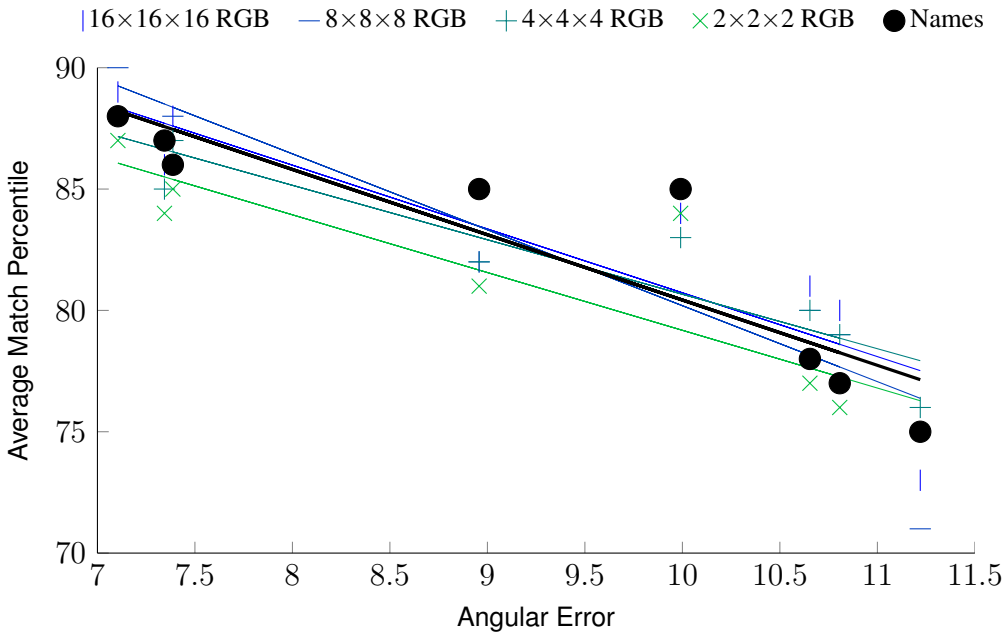
images as used for the name histograms.

To test this, and to validate the legitimacy of this simple correction approach, we introduce another image database: *The Amsterdam Library of Object Images (ALOI)* (Geusebroek et al., 2005). This expansive dataset has been constructed to fill many needs within the field of colour research, and as such it includes a database of objects photographed under varying illumination. This is a much larger dataset than the others described here, containing images of one thousand different objects viewed under twelve illumination conditions (ranging from 2175K to 3075K, as seen in fig. 5.4), and so represents a more real-world scale of image database. We carried out our experiment with the entire set of objects, using the 3075K illuminant as our canonical illuminant. The camera used in constructing the dataset was white-balanced to this illuminant, and so 3075K appears white while the other illuminants towards 2175K appear more reddish.

In figs. 5.5a and 5.6a we see the same general trends for the chromaticity and name-based histograms, as well as for RGB histograms in figs. 5.5b and 5.6b. It is expected that the RGB histograms should outperform the chromaticity histograms (under the assumption that the "exposure correction" is sufficient), as the additional dimensionality of the data representation allows for greater discriminatory power among the database of object images. What is surprising here is that the name-based histograms perform

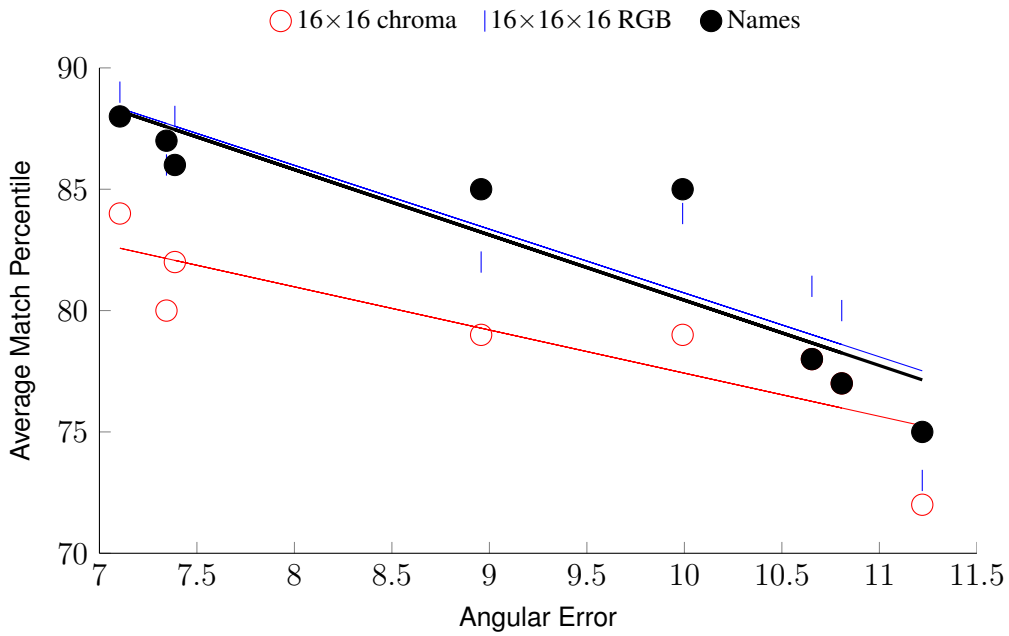


(a) Chromaticity vs colour names



(b) RGB vs colour names

Figure 5.5: Object recognition performance for chromaticity-, RGB-, and colour-name-based histograms for SFU Object Recognition dataset (Funt et al., 1998)

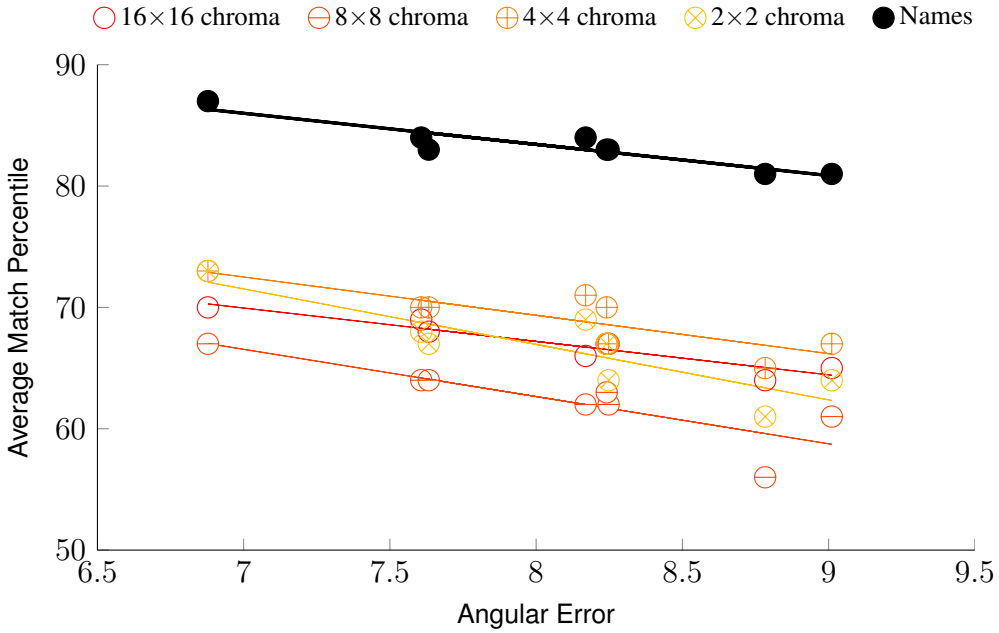


(c) Best-performing chromaticity and best-performing RGB vs colour names

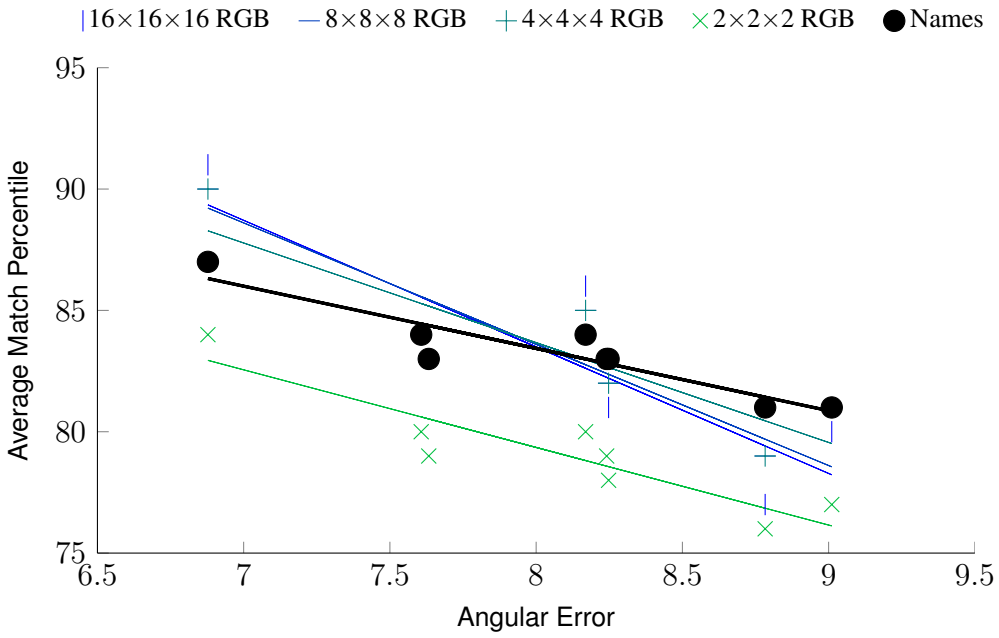
Figure 5.5: Object recognition performance for chromaticity-, RGB-, and colour-name-based histograms for SFU Object Recognition dataset (Funt et al., 1998) *(cont.)*

similarly well and, under the conditions of less accurate illuminant estimation, often outperform the RGB histograms.

Encouragingly, performance of the name-based histograms is on the same order as the best colour histogram approaches, but we use just eleven colour names instead of the $8\times8\times8 = 512$ or $16\times16\times16 = 4096$ bins used for the traditional colour histograms. This is in contrast to the experiments of Funt et al. (1998), as the results they reported were in part due to discarding intensity information. Even when we do encode intensity we need many more bins for colour histograms compared with a name based signature. The power of using colour names is therefore established.

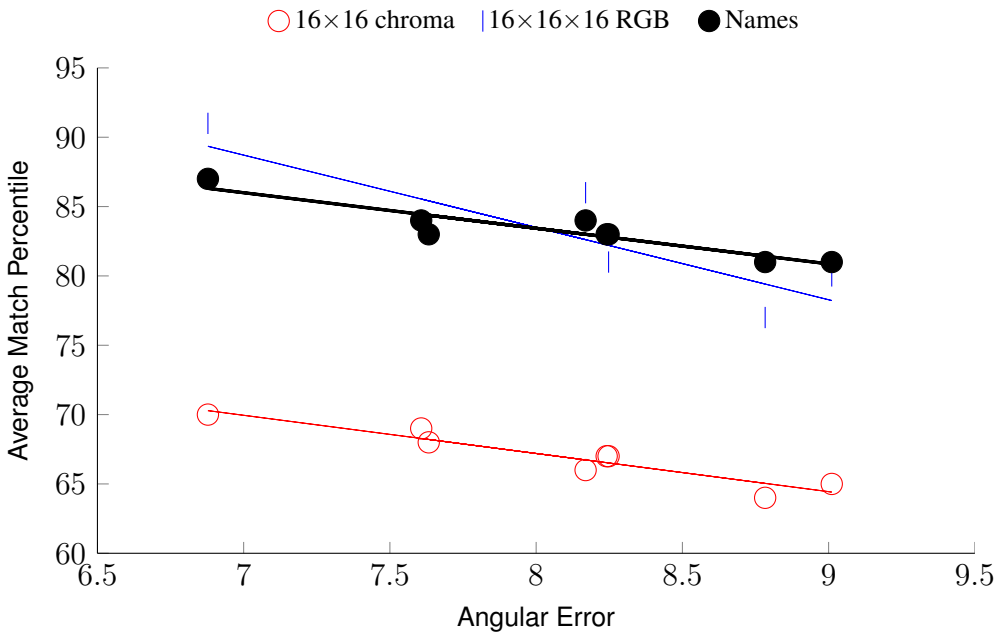


(a) Chromaticity vs colour names



(b) RGB vs colour names

Figure 5.6: Object recognition performance for chromaticity-, RGB-, and colour-name-based histograms for ALOI dataset (Geusebroek et al., 2005)



(c) Best-performing chromaticity and best-performing RGB vs colour names

Figure 5.6: Object recognition performance for chromaticity-, RGB-, and colour-name-based histograms for ALOI dataset (Geusebroek et al., 2005) *(cont.)*

5.5 Query by Colour Name

If we consider specifically the case of image search, image descriptors that are compact and resilient to errors in illuminant estimation are particularly valuable. The value of this compact representation becomes more important as the size of the image search corpus becomes larger. This is also a favourable representation for this particular application as it can more closely model how a human may describe an image. For example, an individual searching for a beachside scene may be able to describe that scene as [40% yellow (for the sand on the beach), 40% blue (for the sea), 10% brown, 10% green (for the customary palm tree in the foreground)]. This very human-like descriptor lends itself excellently as a key with which to query an image database, and it can be used directly with the histogram intersection technique described above.

