Modeling and measuring image quality for gamut mapping

Dissertation

zur Erlegung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik

der Friedrich-Schiller-Universität Jena

von M.Sc. Zofia Barańczuk - Turska

geboren am 16.06.1982 in Polen

Gutachter

1. Prof. Dr. Stefan Funke

2. Prof. Dr. Joachim Giesen

3. Dr. Peter Zolliker

Tag der öffentlichen Verteidigung: 20.10.2010, Jena

Zofia Barańczuk - Turska

Education and Work Experience

- EMPA Employee, PhD student at Friedrich Schiller University of Jena, 11.2008-11.2010
- Maths teacher in High School no. 40 in Warsaw, 2006-2007
- M.Sc., Mathematics, "Stochastic models of autoregulation in genetic networks", 12.2006
- B.S., Computer Science, "Common Internet shopping platform", 6.2005
- Warsaw University, Maths & Computer Science Faculties, 10.2001-12.2006
- High School no. 1 in Bialystok, 1997-2001

List of publications and talks

[1] Z. Barańczuk. Image quality measures and individualized conjoint analysis for evaluating gamut mapping algorithms. Invited talk on Gjøvik Color Imaging Symposium, Juni 2009.

[2] Z. Barańczuk, J. Giesen, K. Simon, and P. Zolliker. Gamut mapping. In Peter Hawkes, editor, *Advances in Imaging and Electron Physics*, volume 160, pages 1–34. Elsevier, 2010.

[3] Z. Barańczuk, P. Zolliker, and J. Giesen. Image quality measure for evaluating gamut mapping. In *17th Color Imaging Conference*, volume 17, pages 21–26, Albuquerque, NM, 2009. IS&T/SID.

[4] Z. Barańczuk, P. Zolliker, and J. Giesen. Image-individualized gamut mapping algorithms. *Journal of Imaging Science and Technology*, 54(3):030201–(7), 2010.

[5] Z. Barańczuk, P. Zolliker, I. Sprow, and J. Giesen. Conjoint analysis of parametrized gamut mapping algorithms. In *16th Color Imaging Conference*, volume 16, pages 38–43, Scottsdale, AR, Nov 2008. IS&T/SID.

[6] G. Cao, M. Pedersen, and Z. Barańczuk. On the use of saliency maps for the detection of print artifacts. In *The 5th European Conference on Colour in Graphics, Imaging and Vision*, Joensuu, Finland, 2010. IS&T/SID.

[7] M. Meili, D. Küpper, Z. Barańczuk, U. Caluori, and K. Simon. Filter methods to preserve local contrast and to avoid artifacts in gamut mapping. In *Color Imaging XV: Displaying, Processing, Hardcopy, and Applications*, volume 7528, page 75280. SPIE, 2010.

[8] I. Sprow, Z. Barańczuk, T. Stamm, and P. Zolliker. Web-based psychometric evaluation of image quality. In *Image Quality and System Performance VI*, page 72420A. SPIE, 2009.

[9] P. Zolliker and Z. Barańczuk. Error estimation of paired comparison tests for Thurstone's Case V. In *The 5th European Conference on Colour in Graphics, Imaging and Vision*, Joensuu, Finland, 2010. IS&T/SID.

[10] P. Zolliker, Z. Barańczuk, I. Sprow, and J. Giesen. Conjoint analysis used for the evaluation of parameterized gamut mapping algorithms. *IEEE Transactions in Image Processing*, 19(3):758–769, 2010.

Contents

Abstract

In modern electronic media, digital images are the major means for image reproduction. Digital data are being presented on different devices with different color reproduction capabilities. Every time digital data has to be transformed from one device to another, one has to adapt the color specification to the output device's capabilities, called output gamut. This process is called gamut mapping.

One of the reasons, why designing a gamut mapping algorithm is a complex problem, is the fact that it is hard to gauge the quality of a gamut mapping algorithm. Traditionally one uses psycho-visual testing to assess the visual performance of algorithms, but this approach is very time-consuming. As an alternative, automated models to measure image quality are being researched, though they cannot precisely reflect human visual preferences.

The thesis consists of two parts. The first part provides theoretical foundations for different aspects of the evaluation of gamut mapping algorithms. We present statistical methods for the evaluation of psycho-visual data, namely, Thurstone's method and conjoint analysis. We describe how to test underlying assumptions and discuss possible methods of error analysis. In addition to psycho-visual testing, we present image quality measures that are useful for evaluating gamut mapping algorithms.

In the second part we present practical applications of the methods introduced in the first part.

We describe a psycho-visual test for the evaluation of parametrized gamut mapping algorithms. In order to handle a huge amount of possible algorithm variants (in our case 1536), conjoint analysis is applied. We verify the assumptions that allow us to deal with this large space of parameters. The results of the conjoint analysis allow us to estimate the importance of different parameters of the algorithms. We also analyze errors for different tests.

Further, we compare individualized and non-individualized evaluation methods (both based on psycho-visual data and image quality measures) using hit rates.

Motivated by the higher accuracy obtained by individualized models we design a meta-algorithm that chooses an optimal (according to the given model) algorithm for every image. We estimate the quality of the meta-algorithm and show that it performs better than any single algorithm considered in the test.

Zusammenfassung

In modernen elektronischen Medien ist die digitale Form von Bildern Standard und bildet die Basis für die Bilderverarbeitung. Da diese Geräte aber sehr verschiedene Farbwiedergabefähigkeiten besitzen, müssen die digitalen Daten jeweils an den Farbumfang (genannt Gamut) des Ausgabegerätes angepasst werden. Diese Anpassung nennt man Gamut Mapping.

Einer der Gründe, die das Design eines Gamut Mapping Algorithmus zu einer so komplexen Angelegenheit machen, ist die Tatsache, dass es sehr schwer ist, die visuelle Qualität eines Gamut Mapping Algorithmus abzuschätzen. Normalerweise führt man psychovisuelle Tests durch, um die visuelle Qualität von Gamut Mapping Algorithmen zu bestimmen. Diese Methode ist aber äusserst zeitaufwändig. Deshalb wäre ein automatisiertes Auswertungsmodell wünschenswert. Solche Modelle werden momentan erforscht - diese können zwar bislang menschliche visuelle Präferenzen nicht genau wiedergeben, aber sie zumindest approximieren.

Diese Arbeit gliedert sich in zwei Teile. Der erste Teil ist eine theoretische Grundlage zur Bewertung von Gamut Mapping Algorithmen. Wir präsentieren statistische Methoden zur Auswertung von psychovisuellen Daten, in Besonderen die Thurstone-Methode und die Conjoint-Analyse. Wir beschreiben, wie man die Modellannahmen überprüfen kann, und diskutieren mögliche Methoden für die Fehlerrechnung. Als eine Alternative zu psychovisuellen Tests präsentieren wir die wichtigsten Bildqualitätsmasse, die für die Bewertung von Gamut Mapping Algorithmen nützlich sind.

Im zweiten Teil wenden wir die im ersten Teil eingeführten Methoden an. Als erstes präsentieren wir einen psychovisuellen Test von parametrisierten Gamut Mapping Algorithmen und dessen Datenauswertung. Wir wenden Conjoint-Analyse an, um die grosse Zahl von möglichen Algorithmen (in unserem Fall 1536) zu beurteilen. Wir untersuchen die Wichtigkeit der verschiedenen Parameter der Algorithmen und analysieren auch Fehler für verschiedene Tests. Des Weiteren vergleichen wir individualisierte und nicht-individualisierte Bewertungsmethoden (basierend auf psychovisuellen Daten oder Bildqualitätsmassen) mit Hilfe von Hit Rates.

Motiviert durch die höhere Genauigkeit der individualisierten Modelle entwickeln wir einen Meta-Algorithmus, der einen optimalen Algorithmus (in Bezug auf das gegebene Modell) für einzelne Bilder wählt. Wir validieren den Meta-Algorithmus und zeigen, dass er besser funktioniert als jeder einzelne Algorithmus den wir in diesem Test berücksichtigt haben.

Acknowledgements

Work for this thesis has been supported by the Hasler Foundation (Hasler Stiftung).

In this place I would like to thank The Gjøvik Color Lab for granting access to one of the data sets.

I would like to thank all people who contributed to this work. I thank my supervisor Joachim Giesen for his guidance and encouragement. I'm sincerely grateful to Peter Zolliker for introducing me to color science and gamut mapping and for always having time for discussions, patience, support and mountain trip tips. Many thanks go to Klaus Simon for his leadership, encouraging me to believe in the impossible, trying to broaden my horizon in history and always believing in me.

I would like to thank Ursina Caluori, Iris Sprow and Matthias Scheller Lichtenauer for their valuable comments on this thesis. I am grateful to my co-authors during my work on PhD and colleagues from the Media Technology Lab at EMPA. It was a great pleasure working and spending time with them. Special thanks to Ursina for bringing joy to my life.

Chapter 1

Introduction

Artists can color the sky red because they know it's blue.
Those of us who aren't artists
must color things the way they really are
or people might think we're stupid.
Jules Feiffer

1.1 Motivation

The visual system is the dominant component of human perception, and an important part of it is the ability to discriminate colors. Consequently, visual information in form of color images is a natural part of modern communication, in particular on the Internet.

In modern electronic media, information is usually processed as digital data. Such data needs a physical device to become visible. Physical devices have restricted color reproduction capabilities, and color scientists refer to the set of colors that a device or a process can present (monitor), reproduce (printer), capture (camera), or store (computer) as the color gamut of the device. For example, due to device limitations a printer is typically not able to reproduce all the colors visible on a display.

The transformation of color data from one device specification to another is called gamut mapping. Typically, the transformation is from an input to an output device specification.

In traditional photo-mechanical reproduction, gamut mapping is implicitly given by the physical behavior of the devices. However, desktop publishing and digital reproduction changed the situation fundamentally as the input is given in digital form, namely, a specification by device independent color coordinates. Gamut mapping is then realized as a colorimetric function in a device independent color space. Traditionally, only the final result of gamut mapping has been evaluated by the user, whereas in a completely digital process it is possible to control and evaluate gamut mappings colorimetrically. Also, gamut mapping is no longer physically determined but has to be designed mathematically.

Consequently, the ICC color management (ICC-CMS) has been introduced in 1993.[1] This nowadays dominating software standard describes how to translate device dependent color coordinates (e.g., RAW-RGB, CMYK) to device independent color spaces (like CIEXYZ or CIELAB) and vice versa. The basic tools are interpolation tables, known as lookup tables (LUT), which are stored

[1] see `www.color.org`

in device characterization files (profiles). This approach works well for all colors reproducible by a device, but out-of-gamut colors are neglected.

Over the last two decades, gamut mapping research evolved from finding the best treatment of out-of-gamut colors to searching for the optimal image dependent mapping in specific environments. Optimality is defined here by visual image quality measures that should keep the mapped image as similar as possible to the original. However, no simple definitions exist to describe similarity or image quality, and thus the definition of the goal of gamut mapping is itself a challenge. This can be addressed by experimental methods assessing human preferences in psycho-visual tests. Human perception can be roughly modeled by perceptual attributes such as lightness, saturation (colorfulness), hue, sharpness, and details (contrast). In his book on psychometric scaling Engeldrum [19] calls these attributes "Customer Perception-Nesses". In order to derive an image quality measure for optimizing gamut mapping algorithms, good approximations of these attributes by physical image parameters (also called 'objective measures') are needed. Furthermore, an appropriate relative weighting of the attributes is important.

Consider the following illustrative example: map each in-gamut color onto itself and each out-of-gamut color to its closest in-gamut color (using an appropriate color space metric). This approach, called clipping, minimizes the average color difference of all pixels in an image. However, colors are always seen in their spatial neighborhood as described by perceptual attributes like sharpness and contrast which are not covered by the colorfulness attribute. Hence, the mapped image is not optimal with respect to these attributes. In Figure 1.1 this is illustrated by a sample image. When the perceived color distances are minimized, i.e., colorfulness is maximized, image details are lost in color saturated regions. On the other hand, details can be preserved by using methods like linear compression, but then the mapped image usually is desaturated. In general, attributes like details and colorfulness cannot be optimally preserved simultaneously. Hence, there is always a trade-off between different attributes involved, and the design of an optimal gamut mapping algorithm needs (1) good psycho-visual test data, (2) appropriate image quality measures, and (3) a versatile gamut mapping framework allowing optimization.

Fig. 1.1: The effect of different gamut mapping strategies: original (left), linear compression (middle), clipping (right).

1.2 Formulation of the present research

Typical gamut mapping studies involve only a few algorithms. However, in order to optimize the performance of an algorithm, we consider many parameters of the algorithm and try to find optimal settings. As parametrization introduces much more algorithms to compare, one has to

choose a method which allows to obtain statistically significant results using a reasonable number of comparisons. We explored the use of conjoint analysis, which is a popular method in market research to test preferences of structured products among customers.

Even though conjoint analysis reduces the required number of comparisons, still a lot of data is needed. Therefore we decided to collect data over the Internet as well. We verified carefully that this approach yields usable data [56].

In order to have an alternative to the time-consuming psycho-visual tests, one wants to assess the quality of the mapped images automatically. Moreover, as already shown e.g. by Hardeberg [32] and what can be extracted from error analysis (compare Section 4.2.2), different algorithms are optimal for different images. Carrying out a psycho-visual test for each new image is not feasible. However, one can approximate the quality of the mapped image computing the difference between a given mapped image and its original using image quality measures (compare Bonnier [3] and Hardeberg [32]). As image quality measures are effective enough to consider individual images in practical applications, one can optimize gamut mapping algorithms by using image-individualization.

1.3 Outline of the thesis

The thesis consists of two parts. The first one (Chapter 2 and Chapter 3) covers the theoretical foundation for our research. The second one (Chapter 4 and Chapter 5) presents applications of the methods described in the first part. The outline of the thesis is as follows:

In Chapter 2 we present in detail methods used to evaluate psycho-visual data: Thurstone's method and conjoint analysis, and the models standing behind these methods. Further, we investigate different methods of computing errors for these models and discuss their appropriateness. We also present methods for validating the models, both for verifying the assumptions of the models and for examining the accuracy of the model comparing consistency between the data and the model. We describe cross validation, which avoids testing the model on the training data.

In Chapter 3 we describe non test-based methods for comparing the quality of gamut mapped images. Firstly, we summarize the basics of color science. This is a background for presenting a set of reference image quality measures, i.e., functions of two images giving as a result a value, which corresponds to the similarity or correlation between these two images.

In Chapter 4 we present a test of parametrized *g*amut *m*apping *a*lgorithms (GMAs). Thanks to the parametrization of the algorithms we can use conjoint analysis to evaluate them. We are not only looking for the optimal algorithm from the whole set, but also discuss the importance of different parameters and their interactions. We also verify fulfilling of the assumptions and analyze errors.

Triggered by the significant image dependency of the optimal parameter settings, we describe in Chapter 5 image-individualized models for evaluating GMAs. The models are based either on psycho-visual data or image quality measures. We test the accuracy of these models using the hit rate and confirm, that the good individualized models are significantly more accurate than non-individualized. We also develop an image-individualized meta-algorithm and validate it.

Part I

Theoretical Foundations

Chapter 2

Psychophysics

Man is the measure of all things.
Protagoras

An important topic in this thesis is the measurement of quality of gamut-mapped images in order to optimize gamut mapping. A possible definition of optimality consists of a visual image quality measure that describes the similarity of a mapped image to its original. However, there are no simple definitions describing similarity or quality of images and thus the definition of the objectives of gamut mapping is itself a challenge. The basic goal of mapping images is that people are not aware that the image was altered at all. There are no automatic methods or models describing human preferences exactly. Moreover, preferences can be different within the group of people. Global evaluation is thus restricted to determine the typical or mean preferences. A traditional way of evaluating gamut-mapped images are experimental methods assessing human preferences in psycho-visual tests. Psycho-visual tests are used for subjective evaluation of the quality of images. The aim of the tests is to compare images with respect to perceived quality.

Such a test is typically conducted in the following way: images mapped differently are shown to observers, who have to answer a specific question concerning the *quality* of the images or the *similarity* to the reference image. A sample user interface for a test is shown in Figure 2.1. There are many possible designs of the test which we will summarize further in this chapter. By evaluating test data one can obtain significant results describing the perceptual quality of the considered images. Often an interval scale for the images is computed from these comparisons, where a scale value is a measure for the quality of an image.

A traditional method of evaluating such data uses Thurstone's Law of Comparative Judgement [57]. However, this method is practical only, if the number of algorithms in consideration is small. For example, already for 20 algorithms, with 10 comparisons for each pair, one would need 1900 comparisons for each considered image. It is too much for practical application, especially because in the test there are typically at least 20 images, which would increase a total number of comparisons to 38000.

Parametrized algorithms usually have more than 20 instances. In this case Thurstone's method can be extend to the multiparameter case using *conjoint analysis*. It allows evaluating the quality of a high number of structured items (having conjoint structure).

In both these methods it is important to compute errors or confidence intervals, as we have to see if differences between algorithms are significant. There are a few methods described in the literature for computing errors for Thurstone's Law of Comparative Judgement. We recapitulate them and present an improved method. We compare the accuracy of these methods. Further we describe how to rescale errors in conjoint analysis.

Fig. 2.1: An example of a user interface for a paired comparison test. The image in the middle is the reference (original) image. Two lower images are two gamut-mapped images. The observer is asked to choose from the mapped images the one more similar to the original.

2.1 Design of psycho-visual tests

In the set up of a psycho-visual test it is important to choose a suitable method. The CIE guideline for testing gamut mapping algorithms [12] try to standardize the way of conducting psycho-visual tests for gamut mapping in order to make the results comparable. They provide specific experimental methods, viewing conditions, and reference algorithms. Three methods of psycho-visual tests are recommended for evaluating the quality of gamut mapping algorithms: paired comparison, rank order and category judgement. The most widely used method is paired comparison. It is the easiest one for observers, especially if differences between estimated images are small. However, the other methods are also useful for specific problems, in particular they are typically less time consuming than paired comparisons. In the following we describe these methods.

2.1.1 Test methods

Paired comparisons

A pair of images from a set A is presented to an observer. He or she is then asked to choose the one that better fulfills instructions of the test. In the gamut mapping case, the instructions usually state that one should choose the more aesthetic image, or the image more similar to the original. In the second case the original image is shown along with the transformed images. Paired comparison tests have been used for example by Farup [24], Bonnier [3] or Morovic [48]. Typically, the data are stored in a frequency matrix $F = (f_{ab})$, where f_{ab} is the number of comparisons where a has been preferred over b ($a \succ b$). In this chapter we focus on processing data obtained using paired comparisons.

Rank order

An observer is asked to rank a set of images from the best to the worst along an attribute defined by instructions, e.g. similarity to the original or aesthetics of the image. Each ordering can be interpreted as $n(n-1)/2$ paired comparisons (for each pair within compared set we know the ordering), so ranking is the more effective assessment method than paired comparison. The gathered data in a rank order test can also be stored in a frequency matrix. Alternative evaluation methods for rank order data is computing average ranks [19], or distance based models as described and applied by Millen et al. [45]. The main disadvantage of this method is the fact, that ordering very similar images can be a difficult task for the observers.

Category judgement

In a category judgement test observers are asked to rate images in categories. Categories can have descriptive names or just be numbers. If the quality of images is considered, category names can be for example: 5. Excellent, 4. Good, 3. Fair, 2. Poor, and 1. Bad. Category judgements have been used for example by Morovic [51]. The main disadvantage of this method is that assigning an image to a given category depends strongly on the subjective interpretation where the borders between the categories are. Different observers also use the scale differently. Hence the most natural evaluation of the data, i.e., taking the mean values, is not reliable. The data must be further processed.

2.1.2 Experimental design

In each of those methods one has to decide, whether the observer should see a reference image while doing the test or just see different transformed images. This choice depends mostly on the aim of the test. In case of gamut mapping the typical task is to find the most similar representation of the given image within the limitations of the target gamut. Hence a reference image is usually shown to observers.

Another issue to consider while planning a test is whether ties are accepted or not. If we let observers state, that there is no difference in quality between images, we can expect the data will be less noisy, as observers make less random choices in case of very similar images. On the other hand, if we use hit rates to validate the models (compare Section 2.4), ties are more difficult to handle, thus they are often omitted.

Experimental design in parametrized tests

When designing a test on multi-parameter items, one has to decide which of the possible items should be compared. For multi-parameter items comparing each set (or each pair) of items is usually impractical. A good overview of multi-parameter test design can be found in Chrzan [7]. He discusses which designs allow to investigate different effects. He also describes the effectiveness of designs, i.e., how many comparisons for different designs result in similar statistical significance. In our studies we used random design. It allows to investigate all effects with the lowest probability of a systematic error. On the other hand, it is less effective than other methods. Therefore, if one is interested only in some basic effect, one should choose a design with a minimum number of required comparisons.

2.2 Discrete choice models

In this section we focus on evaluating choice data concerning unstructured items obtained by the paired comparison method.

Given a set A of n stimuli, e.g., gamut mapping algorithms, and choice data of the form $a \succeq b$ with $a, b \in A$. We know the frequency f_{ab} (F-matrix) that stimulus a is preferred over stimulus b (number of times a is preferred over b). We consider the proportion q_{ab} that stimulus a is preferred over stimulus b

$$q_{ab} = \frac{f_{ab} + \delta}{f_{ab} + f_{ba} + 2\delta}. \tag{2.1}$$

as an indirect measure for the distance of the "qualities" (scale values) v_a of a and v_b of b, respectively[1].

Discrete choice models build on the assumption that the observers' choices are outcomes of random trials: confronted with the two options $a, b \in A$ an observer assigns subjective scale values $u_a = v_a + \epsilon_a$ and $u_b = v_b + \epsilon_b$, respectively, to the stimuli. The error terms ϵ_a and ϵ_b are drawn independently from the same distribution. The observer then prefers the stimulus with the larger scale value. Hence the probability p_{ab} that a is preferred over b is given as

$$\begin{aligned} p_{ab} &= Pr[u_a \geq u_b] \\ &= Pr[v_a + \epsilon_a \geq v_b + \epsilon_b] = Pr[v_a - v_b \geq \epsilon_b - \epsilon_a] \quad . \end{aligned} \tag{2.2}$$

Here we discuss two discrete choice models that differ in the choice of distribution for ϵ_a. We consider normal distributions (Thurstone's (probit) model [57]), and Gumbel distributions (Bradley-Terry's (logit) model [5]). In both cases, the scale values v_a can be computed by a least squares approach from the probability p_{ab} which can be estimated by the proportion q_{ab} that a is preferred over b.

2.2.1 Thurstone's Law of Comparative Judgement

In Thurstone's model [57], the error terms ϵ_a are drawn from a normal distribution $N(0, \sigma^2)$ with expectation 0 and variance σ^2. Here we consider Thurstone's Case V model, where the variances for all stimuli are assumed to be equal. The difference $\epsilon_b - \epsilon_a$ is also normally distributed with expectation 0 and variance $2\sigma^2$ and thus

$$\begin{aligned} p_{ab} &= Pr[u_a \geq u_b] = Pr[\epsilon_b - \epsilon_a \leq v_a - v_b] \\ &= \frac{1}{\sqrt{4\pi\sigma^2}} \int_{-\infty}^{v_a - v_b} e^{-\frac{x^2}{4\sigma^2}} dx = \Phi\left(\frac{v_a - v_b}{\sqrt{2}\sigma}\right) \quad , \end{aligned} \tag{2.3}$$

where Φ is the cumulative distribution function of the standard normal distribution

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \quad . \tag{2.4}$$

This is equivalent to

$$v_a - v_b = \sqrt{2}\sigma\Phi^{-1}(p_{ab}) \quad . \tag{2.5}$$

Using the proportion q_{ab} that a is preferred over b, we set z_a as an approximation for $v_a - v_b$

$$z_{ab} = \sqrt{2}\sigma\Phi^{-1}(q_{ab}) \quad [Z - \text{matrix}] \quad . \tag{2.6}$$

[1] We introduced the bias correction δ in order to eliminate numerical problems for pairs of items, which have zero entries in the frequency matrix. In the data analysis of this thesis we used $\delta = 0.2$. For a discussion of different bias correction formulas see also Engeldrum [19].

2.2.2 Bradley-Terry's model

In Bradley-Terry's model [5] the error terms ϵ_a are drawn from a standard Gumbel distribution, i.e., the Gumbel distribution with location parameter $\mu = 0$ and scale parameter $\beta = 1$. Since the difference of two independent Gumbel distributed random variables is logistically distributed with mean 0 and standard deviation $\sqrt{2}\beta$, we have

$$\begin{aligned} p_{ab} &= Pr[u_a \geq u_b] = Pr[\epsilon_b - \epsilon_a \leq v_a - v_b] \\ &= \frac{1}{1 + e^{-\frac{(v_a - v_b)}{\sqrt{2}\beta}}} = \frac{e^{\frac{v_a - v_b}{\sqrt{2}\beta}}}{1 + e^{\frac{v_a - v_b}{\sqrt{2}\beta}}} = \frac{e^{\frac{v_a}{\sqrt{2}\beta}}}{e^{\frac{v_a}{\sqrt{2}\beta}} + e^{\frac{v_b}{\sqrt{2}\beta}}}. \end{aligned} \tag{2.7}$$

This implies

$$\frac{e^{\frac{v_a}{\sqrt{2}\beta}}}{e^{\frac{v_b}{\sqrt{2}\beta}}} = \frac{p_{ab}}{1 - p_{ab}}, \tag{2.8}$$

which is equivalent to

$$v_a - v_b = \sqrt{2}\beta \ln\left(\frac{p_{ab}}{1 - p_{ab}}\right). \tag{2.9}$$

As we did for Thurstone's model we set

$$z_{ab} = \sqrt{2}\beta \ln\left(\frac{q_{ab}}{1 - q_{ab}}\right) \quad [Z - \text{matrix}]. \tag{2.10}$$

2.2.3 Computing scale values using least squares

From either Thurstone's or Bradley-Terry's model we get an estimate z_{ab} for the difference of the scale values v_a and v_b. We compute the v_a's as the least square approximation for the z_{ab}'s (all equally weighted) in a least squares sense. Let $\hat{v}_a$ be the approximation of v_a. We want to minimize the residual

$$r(v_a | a \in A) = \sum_{a,b \in A; b \neq a}^{n} (\hat{v}_a - \hat{v}_b - z_{ab})^2. \tag{2.11}$$

A necessary condition for the minimum of the residual is that all partial derivatives vanish, which gives

$$\frac{\partial r}{\partial v_a} = 2 \sum_{b \in A; b \neq a} (\hat{v}_a - \hat{v}_b - z_{ab}) = 0. \tag{2.12}$$

Hence

$$n\hat{v}_a = \sum_{b \in A} \hat{v}_b + \sum_{b \in A; b \neq a} z_{ab}. \tag{2.13}$$

If we normalize by setting

$$\sum_{b \in A} \hat{v}_b = 0, \tag{2.14}$$

then the values that minimize the residual are given as

$$\hat{v}_a = \frac{1}{n} \sum_{b \in A; b \neq a} z_{ab}. \tag{2.15}$$

2.2.4 Generalized evaluation using least squares

We assume that the entries of the Z-matrix z_{ab} are noisy measurements of the true differences $v_a - v_b$. If we assume that the noise term ϵ is independent of the pair (a, b), then we have

$$Z = X\hat{v} + \epsilon, \tag{2.16}$$

where X is the $\binom{n}{2} \times n$-matrix (indexed by the pairs/elements in A) that has the entries

$$x_{ab,i} = \begin{cases} 1 & i = a \\ -1 & i = b \\ 0 & i \neq a, b \end{cases} \tag{2.17}$$

If we further assume that ϵ is normally distributed with mean 0 and variance σ^2, then the likelihood function for the scale values

$$L(\hat{v}) = L(\hat{v}; X, Z) = p[Z|X; \hat{v}], \tag{2.18}$$

has the form

$$L(\hat{v}) = \prod_{(a,b)\in\binom{A}{2}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(z_{ab} - \hat{v}^T x_{ab,.})^2}{2\sigma^2}\right). \tag{2.19}$$

This function (or the logarithm of it) is maximized at $\hat{v}$ that minimizes

$$\sum_{b\in A; b\neq a} (\hat{v}_a - \hat{v}_b - z_{ab})^2, \tag{2.20}$$

i.e., the residual function that we maximized in the last subsection (eq. 2.11) under the additional scaling constraint $\sum_{b\in A} \hat{v}_b = 0$.