To test this scenario, we performed an experiment using histogram intersection to

(a) Teddy Bear with Santa Hat *(b) Juggling Ball* *(c) Rubber Duck* *(d) Snowman Ornament*

(e) Santa Ornament *(f) Papa Smurf* *(g) Diet Coke Can* *(h) Marmite Jar*

Figure 5.7: Subset of ALOI dataset (Geusebroek et al., 2005) used for query-by-colour-name experiment (Shown larger in fig. C.1)

query the ALOI image dataset with colour name histograms described by human volunteers. We first selected an appropriate test set of objects to search for. We chose these test objects so that human labellers could gain a colour understanding of the object without first having to see the corresponding image and without being explicitly told any colour names. For example "rubber duck" and "Marmite jar" convey an implicit understanding of the colour of the object without the need to show an example image to the participant, while "coffee cup" is ambiguous and "red toy car" explicitly expresses a colour name. The set of test objects that we felt adequately met these conditions (for UK-resident British-English-speaking participants) amounted to eight objects, shown in fig. 5.7. While this test set may be small, it still serves as a useful demonstration.

We asked participants to describe the colour distribution of the test objects given only the object name – for example [90% yellow, 8% orange, 2% black] given the name "rubber duck". We then used these descriptions to create a normalised colour histogram vector from the eleven standard colour terms described above, which we used to query the entire 1000-object database (under only the 3075K illuminant). Using only these descriptors, we were able to achieve an average match percentile of 88% across eight participants. Given the object name "Juggling Ball", six of the eight participants gave

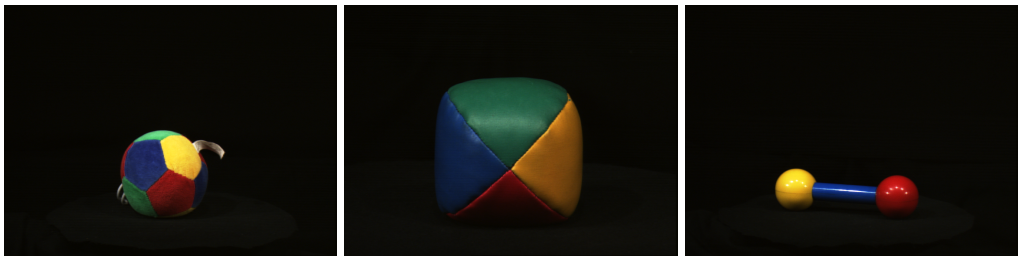


Figure 5.8: Top three search results using colour name descriptor [25% blue, 25% green, 25% red, 25% yellow], which was the most common human labelling for the prompt "Juggling Ball"

the same descriptor – [25% blue, 25% green, 25% red, 25% yellow]; fig. 5.8 contains the top three results after querying the ALOI dataset using that descriptor.

These results are again encouraging. Previous authors (Mehtre et al., 1995) have noted that histograms based on colour names can be powerful descriptors for object recognition, and we now show that a very simple eleven component colour name histogram is a powerful descriptor for both object recognition as well as human-guided image search. It is not clear that such a verbal query can be handled using traditional colour histograms.

5.6 Conclusion

We have shown that histograms derived from colour names allow for greater recall in the task of object recognition than chromaticity-based histograms, and comparable recall to full three-dimensional RGB histograms with the advantage of a more compact representation. While this could be used to improve the task of object recognition under varying illumination, this task has already been effectively solved by other means (Funt and Finlayson, 1995; Healey and Slater, 1994). However these findings open questions into the power of colour names as image descriptors. In the results shown here, we attain similar, if not better, performance to full three-dimensional RGB histograms using only an eleven component vector of colour names. This has significant implications for

applications such as image search; if each image in a corpus can be effectively represented by a much smaller descriptor (11 bins for names vs. $16\times16\times16 = 4096$ bins for the best-performing RGB histogram), the overall storage size of the entire corpus can be significantly reduced. Berens et al. (2000) showed the value of smaller histogram sizes by compressing traditional chromaticity histograms. We have also shown that this representation offers a powerful descriptor for human understanding of image content. The use of colour-name-based histograms as image descriptors offers clear advantages in terms of the required storage space, resilience to illumination changes, and relevance to human understanding.

Chapter 6

Constraint Propagation for Illumination Invariance

We have seen in chapter 5 that data sourced from the web can be successfully used to develop algorithms with practical applications in the field of image processing. In that chapter, we learned that colour names provide an incisive representation of the colour content of images, and that such a representation has useful practical applications in image indexing for querying by machines and humans alike. Colour names were also shown to be somewhat resilient to errors in illuminant estimation accuracy produced by a suite of commonly used illuminant estimation algorithms.

This chapter seeks to circumvent the illuminant estimation step utilised in chapter 5. Our objective is to develop an algorithm capable of computational colour naming, but which can accept an image taken under any illumination condition and produce the same output – i.e. identify a surface as appearing 'red' under a canonical white illuminant, regardless of whether the actual scene illuminant is daylight, fluorescent etc.

We seek to solve for colour constancy in a restricted sense – by determining illumination-invariant image descriptors. We seek only to recover correct colour names, not the full three-dimensional aspect of colour.

6.1 Introduction

The results in chapter 5 show that colour names, based upon human designations of RGB values to categorical colour name labels, can be very useful for object indexing. Moreover colour names have the desirable property of being more resilient to changes in illumination than some other quantisations of colour space. However, this resilience is not absolute. As illumination conditions deviate further from the canonical conditions (or equivalently, as illuminant estimation error increases), pixel values become miscategorised – for example as illumination changes from a reference white to a tungsten illuminant, surfaces initially categorised as 'white' may become 'red'.

Other authors (Finlayson and Hordley, 2001) have carried out related work in object indexing by dispensing with the traditional RGB, or two-dimensional chromaticity, histograms and instead constructing histograms from a one-dimensional illumination-invariant descriptor. The results generated by this approach were very favourable, as compared to post-illumination-estimation chromaticity histograms.

So then, is it possible to calculate an illumination invariant colour descriptor which is based on colour names? Such a descriptor would retain the benefits of the illumination invariant descriptor used by Finlayson and Hordley (2001), but would also have the benefits of being suitable for queries formulated by humans, as described in chapter 5.

As presented in section 6.3, we formalise this as a discrete relaxation problem, where the ratios of neighbouring colours propagate and constrain the colour names in image regions.

6.2 Background

Much research has been undertaken to better understand perceptually-relevant colour names (Benavente et al., 2008, 2012; Heer and Stone, 2012; Moroney, 2003), and there is much experimental psychology literature attributing colour names (Olkkonen et al.,

2009), as well as several other perceptual phenomena such as contrast (Foster, 2003; Land, 1977), mutual reflections (Kraft and Brainard, 1999) and colour memory (Hansen et al., 2006, 2007), as contributors to human colour constancy (Hurlbert, 1999). However exploiting this link between perceptually important phenomena and illuminant estimation remains a sparsely explored area. One notable contribution is that of Vazquez-Corral et al. (2012), which favours illuminant estimates which enable the colours in an image to be better 'anchored' to basic colour terms. That this approach generates favourable estimates shows that the perceptual importance of colour names can help with the illuminant estimation problem.

The method we seek, however, is not strictly an illuminant estimation approach: we are seeking a colour constancy algorithm – i.e. an algorithm which will allow us to make constant colour name designations to surfaces regardless of illumination conditions. We do not hope to recover the scene illuminant, nor the corresponding pixel values for surfaces under a canonical illuminant. In this regard, the work of Finlayson and Hordley (2001) is relevant. This work notes that much of the literature in modern illuminant estimation accepts the argument of Maloney and Wandell (Maloney and Wandell, 1986; Wandell, 1987), namely that the recovery of full three-dimensional RGB illuminant estimates may be over-ambitious, and authors have instead shifted the target toward the recovery of constant chromaticity recovery. Finlayson and Hordley (2001) take the reductionist approach of two dimensional recovery one step further and attempt to recover a one-dimensional descriptor. We hope to take this even further still and recover a constant singular categorical descriptor for each surface.

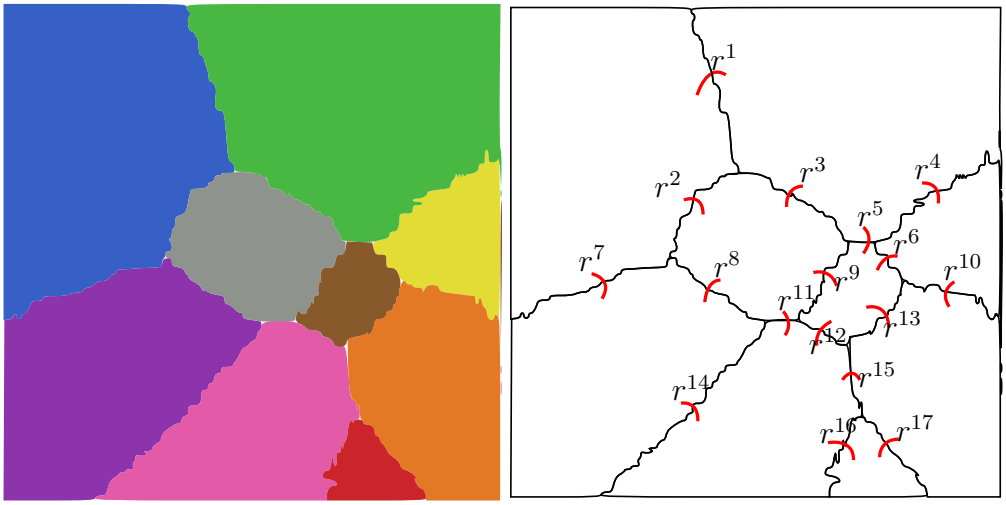
To meet this aim, we exploit the same property of the diagonal model as Finlayson and Hordley (2001), albeit in a different way. As discussed in section 2.1, if illumination change can be modelled as a diagonal transform, it then follows that the ratios of neighbouring pixel values will remain constant across all illumination changes (save for clipping and quantisation errors). Consider two pixel values $\underline{\rho}^x$ and $\underline{\rho}^y$, corresponding

to two surfaces x and y. The diagonal model suggests that any illumination change can be modelled as a separate multiplicative operation to all elements ρ_k^x (and correspondingly for ρ_k^y) for all k. This means that for any illuminant E, the k^{th} element of $\underline{\rho}^x$ will be scaled by the same factor as the k^{th} element of $\underline{\rho}^y$, and so their ratio will remain constant:

$$r = \underline{\rho}^{E,y} \big/ \underline{\rho}^{E,x} \; \forall \, (E) \,, \tag{6.1}$$

where the division operator, in this case, represents an element-wise operation. This *ratio constraint* is employed by several methods in the literature (Funt and Finlayson, 1995; Nayar and Bolle, 1993). However, care must be taken not to apply this observation too liberally: as noted by several authors (Barnard et al., 1997; D'Zmura, 1992; Finlayson et al., 1995; Tsukada and Ohta, 1990) illumination is not usually uniform across an image, there usually exists a gradient or *illumination field*. For example, consider a single flat matte surface imaged frontally: if the illumination source in this scene is placed to the left of the surface, then there will be an illumination gradient from left to right across the image. Further still, many scenes contain more than one illumination source, and the illumination conditions across the scene vary as a mixture of the competing illumination sources. Barnard et al. (1997); D'Zmura (1992); Finlayson et al. (1995); Tsukada and Ohta (1990), as well as identifying this issue, actually exploit it to add a further constraint to their illuminant estimation schemes. For our purposes, we shall not be exploiting this phenomena, but we will take measures to mitigate its effects by only considering the ratios between neighbouring pixels – as the illumination field changes gradually across a scene, the differences in illumination conditions between one pixel and its neighbours are largely negligible.

At this point it also important to remember, as noted in section 2.1, that the ratio constraint derived from the diagonal model does not always hold in RGB or camera-native colour spaces. In light of this, all the processing in this chapter is performed



(a) Segmented input image *(b) Network of ratio constraints*

Figure 6.1: Network of local ratio constraints across entire image

in *Sharp space* as defined by Süsstrunk (2005). This colour space is based upon a chromatic adaptation transform designed to optimise the ratio constraint. Later in this chapter we will, for the sake of simplicity, continue to refer to "RGB triplets" and the like, but in these cases we are in fact referring to points in Sharp space.

6.3 Method

Employing the ratio constraint discussed above, distributed across the whole image, gives us a network of local ratio constraints to be satisfied as depicted in fig. 6.1. How then do we convert these ratio constraints into invariant colour names?

To begin, let us first reintroduce some of the discrete relaxation nomenclature seen in section 2.9. Recall that we have

$$U = \{u_1, \ldots, u_n\}, \tag{6.2}$$

which is a collection of n objects, to each of which we seek to assign one of m labels

$$\Lambda = \{\lambda_1, \ldots, \lambda_m\}. \tag{6.3}$$

In this case, U is the collection of n pixels in our image and Λ is the collection of eleven colour names (we have, throughout this and the previous chapter been working with the eleven basic colour terms defined by Berlin and Kay (1969), but the same principles hold if we were to extend to a larger quantity of labels). To solve this problem with discrete relaxation we need to construct a matrix $\mathbf{L}$, an $n \times m$ binary matrix which encodes our candidate labels λ_m for each u_n, as well as our set of compatibility matrices R.