The assumption that the variance σ^2 of the error term is independent of the pair (a, b) may not be realistic in particular if not all pairs have been compared equally often. We can actually estimate this variance σ_{ab}^2 for every pair (a, b) using error analysis (compare Section 2.5). We can normalize the least squares approach to accommodate the difference in variance by multiplying with the following $\binom{n}{2} \times \binom{n}{2}$ diagonal matrix N

$$N = \mathrm{diag}\left(\frac{1}{\sigma_{ab}} \middle| (a, b) \in \binom{A}{2}\right). \tag{2.21}$$

That is, we assume

$$NZ = NX\hat{v} + \epsilon, \tag{2.22}$$

where ϵ is normally distributed with mean 0 and variance σ^2 (independent of the pair (a, b)). The maximum (log)likelihood solution is then given as

$$\hat{v} = ((NX)^T NX)^{-1}(NX)^T Z. \tag{2.23}$$

Again we can translate the scale by setting $\hat{v}_a := \hat{v}_a - \frac{1}{n}\sum_{b\in A} \hat{v}_b$ resulting $\sum_{b\in A} \hat{v}_b = 0$.

In order to enforce our constraint $\sum_{b\in A} \hat{v}_b = 0$ we add an additional equation with our constraint. We define $\bar{V}$, $\bar{X}$, $\bar{Z}$ as follows:

$$\bar{Z} = [Z, 0]^T \tag{2.24}$$

$$\bar{X} = [X, [1, 1, \ldots 1]]^T \tag{2.25}$$

$$\bar{N} = [N, \frac{1}{\sigma_c}]^T \tag{2.26}$$

where σ_c defines the weight of that equation and can be interpreted as the uncertainty of the constraint. The resulting equation for $\hat{v}$ is

$$\hat{v} = ((\bar{N}\bar{X})^T \bar{N}\bar{X})^{-1} (\bar{N}\bar{X})^T \bar{Z}. \tag{2.27}$$

Let us notice that the solution for $\hat{v}$ is independent from the choice of σ_c. The strengths of this linear regression method is, that it is also valid for incomplete paired comparisons. If unit weights are used in $\bar{N}$ the solution corresponds to the incomplete matrix solution to Thurstone's Case V, published by Morrissey [52] as well as Gulliksen [31].

2.2.5 Mosteller's Test

Mosteller's test is used to verify the assumption that the scale values are uncorrelated, equally distributed variables (either normally in case of Thurstone's model, or Gumbel distributed in case of Bradley-Terry's model). A description of Mosteller's test can be found in Engeldrum [19] or in the original article by Mosteller [53]. The goal of Mosteller's test is to compare the computed scale values $\hat{v}_a$, or more precisely their differences $\hat{v}_a - \hat{v}_b$ to the observed proportions q_{ab}. We use the respective distribution function, i.e., either Equation 2.5 in case of Thurstone's model, or Equation 2.9 in case of Bradley-Terry's model to compute probabilities $\hat{p}_{ab}$ from the differences $\hat{v}_a - \hat{v}_b$. Then we transform both q_{ab} and $\hat{p}_{ab}$ into angles θ_{ab} and $\hat{\theta}_{ab}$, respectively, using the arcsine transformation given by

$$\theta_{ab} = \sin^{-1}(2q_{ab} - 1) \quad \text{and} \quad \hat{\theta}_{ab} = \sin^{-1}(2\hat{p}_{ab} - 1) \tag{2.28}$$

The arcsine transformation converts binomially distributed frequencies into asymptotically normally distributed variables with variance $1/m_{ab}$, where m_{ab} is the number of choices between stimulus a and stimulus b. Our hypothesis is that θ_{ab} is normally distributed with expectation $\hat{\theta}_{ab}$ and variance $1/m_{ab}$ for all $a < b$. As test statistic we use

$$\chi^2 = \sum_{a<b} m_{ab}(\theta_{ab} - \hat{\theta}_{ab})^2. \tag{2.29}$$

If the hypothesis is true then the test statistic χ^2 is approximately χ^2-distributed with $\binom{n-1}{2}$ degrees of freedom. Thus, at level α we have to compare our test statistic to the $1-\alpha$ quantile of the χ^2-distribution with $\binom{n-1}{2}$ degrees of freedom.

2.3 Conjoint analysis

Conjoint analysis has its origins in mathematical psychology [41] and is typically applied in market research to investigate customers' preferences. It can be used to evaluate preferences on any structured set of items in a psycho-visual test.

2.3.1 Conjoint structure

In the previous part of this chapter we considered a set of unstructured stimuli. Now we assume that the set A of stimuli is structured, namely we assume it is a parameterized domain. We call a domain parameterized, if it is given as a Cartesian product $A = A_1 \times \ldots \times A_m$ of parameter sets $A_1, \ldots, A_m$. Every element a of A is a vector $a = (a_1, \ldots, a_m)$, where $a_k \in A_k$. The elements a_k are called parameter levels [2], m is the number of parameters.

[2] For example let considered items be different yogurts. Then possible set of parameters can be as follows:

One goal of conjoint analysis is to determine how much every parameter level contributes to the observed outcome of a (preferential) choice measurement—this contribution is called the part-worth of the parameter level. As for the discrete choice models we assume that we have a set of choice data on the stimulus set A. We want to estimate the part-worth of all the parameter values from the choice data.

Giesen [26] extended the least squares approach described for discrete choice models to the multi-parameter (conjoint) case. The extension entails to apply the least squares method for each parameter using Thurstone's model, which provides an initial set of part-worths. Part-worths are computed independently for each parameter, hence rescaling of these values is needed to make the scales of the different parameters comparable. The overall value of a stimulus in A is then obtained by summing up the re-scaled part-worths of the parameter levels present in the object. In the following, we first provide more details on the re-scaling approach.

2.3.2 Re-scaling for Thurstone's model

Note that Thurstone's model has a free parameter σ that we (without loss of generality) set to 1. Now, let $v_{k_1}, \dots v_{k_{n_k}}$ be the scale values computed by using Thurstone's model for every parameter A_k with levels $a_{k_1}, \dots a_{k_{n_k}}$. To get the scale value of a stimulus in A we sum up all the scale values (part-worth) of the parameter levels present in the stimulus, i.e., we assume a linear model. But for the linear model to be meaningful, the scale values for different parameters need to be on comparable scales.

To make the scales for different parameters comparable we normalize them by the following normalization procedure: for any set of parameter's levels A_k the scale values $v_{k_1}, \dots v_{k_{n_k}}$ are normally distributed with variance σ_{k1}^2 and expected value 0. The σ_{k1}^2 values are themselves drawn from another normal distribution with expected value 0 and variance σ_{k2}^2. Hence, quality values for the levels of parameter A_k are drawn from a normal distribution N_k with variance $\sigma_{k1}^2 + \sigma_{k2}^2$ and expected value 0 (as the convolution of two normal distributions with expected value 0 and variances σ_{k1}^2 and σ_{k2}^2, respectively). The value σ_{k1}^2 is the same for all $a_{k_a} \in A_k$ and will be chosen such that the quality values for different parameters are comparable. We assume, that N_k are the same for all $k = 1 \dots m$. This assumption is quite reasonable since the scale values for the different parameters are all computed from the same database.

Hence, the value $\sigma_{k1}^2 + \sigma_{k2}^2$ is independent of $k = 1, \dots, m$. Now we want to find a re-scale value λ_k^2, such that $\lambda_k^2(\sigma_{k1}^2 + \sigma_{k2}^2) = \text{const}$, where the constant does not depend on k. Without loss of generality we can set the constant to 1 which gives $\lambda_k^2(\sigma_{k1}^2 + \sigma_{k2}^2) = 1$ for $k = 1, \dots, m$, or $\lambda_k^2 + \lambda_k^2\sigma_{k1}^2 = 1$ if we assume $\sigma_{k2}^2 = 1$. (As mentioned before, for the computation on the parameter level we had set the value of σ_{k2}^2 to 1.) In the following we fix the parameter k and drop it from the index. We can estimate σ_1 from the scale values computed from Thurstone's model scaled by λ, i.e., λv_a, where v_a is the scale value of level a, as follows[3]:

$$2\sigma_1^2 = \lambda \frac{\sum_{a=1}^{n}\sum_{b=1}^{n}(v_a - v_b)^2 f_{ab}}{2\sum_{a=1}^{n}\sum_{b=1}^{n} f_{ab}}, \tag{2.30}$$

- SIZE= {0.175l, 0.2l, 0.25l}
- PACKAGE COLOR= {white, green, blue}
- PRICE = {1\$, 1.2\$, 1.5\$}
- TASTE= {Natural, Vanilla, Strawberries}

The profile of the yogurt is then a vector in SIZE × PACKAGE COLOR × PRICE × TASTE, for example (0.175l, white, 1.5\$, Natural).

[3] Note that $u_a - u_b = v_a - v_b + \epsilon_a - \epsilon_b$, where $\epsilon_a - \epsilon_b$ is normally distributed with zero mean and variance $2\sigma_2^2$, and also note that we compute the v_a setting $\sigma_2^2 = 1$ normalizing by $\frac{1}{n}\sum_{a=1}^{n} v_a = 0$.

where f_{ab} is the frequency the a'th level is preferred over the b'th level. Plugging the resulting estimate for σ_1^2 into $\lambda^2 + \sigma_1^2\lambda^2 = 1$, we get:

$$\lambda = \frac{1}{\sqrt{1 + \frac{\sum_{a=1}^{n}\sum_{b=1}^{n}(v_a - v_b)^2 f_{ab}}{2\sum_{a=1}^{n}\sum_{b=1}^{n} f_{ab}}}} \tag{2.31}$$

Now for the fixed parameter k we re-scale the scale values $v_{k_1}, \ldots, v_{k_{n_k}}$ by the value of λ_k as defined above. The normalized scale values of the parameter levels are our part-worths that we assume to contribute linearly to the quality of a stimulus, i.e., the scale value of a stimulus $(a_1, \ldots a_m) \in A$ is $\sum_{k=1}^{m} \lambda_k v_k$ which is the sum of the part-worths of the parameter levels present in the stimulus.

2.3.3 Linearity assumption in conjoint analysis

Here we describe how to test the linearity assumption that we make for conjoint analysis. Let A_1 and A_2 be two parameters, let $C = A_1 \times A_2$ be the parameter that results from combining A_1 and A_2, and let $c_1, \ldots, c_k$ be its levels. We compute scale values for the levels of C in two different ways. First, for every level $c_a = (a_{i1}, a_{i2})$ with $a_{i1} \in A_1$ and $a_{i2} \in A_2$ we add up the comparable scale values for a_{i1} and a_{i2} that we compute as described before. Let $v_1, \ldots, v_k$ be the resulting scale values. Second, we apply Thurstone's method directly to the combined parameter C and make the resulting scale values comparable with the scale values of all levels of the parameters different from A_1 and A_2. This results in scale values $v'_1, \ldots, v'_k$. If additivity holds, then we expect that $v_a \approx v'_a$. Thus, our hypothesis is that $v_a = v'_a$ for all $1 \leq a \leq k$. As a test statistic we use

$$\chi^2 = \sum_{a=1}^{k} \frac{(v_a - v'_a)^2}{\sigma_a^2 + \sigma_a'^2}, \tag{2.32}$$

where σ_a and σ'_a are computed by error propagation from the errors of the observed frequencies. If the hypothesis is true then the test statistic χ^2 is approximately χ^2-distributed with $k-1$ degrees of freedom. The hypothesis is rejected at a significance level of α if $\chi^2 > \chi^2_{1-\alpha,k-1}$ where $\chi^2_{1-\alpha,k-1}$ is the $1-\alpha$ quantile of the χ^2-distribution with $k-1$ degrees of freedom.

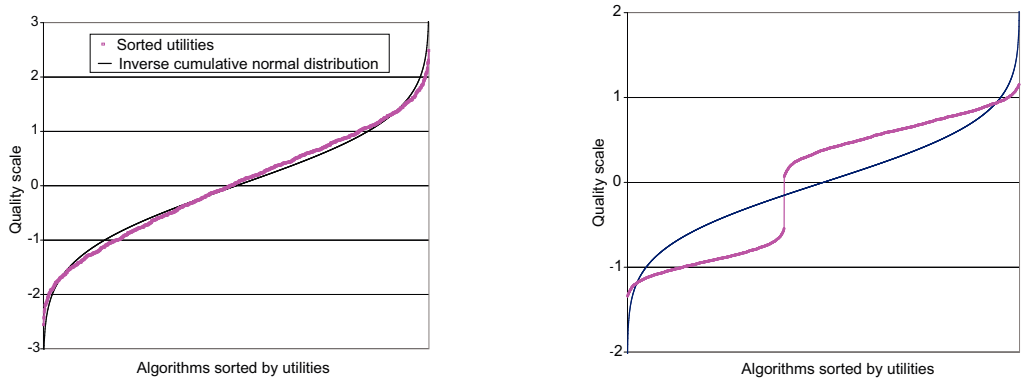
2.3.4 Distribution of scale values in conjoint analysis

For our approach towards conjoint analysis it is important, that there is no dominant parameter. In this case the assumption that the scale values of the whole set of items follow the normal distribution would not hold. Hence, our re-scaling method would not be eligible. One possible way of testing if there is no determining parameter is looking at the ordered scale values. If there is no dominant parameter the curve should resemble inverse cumulative distribution function of the normal distribution. In other case we would see separate parts of the curve, compare Figure 2.2.

2.4 Individualized models

We test gamut mapping algorithms in a psycho-visual test by evaluating different images mapped using these algorithms. The obtained data can be evaluated for the whole groups of images or for subgroups or even individual images. The same concerns observers –we can evaluate the whole group together, split them into smaller groups or even evaluate individual persons separately.