To implement our method we need a mapping of RGB values ($\underline{\rho}$) to colour names (Λ) – we re-used the Gaussian mixture colour naming model introduced in chapter 5 (hereafter referred to as the *GMM*), which essentially provides a large lookup table mapping RGB values to categorical colour names. To construct $\mathbf{L}$ and R, consider again our two surfaces x and y, which under unknown illumination give rise to the pixel values $\underline{\rho}^x$ and $\underline{\rho}^y$, related by constant ratio r. The GMM allows us to fix $\hat{\underline{\rho}}^x$ to some arbitrary value and ascertain the name given to that value. Then we can set $\hat{\underline{\rho}}^y = r\hat{\underline{\rho}}^x$ and use the GMM once again to determine the name given to the new value of $\hat{\underline{\rho}}^y$. This allows us, at a high level, to ask 'if x is "red", then what colour must y be?'. To rephrase this in the language of discrete relaxation: 'which pairs of colour names are *compatible* with the ratio r?'. Of course there are many discrete pixel values which are designated the name "red", and so there may be several compatible colour names for y under the assumption that x is "red". If we repeat this postulation for all possible values of $\hat{\underline{\rho}}^x$, we can generate a complete compatibility matrix $R_{x,y}$ for x and y, where the compatibility matrix is an 11×11 binary matrix where each column represents a putative colour name for x and, correspondingly, each row encodes the possibilities for y. We could construct this matrix by naïvely iterating over all possible values in the RGB cube for $\hat{\underline{\rho}}^x$ but, as will be discussed below, a more efficient implementation is possible.

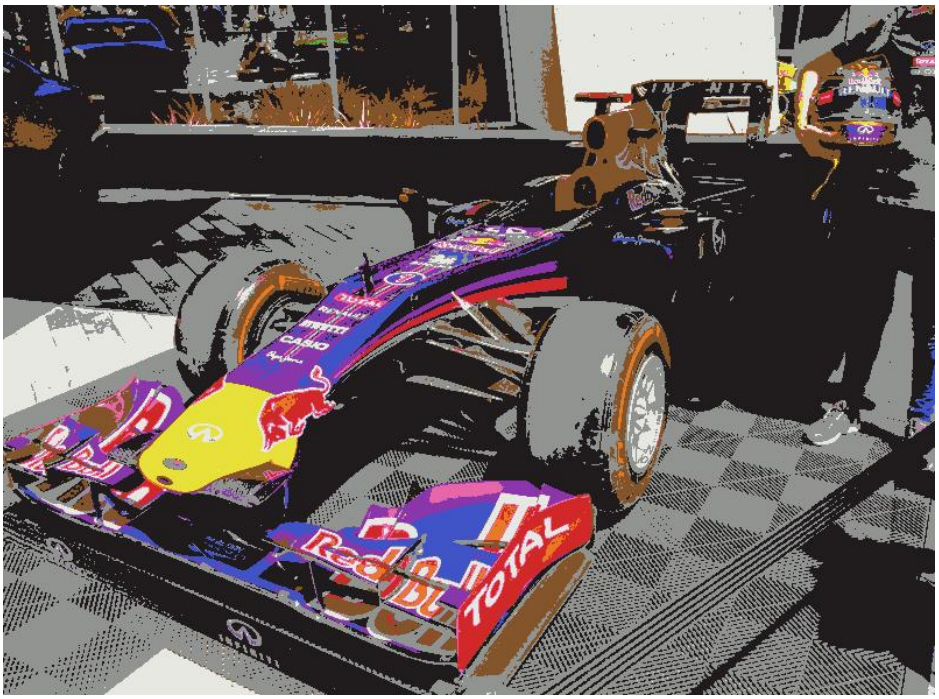
With a collection of compatibility matrices R constructed in this way, we have the basis for a problem which can be solved by discrete relaxation as described in section 2.9 (for now we initialise $\mathbf{L}$ to be all ones, but we will revisit this later). The sections below discuss further refinements and optimisations but, at its core, our method can be described by this simple application of discrete relaxation to a network of local ratio constraints.

6.3.1 Segmentation

Clearly, if we were to repeat the process of generating compatibility matrices for every pair of pixels in the image, the computational cost would be high. Fortunately, however, this is not necessary. Colour names are seldom finely scattered across an image – as can be seen in fig. 6.2, colour names appear in patches, corresponding with the areas covered by the surfaces in the image. Because of this, we can segment our input images and deal with the ratios between average pixel values of patches in the image, instead of between individual pixels. Therefore, the collection of objects U, as defined in the previous section, is now the collection of image segments, or patches, to which we wish to assign colour names.

As the patches in the image can be fairly large, we have to be wary of the nonuniformity of the illumination conditions, as described in section 6.2. If we have a surface occupying a large area of the image which is all "blue", it is entirely possible (indeed likely) that the pixel values at one side of this patch are a different "blue" to those at the opposing side, due to the varying illumination across the surface. To tackle this, when computing our ratios we should consider only those pixels close to the border between the two patches we are considering (see fig. 6.3). Therefore when considering two neighbouring patches x and y, $\underline{\rho}^x$ is now defined to be the average of the pixel values in x which are close to the border with y, and vice-versa with $\underline{\rho}^y$.

(a) Original image



(b) Image labelled by GMM

Figure 6.2: Distribution of colour names in image

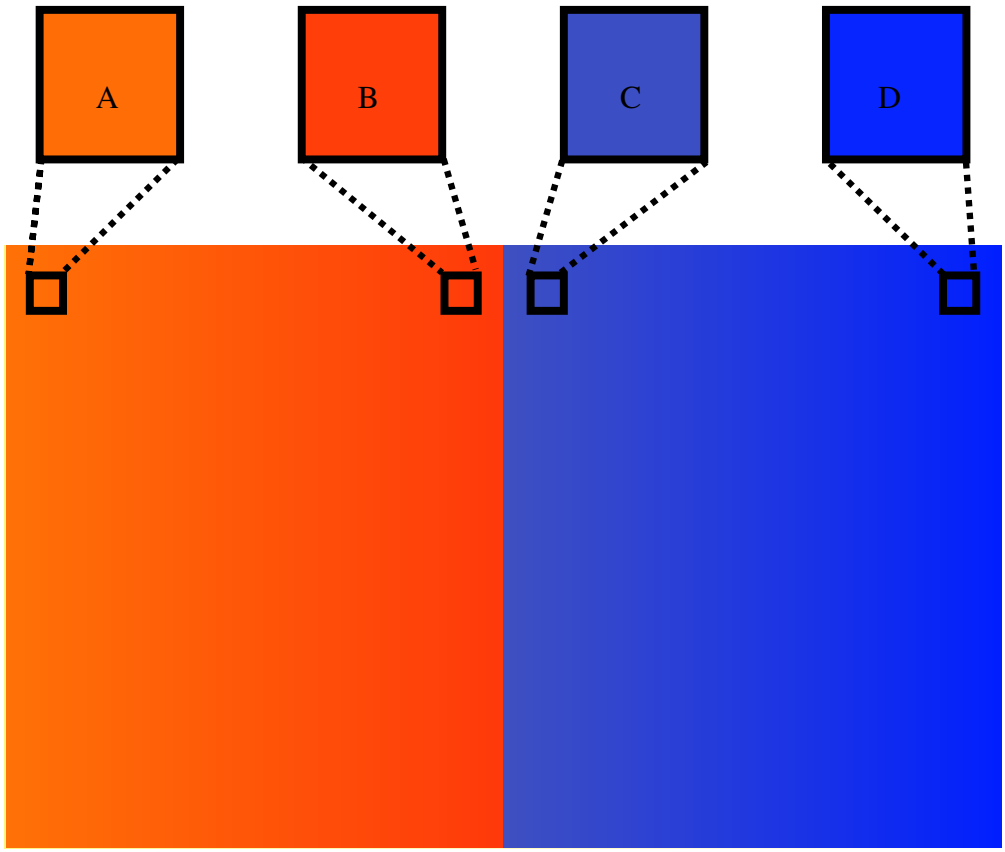


Figure 6.3: Pixels close to the border between two colour patches are less susceptible to spatially varying illumination – pixels B and C have similar lighting conditions, while A and D do not

There are many approaches to image segmentation (Achanta et al., 2012; Cheng et al., 2001; Fu and Mui, 1981). We performed experiments with mean-shift segmentation and several "superpixel" approaches, but found best results with a simple alternative: we cluster the pixels in the image by the colour names designated to them by the GMM. While this may seem unintuitive for reasons described in earlier sections (the GMM applied to an image under unknown illumination conditions may label a surface "pink" which under white light is actually "white"), the actual labels designated by the GMM are, at this stage, unimportant - we are only using this information to identify differently coloured patches in the scene.

6.3.2 Additional Constraints

A problem with the method as introduced so far is that the labels designated to a scene by the GMM under any illuminant E, are just as valid as those under any other illuminant. In other words, for a pair of pixels which are labelled as "white" and "yellow" under a canonical white illuminant, and "pink" and "orange" under some other illuminant, the method cannot distinguish whether the first of those pixels should be definitively labelled as "white" or "pink", nor the second as "yellow" or "orange". Indeed every labelling generated by the GMM under any possible illuminant is compatible with the ratio constraint, and so on its own the ratio constraint offers insufficient utility to meet our objective of calculating illumination-invariant descriptors.

To address this problem, we introduce two further constraints to the generation of the compatibility matrices, as well as a new unary constraint.

Gamut Constraints

Our first additional constraint on the generation of compatibility matrices is built upon the observation that, under a finite set of known plausible illuminants, the plausible pixel values generated by a single surface is also finite. As such, we shall refer to this constraint as the *plausible illumination* constraint. Consider again the diagonal model, where any illuminant can be characterised by a diagonal matrix E, and where any pixel value $\underline{\rho}^x$ observed under that illuminant can be mapped back to its corresponding value $\hat{\underline{\rho}}^x$ under a canonical white illuminant by E^{-1}. If we have a representative set of plausible illuminants, we can generate a corresponding set of all the mappings back onto the canonical illuminant, which we can characterise by its convex hull $\mathcal{E}^{-1}$. If we now have a pixel value $\underline{\rho}^x$ as observed under an unknown illuminant, we can multiply it by each point on $\mathcal{E}^{-1}$ to generate a new convex set $\mathcal{X}$ which characterises all the possible surfaces (as observed under the canonical illuminant) which could give rise to this observed pixel value under unknown illumination conditions.

As suggested toward the beginning of this section, we could generate the compatibility matrix for x and y by naïvely iterating over the entire RGB cube as candidate values for $\underline{\hat{\rho}}^x$, and multiplying each by r to give candidate values for $\underline{\hat{\rho}}^y$. However, to satisfy our objective of designating illumination-invariant descriptors (λ_x, λ_y) to x and y, we seek only the candidate values $\underline{\hat{\rho}}^x$ and $\underline{\hat{\rho}}^y$ which can be plausibly manifested by the surfaces x and y under the canonical illuminant. We now know that, under the canonical illuminant, only a subset of the RGB cube, $\mathcal{X}$, represents surfaces which can give rise to the observed value $\underline{\rho}^x$. Also, since this set is characterised by its convex hull, we need only multiply the points on the hull defining $\mathcal{X}$ by r to give a corresponding convex hull $\mathcal{Y}'$ for all the corresponding values (as defined by the ratio constraint) for $\underline{\hat{\rho}}^y$. Furthermore, we know from the same observation that $\underline{\rho}^y$ can only be manifested from some set of surfaces $\mathcal{Y}$, and since $\mathcal{Y}$ is related to $\underline{\rho}^y$ by the same multiplicative operation that relates $\mathcal{X}$ to $\underline{\rho}^x$, $\mathcal{Y}$ and $\mathcal{Y}'$ are equal.

The introduction of the plausible illumination constraint means that $\mathcal{X}$ characterises all the candidate pixel values for $\underline{\hat{\rho}}^x$ which satisfy the ratio constraint *under the canonical illuminant*, and similarly with $\mathcal{Y}$ for $\underline{\hat{\rho}}^y$.

We can now introduce a second, albeit similarly motivated, gamut constraint which we shall call the *plausible surface* constraint. The constraint described above is built on the observation that, under a characteristic collection of known plausible illuminants, an observed pixel value under unknown illumination (within the plausible set) can only be generated by a finite set of surfaces. This idea is parallel to the gamut constraint traditionally applied in illuminant estimation techniques based on gamut mapping (as described in section 2.2.5) – that, with a known set of plausible surface reflectance functions, the set of pixel values that can be observed under a single illuminant is a subset of the entire RGB cube. We too can exploit this observation by generating a gamut $\mathcal{S}$ of all pixel values that can be observed under the canonical illuminant. Then, when building compatibility matrices, we can intersect $\mathcal{X}$ and $\mathcal{Y}$ with $\mathcal{S}$ to further constrain the search

space for candidate values for $\underline{\hat{\rho}}^x$ and $\underline{\hat{\rho}}^y$.

We first intersect $\mathcal{X}$ with $\mathcal{S}$ to give $\mathcal{X}'$, the set of candidate values for $\underline{\hat{\rho}}^x$ which satisfy the ratio constraint, the plausible illumination constraint, and the plausible surface constraint (only for $\underline{\rho}^x$ at this stage). We now multiply $\mathcal{X}'$ by r to give $\mathcal{Y}'$. Note that, contrary to what was noted earlier in this section, $\mathcal{Y}'$ is now not equal to $\mathcal{Y}$ – due to the intersection of $\mathcal{X}$ with $\mathcal{S}$, $\mathcal{Y}'$ is now a subset of $\mathcal{Y}$. Further intersecting $\mathcal{Y}'$ with $\mathcal{S}$ gives $\mathcal{Y}''$, the set of candidate values for $\underline{\hat{\rho}}^y$ which satisfy the ratio constraint, the plausible illumination constraint, and the plausible surface constraint for *both* $\underline{\rho}^x$ and $\underline{\rho}^y$. Finally multiplying $\mathcal{Y}''$ by r^{-1} gives $\mathcal{X}''$, the final set of plausible values for $\underline{\hat{\rho}}^x$ which satisfy all constraints. By passing corresponding pairs of candidate ($\underline{\hat{\rho}}^x$, $\underline{\hat{\rho}}^y$) values from $\mathcal{X}''$ and $\mathcal{Y}''$ to the GMM, we can build our final compatibility matrix for x and y.