We can naturally evaluate any subset of the data in the same way as the whole data set. However, splitting the data implies reducing the available number of comparisons. To alleviate this issue,



a b

Fig. 2.2: a) An example of the possible distribution of scale values without a dominating parameter b) An example of the possible distribution of scale values with a dominating parameter.

we can use a linear combination of group-wise (sv_{ind}) and general scale value (sv_{gen}).

$$\mathrm{sv} = \alpha \cdot \mathrm{sv}_{gen} + (1 - \alpha) \cdot \mathrm{sv}_{ind} \tag{2.33}$$

The α is chosen to maximize the hit rate (compare Section 2.6) on the test set.

2.5 Computing errors for discrete choice models

A good error estimation is needed in order to gauge the statistical significance of differences between scale values. Even though a thorough treatment of the involved statistical errors is often neglected. Morovic [48] gives a simple formula to estimate confidence intervals. It basically scales with the square root of the number of observations N per pair of stimuli but not with the number of stimuli n. This methods has been used in many psycho-visual studies on gamut mapping, though only a few of them cite the used formula explicitly [32]. The CIE-guidelines [12] only give a reference to Morovic's thesis [48] concerning confidence intervals for paired comparisons.

In a recent paper, Montag [46] has investigated the dependency on the number of stimuli n using Monte Carlo simulations and has derived an empirical formula where the estimated error scale with the square root of the product of N and n. Standard books on psycho-visual scaling [2; 14] give a more detailed description for error estimations, however their results are tied to their specific data analysis and a direct application to the standard evaluation of Thurstone's Case V case is not obvious.

Here we will give a direct derivation of an error estimation for Thurstone's Case V and its rescaling for conjoint analysis. It is based on error propagation. We will use Monte Carlo simulations to compare the results with the other commonly used methods and to find the region of applicability as a function of the number of observations N, the number of stimuli n, and the z-scale value range.

2.5.1 Morovic's error estimation

Morovic [48, Chapter 4] gives the following formula to estimate the 95 percent confidence interval:

$$CI_S = 1.96 \frac{\sigma}{\sqrt{N}}. \tag{2.34}$$

With $\sigma = 1/\sqrt{2}$ we can compute the underlying error estimate for the scale values:

$$E_m = \frac{1}{\sqrt{2N}} \tag{2.35}$$

2.5.2 Montag's error estimation

Montag [46] has published an empirical formula based on Monte Carlo simulations. It shows the expected approximate dependency of the estimated error with the square root of the product of N and n.

$$E_e = b_1(n - b_2)^{b_3}(N - b_4)^{b_5} \tag{2.36}$$

with $b_1 = 1.76$, $b_2 = -3.08$, $b_3 = -0.613$, $b_4 = 2.55$ and $b_5 = -0.491$.

2.5.3 Analytic error estimation

Here we derive an analytic error estimate for Thurstone's method. The basic approach is to estimate the error in the image choice process and then propagating the error through the data evaluation steps: this process of choosing one image from the pair of images can be modeled as a Bernoulli trial with success probability p_{ab}. The standard deviation for p_{ab} equals the standard deviation for a Bernoulli variable in N trials

$$\sigma_{p_{ab}} = \sqrt{\frac{p_{ab}(1 - p_{ab})}{N}} \tag{2.37}$$

As we approximate p_{ab} by the empirical value q_{ab} the estimated error $E_{q_{ab}}$ for the proportion q_{ab} can be written as

$$E_{q_{ab}} = \sigma_{q_{ab}} = \sqrt{\frac{q_{ab}(1 - q_{ab})}{f_{ab} + f_{ba} + 2\delta}}. \tag{2.38}$$

To compute the errors of the entries z_{ab} in the Z matrix, we propagate the error using equation (2.6)

$$E_{z_{ab}} = E_{q_{ab}}\sqrt{2}\sigma\frac{d}{dq_{ab}}\Phi^{-1}(q_{ab}). \tag{2.39}$$

Using equation (2.15) the errors of the scale values v_a are computed as

$$E_{v_a} = \frac{1}{n}\sqrt{\sum_{b \in A; a \neq b} E_{z_{ab}}^2}. \tag{2.40}$$

2.5.4 Approximation of errors

An approximate error estimate can be derived for Thurstone's Case V if the probabilities p_{ab} are not far from 1/2. Then their standard deviation is

$$E_{q_{ab}} \approx const = \sqrt{\frac{1}{4N}} \tag{2.41}$$

and the error of the Z-matrix elements z_{ab} can be approximated as

$$E_z \approx \sqrt{2}\sigma\sqrt{\frac{1}{4N}}\frac{d}{dq}\Phi^{-1}(q = 0.5) = \sqrt{2}\sigma\sqrt{\frac{\pi}{2N}} \tag{2.42}$$

independent of a and b. Assuming $\sigma = 1/\sqrt{2}$ the error for the scale values v_a is approximatively

$$E_v \approx \frac{1}{n}\sqrt{\frac{\pi(n-1)}{N}} \tag{2.43}$$

independent of a.

2.5.5 Experimental error estimation

Experimental error estimation is based on minimal assumptions. It samples the error by dividing the choice data randomly into two groups. For both groups, scale values are individually computed and errors are estimated from the differences of the values obtained from both groups. This process is repeated several times and the results are averaged to increase the accuracy of the error estimation. If all model assumption are fulfilled, this error estimation should deliver the same values as obtained from analytic error estimation.

A special option of this method is to test for heterogeneity among the observers (or among the images). In this case data from individual observers (or individual images) are randomly divided into two groups and errors are computed from the average difference of the scale values between the two groups. Hence error estimation using such biased samplings allows us to test whether the choices depend on individual observers (or images).

2.5.6 Simulation

We used Monte Carlo simulation in order to compare the different error estimates and to investigate their validity as a function of the number of observations N, the number of stimuli n and the scale value range. For all simulations, we assumed a psychological continuous scale that conforms to Thurstone's Case V, i.e., the discriminal differences follow a Gaussian distribution of equal width and no correlation exists between two stimuli a and b. Furthermore, we assumed no correlation in the responses of an individual observer nor in responses for an individual image. Thus we assumed ideal conditions for the simulated experiments.

Simulation for small scale values

In the first experiments we used only stimuli with small scale value differences compared to the width of the distribution of the discriminal process. We used the following number of stimuli $n = [4, 7, 10, 15]$ and the number N of observations per pair of stimuli was in the range of [10...60]. The n scale values were assumed to be uniformly distributed in the range $[-0.25... + 0.25]$. Every experiment was repeated 10'000 times. The simulated error E_s of the scale values is calculated using the standard deviation of the scale values from the experiments. Two different values for the bias correction δ, $\delta = [0.1, 0.5]$, were used in the simulation.

In Figure 2.3, the experimental error is compared to the different error estimates E_m (eq. 2.35), E_e (eq. 2.36) and E_v (eq 2.43). The error E_m generally overestimates the simulated error. The overestimation increases with the number n of stimuli. The error estimates E_e and E_v are in good agreement with the simulated error for all investigated combinations of n and N. The differences between E_e and E_v get smaller with higher N. For small N, the differences are mainly due to the use of different bias correction values δ. Simulated errors using a bias correction $\delta = 0.1$ follow Montag's error estimation E_e, while a use of $\delta = 0.5$ is in better agreement with E_v.

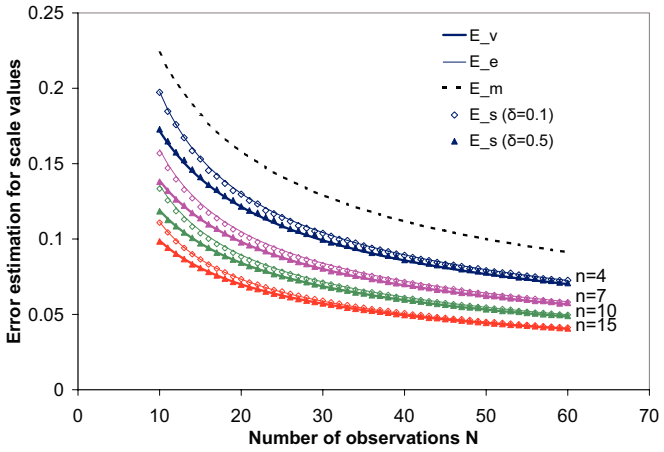


Fig. 2.3: Simulated scale value errors E_S compared to error error estimate E_m (eq. 2.35), E_e (eq. 2.36) and E_v (eq 2.43).

Simulation of individual errors

The estimated error E_{v_a} in eq. 2.40 shows a dependency on the scale value. In a second simulation, this dependency has been investigated. We used an experiment with the following numbers of stimuli $n = [4, 8, 16]$. The n scale values were assumed to be uniformly distributed in the range $[-1.0... + 1.0]$. The number of observation was fixed to a rather large number $N = 200$ to avoid a significant influence of the bias offset. The results are shown in Figure 2.4. The simulated error E_s is increasing with higher scale values. This increase is nicely reproduced by the error estimation E_{v_a} given in eq. 2.40. The error approximation E_v gives a lower limit for the simulated error. The same is true for E_e which for this high number N is basically identical to E_v.

2.5.7 Accuracy of error estimation

A third series of simulation experiments has been performed to test the accuracy of the four error estimates E_m, E_e, E_v and E_{v_a} as a function of N, n and the scale value range. For this purpose, we investigate the relative accuracy in percentage of the error compared to the simulated error. Within one specific experiment, the maximum relative deviation of an estimated error from the simulated error was taken as a measure for the accuracy of an estimated error. 40 different scale value ranges were used up to $[-2.0 \ldots 2.0]$, N was in the range $[2 \ldots 100]$ in steps of 2 and the number of stimuli n was in $[3; 4; 5; 8; 12; 16]$. For every experiment, scale values for n stimuli were selected as follows: n values x_a were randomly chosen in the range $[-1.0 \ldots +1.0]$. Then, these values were scaled such that the difference between the minimum and maximum scale corresponded to the target scale value range. For every combination of N, n and scale value ranges, the experiment was repeated at least 2000 times and simulated errors E_s, average scale value errors E_{v_a}, approximate errors E_m and E_e were calculated. In Figure 2.1, we show the results for 3, 5 and 8 stimuli. The error estimation E_m is accurate in a range of 20% of simulated error only for $n = 3$. It is not accurate for larger number of stimuli. This confirms, that the error estimation E_m should, if at all, be used only for a small number of stimuli, i.e., smaller than five. The error estimations E_e and E_v have regions with high accuracy for all numbers of stimuli, but only for small and moderate scale values up to about 1.0. For these scale value ranges, the error estimates for all scale values are

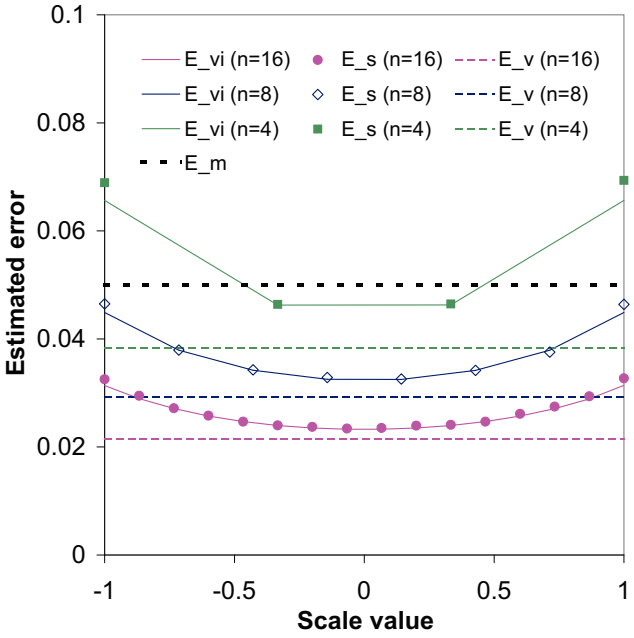


Fig. 2.4: Simulated scale value errors as a function of z-scale compared to error estimations.

approximately equal and E_e (as well as E_v) gives a simple, quick and accurate error estimate. The best error estimate is given by E_{v_a}. The accuracy is better than 10% for all numbers of stimuli n and numbers of observations N up to a limiting scale value range. The limiting scale value range basically scales with the square root of N. The limit is reached when one or more expected entries in the frequency matrix are close to or smaller than one. Interestingly the accuracy of the error estimations E_e, E_{v_a} and E_v depends only marginally on the number of stimuli n.

The simple error estimation given by Morovic [48] generally overestimates the error and is approximately correct only for small numbers of stimuli ($n \approx 3$). It is not suitable as a general error estimation method.

In many cases, the approximative error E_v as well as the empirical error estimation E_e given by Montag [46] are sufficiently accurate. The advantage of deriving the error approximation analytically is, that an adaption to other discriminal distributions such as the logistic distribution used by Bradley-Terry [5] is straight forward. Only the inverse cumulative distribution function Φ^{-1} and its derivative have to be replaced by the appropriate functions. Furthermore note that an empirical formula is always restricted to the parameter range used in the fitting process. It is not clear, whether the formula given by Montag can be extended to $n < 4$ or to large N.

The limiting factor in the error estimation is the expected number of judgements for the least probable entry in the frequency matrix f_{ab}. This is due to the fact that the entries z_{ab} in the Z-matrix are averaged and the error of the entry with the highest error also has the highest contribution to the error of the scale value v_a. A weighted linear regression method such as described by Bock and Jones [2, Chapter 6] could give more reliable scale values and error estimates for the case of a large scale value range where the proposed error estimation reaches its limit.