Unary Constraint

Recall from section 2.9 that the discrete relaxation algorithm allows for unary constraints to be imposed. To exemplify this, we can pose the objective of our illumination invariant colour naming algorithm more concretely. Suppose we have an image which, after segmentation (as described above), depicts twenty coloured patches. We are seeking to label each of those twenty patches with one of eleven colour names. To meet this aim, we construct a 20×11 binary matrix $\mathbf{L}$, where each row corresponds with a colour patch in the image, and each column with a candidate colour name. A 'one' in a location in $\mathbf{L}$ indicates that the corresponding colour patch can be *consistently* labelled with the colour name associated with that column. With the algorithm we have introduced so far, $\mathbf{L}$ is initialised to be all ones and, by means of a discrete relaxation algorithm, is 'pruned' in accordance with the compatibility matrices.

However, $\mathbf{L}$ does not have to be initialised to all ones – we can encode unary constraints by pruning some of the matrix before running the relaxation algorithm. We have introduced above the notion of a priori knowledge of plausible surfaces and plaus-

ible illuminants for scenes. With both of these, it is possible to enumerate the plausible pixel values that can be generated by any surface under any illuminant by taking the outer product of the two sets.

Suppose that $\underline{\rho}_{i,j}$ is the pixel value generated by surface i under illuminant j and that

$$\underline{\rho}_{i,j} \leftarrow \lambda_k \tag{6.4}$$

means that under illuminant j, surface i is designated the colour name label at index k by the GMM. For each plausible surface i, let us construct a set of the possible colour names which that surface can be given under every illuminant j (*excluding* the canonical illuminant), which we can represent as a vector:

$$p_k^{i,j} = \begin{cases} 1 & \text{if } \underline{\rho}_{i,j} \leftarrow \lambda_k \\ 0 & \text{otherwise} \end{cases} . \tag{6.5}$$

Now consider the special case of surface i under the canonical illuminant, which we shall call c. We determine the index of the colour name label given to a surface i under c to be λ_c, i.e.

$$\underline{\rho}_{i,c} \leftarrow \lambda_c. \tag{6.6}$$

We can now aggregate all these sets across all surfaces i and all illuminants j (excluding c) to generate an 11×11 matrix P:

$$\mathbf{P}_{c,k} = \mathbf{P}_{c,k} \vee p_k^{i,j}, \tag{6.7}$$

where $\vee$ indicates a logical "OR" operation. $\mathbf{P}$ now provides a lookup table of plausible labels under unknown illuminants, given a label under the canonical illuminant.

Accordingly, $\mathbf{P}^{\intercal}$ provides a lookup of plausible labels under the canonical illuminant, given a label from a surface under unknown illumination. For example, we may find that a surface which is labelled "blue" under unknown illumination could have a corresponding label of "purple" or "green" under the canonical illuminant, but can never come from a surface which is labelled "red" under the canonical illuminant.

With this information, we can construct $\mathbf{L}$ in such a way that we encode the unary constraint that if a surface is designated some colour name label λ under unknown illumination, it can only be plausibly labelled with a subset of Λ under the canonical illuminant. Formally

$$\underline{l}_i = \underline{p}_\lambda. \tag{6.8}$$

This initialisation of $\mathbf{L}$ constrains the result of our algorithm in a similar fashion to that of the plausible illumination gamut constraint, but does so in a more discrete way. Additionally, the pre-pruning of L means that fewer possible labellings are passed to the discrete relaxation algorithm and so there is less processing to complete. The addition of this unary constraint, therefore, reduces the overall computational cost of the algorithm – $\mathbf{P}$ does not need to be recalculated for each image.

6.3.3 Summary of Method

The steps comprising the method have been discussed above in an order which has allowed us to introduce foundational ideas first, before considering additional constraints which build upon them. For clarity, we summarise the method here in chronological order:

Pre-processing Steps

1. Use a collection of plausible illuminants to generate a gamut $\mathcal{E}^{-1}$, with which to enforce the plausible illumination constraint.

2. Use a collection of plausible surface reflectance functions to generate a gamut $\mathcal{S}$, with which to enforce the plausible surface constraint.

3. Use the collections of plausible surfaces and illuminants to generate $\mathbf{P}$ – with which to enforce the unary constraint.

Processing an Individual Image

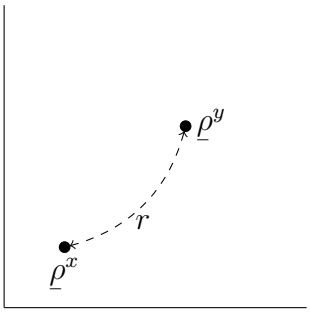
1. Begin with an image under unknown illumination conditions (within the plausible set defined above).

2. Segment the image to identify distinct image patches.

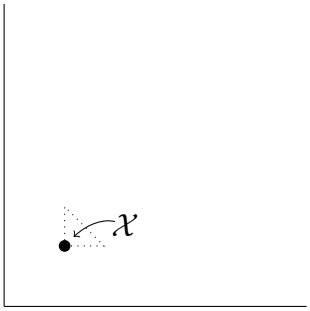
3. Use $\mathbf{P}$ and the colour names designated by the GMM to neighbouring patches to generate $\mathbf{L}$.

4. For each pair of x, y of neighbouring patches, generate $R_{x,y}$ as in fig. 6.4.

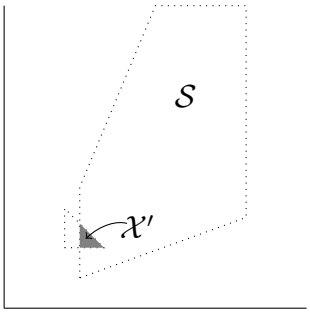
5. Use $\mathbf{L}$ and R to label the image patches with illumination-invariant colour name labels using discrete relaxation.



(1) Calculate r from the observed values of $\underline{\rho}^x$ and $\underline{\rho}^y$

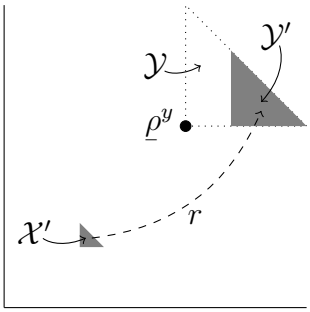


(2) Satisfy the plausible illumination constraint by multiplying $\underline{\rho}^x$ by $\mathcal{E}^{-1}$ to give $\mathcal{X}$, the plausible candidate values for $\hat{\underline{\rho}}^x$ under the canonical illuminant

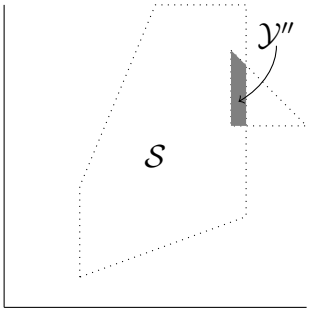


(3) Intersect $\mathcal{X}$ with $\mathcal{S}$ to give $\mathcal{X}'$, which satisfies the plausible surface constraint for $\underline{\rho}^x$

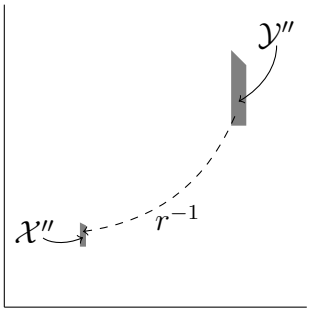
Figure 6.4: Method summary: construction of compatibility matrices. Figures are shown in two dimensions for visual clarity – this is a three-dimensional process in reality



(4) Multiply each point on $\mathcal{X}'$ by r to give $\mathcal{Y}'$ which, since it is a subset of $\mathcal{Y}$, satisfies the plausible illumination constraint for $\underline{\rho}^y$



(5) Intersect $\mathcal{Y}'$ with $\mathcal{S}$ to give $\mathcal{Y}''$, which satisfies all constraints for $\underline{\rho}^y$



(6) Multiply each point on $\mathcal{Y}''$ by r^{-1} to give $\mathcal{X}''$, which satisfies all constraints for $\underline{\rho}^x$. Pass pairs of candidate values from $\mathcal{X}''$ and $\mathcal{Y}''$ to the GMM to ascertain corresponding pairs of colour names (λ_x, λ_y). For each pair i, set $R_{x,y}(\lambda_x^i, \lambda_y^i) = 1$

Figure 6.4: Method summary: construction of compatibility matrices *(cont.)*

6.4 Experiments

To test the algorithm, we conducted two experiments. Firstly we tested the fundamentals of the algorithm using synthetic data, where we can force the assumptions of the diagonal model to hold. Then we moved to real-world data and, ultimately, tested the core objective of the algorithm – can it deliver illumination-invariant descriptors of image content which are sufficient to enable both machine object indexing, and human querying by colour name?

6.4.1 Synthetic Data

For our tests with synthetic data we used the collections of measured real-world illuminant spectral power distributions and surface reflectance spectra published in Barnard et al. (2002). For the canonical illuminant, we used a synthesised pure white illuminant, i.e. the E term from eq. (2.6) is the identity matrix. With these collections we generated the gamuts necessary for the constraints described in section 6.3.2. We used the CIE 1931 colour matching functions (Smith and Guild, 2002) to calculate sRGB co-ordinates, which were then converted into Sharp space (and, accordingly, into CIE L*a*b* as required by the GMM).

After completing the preprocessing necessary for enforcing constraints, one hundred experimental runs were completed, as per the following procedure (summarised in fig. 6.5):

1. Randomly select twenty-five surfaces from the surface collection.
2. Arrange the chosen surfaces into a 5×5 grid of image patches.
3. Render the grid under the canonical (white) illuminant and label the resulting image with the GMM.

4. Randomly select an illuminant from the illuminant collection and re-render the grid under that illuminant.

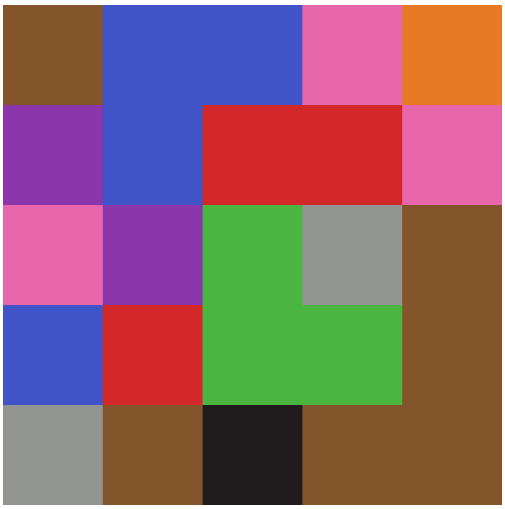
5. Label the re-rendered image with our new method.

6. Compare the labels generated in steps 3 and 5, and note the number of patches labelled identically.

Unfortunately, in carrying out this experiment we found that the constraints introduced above are, in many cases, insufficient to enable the algorithm to converge to a unique solution for every colour patch. What this means is that for many colour patches (approximately 25% in this experiment) the algorithm is unable to prune the possible colour names to a single, unique, label – or, more formally, several labellings are simultaneously consistent for these colour patches. This can be seen in fig. 6.5d – the patches containing red crosses represent those for which multiple labels are consistent. Fortunately, for the majority of patches ($>$80%), the algorithm is able to reduce the candidate set of eleven labels to one, two, or three possible consistent labels. Furthermore, these labels will usually be those that occupy neighbouring areas of the CIE L*a*b* colour space – i.e. a patch which should be labelled "blue" might be assigned ["blue", "purple"], but a patch that should be "red" will, generally, not be assigned ["red", "green'].

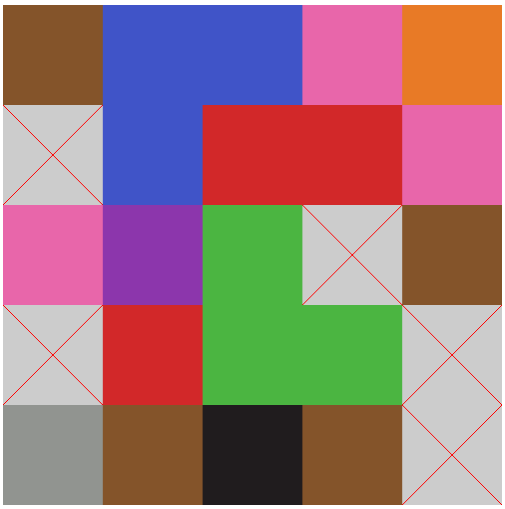
In light of this finding, before investigating correctness it would be useful to report the number of solutions that the algorithm converges to – fig. 6.6 provides these data for the synthetic data experiment. A subset of the colour patches do indeed converge to a unique answer – of the twenty-five patches in each experiment, the mean number to converge to a unique answer was 6.5. Also noteworthy is that we see a non-zero number of patches which are reduced to zero candidate labellings. This means that, for these patches, there are no possible labellings which are consistent according to the constraints introduced above. This is perhaps surprising, as we are dealing with syn-

(a) Surface selection rendered under canonical white illuminant



(b) Ground-truth colour name labelling as designated by GMM

(c) Surface selection re-rendered under test illuminant



(d) Colour name labelling as designated by our algorithm (crosses indicate patches that did not converge to a single solution)

Figure 6.5: Synthetic images (a, c) rendered using surface and illuminant spectra measured by Barnard et al. (2002), and colour-name-labelled counterparts (b, d)

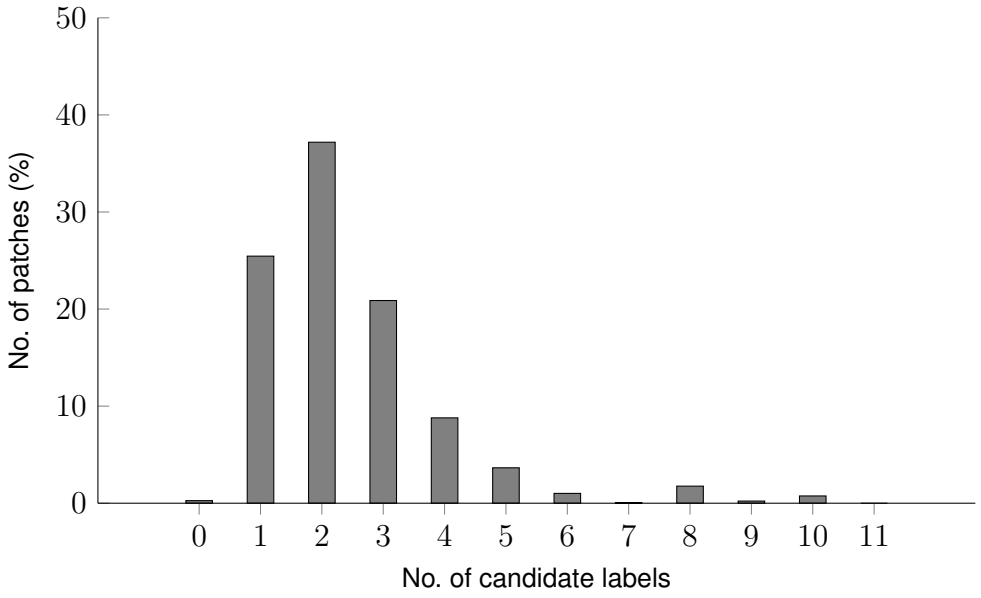


Figure 6.6: Number of simultaneously consistent labellings per colour patch, for synthetic data

thetic data in which the assumptions of diagonal model have been artificially enforced, but a possible explanation is discussed in section 6.5.

Of those patches which do converge to a unique answer (mean 6.5 out of twenty-five patches), a mean of 6.1 are, encouragingly, labelled correctly. Furthermore, across all patches the correct label is erroneously discarded only 3% of the time.

In summary of these results from synthetic data, we observe that the algorithm is able to prune the candidate set of eleven labels to a single, unique, label on approximately 25% of occasions, and to three or fewer consistent labels on approximately 84% of occasions. If, under this caveat, we define correctness to be whether or not the correct label remains in the reduced candidate set after the algorithm has run, we see that 97% of patches are correctly assigned. Of the patches for which the algorithm does converge to a unique solution, 93% are correctly labelled.

6.4.2 Consistent Labellings for Object Recognition

If we use the same name-based histogram intersection scheme from chapter 5, it may seem from the results reported above that our algorithm will be insufficient to meet our objective of calculating colour-name-based illumination-invariant descriptors for object indexing. If the algorithm is unable to definitively assign a colour patch with a single unique label, and instead we are left with, for example, [“blue”, “purple”] as the consistent labels, we would be unsure whether to count the corresponding pixels in the “blue” or the “purple” histogram bins. However, under the assumptions introduced in section 6.3, this non-unique labelling should be similarly manifested under any illumination conditions – i.e. under the assumptions of the diagonal model and our gamut constraints, a patch which our algorithm labels [“blue”, “purple”] under some illuminant E, should also be labelled [“blue”, “purple”] under all other illuminants. If then, we can encode this non-unique but consistent labelling into our histograms used for object indexing, we should still be able to achieve illumination invariance.

An approach to this could be to add further histogram bins for each possible intersection of labels – so there would be eleven bins for the basic colour terms, plus an additional bin for [“blue”, “purple”], one for [“red”, “yellow”] and so on. However, after all intersections had been enumerated, especially if we went so far as to encode all intersections of up to eleven names, we would have many bins in each histogram. One motivating advantage of the eleven-component vector representing a histogram of colour names is its compactness, and so this approach would be undesirable. A simpler approach is to retain the use of an eleven-component vector, and simply count the [“blue”, “purple”] labelling in both the “blue” and “purple” bins. Below we show that, in spite of its apparent naïveté, this approach is sufficient to deliver positive results.

To test this, we repeated the approach taken in chapter 5, but using colour name histograms constructed using our new algorithm. Once again we used the SFU object recognition dataset (Funt et al., 1998) to test object recognition performance – we used

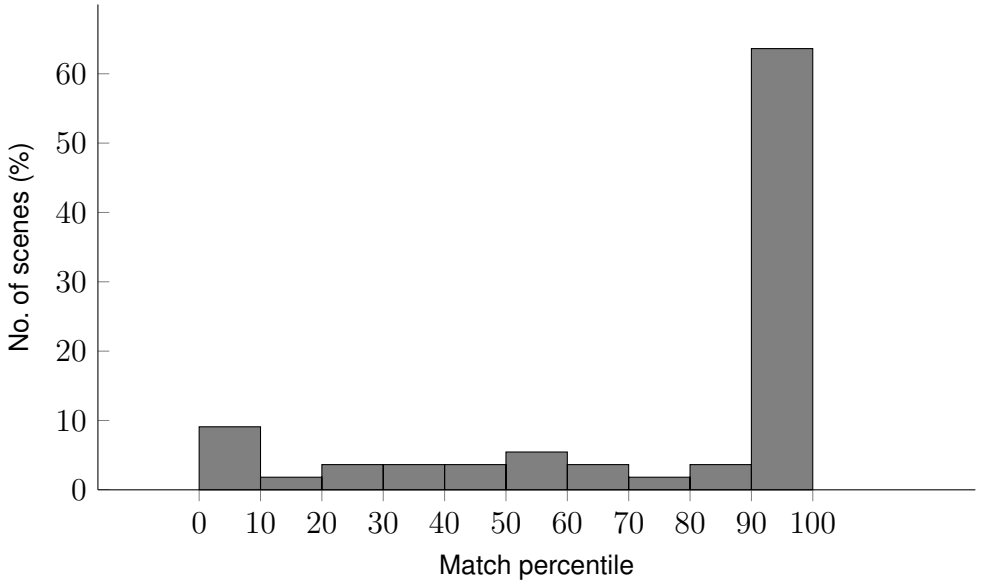


Figure 6.7: Distribution of match percentiles for new method, using the SFU object recognition dataset (Funt et al., 1998)

the "syl-cwf" illuminant as the canonical illuminant. Across all objects and illuminants (leaving aside the canonical) we were able to achieve a mean match percentile of 76%. This is perhaps somewhat disappointing in comparison to the 96% achieved by Finlayson and Hordley (2001), however we achieve a median match percentile of 100%. Indeed for 64% of scenes (across all objects and illuminations), a match percentile of 100% is achieved. This reveals that the mean is heavily influenced by a minority of poorly-performing scenes, as shown by the distribution in fig. 6.7.

Upon closer inspection it is revealed that our method delivers consistently poor results for the "javex" object in particular. The mean match percentile for this scene is 52% (a match percentile of 50% would be expected by random chance). A visual inspection (see fig. 6.8), suggests that the reason for the poor performance observed for this object in particular may simply be because it is colour deficient – i.e. there is not a high enough degree of colour diversity in the scene. As an optimisation to our algorithm we avoid processing small colour patches by means of a spatial threshold. For the "javex" scene the red area was insufficiently large and was ignored by our algorithm, meaning that

Figure 6.8: "Javex" object from SFU object recognition dataset (Funt et al., 1998)

the only remaining colour patches were either blue or white (the black backgrounds of the images were masked out prior to any further processing). These two colour classes offered little constraint to the algorithm and so it was unable to successfully prune many candidate labellings.

6.4.3 Consistent Labellings for Query by Colour Name

The results presented so far are promising, but it may seem that the method described above of counting labels into multiple histogram bins may negatively affect the ability to query the colour name histograms with human-generated queries. To test this we repeated the experiment using the ALOI dataset (Geusebroek et al., 2005) described in

section 5.5, but this time using histograms constructed as described above, and using our new labelling algorithm.

At first we repeated the experiment with only the 3075K illuminant (as per section 5.5) and, encouragingly, achieved a mean match percentile of 81% across all objects. This is slightly lower than the 88% reported in section 5.5, which may in part be attributable to the construction of the histograms, but is generally a positive result. However, we then repeated the experiment with every other illuminant in the dataset (the objects are imaged under a range of twelve illumination conditions from 2175K to 3075K – the camera was white balanced to 3075K which results in the apparent illumination conditions varying from reddish to white) and, remarkably, the mean match percentile across all illuminants was 82%.

6.5 Discussion

The results presented above are indeed encouraging. Although the illumination-invariant object recognition match percentiles are not as high as those achieved by Finlayson and Hordley (2001) (75% as opposed to their 96%), that we can perform illumination-invariant object recognition whilst also meeting our second objective of preserving perceptually relevant descriptors is noteworthy. However, there are two particular issues outstanding which merit some further examination. The first is that, for some patches, the algorithm discards the correct colour name label and/or reduces the plausible set to zero candidates (as seen in section 6.4.1), i.e. there is no consistent labelling. The second, conversely, is that the algorithm is sometimes not able to reduce the candidate set of plausible labels to a useful extent (as seen with the "javex" scene in section 6.4.2) – there are multiple consistent labellings. We address these two issues separately below.

6.5.1 Inconsistent Labellings

Firstly, let us address the issue of over-pruning the candidate set of plausible labels such that no labelling is deemed consistent. By design, discrete relaxation algorithms will not discard a consistent labelling. This issue then is not that the algorithm discards a valid labelling, but that there is no valid labelling of the data. This is due to a failure of the assumptions we introduced in sections 6.2 and 6.3. In particular, this is likely to be a failing of the ratio constraint. In utilising this constraint, we tacitly accept that the diagonal model of image formation is correct. As mentioned in section 2.1, while this model generally holds (and in particular it should suffice while using the Sharp space introduced in section 6.2), it can be imprecise. In many computational approaches these failings are not noticeable; for example the approaches of many classical illuminant estimation schemes discussed in section 2.2 means that these slight errors are usually hidden by averaging pixel values. However, the construction of the algorithm described above means that a slight error can cascade and be manifested as a failure of the algorithm.

Of particular note is that this situation arises in section 6.4.1, where the assumptions of the diagonal model have been enforced (i.e. the synthetic rendering of a scene under some illuminant E is done by representing E as a diagonal matrix). The reason for this is that, in order to increase the realism of the image synthesis, the rendered images are quantised to eight bit integer values. This step is included in the model as an acknowledgment that real-world image data is usually sourced from JPEG images, or perhaps TIFF or RAW data which can handle greater bit depth, as opposed to arbitrary-precision data. In so doing we can explicitly contravene the ratio constraint. For example, consider two pixel values $\underline{\rho}^x$ and $\underline{\rho}^y$, for which we shall only examine the scalar values for a single colour channel k: ρ_k^x and ρ_k^y. Under some illuminant E_a:

$$\rho_k^{E_a,x} = 7,$$
$$\rho_k^{E_a,y} = 20.$$

Under some other illuminant E_b, for which the irradiance in the part of the spectrum covered by sensor channel k is half that of E_a, we would expect

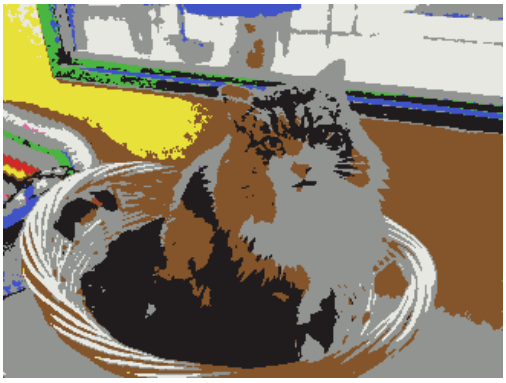
$$\rho_k^{E_b,x} = 3.5,$$
$$\rho_k^{E_b,y} = 10.$$

However, due to quantisation $\rho_k^{E_b,x} = 4$, and thus the ratio constraint is violated:

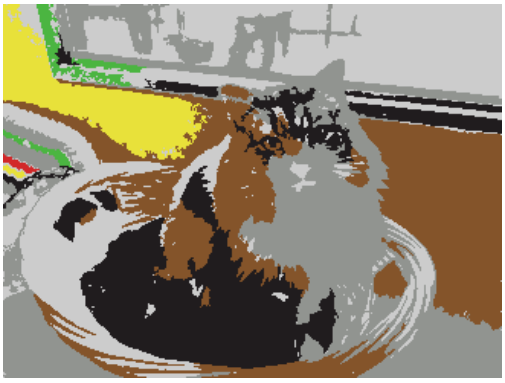
$$\frac{\rho_k^{E_a,x}}{\rho_k^{E_a,y}} \neq \frac{\rho_k^{E_b,x}}{\rho_k^{E_b,y}}. \tag{6.9}$$

Generally, this inequality does not incur too much of a penalty. The construction of the compatiblity matrices is not done by enumerating all possibilities $\underline{\rho}^x$ and $\underline{\rho}^y$ and then checking if their ratios are equal to r, but by deriving the possible values for $\underline{\rho}^y$ by multiplying the values for $\underline{\rho}^x$ by r. By doing this we explicitly encode the observed error into our compatibility matrices. Usually this is of no consequence, the errors are so small and the plausible pixel values so numerous that there is no ill effect. However, on the infrequent occasion that the error is large enough, or that the plausible pixel values for $\underline{\rho}^y$ are very close to a border between colour names, this can mean that a colour name pair is deemed inconsistent when it should in fact be consistent. In this scenario, the error can cascade leading to ultimate failure of the algorithm.