In all our simulations, we assumed the ideal Case V of Thurstone's Law of Comparative Judgement. Note that besides scale values, also estimation of their errors is valid only if the underlying

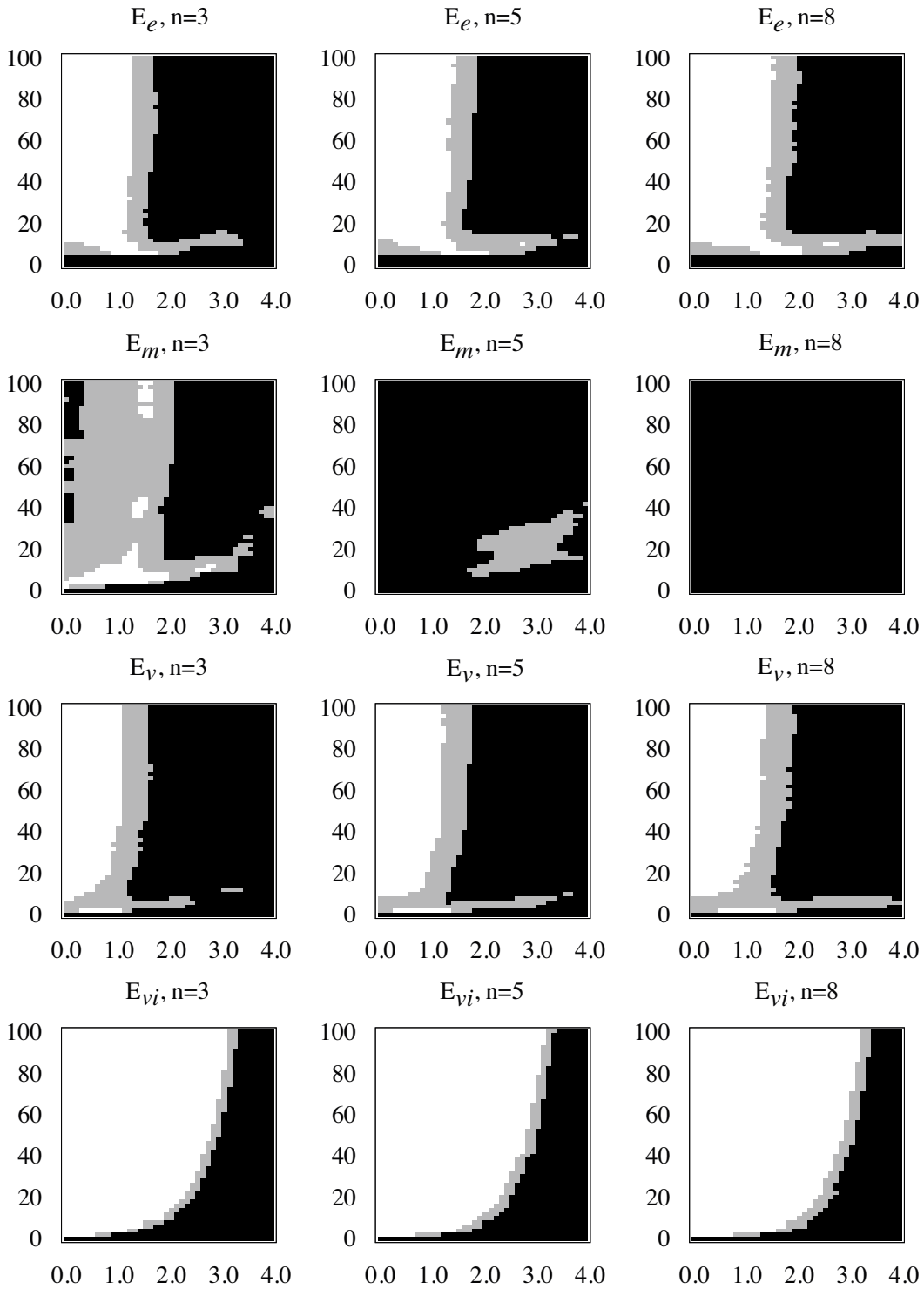


Table 2.1: Map for the validity range of error estimates E_e, E_m, E_v and E_{v_a}. White areas correspond to computed errors within 10% from the real one, gray areas to the computed error between 10% and 20% off from the real error and black areas to computed error more than 20% off.

assumptions hold. If the Mosteller's test [53] fails or if there are indications of other correlations in the data, the error estimation must also be questioned.

2.5.8 Re-scaling errors in conjoint analysis

In conjoint analysis, part-worth values v_k re-scaled by coefficients λ_k for single parameters are summed up resulting in a scale value for a given item. Errors E_{v_k} for those part-worths are scaled by the respective coefficient λ_k. A square root of the squared sum of those errors provides then an error of a scale value for a given item.

2.6 Cross validation

Each model designed for evaluating image quality requires validation. We want to assess how accurately a predictive model will perform. Our validation procedure estimates how well a given model aligns with observers' ratings which we obtained in psycho-visual tests. As mentioned before, the psycho-visual test data are of the form: given an original image and two images obtained by applying different gamut mapping algorithms, a user chooses the one that reproduces the original image better in his/her opinion. We validate a model by the percentage of correctly predicted observer choices. This validation measure is known as hit rate. When computing hit rates for Thurstone's method we need to be careful that we do not validate the method on the same test data that we used to derive the model—remember that Thurstone's method is, in contrast to image quality measures described in Chapter 3, based on observers' data. To circumvent this problem, one uses cross validation, i.e., part of the data is used to derive a model and the remaining part to validate it. In the following, we provide more details on how to compute hit rates and use cross validation.

2.6.1 Hit rate

For each paired comparison in a psycho-visual test we know the choice of the observer. In some tests we allowed ties, i.e., neither of the two options is preferred. We omit such ties from further analysis. Let C be the set of non-tied observer choices. For an image quality measure or scale values we always predict the choice with the higher value for this measure on the elements in C. Let $S \subseteq C$ be the subset of correctly predicted choices, then the hit rate is defined as

$$HR = \frac{|S|}{|C|}, \tag{2.44}$$

where $|S|$ and $|C|$ are numbers of elements in the sets S and C, respectively.

2.6.2 Cross validation algorithm

There are a few slightly different methods of applying cross-validation [33]. Here we will discuss one of them, which we were using in our applications. For that the set C of non-tied observer choices is partitioned randomly into N subsets of equal size. Out of the N subsets, each is once retained for validating the model, and the remaining $N-1$ subsets are used as training data. The whole process is repeated n times. The mean hit rate over all nN validation sets is used as the validation quality measure.

For the individualized variant of Thurstone's method (compare Section 2.4), we carried out a double cross validation, i.e., we use $N-2$ of the N subsets as training set, one as optimization

set, and the remaining one for validation. We compute general and individual scale values by Thurstone's method on the training set. Then we optimize the weights for the linear combination of the population and individualized scale values using the optimizing set. Finally, we use the hit rate on the validation set. We repeat this process n times and use the mean of the hit rates as validation quality measure.

2.7 Equivalence of data sets

Note that a χ^2-test similar to the one used in Mosteller's test can be used to test for significant differences between frequency matrices for different viewing scenarios. Let q_{ab} and q'_{ab} be proportion matrices on different data sets obtained from the same population of observers, m_{ab} and m'_{ab} be the numbers of comparisons between items a and b on different sets. As data sets with different numbers of judgements are compared, their variances have to be adjusted. The difference of independent normal distributions with variances $\frac{1}{m}$ and $\frac{1}{m'}$, respectively, is a normal distribution with variance $\frac{1}{m} + \frac{1}{m'}$. Our hypothesis is that both distributions have the same mean. As test statistic for non-parametrized case we use

$$\chi^2 = \sum_{b<a} \left(\frac{m_{ab} \cdot m'_{ab}}{m_{ab} + m'_{ab}}\right) \cdot \left(\sin^{-1}(2q_{ab} - 1) - \sin^{-1}(2q'_{ab} - 1)\right)^2 \tag{2.45}$$

In the parametrized case, we apply the similar statistic for all P parmeters together.

$$\chi^2 = \sum_{p=1}^{P} \sum_{b<a} \left(\frac{m_{ab} \cdot m'_{ab}}{m_{ab} + m'_{ab}}\right) \cdot \left(\sin^{-1}(2q_{ab} - 1) - \sin^{-1}(2q'_{ab} - 1)\right)^2 \tag{2.46}$$

It should be noted that now the entries of all P proportion matrices are considered together. The number of degrees of freedom is the number of elements in the sum.

Chapter 3

Image quality measures

Models are to be used, not believed.
H. Theil

The psycho-physical methods described in the previous chapter are often used to gather data, from which a model of human preferences of images (perceptual image quality) can be derived. However, these methods are very time-consuming. Furthermore, an extrapolation to changed settings and new images is problematic. Hence it would be helpful if we could approximately predict human preferences on the basis of a mathematical model. The class of such models is usually called image quality measures and the development of such models is an active research field [15].

We can distinguish image quality measures based on different criteria. A typical classification is no-reference and reference models.

Reference (or similarity) measures require an original (reference) image and the goal of the model is to find the difference between the reference and the transformed image. Hence, these models are also called *image similarity measures*, or *image quality metrics*. In this work, we do not label these models as metrics, as some of them do not fulfill the mathematical definition of a metric and the name could be confusing. We will focus here on these models, as the main subject of this thesis is gamut mapping where one usually aims at producing an image as similar to the original as possible and not as good (aesthetic, natural) as possible. Within reference measures, one can also focus on different properties. Typical distortions in gamut mapping are loss of saturation, loss of details and artifacts (halo, contouring). It would be useful to find a measure, which evaluates all those distortions and weighs them similarly as human observers would do, but this is far from easy.

No-reference models try to assess the quality of images on the basis of the considered image only. Within no-reference models we could try to assess general quality, aesthetics or naturalness. One can also gauge some more specific properties, e.g. sharpness, noisiness or some other properties of an image.

In this chapter, we review image quality measures in the context of gamut mapping. Before that, we summarize basics of human vision, color perception and color spaces, which are important to understand the objective evaluation of the distance between images.

3.1 Fundamentals of color science

3.1.1 Human vision

Let us notice first, that color is not a physical parameter but a sensation [54; 55; 59]. If we want to measure color using an objective, mathematical model, we need a quantitative color specification. In order to derive a quantitative color specification, it is necessary to correlate a perceived light stimulus intensity with the magnitude of sensation evoked by it. Some aspects, in particular lightness, follow general principles like Weber-Fechner's law or Steven's power function [62], but the most interesting one, chroma, does not. Additionally, human sensation is strongly dependent on the current viewing conditions. For that reason, it is not surprising that the development of device independent color spaces is a tedious process that cannot be considered completed. However, before going into details let us briefly discuss the retina and its influence on vision and color spaces.

3.1.2 Color spaces

The retina includes several layers of neural cells, especially two kinds of photo-receptors, the rods and cones. The rods contribute mostly to vision at low luminance levels (i.e., less than 1 cd/m^2) while the cones serve vision at higher levels. The cones can be subdivided into three types: L (long-wavelength), M (middle-wavelength), and S (short-wavelength) according to their peak of spectral responses. Their differences in spectral sensitivity are the basis of color vision and can be modeled by RGB-color spaces. But this trichromatic approach of color vision, as described by Young [63], Helmholtz [58], and Maxwell [42], cannot sufficiently explain effects like opponent colors [62, p. 446] and perceived color differences. To understand these effects, the neuronal structure of retina cells that are organized into receptive fields needs to be considered. The input signals from the photo-receptor cells are integrated, concentrated and modified in several neuronal processes and, finally, transmitted to the brain. The resulting output signal contains a luminance, a red-green and a yellow-blue component. A color space that reflects this representation is the CIELAB space.

In the following we are going to describe the aforementioned color spaces in more detail. Basis of their description are quantitative color descriptions that are addressed in colorimetry. Colorimetry is the branch of color science, which quantifies and describes human color perception, see Wyszecki and Stiles [62, p. 117], specifying numerically the colors of physically defined visual stimuli at fixed viewing conditions.

RGB and XYZ color spaces

In 1853, Grassmann [29] showed that empirical color matchings satisfy a mathematical structure introduced by himself some years earlier [28]. This structure is nowadays known as a vector space. In case of color, we observe a three-dimensional vector space, accordingly, colors can be specified as tristimulus values. There is a strong correlation between the mathematical structure and the underlying physics of light stimuli. A vector can be understood as light source, its length as intensity and the addition of vectors as the physical mixture of the corresponding light sources. For fixed viewing conditions, this approach was carefully tested and documented in 1928 by Wright [61] and 1931 by Guild [30], respectively. This resulted in the introduction of the standardized color spaces CIERGB and CIEXYZ [8], for short RGB and XYZ.

Primary colors are the colors of three reference lights by whose additive mixture nearly all other colors may be produced. The primaries of RGB are defined as colors of monochromatic light at wavelength $\lambda = 435.1$ (**B**), 546.1 (**G**) and 700 (**R**) nm. Then, tristimulus values for monochromatic light at a given wavelength λ have been determined and documented as the color matching functions

$\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$. These functions allow the calculation of the tristimulus value (R, G, B) of an arbitrary light stimulus with given spectral power distribution $\Phi(\lambda)$ as

$$R = k \int \Phi(\lambda)\,\bar{r}(\lambda)\ \mathrm{d}\lambda, \quad G = k \int \Phi(\lambda)\,\bar{g}(\lambda)\ \mathrm{d}\lambda, \quad B = k \int \Phi(\lambda)\,\bar{b}(\lambda)\ \mathrm{d}\lambda \tag{3.1}$$

where k means a normalizing constant. Then, the psychophysical color $\mathbf{Q}$ can be represented as follows: $\mathbf{Q} = R\mathbf{R} + G\mathbf{G} + B\mathbf{B}$.

More popular than RGB is XYZ which is mathematically derived from RGB by a change of the vector space base, i.e., by a base transformation matrix:

$$\begin{pmatrix} \bar{x}(\lambda) \\ \bar{y}(\lambda) \\ \bar{z}(\lambda) \end{pmatrix} = 5.6508 \begin{pmatrix} 0.490 & 0.310 & 0.200 \\ 0.177 & 0.812 & 0.011 \\ 0.000 & 0.010 & 0.990 \end{pmatrix} \begin{pmatrix} \bar{r}(\lambda) \\ \bar{g}(\lambda) \\ \bar{b}(\lambda) \end{pmatrix} \tag{3.2}$$

The choice of the XYZ-base seems arbitrary but optimizes some technical constraints, for instance, the Y-coordinate is identical to the CIE spectral luminance efficiency function also known as photopic luminance efficiency function $V(\lambda)$.

Many other well-known RGB-color spaces like sRGB, Adobe 98-RGB, ECI-RGB are also derived from RGB and have to be understood as mathematically equivalent. Contrary to these color spaces, CMYK typically denotes a device dependent specification describing the amount of ink (**c**yan, **m**agenta, **y**ellow and blac**k**) placed in a print raster cell.

The CIEXYZ-system represents the average ability of humans to discriminate colors in a particular viewing conditions, sometimes called standard or normal observer.

CIELAB color space

Unfortunately, the Euclidean distance in XYZ-space does not match with perceived color distance, and thus XYZ is not well suited for gamut mapping. In 1976, two color spaces, CIELUV and CIELAB, have been recommended by the CIE [9] which approximately correlate with the perceived lightness, chroma and hue of a stimulus. Although originally both spaces were recommended, CIELAB is almost universally used today, in particular for reflective color measurements. In CIELAB the psychometric lightness L^* defined as

$$L^* \;=\; L^*(Y) \;\stackrel{\text{def}}{=}\; \begin{cases} 116\,\sqrt[3]{\dfrac{Y}{Y_0}} - 16 & \text{for} \quad 0.008856 \le \dfrac{Y}{Y_0} \le 1 \\ 903.29\,\dfrac{Y}{Y_0} & \text{for} \quad 0 \le \dfrac{Y}{Y_0} \le 0.008856 \end{cases} \tag{3.3}$$

where (X_0, Y_0, Z_0) stands for the reference white. This definition agrees with Stevens power function and also roughly with the Weber-Fechner-law. Then, CIELAB contains the a^* (red-green) and b^* (yellow-blue) coordinates:

$$a^* \;=\; 500\left[f(\frac{X}{X_0}) - f(\frac{Y}{Y_0})\right] \tag{3.4}$$

$$b^* \;=\; 200\left[f(\frac{Y}{Y_0}) - f(\frac{Z}{Z_0})\right] \tag{3.5}$$

where

$$f(w) \;=\; \begin{cases} \sqrt[3]{w} & \text{for} \quad w > 0.008856 \\ 7.787\ w + \frac{16}{116} & \text{otherwise} \end{cases} \tag{3.6}$$

According to the uniformity of CIELAB, color differences are understood as Euclidean distances and denoted as ΔE_{ab}. Over the years, some improvements for color differences have been introduced by the CIE, especially ΔE_{94} [10], and ΔE_{00} [11], that are modifications of ΔE_{ab} along with stricter specifications of the viewing conditions.

A color gamut includes all colors which can be rendered by a device (e.g. a monitor or a printer) or that are contained in a given image. In CIELAB color space we can visualize different gamuts. In particular, we can visualize the difference in source and destination gamut quantitatively. In Figure 3.1 typical printing gamuts are shown as colored objects and compared to a standard sRGB source gamut (as it is typical for a monitor). We can also compute the differences between colors in different gamuts in CIELAB color space.

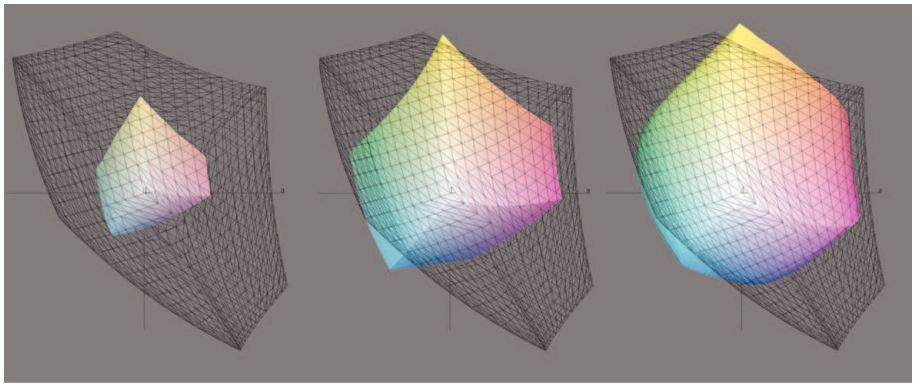


Fig. 3.1: Visualization of typical printer gamuts (colored objects) in comparison to a standard sRGB-gamut (wire frame: newspaper gamut (left), coated offset paper (middle), inkjet printer (right).

3.1.3 Color and image appearance models

Since viewing conditions play an important role, appropriate models are necessary in particular when colors or images are compared under different conditions. Fairchild explains this in his textbook [21, p.1]:

> Color appearance models aim to extend basic colorimetry to the level specifying the perceived color of stimuli in a wide variety of viewing conditions.

A first step towards a color appearance space is the compensation of the reference white in the CIELAB color space. This allows a first order compensation for the adaption of the human viewing system to the lighting condition. However, the CIELAB model lacks compensation of other viewing parameters such as the surround condition. It also has some limitations in its color difference metric. This led to the development of new color appearance models, such as CIECAM97 and CIECAM02 [13]. The CIECAM02 model allows to calculate dependencies for the six technically-defined dimensions of color appearance: brightness, lightness, colorfulness, chroma, saturation, and hue. The model has a set of parameters to compensate for the relevant influence of surround conditions:

- The surround ratio of the absolute luminance of the reference white measured in the surround field to the display area S_R.

- A factor F determining the degree of adaptation of an eye to the lightness
- A parameter c compensating the impact of the surrounding
- The chromatic induction factor N_c.

These parameters are defined for a set of typical viewing conditions: "Average" for viewing surface colors, "Dim" for viewing television, and "Dark" for using a projector in a dark room.

One major shortcoming of color appearance models is that they do not directly account for spatial and temporal properties of human vision. They basically treat each pixel in an image as a completely independent stimulus. Thus a new class of models, named image appearance models have been developed and still are under development. An interesting approach to include spatial aspects of image appearance is Land's retinex model [40]. Retinex models are quite successfully used in computer vision, however, it has been shown that they do not accurately model human color perception [35]. Recently, a more sophisticated model, the iCAM-Framework was proposed by Fairchild and Johnson [22]. The basic idea is to extract image appearance components from four different images:

- A high frequency color image;
- A low frequency color image;
- A low frequency image gray scale;
- A low frequency image within its surrounding.

3.2 Image quality measures overview

There are plenty of image quality measures described in the literature [15]. Different models are suitable for different applications. Here we will concentrate on the measures relevant for gamut mapping, which is a relatively easier task, as the typical problems in gamut mapping are mostly the preservation of color and local contrast. Halo artifacts or continuity artifacts can also be a problem. But noise or compression artifacts are hardly created during the gamut mapping process.

In this section, we review image quality measures that can be useful for comparing images obtained in gamut mapping. Most of the measures described here are one-dimensional and designed for grayscale images. However, they are also applicable for color images, either by choosing one color coordinate or by computing the given image quality measure for all coordinates and then combining them. When deciding for only one coordinate, it is probably best to choose lightness.

Here we always compare two images X and Y with $n \times m$ pixels. At the pixels $x_{ij} \in X$ and $y_{ij} \in Y$, respectively, we consider color coordinates. Mostly we are using the lightness coordinate L in CIELAB color space. If not stated otherwise, we do not distinguish in our notation between a pixel and the color coordinate considered at this pixel. Here, we divide reference image quality measures into two groups: pointwise measures and structural measures.

3.2.1 Pointwise measures

Pointwise measures are based on the differences of pixel colors from two images at the same location. These distance at corresponding pixels are averaged or summed up resulting in the distance between two images. The main advantage of pointwise measures is that they are easy and fast to compute. The disadvantage of these measures is that they do not include the structure of the images. An example showing imperfections of an exemplary pointwise measure is shown in Figure 3.2.

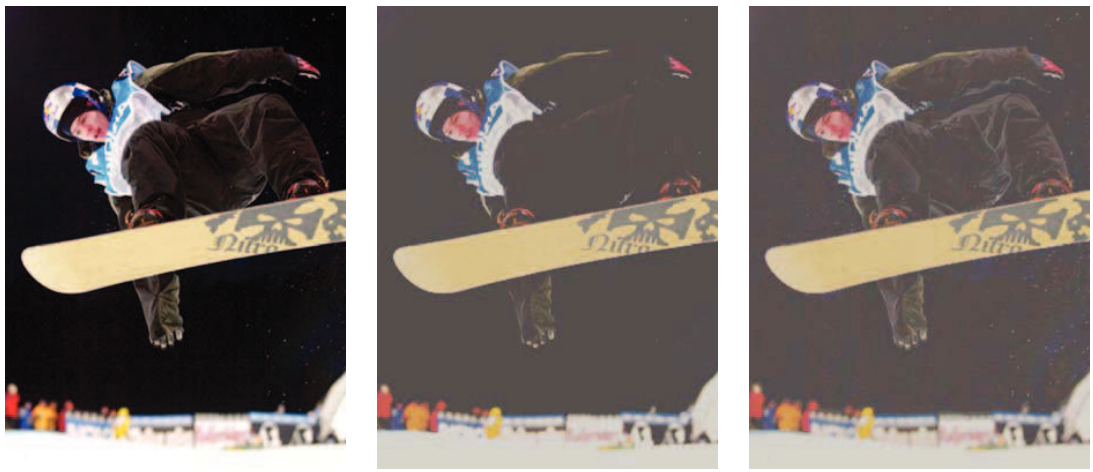


Fig. 3.2: The original image (on the left) and two gamut mapped images (in the middle and on the right). For the image in middle we have a pointwise measure $(Q_{\Delta E}) = 24.65$ and a structural measure $(Q_{\Delta LC}) = 0.341$ using HPminDE algorithm [12] without detail enhancement, and on the right we have the pointwise measure $(Q_{\Delta E}) = 27.00$ and the structural $(Q_{\Delta LC}) = 0.318$ using HPminDE with details enhancement. For the image in the middle $Q_{\Delta E}$ is smaller than for the image on the right, but the middle image has lost a lot of details and has the larger perceptual distance from the original (left image).

Mean Square Error (MSE)

The mean square error is the squared pointwise difference between the images X and Y. The corresponding image quality measure Q_{MSE} is defined as

$$Q_{\text{MSE}}(X,Y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-y_{ij})^2, \tag{3.7}$$

where x_{ij} and y_{ij} are L coordinates in the CIELAB color space for the points in images X and Y respectively.

Euclidean distance in CIELAB color space (ΔE)

ΔE is another pointwise distance measure, similar to MSE. The difference is, that it is computed in 3-D color space instead of in a one-dimensional projection. It is defined as the Euclidean distance in CIELAB color space between corresponding pixels in two images. That is, locally at a pixel $x \in X$ and the corresponding pixel $y \in Y$ the ΔE distance is defined as:

$$\Delta E(x,y) = \sqrt{(L_x - L_y)^2 + (a_x - a_y)^2 + (b_x - b_y)^2} \tag{3.8}$$

As our image quality measure $Q_{\Delta E}$ we take the average ΔE over the pixels of the two images, i.e.,

$$Q_{\Delta E}(X,Y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\Delta E(x_{ij},y_{ij}). \tag{3.9}$$

ΔE is a popular image quality measure since it is easy to compute and has a natural interpretation. In principle it could be replaced by any of the more sophisticated color distance measures such as CIECAM02 [13; 47] or ΔE_{94} [10; 43].

3.2.2 Structural measures

Structural measures consider not only differences at the level of single points, but take the neighborhood of the points in account. The size of the neighborhood is often a free parameter of those measures, as different sizes are optimal for different settings, e.g. for different viewing distances. With the use of structural measures, it is possible to assess preservation of details. In Figure 3.2 one can see an example, where a structural measure (ΔLC) correlates better with perceived distances than a pointwise measure.

Next we describe a few approaches from research on structural image quality measures.

Laplacian Mean Square Error (LMSE)

The Laplacian Mean Square Error [20] is a local measure for the difference in two images. We compute the following quantities at each pixel (with indices $2 \leq i \leq n-1$ and $2 \leq j \leq m-1$) of X and Y, respectively:

$$\begin{aligned} L(x_{ij}) &= x_{(i+1)j} + x_{(i-1)j} + x_{i(j+1)} + x_{i(j-1)} - 4x_{ij} \\ \text{and} & \\ L(y_{ij}) &= y_{(i+1)j} + y_{(i-1)j} + y_{i(j+1)} + y_{i(j-1)} - 4y_{ij} \end{aligned} \tag{3.10}$$

The image quality measure Q_{LMSE} is then defined as

$$Q_{\mathrm{LMSE}}(X,Y) = \frac{1}{(n-2)(m-2)} \sum_{i=2}^{n-1} \sum_{j=2}^{m-1} \left(L(x_{ij}) - L(y_{ij})\right)^2 . \tag{3.11}$$

Structural similarity index (SSIM)

The *S*tructural *SIM*ilarity index (SSIM) was introduced by Wang et. al. [60] and is defined on quadratic image patches of size $k \times k$ at the same location within the images X and Y. Let $P_X \subset X$ be such a patch and P_Y the corresponding patch for Y. We compute the following quantities for the patches:

$$\begin{aligned} \bar{P}_X &= \frac{1}{k^2} \sum_{x \in P_X} x, \quad \bar{P}_Y = \frac{1}{k^2} \sum_{y \in P_Y} y, \\ {\sigma_{P_X}}^2 &= \frac{1}{k^2 - 1} \sum_{x \in P_X} \left(x - \bar{P}_X\right)^2 , \\ {\sigma_{P_Y}}^2 &= \frac{1}{k^2 - 1} \sum_{y \in P_Y} \left(y - \bar{P}_Y\right)^2 , \\ \sigma_{P_X P_Y} &= \frac{1}{k^2 - 1} \sum_{i=1}^{k^2} \left(x_i - \bar{P}_X\right)\left(y_i - \bar{P}_Y\right) . \end{aligned} \tag{3.12}$$

The structural similarity index is then defined as

$$\mathrm{SSIM}(P_X, P_Y) = \frac{\left(2\bar{P}_X \bar{P}_Y + c_1\right)\left(2\sigma_{P_X P_Y} + c_2\right)}{\left(\bar{P}_X^2 + \bar{P}_Y^2 + c_1\right)\left(\sigma_{P_X}^2 + \sigma_{P_Y}^2 + c_2\right)}, \tag{3.13}$$

with two constants c_1 and c_2 depending on resolution of the image and quantization.