(a) Original image



(b) Image labelled by GMM



(c) Image labelled by discrete relaxation method

Figure 6.9: Real image labelled by GMM, and by new method

6.5.2 Multiple Consistent Labellings

The second issue, which we see in section 6.4.2, is that sometimes the algorithm has insufficient constraint to successfully prune the candidate labellings. This is particularly apparent for the "javex" scene discussed above, but is representative of a broader problem with the algorithm: even under optimum conditions, the algorithm does not converge to a unique answer for the majority of image patches. As seen by the favourable object indexing and human-based-querying results above, this is not necessarily a large problem for these particular applications. However, if we wanted to use this

algorithm to assign definitive illumination-invariant colour name labels – perhaps for the purpose of visual display as in fig. 6.2b – we would not usually be able to do so. Figure 6.9 shows a real image for which the method has produced a result suitable for this purpose, but even in this example there are still some patches (the cat's nose, some parts of the basket, and the clipped pixels in the window) where a unique solution has not been possible. This problem could perhaps be addressed by the addition of further constraints to the method, however no further constraints are immediately apparent that would not have a drastic effect on computational cost.

6.6 Conclusion

Chapter 5 demonstrated that colour names can be useful for object indexing, and also as a key by which to index images for querying by colour name. However the results in chapter 5 were dependent upon images first being colour corrected by means of an illuminant estimation algorithm. Also, as illuminant estimation accuracy degraded, so then did the utility of colour names for this purpose. In this chapter we sought to overcome this problem, and in so doing obviate the need for the illuminant estimation step altogether, by developing an algorithm capable of assigning colour name labels to images regardless of the prevailing illumination conditions in the scene. Crucially, these labels should be perceptually relevant, and indeed are derived from human designations of colour names to surface colours under a canonical white illuminant. In short, given a surface which appears white under a pure white illuminant, but imaged under a tungsten illuminant which makes it appear reddish, the algorithm should return "white". Further, it should be capable of making this designation reliably enough that we can still successfully perform machine object indexing and querying by human-generated labels.

By utilising some commonly used observations, namely that image ratios are in-

variant to illumination changes, and the gamut constraints typically applied in gamut-mapping-based illuminant estimation algorithms, and combining them into an algorithm based on classical boolean discrete relaxation, we were able to meet these objectives. Our algorithm is capable of labelling objects in scenes with unknown illumination with perceptually-relevant colour names such that they can be successfully retrieved from a database of object images by using the object indexing approach of Swain and Ballard (1991), whilst simultaneously being easily indexed by colour-name-based descriptors given by human queries.

The method introduced does have some caveats, and there are failure cases. Further, the approach does not outperform other existing approaches to either of these two objectives in isolation: illuminant-invariant object indexing is better solved by other approaches (e.g. Finlayson and Hordley (2001)), and human-based querying by colour name is better served by the simple approach described in chapter 5. However, that the approach described here performs favourably for both of these objectives in unison is quite remarkable.

Chapter 7

Final Conclusions and Future Work

This thesis has covered several subtopics in the field of colour science and photographic imaging. While the topics themselves are somewhat disparate, this thesis has followed the story of the acquisition and validation of large-scale web-based data, an example application of such data, and finally onto derivations of that application into new algorithms for photographic imaging.

We saw in chapter 3 that web-based paired comparison experiments can deliver acceptable and valuable results so long as certain conditions are met. Namely care must be taken over image presentation to ensure a consistent experience for as many observers as is feasible, the images used for comparison must be carefully selected so as to not introduce bias in the context of web-based experiments, and careful attention must be given to the phrasing of any prompts given to observers. We showed that, when these steps are taken, we can achieve results that are concordant, to a highly significant degree, to those results acquired from lab-based experiments. Furthermore, we identified potential reasons for why historical attempts at similar experiments have failed to deliver such results. These results are promising, and indeed valuable. Web-based data collection is potentially revolutionary for psychophysical experimentation, so long as it is performed correctly. The experiments in chapter 3, however, cover but one particular

experimental paradigm – paired comparisons. Future work should extend into various other experimental paradigms. Ultimately, it would be desirable to construct an open experimental platform offering many experimental paradigms, upon which researchers could easily develop web-based experiments with little effort or monetary expense. Such a platform would offer much utility as, unfortunately, it seems that, while other fields are capitalising on web-based research and crowd-sourced data, the colour research community has made only limited use of this approach.

After observing a need that we, and other experimenters, have for a measure of completeness for paired comparison experiments, in chapter 4 we proposed a statistical tool to quantify the stability or otherwise of paired comparison data. We proposed a method based on the notion of simulated anomalous observers, that is observers whose contributed data would cause maximal perturbation to the currently acquired data. This tool can be applied to provide an estimate of whether or not sufficient observers have completed an experiment, such that reliable results have been gathered and valid conclusions can be derived therefrom.

In chapter 5 we began our exploration of potential applications for web-based data, using freely-available data from an existing experiment (Munroe, 2010). We investigated how computational colour naming (using a model built using the web-sourced data) was affected by illuminant estimation accuracy, and whether colour names could be usefully applied in the context of the object recognition framework of Swain and Ballard (1991). We found that colour names represent a particularly useful quantisation of colour space, offering similar or greater utility than traditional colour histograms, particularly under the conditions of inaccurate illuminant estimation. Furthermore, we demonstrated that histograms constructed from the colour names present in an image not only represent useful image descriptors from a machine object indexing perspective, but that they are also highly perceptually relevant. This allows the same representation to be used to index an image corpus for searching by human-generated queries.

We noted that the colour-name-based descriptors used in chapter 5 were constructed by first correcting images using traditional illuminant estimation algorithms and, in chapter 6, sought to eliminate this need by directly deriving perceptually-relevant illumination-invariant image descriptors based on colour names. We developed an algorithm, using well established observations about the diagonal model of image formation, that was able to deliver such descriptors via the application of a classical boolean discrete relaxation approach. The method was able to assign colour-name-based descriptors to image patches as they would appear under a canonical illuminant, regardless of the actual illumination conditions of the image. These descriptors retained the utility for object indexing and human-based querying as seen in chapter 5, but also provided for illumination invariance. While the results for these specific applications of this approach were satisfactory, the descriptors delivered were unsuitable for other applications such as colour name labelling for display purposes. This was due to the fact that, in many cases, the algorithm was unable to definitively assign image patches with a single, unique, label – e.g. for a blue colour patch the algorithm may converged to [“blue”, “purple”], as opposed to the singular “blue”. Future work could extend to discovering additional constraints to alleviate this issue. Further, the approach described used a boolean discrete relaxation approach – there may also be merit in a probabilistic approach.

This thesis describes contributions which span several topics. The topics discussed in the earlier chapters are fast-moving, and we believe that the contributions made are both valuable and timely. Meanwhile, the contributions made in the later chapters pertain to more mature and slow-moving topics. The foundations that we build upon are, in some cases, decades old, and as such represent well-tested and well-understood techniques and observations. We believe that our contributions provide valuable new ways of exploiting these foundations, and that the results delivered by doing so are particularly noteworthy.

Appendix A

High Dynamic Range Dataset

(b) Belgium
Dani Lischinski

(a) Atrium Night
Karol Myszkowski

Figure A.1: High dynamic range image dataset

(d) Clock Building
Greg Ward

(c) Bristol Bridge
Greg Ward

Figure A.1: High dynamic range image dataset *(cont.)*

(f) Foyer
Harlan Hambright

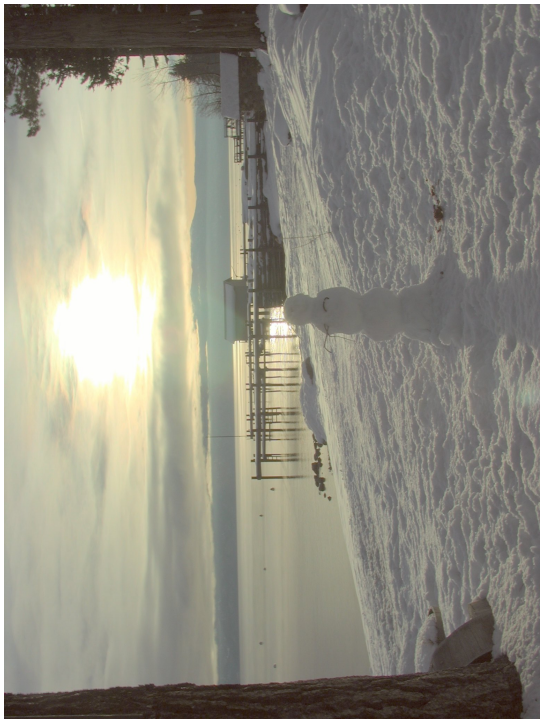
(e) Fog
Jack Tumblin

Figure A.1: High dynamic range image dataset *(cont.)*

(h) *Memorial*
Paul Debevec

(g) *Indoor*
Jacques Joffre

Figure A.1: High dynamic range image dataset *(cont.)*



(j) Tahoe
Greg Ward

(i) Synagogue
Dani Lischinski

Figure A.1: High dynamic range image dataset *(cont.)*

(k) *Tinterna*
Greg Ward

(l) *Tree*
Industrial Light and Magic

(m) *Venice*
Gian Luca Brizi

Figure A.1: High dynamic range image dataset (*cont.*)

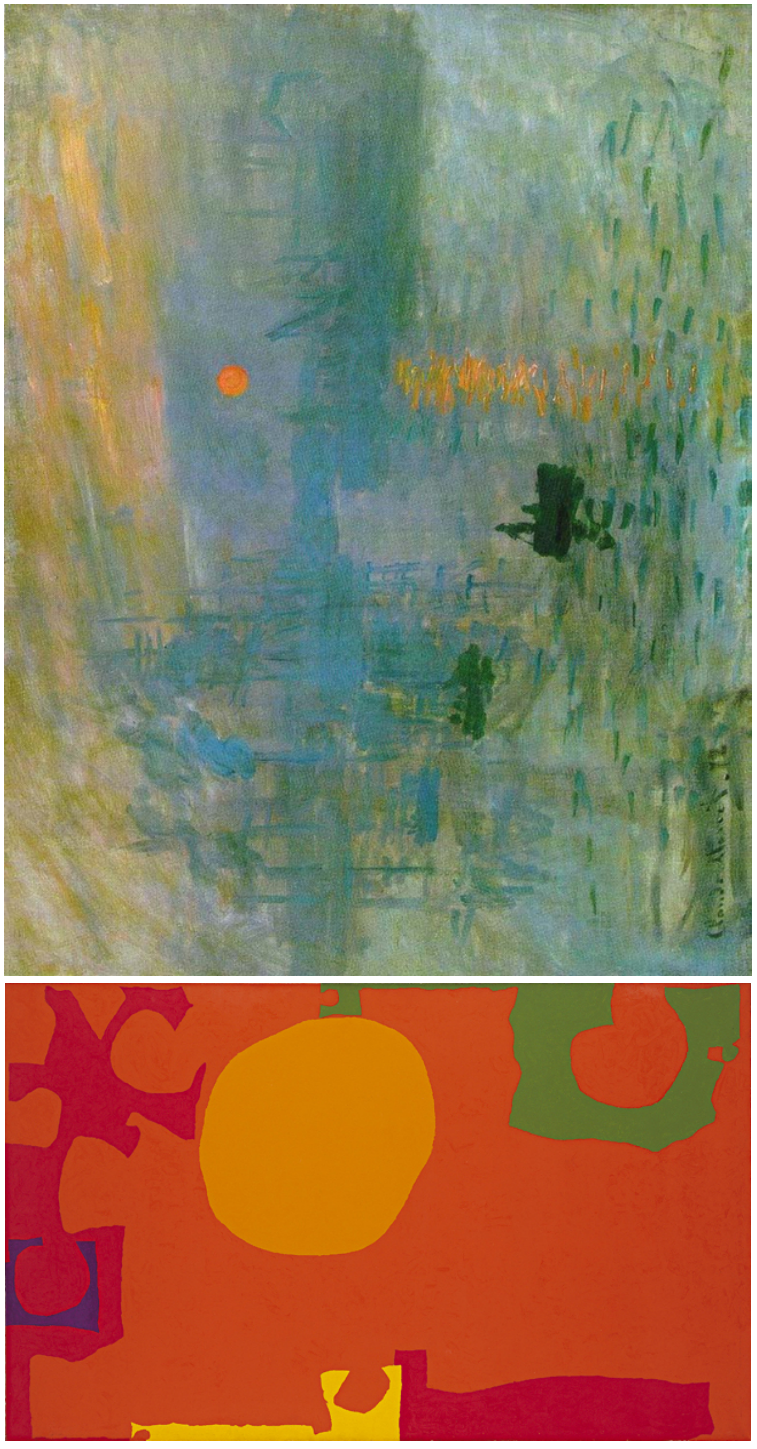
Appendix B

Colour to Greyscale Dataset

(a) Girl
Eastman Kodak Company

(b) Hats
Eastman Kodak Company

Figure B.1: Colour to greyscale image dataset



(c) Heron
'Rumbold Vertical Three: Orange Disc in Scarlet with Green'
Patrick Heron