Using the structural similarity index an image quality measure $Q_{\mathrm{SSIM}}(X, Y)$ can be defined as the structural similarity index averaged over all possible $k \times k$ patches in the images X and Y. The resulting value is in the range $[-1, 1]$, and the higher the Q_{SSIM} value, the more similar are the compared images.

Discrete wavelet transform (DWT)

The discrete wavelet transform image quality measure has been defined by Gayle [25]. Images X and Y are compared as follows: a discrete wavelet transform is applied to the luminance layer of image X and Y, respectively. Let M_X^f be the magnitudes of the discrete wavelet transform coefficients obtained for X and frequency band f, and let M_Y^f be the corresponding magnitudes for image Y. From M_X^f and M_Y^f the absolute values of differences

$$d_i^f(X, Y) = \left| M_{Xi}^f - M_{Yi}^f \right|, \quad i = 1, \ldots, \left| M_X^f \right| = \left| M_Y^f \right|. \tag{3.14}$$

are computed for each frequency band. Let $\sigma_f(X, Y)$ be the standard deviation of the differences $d_i^f(X, Y)$ for frequency band f. Now, the $Q_{\mathrm{DWT}}(X, Y)$ image quality measure is defined as the mean of the $\sigma_f(X, Y)$ for all the frequency bands.

Q_{DWT}.

Difference in local contrast (ΔLC)

The image quality measure $Q_{\Delta LC}$ is based on a local contrast measure. Here the Michelson contrast [44] is used as a measure of local contrast. We compute it on a $k \times k$ patch $P_X \subset X$ of the image X as follows:

$$LC(P_X) = \frac{x_{max} - x_{min}}{x_{max} + x_{min}}, \tag{3.15}$$

where x is a luminance coordinate in XYZ color space (at pixel $x \in P_X$), and x_{max} and x_{min} are the highest value and the lowest value, respectively, of this intensity on the patch P_X. Analogously, we can compute the value $LC(P_Y)$ for the corresponding patch P_Y in image Y, and define

$$\Delta LC(P_X, P_Y) = |LC(P_X) - LC(P_Y)|. \tag{3.16}$$

The image quality measure $Q_{\Delta LC}(X, Y)$ is then finally defined as the measure ΔLC averaged over all possible $k \times k$ patches in images X and Y.

Linear combination of ΔLC and ΔE

As we mentioned before, in images transformed during the gamut mapping process the most important factors are usually color preservation, detail preservation and avoiding artifacts. Here we describe a measure, which combines two of these factors, namely color preservation (using ΔE) and detail preservation (ΔLC). The measure $\Delta_{E,LC}$ is a linear combination of these factors.

$$Q_{\Delta_{E,LC}} = \alpha \cdot Q_{\Delta LC} + (1 - \alpha) \cdot Q_{\Delta E} \tag{3.17}$$

The coefficient α is chosen to maximize hit rates using cross validation.

Part II

Applications

Chapter 4

Evaluation of gamut mapping algorithms

However beautiful the strategy,
you should occasionally look at the results.
Sir Winston Churchill

As shown in the previous chapter, gamuts of different devices differ strongly from each other. For instance, because of device limitations, a printer is typically not able to reproduce all the colors visible on a display. As an illustration, we show in Figure 4.1 an original image and its reproduction by different printer gamuts, where all out-of-gamut colors are reproduced by white. When reproducing color images on different devices, one has to adapt these images to device gamut limitations. The process of adapting the colors to device limitations is called gamut mapping.

In this chapter, we describe psycho-visual tests that can be used for the comparison of GMAs and also in the development of GMAs. Our approach builds on the insight that gamut mapping can be seen as a highly parameterized problem. There are many, sometimes competing parameters relevant for gamut mapping: first of all the preservation of hue, lightness and saturation. Also, in the realization of GMAs, we have a choice of working color space, the mapping direction, compression type (clipping, linear, nonlinear compression) and an application depending source gamut description. We use psycho-visual tests—paired comparisons—to determine an optimal parameter setting. The data elicitation phase of our test is the same as in traditional psycho-visual tests conducted to compare different gamut mapping algorithms. In particular, the number of paired comparisons per observer is not larger, and the number of observers can be kept reasonable, although the potential number of mapping algorithms that can be compared is much larger. The difference to traditional psycho-visual tests comparing GMAs is in the way, how we analyze the elicited data. As mentioned in the Chapter 2, Thurstone's method is not efficient enough for testing multi-parameter algorithms. Hence, here we use conjoint analysis that essentially fits a linear model [26] to the data by assigning a part-worth value to each parameter level. In addition to quality values of a parameter setting, we are also interested in extending and testing the underlying model, including parameter interdependencies, choice models, and the influence of individual images and observers.

We should point out that we are not the first who systematically include observer experiments in the development of GMAs, see for example the work done by Kang et al. [37]. Multivariate analysis techniques also have been used in image processing to gauge the importance of parameters, see for example the book of Keelan [38]. Here, the scaling between different parameters is ensured with the use of Just Noticeable Differences (JNDs).

Fig. 4.1: Demonstration of in-gamut colors for typical printing gamuts. Colors not in gamut are left white: Original sRGB (a), photo paper (b), coated offset paper (c), and newspaper (d).

4.1 Conjoint analysis for evaluating parametrized GMAs.

4.1.1 Algorithms

In our study, we consider one master algorithm with free parameters. The master algorithm is quite simple, it maps any color point in the source gamut along a line segment connecting a focal point and the color point into the destination gamut. Additionally, we consider the influence of detail enhancement and working color space. Furthermore, we want to compare the influence of those gamut mapping parameters with typical color and lightness operations on an image and with parameters of the destination gamut. In the following, we present the parameters which we have studied. We always used sRGB as source gamut, i.e., we did not consider the source gamut as a parameter.

Compression. This parameter describes how our master algorithm moves a color point along the line segment. We have tested different strategies: linear compression, clipping and sigmoidal compression algorithms. In order to parameterize the sigmoidal compression we used a weighted average of linear and non-linear compression [64]. The scale factor β is computed as

$$\beta = \alpha \cdot D \cdot \tanh\left(\frac{S}{D} \cdot \tanh^{-1}\left(\frac{X}{S}\right)\right) + (1-\alpha) \cdot \frac{X \cdot D}{S}, \tag{4.1}$$

where

- X is the distance of the focal point to the color point that needs to be mapped,
- S is the distance of the focal point to the source gamut boundary,
- D is the distance of the focal point to the destination gamut boundary, and
- α is a weighting factor in the range $0 < \alpha \leq 1$.

Details. Reconstructing details can essentially improve the quality of the mapped image [4; 23; 65]. We used a detail enhancement procedure independent of the master algorithm. But we can interpret it as a parameter of the master algorithm in the sense that we can apply detail enhancement in varying degrees to the results obtained from the master algorithm. We use the detail enhancement method based on edge-preserving smoothing filters described by Zolliker [65] with different weighting factors r. The other parameters were kept at default values: $\sigma_c = 20$ and $\sigma_s = 4\%$.

Color space. Note that our master algorithm can be applied in many color spaces and this parameter describes the choice of working color space. In our study we used either IPT [18] or CIELAB [10] color space.

Color and Lightness. Another free parameter of our master algorithm is the choice of focal point. The idea is to produce well defined color and density shifts in the mapped image by varying the focal point. A natural choice for the focal point is close to the mid point of the gray axis in the destination gamut. Moving the focal point on the lightness axis results in a overall lightness change of the mapped image. The amount of the lightness change of a specific pixel decreases from a maximum for the colors close to the focal point towards zero at the gamut boundary. Similarly, a shift of the focal point in the chroma plane results in a color shift of the mapped image.

Hue. In order to study the influence of hue shifts the color of all pixels were shifted in hue by a defined angle prior to applying the master algorithm.

Gamut Size. To gauge the importance of the destination gamut we also tested a gamut size parameter. This is not actually a free parameter of our master algorithm, but we included it, because it allows us to estimate the relative importance of the destination device capabilities compared to the free parameters of the master algorithm. We tested four different destination gamuts. The smallest was ISO-Newspaper, the largest ISO-Coated. The remaining gamuts were created from the two as weighted average.

Gamut Shift. This parameter describes a shift of the destination gamut in the working color space.

Gamut Rotation. Another parameter that we considered is a rotation of the destination gamut in the working color space.

4.1.2 Setup of the conjoint studies

Two conjoint studies were defined based on the above parameters. After preliminary evaluating the results of *Study 1* it was realized, that the gain in information from two of the parameters (*Gamut Shift* and *Gamut Rotation*) was marginal. Thus, in *Study 2*, those two parameters were replaced by new parameters (*Color/Lightness* and *Hue/Color Space*).

In order to keep the number of possible combinations reasonably small, some of the parameters were combined into one parameter and the number of considered levels was reduced. Lightness and Color were combined into one parameter with only six levels. Because the main difference of CIELAB and IPT color space are hue conservation issues, the parameters Color Space and Hue parameters were also combined into one parameter. In *Study 1*, only neighboring Gamut Size levels

were used in the comparisons. This restriction was removed in *Study 2* in order to better test the distribution assumption (compare Sections 2.3.4, 2.2.5) in the evaluation model.

The used parameters and levels for the two studies are summarized in Table 4.1. For both studies, every image had 1536 possible mapping combinations. The five parameters had a total of 22 levels.

Parameter	Level	Level description	Study
Gamut Size	S1	Newspaper	1 2
	S2	2/3 Newsp. 1/3 ISOcoat.	1 2
	S3	1/3 Newsp. 2/3 ISOcoat.	1 2
	S4	ISOcoated	1 2
Compression	C1	Linear	1 2
	C2	Sigmoidal $\alpha = 0.5$	1 2
	C3	Sigmoidal $\alpha = 0.8$	1 2
	C4	Clipping	1 2
Details	S1	no enhancement	1 2
	S2	weighting factor $r = 0.5$	1 2
	S3	weighting factor $r = 1.0$	1 2
	S4	weighting factor $r = 1.5$	1 2
Gamut Shift	Sh0	no shift	1
	Sh+	shift $(+5, 0, 0)$	1
	Sh-	shift $(-5, 0, 0)$	1
	ShC1	shift $(+0, 3, 0)$	1
	ShC2	shift $(0, -1.5, -1.5\sqrt{3})$	1
	ShC3	shift $(0, -1.5, +1.5\sqrt{3})$	1
Gamut Rotation	IPT-R0	IPT, no rotation	1
	IPT-R-	IPT, hue -0.1 radians	1
	IPT-R+	IPT, hue +0.1 radians	1
	Lab-R0	CIELAB, no rotation	1
Color/Lightness	L0	no shift	2
	L+	shift $(+5, 0, 0)$	2
	L-	shift $(-5, 0, 0)$	2
	Col1	shift $(+0, 3, 0)$	2
	Col2	shift $(0, -1.5, -1.5\sqrt{3})$	2
	Col3	shift $(0, -1.5, +1.5\sqrt{3})$	2
Hue/Color space	IPT-H0	IPT, no rotation	2
	IPT-H-	IPT, hue -0.1 radians	2
	IPT-H+	IPT, hue +0.1 radians	2
	Lab-H0	CIELAB, no rotation	2

Table 4.1: Parameters and their levels used in *Study 1* and *Study 2*.

4.1.3 Test setup

In this section, we describe how we collected paired comparison data in a psycho-visual test to analyze our master gamut mapping algorithm. In every paired comparison, we presented an original image and two images mapped by different incarnations of our master algorithm on a LCD screen.

The original image was presented in the upper half of the screen and two mappings below the original side by side, compare Figure 2.1 in Chapter 2. Observers were asked to make their choice according to the following instruction: *"Choose the best representation of the original. If you see no difference, click the original".*

The two mappings were chosen at random from our parameter space. As mentioned in Section 2.1.2 random designs may not be the most efficient ones, but avoiding systematic errors had a higher priority for us than efficiency. For *Study 1* we had the constraint, that gamut size levels in compared images differ no more than one consecutive (in the natural order) level since larger differences in gamut size essentially determine the choice.

The observers who participated in our test had to choose the mapped image that better reproduces the original. For their choices, the observers used a mouse to click on the corresponding image. If no difference could be seen, the original had to be selected in order to avoid a forced choice.

Test sets

The same visual study was carried out within different environments. While the **laboratory** set-up was carried out in a controlled environment, adjusted closely to the CIE viewing standard, the **web**-based test was carried out by observers on their own systems. We also had a group of observers, who made the test both in laboratory and web environment. This resulted in obtaining Cross Link Laboratory and Cross Link Web data sets.

The key properties of the data sets used in this study are summarized in Table 4.2.

Identification	Study	Type	nr of observers	nr of pairs
Lab-1	1	lab	70	3500
Web-1	1	web	590	25108
Cross Link-Lab	1	lab	41	1440
Cross Link-Web	1	web	41	1440
Lab-2	2	lab	24	2100
Web-2	2	web	96	5358

Table 4.2: Discussed test sets.

Next we give a summary of the set setup:

Laboratory setup. For the test we used LCD displays. An 22" Eizo CG 241W-BK monitor calibrated to show sRGB-colors was used to display test images. The ambient illumination measured in the middle of the switched off monitor was at 40 lx. Monitor flaps around the screen prevented flare. The monitor's background was set to a neutral gray.

Web setup. For the Internet based part, we had to consider a variety of viewing conditions and displays (concerning brightness, size, resolution, white point and color gamut). Therefore, additional information was collected from the web study participants concerning their employed system (ambient illumination, display type and size, Internet browser and operating system). We used JPEG images with very low compression and a maximal width or height of 400 pixels which resulted in about 150 KB per image. The resulting test pages were verified to be presentable on common operating systems, browsers and even most of laptop displays.