(d) Monet
'Impression, Sunrise'
Claude Monet

Figure B.1: Colour to greyscale image dataset *(cont.)*

(e) Parrot
Unknown origin

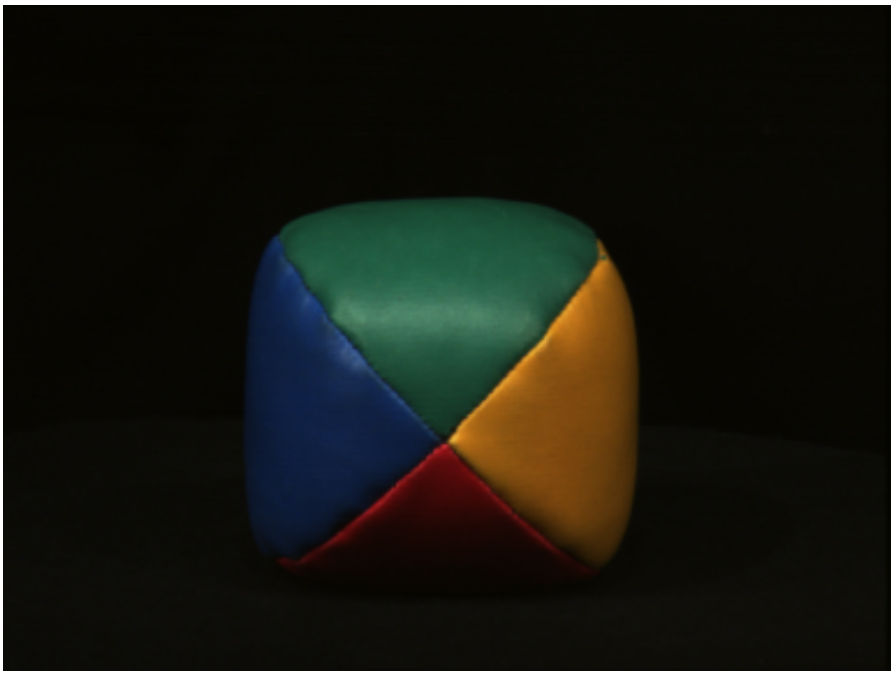
(f) Poppies
Unknown origin

Figure B.1: Colour to greyscale image dataset *(cont.)*

Appendix C

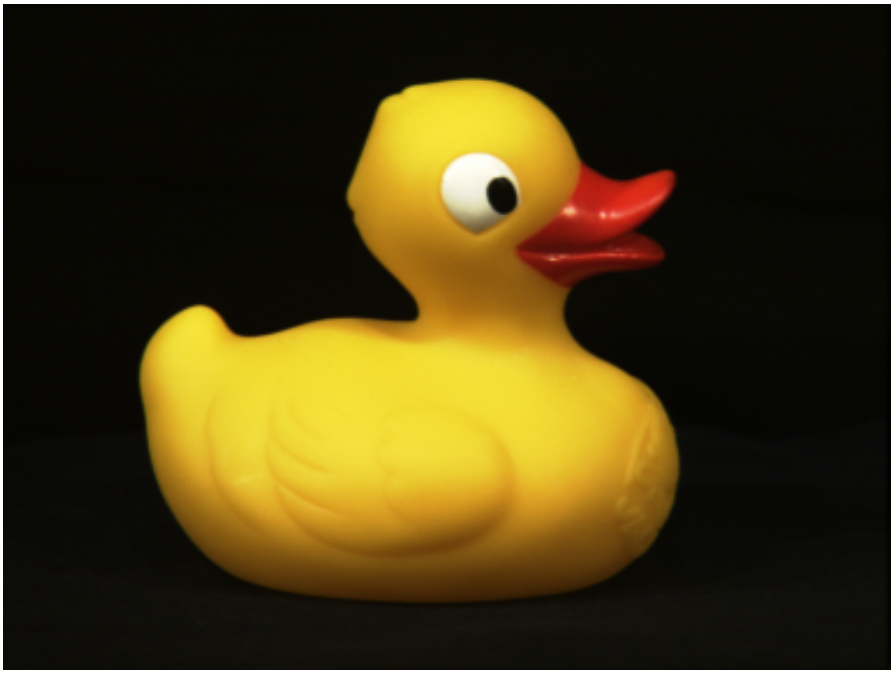
Subset of ALOI Dataset

(a) Teddy Bear with Santa Hat



(b) Juggling Ball

Figure C.1: Subset of ALOI dataset (Geusebroek et al., 2005) used for query-by-colour-name experiment



(c) Rubber Duck

(d) Snowman Ornament

Figure C.1: Subset of ALOI dataset (Geusebroek et al., 2005) used for query-by-colour-name experiment *(cont.)*

(e) Santa Ornament

(f) Papa Smurf

Figure C.1: Subset of ALOI dataset (Geusebroek et al., 2005) used for query-by-colour-name experiment *(cont.)*

(g) Diet Coke Can

(h) Marmite Jar

Figure C.1: Subset of ALOI dataset (Geusebroek et al., 2005) used for query-by-colour-name experiment *(cont.)*

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S., 2012, "SLIC superpixels compared to state-of-the-art superpixel methods." *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–82.

Alsam, A. and Kolås, Ø., 2006, "Grey colour sharpening," in *14th Color and Imaging Conference Final Program and Proceedings*, pp. 263–267.

Bala, R. and Eschbach, R., 2004, "Spatial color-to-grayscale transform preserving chrominance edge information," *signal*, vol. 100, p. 4.

Barnard, K., Finlayson, G., and Funt, B., 1997, "Colour constancy for scenes with varying illumination," *Computer Vision and Image Understanding*, vol. 65, pp. 311–321.

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

Beis, J. and Lowe, D., 1994, "Learning indexing functions for 3-D model-based object recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994*.

Beis, J. and Lowe, D., 1997, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997*.

Benavente, R., Vanrell, M., and Baldrich, R., 2008, "Parametric fuzzy sets for automatic color naming." *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 25, no. 10, pp. 2582–93.

Benavente, R., Vanrell, M., Schmid, C., Baldrich, R., Verbeek, J., Larlus, D., and Van De Weijer, J., 2012, "Color Names," in Gevers, T., Gijsenij, A., van de Weijer, J., and Geusebroek, J.-M., eds., *Color in Computer Vision*, vol. 25, pp. 11–43, Wiley.

Berens, J., Finlayson, G., and Qiu, G., 2000, "Image indexing using compressed colour histograms," in *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 147, p. 349.

Beretta, G. B. and Moroney, N. M., 2012, "Harmonious colors: from alchemy to science," *IS&T/SPIE Electronic Imaging*, vol. 8292, pp. 82920I–82920I–7.

Berlin, B. and Kay, P., 1969, "Basic colour terms," *University of California Press*, vol. 19, p. 23.

Berretti, S., Bimbo, A. D., and Pala, P., 2000, "Retrieval by shape similarity with perceptual distance and effective indexing," *IEEE Transactions on Multimedia*, vol. 2, pp. 225–239.

Birnbaum, M. H., 2004, "Human research and data collection via the Internet," *Psychology*, vol. 55, no. 1, p. 803.

Borges, C., 1991, "Trichromatic approximation method for surface illumination," *JOSA A*, vol. 8, no. 8, pp. 1319–1323.

Boykov, Y., Veksler, O., and Zabih, R., 2001, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239.

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

Chen, T., Chen, L.-H., and Ma, K.-K., 1999, "Colour Image Indexing Using SOM for Region-of-Interest Retrieval," *Pattern Analysis & Applications*, vol. 2, pp. 164–171.

Cheng, H. D., Jiang, X. H., Sun, Y., and Wang, J., 2001, "Color image segmentation: Advances and prospects," *Pattern Recognition*, vol. 34, pp. 2259–2281.

Chong, H., Gortler, S., and Zickler, T., 2007, "The von Kries Hypothesis and a Basis for Color Constancy," *2007 IEEE 11th International Conference on Computer Vision*.

Chuang, J., Stone, M., and Hanrahan, P., 2008, "A probabilistic model of the categorical association between colors," in *Proceedings of the Sixteenth Color and Imaging Conference*, pp. 6–11, Society for Imaging Science and Technology.

Connah, D., Finlayson, G. D., Bloj, M., Science, C., Anglia, E., Colour, B. O., and Sciences, L., 2007, "Seeing beyond luminance: A psychophysical comparison of techniques for converting colour images to greyscale," in *15th Color Imaging Conference: Color, Science, Systems and Applications*, pp. 336–341.

Darrodi, M. M., 2012, "Models of Colour Semiotics," Ph.D. thesis, University of Leeds.

David, H. A., 1988, *The method of paired comparisons*, Oxford University Press, New York, 2nd ed., ISBN 0195206169.

Drago, F., Martens, W. L., Myszkowski, K., and Seidel, H.-P., 2002, "Perceptual evaluation of tone mapping operators with regard to similarity and preference," Tech. rep., Max-Planck-Institut für Informatik.

Drago, F., Myszkowski, K., Annen, T., and Chiba, N., 2003, "Adaptive logarithmic mapping for displaying high contrast scenes," *Computer Graphics Forum*, vol. 22, no. 3, pp. 419–426.

Drew, M. S. and Funt, B. V., 1992, "Natural metamers," *CVGIP: Image Understanding*, vol. 56, no. 2, pp. 139–151.

Duan, J., Bressan, M., Dance, C., and Qiu, G., 2010, "Tone-mapping high dynamic range images by novel histogram adjustment," *Pattern Recognition*, vol. 43, no. 5, pp. 1847–1862.

Durand, F. and Dorsey, J., 2002, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3, pp. 257–266.

D'Zmura, M., 1992, "Color constancy: surface color from changing illumination," *Journal of the Optical Society of America A*, vol. 9, no. 3, pp. 490–493.

Engeldrum, P. G., 2000, *Psychometric scaling: a toolkit for imaging systems development*, Imcotek Press, Winchester, Mass., ISBN 0967870607.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., 2010, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338.

Fattal, R., Lischinski, D., and Werman, M., 2002, "Gradient domain high dynamic range compression," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 249–256.

Felzenszwalb, P. F. and Huttenlocher, D. P., 2005, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, pp. 55–79.

Finlayson, G., 1996, "Color in Perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 1034–1038.

Finlayson, G., Funt, B., and Barnard, K., 1995, "Color constancy under varying illumination," *Proceedings of IEEE International Conference on Computer Vision*.

Finlayson, G. and Hordley, S., 1999, "Selection for gamut mapping colour constancy," *Image and Vision Computing*, vol. 17, pp. 597–604.

Finlayson, G. D., Drew, M. S., and Funt, B. V., 1994, "Spectral sharpening: sensor transformations for improved color constancy." *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 11, pp. 1553–1563.

Finlayson, G. D., Hordley, S., and Hubel, P. M., 2002a, "Illuminant estimation for object recognition," *Color Research & Application*, vol. 27, no. 4, pp. 260–270.

Finlayson, G. D. and Hordley, S. D., 2001, "Color constancy at a pixel," *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 18, pp. 253–264.

Finlayson, G. D., Hordley, S. D., and Hubel, P. M., 2002b, "Color by correlation: A simple, unifying framework for color constancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1209–1221.

Finlayson, G. D. and Morovic, P. M., 2000, "Metamer Constrained Color Correction," *JIST (The Journal of Imaging Science and Technology)*, vol. 44, pp. 295–300, 379–380.

Finlayson, G. D. and Schaefer, G., 2001, "Solving for Colour Constancy using a Constrained Dichromatic Reflection Model," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 127–144.

Finlayson, G. D. and Trezzi, E., 2004, "Shades of gray and colour constancy," in *Proceedings of the Twelfth Color Imaging Conference*, 1, pp. 37–41.

Flickner, M., Sawhney, H., and Niblack, W., 1995, "Query by image and video content: The QBIC system," *Computer*, vol. 28, no. 9, pp. 23–32.

Forsyth, D. A., 1992, "A novel algorithm for color constancy," in *Color*, vol. 36, p. 271, Jones and Bartlett Publishers, Inc.

Foster, D. H., 2003, "Does colour constancy exist?" *Trends in Cognitive Sciences*, vol. 7, pp. 439–443.

Fredembach, C. and Finlayson, G., 2008, "Bright chromagenic algorithm for illuminant estimation," *Journal of Imaging Science and Technology*, vol. 52, no. 4, pp. 40901–40906.

Fu, K. and Mui, J., 1981, "A survey on image segmentation," *Pattern Recognition*, vol. 13, pp. 3–16.

Funt, B., Barnard, K., and Martin, L., 1998, "Is machine colour constancy good enough?" in *Proceedings of the 5th European Conference on Computer Vision*, pp. 445–459.

Funt, B., Cardei, V., and Barnard, K., 1996, "Learning color constancy," in *Proc. IS&T/SID Fourth Color Imaging Conference: Color Science, Systems and Applications*, pp. 58–60, Citeseer.

Funt, B. and Finlayson, G., 1995, "Color constant color indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 522–529.

Funt, B. V., Drew, M. S., and Ho, J., 1991, "Color constancy from mutual reflection," *International Journal of Computer Vision*, vol. 6, pp. 5–24.

Gershon, R., Jepson, A. D., Tsotsos, J. K., and Toronto, T., 1987, "From [R,G,B] to surface reflectance: computing color constant descriptors in images," in *IJCAI'87: Proceedings of the 10th international joint conference on Artificial intelligence*, pp. 755–758, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Geusebroek, J., Burghouts, G., and Smeulders, A., 2005, "The Amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112.

Google Inc, 2009, "Google Browser Size," URL `browsersize.googlelabs.com`.

Grundland, M. and Dodgson, N. A., 2007, "Decolorize: Fast, contrast enhancing, color to grayscale conversion," *Pattern Recognition*, vol. 40, no. 11, pp. 2891–2896.

Hansen, T., Olkkonen, M., Walter, S., and Gegenfurtner, K. R., 2006, "Memory modulates color appearance." *Nature neuroscience*, vol. 9, pp. 1367–1368.

Hansen, T., Walter, S., and Gegenfurtner, K. R., 2007, "Effects of spatial and temporal context on color categories and color constancy." *Journal of vision*, vol. 7, p. 2.

Harris, M. D., 2011, "Colourwar," URL `http://colourwar.com/`.

Hassan, E., Chaudhury, S., and Gopal, M., 2009, "Shape descriptor based document image indexing and symbol recognition," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 206–210.

Healey, G. and Slater, D., 1994, "Global color constancy: recognition of objects by use of illumination-invariant properties of color distributions," *Journal of the Optical Society of America A*, vol. 11, p. 3003.

Heer, J. and Stone, M., 2012, "Color naming models for color selection, image editing and palette design," *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pp. 1007–1016.

Henderson, T., 1990, *Discrete Relaxation Techniques*, Oxford University Press, Inc., New York, NY, USA, ISBN 0-19-504894-6.

Hummel, R. A. and Zucker, S. W., 1983, "On the foundations of relaxation labeling processes." *IEEE transactions on pattern analysis and machine intelligence*, vol. 5, pp. 267–287.

Hurlbert, A., 1999, "Colour vision: Is colour constancy real?" *Current Biology*, vol. 9, no. 15, pp. R558—-R561.

ISO, 2009, "ISO 3664:2009 Graphic technology and photography: Viewing conditions," URL `www.iso.org`.

Jiang, J., Frey, F., Farnand, S., and Frey, J., 2011, "Evaluating the Perceived Quality of Soft-Copy Reproductions of Fine Art Images with and without the Original Present," in *Nineteeth Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications*, pp. 276–284.

Kawakami, R. and Ikeuchi, K., 2009, "Color estimation from a single surface color," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 635–642, IEEE Computer Society, Los Alamitos, CA, USA.

Kawrykow, A., Roumanis, G., Kam, A., Kwak, D., Leung, C., Wu, C., Zarour, E., Sarmenta, L., Blanchette, M., and Waldispühl, J., 2012, "Phylo: A citizen science approach for improving multiple sequence alignment," *PLoS ONE*, vol. 7.

Kay, P. and Regier, T., 2003, "Resolving the question of color naming universals," *Proceedings of the National Academy of Sciences*, vol. 100, no. 15, pp. 9085–9089.

Keener, J. P., Review, S., and Mar, N., 1993, "The Perron-Frobenius theorem and the ranking of football teams," *SIAM review*, vol. 35, no. 1, pp. 80–93.

Kendall, M. G., 1938, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93.

Kendall, M. G. and Smith, B. B., 1940, "On the method of paired comparisons," *Biometrika*, vol. 31, no. 3/4, pp. 324–345.

Kraft, J. M. and Brainard, D. H., 1999, "Mechanisms of color constancy under nearly natural viewing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 1, p. 307.

Lakhani, K. R., Boudreau, K. J., Loh, P.-R., Backstrom, L., Baldwin, C., Lonstein, E., Lydon, M., Maccormack, A., Arnaout, R. a., and Guinan, E. C., 2013, "Prize-based contests can provide solutions to computational biology problems." *Nature biotechnology*, vol. 31, pp. 108–11.

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

Land, E. H. and McCann, J. J., 1971, "Lightness and retinex theory," *Journal of the Optical society of America*, vol. 61, no. 1, pp. 1–11.

Larson, G. W., Rushmeier, H., and Piatko, C., 1997, "A visibility matching tone reproduction operator for high dynamic range scenes," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 3, no. 4, pp. 291–306.

Ledda, P., Chalmers, A., Troscianko, T., and Seetzen, H., 2005, "Evaluation of tone mapping operators using a high dynamic range display," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 640–648.

Lee, H. C., 1986, "Method for computing the scene-illuminant chromaticity from specular highlights." *Journal of the Optical Society of America. A, Optics and image science*, vol. 3, pp. 1694–1699.

Mahmoudi, F., Shanbehzadeh, J., Eftekhari-Moghadam, A.-M., and Soltanian-Zadeh, H., 2003, "A new non-segmentation shape-based image indexing method," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 3.

Maloney, L. T. and Wandell, B. A., 1986, "Color constancy: a method for recovering surface spectral reflectance," *JOSA A*, vol. 3, no. 1, pp. 29–33.

Mantiuk, R., Daly, S., and Kerofsky, L., 2008, "Display adaptive tone mapping," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3, p. 68.

Mehtre, B. M., Kankanhalli, M. S., Narasimhalu, A. D., and Man, G. C., 1995, "Color matching for image retrieval," *Pattern Recognition Letters*, vol. 16, no. 3, pp. 325–331.

Mei, Y., 2010a, "High Dynamic Range Image Comparison," URL `http://hdri.cs.nott.ac.uk/`.

Mei, Y., 2010b, "High Dynamic Range Image Comparison - TOP 10," URL `http://hdri.cs.nott.ac.uk/v1/top10.php`.

Moroney, N., 2003, "Unconstrained web-based color naming experiment," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 5008, pp. 36–46.

Moroney, N. and Beretta, G., 2011, "Validating large-scale lexical color resources," in *Midterm Meeting of the International Colour Association (AIC)(June 2011)*.

Morrissey, J. H., 1955, "New method for the assignment of psychometric scale values from incomplete paired comparisons." *Journal of the Optical Society of America*, vol. 45, no. 5, pp. 373–378.

Mosteller, F., 1951, "Remarks on the method of paired comparisons: III. A Test of Significance for Paired Comparisons When Equal Standard Deviations and Equal Correlations are Assumed," *Psychometrika*, vol. 16, no. 2, pp. 207–218.

Munroe, R., 2010, "Color Survey Results," URL `http://blog.xkcd.com/2010/05/03/color-survey-results/`.

Mylonas, D., Stutters, J., Doval, V., and MacDonald, L., 2013, "Colournamer, a synthetic observer of colour communication," in *Proceedings of the Twelfth International AIC (Association Internationale de la Couleur) Congress*, Newcastle Upon Tyne, England.

Nayar, S. K. and Bolle, R. M., 1993, "Computing reflectance ratios from an image," *Pattern Recognition*, vol. 26, pp. 1529–1542.

Olkkonen, M., Hansen, T., and Gegenfurtner, K., 2009, "Categorical color constancy for simulated surfaces," *Journal of Vision*, vol. 9, pp. 1–18.

Oren, M. and Nayar, S., 1995, "Generalization of the Lambertian model and implications for machine vision," *International Journal of Computer Vision*, vol. 14, pp. 227–251.

Pearson, E. S. and Hartley, H., 1966, *Biometrika tables for statisticians*, vol. 1, Cambridge University Press, 3rd ed.

Pelillo, M., 1997, "The dynamics of nonlinear relaxation labeling processes," *Journal of Mathematical Imaging and Vision*, vol. 7, pp. 309–323.

Qiu, G., 2002, "Indexing chromatic and achromatic patterns for content-based colour image retrieval," *Pattern Recognition*, vol. 35, pp. 1675–1686.

Qiu, G. and Duan, J., 2005, "Hierarchical tone mapping for high dynamic range image visualization," in *Visual Communications and Image Processing*, vol. 5960, pp. 2058–2066, Citeseer, Spie.

Qiu, G., Mei, Y., and Duan, J., 2011, "Evaluating HDR Photos Using Web 2.0 Technology," in *IS&T/SPIE Electronic Imaging*, vol. 7867, p. 24.

Ramanath, R., Snyder, W. E., Yoo, Y., and Drew, M. S., 2005, "Color image processing pipeline," *Signal Processing Magazine, IEEE*, vol. 22, no. 1, pp. 34–43.

Rasche, K., Geist, R., and Westall, J., 2005a, "Detail preserving reproduction of color images for monochromats and dichromats," *Computer Graphics and Applications, IEEE*, vol. 25, no. 3, pp. 22–30.

Rasche, K., Geist, R., and Westall, J., 2005b, "Re-coloring images for gamuts of lower dimension," *Computer Graphics Forum*, vol. 24, pp. 423–432.

Rasmussen, D., 2008, "Online image quality surveys based on response time," *Electronic Imaging 2008*, vol. 6808, pp. 1–12.

Reinhard, E. and Devlin, K., 2005, "Dynamic range reduction inspired by photoreceptor physiology," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 11, no. 1, pp. 13–24.

Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., 2002, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 267–276.

Roberson, D., Davidoff, J., Davies, I. R. L., and Shapiro, L. R., 2005, "Color categories: Evidence for the cultural relativity hypothesis," *Cognitive Psychology*, vol. 50, no. 4, pp. 378–411.

Roberson, D., Davies, I., and Davidoff, J., 2000, "Color categories are not universal: replications and new evidence from a stone-age culture." *Journal of experimental psychology. General*, vol. 129, no. 3, pp. 369–398.

Saunders, D., Yoshida, K., Sambles, C., Glover, R., Clavijo, B., Corpas, M., Bunting, D., Dong, S., Clark, M., Swarbreck, D., Ayling, S., Bashton, M., Collin, S., Hosoya, T., Edwards, A., Crossman, L., Etherington, G., Win, J., Cano, L., Studholme, D., Downie, J. A., Caccamo, M., Kamoun, S., and MacLean, D., 2014, "Crowdsourced analysis of ash and ash dieback through the Open Ash Dieback project: A year 1 report on datasets and analyses contributed by a self-organising community." Tech. rep., The Sainsbury Laboratory, Norwich, doi:10.1101/004564, URL `http://biorxiv.org/content/early/2014/04/25/004564.abstract`.

Schettini, R., Ciocca, G., and Zuffi, S., 2002, "A survey of methods for colour image indexing and retrieval in image databases," in *Colour Image Science: Exploiting Digital Media*, vol. 54, pp. 183–212, Wiley, ISBN 978-0471499275.

Shafer, S. a., 1985, "Using color to separate reflection components," *Color Research and Application*, vol. 10, pp. 210–218.

Smith, T. and Guild, J., 2002, "The C.I.E. colorimetric standards and their use," *Transactions of the Optical Society*, vol. 33, pp. 73–134.

Socolinsky, D. A. and Wolff, L. B., 2002, "Multispectral image visualization through first-order fusion," *Image Processing, IEEE Transactions on*, vol. 11, no. 8, pp. 923–931.

Sprow, I., Baranczuk, Z., Stamm, T., and Zolliker, P., 2009, "Web-based psychometric evaluation of image quality," *Image Quality and System Performance VI*, vol. 7242, p. 72420A.

Stokes, M., Anderson, M., Chandrasekar, S., and Motta, R., 1996, "A standard default color space for the internet-sRGB," *Microsoft and Hewlett-Packard Joint Report*.

Süsstrunk, S., 2005, "Computing Chromatic Adaptation," Ph.D. thesis, University of East Anglia.

Swain, M. J. M. and Ballard, D. H. D., 1991, "Color indexing," *International journal of computer vision*, vol. 32, pp. 11–32.

Tan, T., Nishino, K., and Ikeuchi, K., 2003, "Illumination chromaticity estimation using inverse-intensity chromaticity space," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 673–682.

Tauber, J., 2009, "typewar," URL `http://typewar.com/`.

Thurstone, L. L., 1927, "A law of comparative judgment," *Psychological review*, vol. 34, no. 4, pp. 273–286.

Tsukada, M. and Ohta, Y., 1990, "An approach to color constancy using multiple images," in *Proceedings of the Third International Conference on Computer Vision*.

Tumblin, J. and Turk, G., 1999, "LCIS: A boundary hierarchy for detail-preserving contrast reduction," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 83–90, ACM Press/Addison-Wesley Publishing Co.

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

Vazquez-Corral, J., Vanrell, M., Baldrich, R., and Tous, F., 2012, "Color Constancy by Category Correlation," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1997–2007.

Čadík, M., Wimmer, M., Neumann, L., and Artusi, A., 2008, "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Computers and Graphics (Pergamon)*, vol. 32, pp. 330–349.

Vrhel, M. and Trussell, H., 1992, "Color correction using principal components," *Color Research & Application*, vol. 17, no. 5, pp. 328–338.

Waltz, D., 1975, "Understanding line drawings of scenes with shadows," in Winston, P. H., ed., *The Psychology of Computer Vision*, pp. 19–91, McGraw-Hill, ISBN 0070710481.

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

Worthey, J. a. and Brill, M. H., 1986, "Heuristic analysis of von Kries color constancy." *Journal of the Optical Society of America. A, Optics and image science*, vol. 3, no. 10, pp. 1708–12.

Yoshida, A., Blanz, V., Myszkowski, K., and Seidel, H.-p. P., 2005, "Perceptual evaluation of tone mapping operators with real-world scenes," in *Proc. SPIE*, vol. 5666, pp. 192–203, Citeseer.

Yu, Z., 2009, "Intrinsic shape signatures: A shape descriptor for 3D object recognition," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, pp. 689–696.

Zuffi, S., Brambilla, C., Eschbach, R., and Rizzi, A., 2008, "Controlled and Uncontrolled Viewing Conditions in the Evaluation of Prints," *Color Imaging XIII: Processing, Hardcopy, and Applications*, vol. 6807, no. 1, p. 680714.

Zuffi, S., Scala, P., Brambilla, C., and Beretta, G., 2007, "Web-based vs. controlled environment psychophysics experiments," *Image Quality and System Performance IV, Proceedings of SPIE*, vol. 6494, no. February 2007.