Observers. Three user groups were considered in this experiment: *lab*, *web* and *cross-link*. Observers in the *lab* user group were recruited from staff of Swiss Federal Laboratories for Material Science and Technology (EMPA), and participants of a FOGRA symposium who were mostly color experts. Each observer had passed the Ishihara test for color deficiency. To recruit observers for

the *web* user group, the Internet test was posted on the homepage of Media Technology Laboratory, EMPA. Students, color specialists and other people were invited to participate via e-mail and Internet user groups. In the *cross link* study, the same observers participated in both environments (identified with a user ID). The study was carried out by students from the Swiss Federal Institute of Technology Zürich and by staff of EMPA. The number of participating people for each study are given in Table 4.2.

Test Images. The image set included the obligatory "Ski" image that is specified by the CIE 156:2004 guidelines [12] and additional ISO images. A total of 99 different images including a wide range of scenes was used in the experiment in order to get good average results and to be able to study the influence of different images on the psycho-visual results. Most of them were taken from royalty free libraries as well as from private stock. The images are presented in Figure 4.8

4.2 Results

4.2.1 Importance of parameters

At first we present the computed part-worths for all different parameters individually. The results for both studies are shown in Figure 4.2. The comparison of the part-worths allows to answer questions like: What is the relative importance of the different parameters? Which levels of the parameters are most preferred?

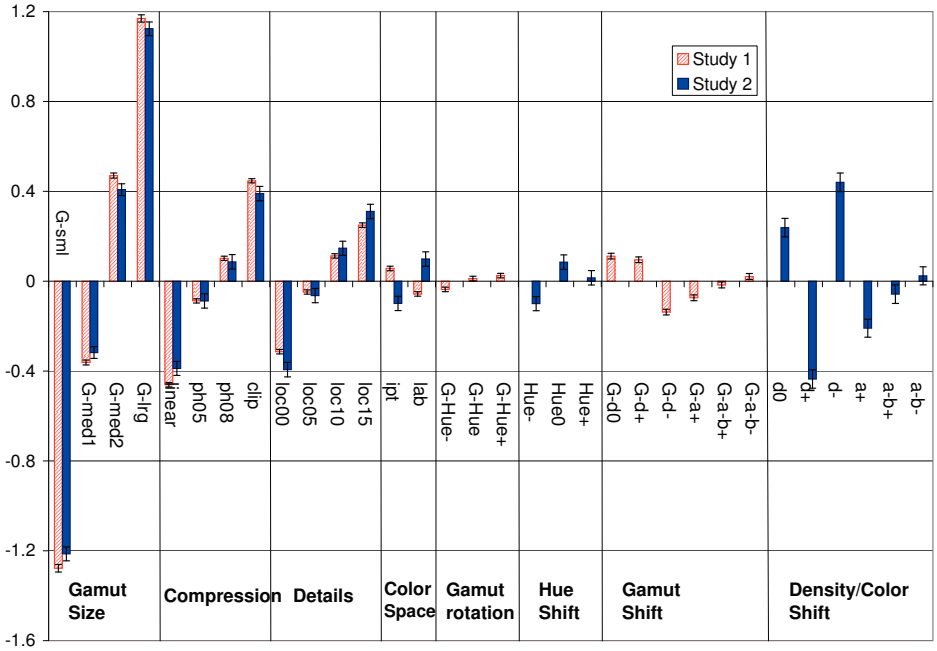


Fig. 4.2: Part-worths for all parameter levels. The light bars (red) show results of *Study 1* and the filled bars (blue) those of *Study 2*. Error bars show one estimated standard deviation computed using analytical error estimation.

At first we look at the importance of parameters, which describes how much each parameter contributes to the quality value on the stimulus level, i.e., the combination of all parameters. For this we use the standard deviation of the part-worths within the parameters. Note that the computed importance depends on the levels chosen for the parameters, e.g., if we choose levels for a parameter that hardly can be distinguished, then the importance of this parameter will be low, though it could be high for a different choice of levels. Hence the choice of levels is an important task in conjoint analysis. The importance of the different parameters is shown in Table 5.1.

	Study 1		*Study 2*		
	Importance	Rank	Importance	Rank	$\sigma_{\langle\Delta E\rangle}$
Gamut Size	0.572	1	0.473	1	3.30
Compression	0.187	2	0.131	4	2.03
Color/Lightness			0.161	2	1.74
Color space	0.042	5	0.064	6	0.62
Gamut shift	0.074	4			0.48
Details	0.130	3	0.150	3	0.19
Hue			0.085	5	0.17
Gamut rotation	0.017	6			0.12

Table 4.3: Importance of parameters for both studies (scaled such that sum is 1). The entries are sorted by the last column, which shows $\sigma_{\langle\Delta E\rangle}$.

In this table we also show the standard deviation $\sigma_{\langle\Delta E\rangle}$ which serves as a first order measurement of the perceived image distance between original and mapped images. $\langle\Delta E\rangle$ is the distance between the transformed image and its original, averaged over the images. The standard deviation $\sigma_{\langle\Delta E\rangle}$ was calculated from $\langle\Delta E\rangle$ values for different levels of a specific parameter taking default levels for all other parameters (S3, C3, ShL0/L0, IPT-R0/IPT-H0, D1). Note that in general the importance of parameters correlates with the average difference $\sigma_{\langle\Delta E\rangle}$. An exception is the *Details* parameter, which shows a very small $\sigma_{\langle\Delta E\rangle}$ despite of its relative importance. This is not surprising, as local contrast conservation can not be measured by a global color distance measure such as $\sigma_{\langle\Delta E\rangle}$.

Gamut size is the most important parameter in both studies, but it is not the only deciding factor. *Compression*, *Details* and in the case of *Study 2* also *Color/Lightness* all can contribute to the quality as much as the change of gamut size of two consecutive size settings.

Clipping emerges as the best method of compression. Linear compression is not well suited. Sigmoidal compression is better the closer it is to clipping. This result shows, that saturation is an important factor for respondents and is in agreement with many gamut mapping studies in the literature [49].

About equally important as compression is detail preservation. The higher the weighting factor r the more preferred it is. Surprisingly, this even holds for an exaggerated detail enhancement with a factor of 1.5. A factor of $r = 1.0$ reconstructs small details of the original image except for colors close to the gamut boundary and due to the edge-preserving filter also for colors close to an edge.

According to the computed part-worths, the preferred color space is IPT in *Study 1* and CIELAB *Study 2*. This is rather unexpected. The advantage of IPT is, that it better preserves hue, especially in blue regions. On the other hand, CIELAB may have advantages over IPT, because most gamut mapping algorithms and their optimizations (e.g. choice of focal point) were elaborated in CIELAB. One possible reason for our conflicting result could be that the hue advantage is relevant mainly in *Study 1*. *Study 2* has explicit color changes larger than the expected hue shift in the CIELAB space and for images with a color cast, the hue advantage may not be important. In fact, a partial

evaluation of the data in *Study 2* disregarding the color shift level (C1, C2, C3) shows an increased part worth of IPT compared to CIELAB. However, in view of the rather small part-worths of the color space parameters, compared to the other parameters, we can not rule out that some systematic shortcomings of our conjoint model could be the reason for the result.

For *Color/Lightness*, the most preferred level is L- followed by L0. As default value of the focus point in the destination gamut, we used the mid point $L = (59, 0, 0)$ between black and white point of the smallest gamut. Because the mid point of the source gamut is $L = (50, 0, 0)$ a neutral gray with the default parameter L0 is mapped to a lighter color than in the original. The results of our study show, that in general darker images (level L-) were preferred for which the mapped neutral gray is closer to that of the original. This indicates, that the mid grays tone should be mapped close to its original, independent of the lightness of the destination black and white. As expected, the color changes clearly have a negative influence on the perceived quality. A color change due to a focus shift of $3\Delta E$ causes a quality decrease of the same order as the differences between sigmoidal compression and clipping, or the difference of two successive details enhancement factors. In a similar manner, hue changes in either direction cause a quality decrease, but the magnitude of the studied hue changes (± 0.1 radians) is only about half of that of the studied color changes.

We do not try to interpret the results of the levels of *Gamut Shift* and *Gamut Rotation*. Their part-worth values are small anyway.

4.2.2 Testing the model

Mosteller's test

We made the assumption on the parameter level, that the part-worth are uncorrelated normally distributed variables with equal variances. We tested these assumptions using Mosteller's test. A description of Mosteller's test can be found in Engeldrum [19] or Mosteller [53]. Results are presented in Table 4.4. Most parameters passed the test at a significance level $\alpha = 0.01$. Only

	χ^2, *Study 1*		χ^2, *Study 2*		$\chi^2_{\alpha=0.01}$
	Probit	Logit	Probit	Logit	
Compression	**22**	**15**	2.1	2.1	11.3
Details	0.3	0.3	4.1	4.0	11.3
Gamut Size			**24.7**	10.0	11.3
Color Space/Hue			2.0	2.0	11.3
Gamut Rotation/Hue	3.2	2.0			11.3
Gamut Shift	9.4	9.4			23.2
Color/Lightness			11	11	23.2

Table 4.4: Mosteller's test for parameter compared to χ^2 with significance level $\alpha = 0.01$ for Gaussian and logistic distribution.

the *Compression* parameter in *Study 1* and the *Gamut Size* parameter in *Study 2* show significant deviations. One possible reason could be the assumption of a Gaussian distribution. A distribution with a wider tail could better explain the data. Evaluations using a logistic distribution show better, but not perfect results. The corresponding numbers are shown in Table 4.4. Note that we could not apply Mosteller's test to the gamut parameter in *Study 1*, as the frequency matrix for the gamut parameter has not enough entries (only specific pairs of gamut levels had been compared).

Equivalence of data sets

For *Study 1*, each data set from the three observer groups was analyzed separately: the two laboratory data sets that were collected once at a symposium and once in the lab, then the control study, where the same observer performed the test on the Internet and in the laboratory environment. Surprisingly, the three sub-tests showed similar results [56]. For *Study 2* the laboratory and the Internet data were also analyzed separately and compared but showed no significant difference. The hypothesis that the results of the studies cannot be distinguished was tested with a χ^2 comparison test. The results support our hypothesis that the results of the studies cannot be distinguished on the base of our data. The results are summarized in Table 4.5.

		χ^2	$\chi^2_{\alpha=0.01}$	deg. of freedom
Study 1	Ctrl-Laboratory - Ctrl-Internet	41	59	36
	Ctrl-Laboratory - Symposium	32	59	36
	Ctrl-Laboratory - Internet	26	59	36
	Symposium - Internet	49	59	36
Study 2	Laboratory - Internet	60	62	39

Table 4.5: χ^2-test for comparison of test sets.

The collection of the large data set using the Internet allows us to draw more precise conclusions about our model parameters.

Linearity

The linearity assumption was tested for each parameter pair. The results of the χ^2-tests for *Study 1* are shown in Table 4.6. The χ^2-values of most parameter pairs did not indicate a deviation from linearity. Two combinations, *Compression-Details* and *Gamut Size-Gamut Shift*, show clearly significant deviations. Interestingly, those two combination also show the largest increase in hit rate when combined parameter levels are used.

Parameter Combinations	χ^2	$\chi^2_{\alpha=0.01}$	Hit rate gain/loss
Compression - Details	**99**	29	**0.35**
Compression - Gamut Size	28	29	-0.05
Compression - Gamut Rotation	15	29	0.09
Compression - Gamut Shift	21	40	-0.13
Details - Gamut Size	23	29	-0.07
Details - Gamut Rotation	14	29	-0.05
Details - Gamut Shift	13	40	0.04
Gamut Size -Gamut Rotation	11	29	0.13
Gamut Size - Gamut Shift	**66**	40	**0.16**
Gamut Rotation - Gamut Shift	14	29	0.02

Table 4.6: χ^2-values for the linearity test for *Study 1*. Significant deviations are shown in bold. The last column shows changes in hit rate. Largest hit rate gains are shown in bold.

A detailed inspection of the combined *Compression-Details* results show that the gain due to high level of *Details* parameter are about half as large for clipping as the gains for other compression levels (see Figure 4.3). A possible explanation is the fact, that detail reconstruction can bring the colors out of gamut again. These colors have to be mapped back into the gamut. Using clipping, much more colors are affected by this second mapping compared to the other compression parameters. The nonlinearity in *Gamut Size-Gamut Shift* can be characterized as an increase of the part-worth for Sh- on the cost of Sh0 with increasing gamut size.

For *Study 2* no significant deviation from linearity could be detected and the hit rate could not be increased for any parameter pair. This is presumably due to the limited size of the data set compared to *Study 1*.

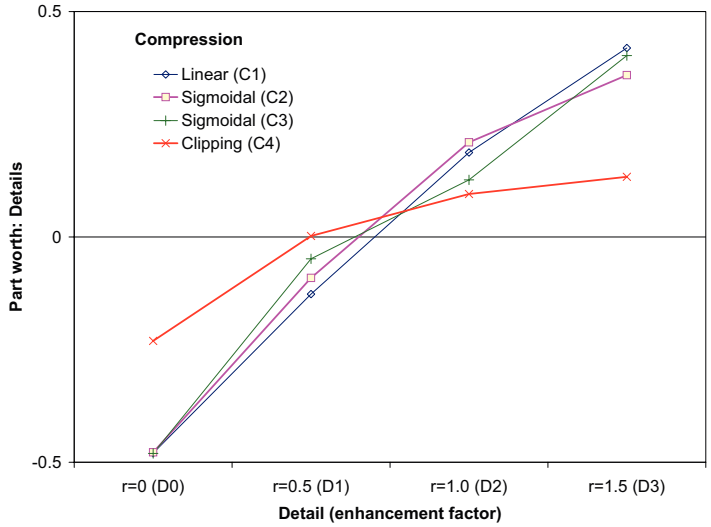


Fig. 4.3: Part worths of *Details* as a function of *Compression*.

Distribution function

At first we qualitatively verify the assumption that the distribution of the sorted scale values of possible parameter combination does not have gaps between successive quality values, i.e., that no parameter is dominant over the other parameters (compare Section 2.3.4). This is visualized in Figure 4.4. In a second step, we can experimentally estimate the average cumulative distribution function: Histograms are collected on all judgments based on their estimated psycho-visual distance. From them, the probability that an observer's judgment agrees with the modeled quality distance can be computed and compared to the cumulative distribution function, see Figure 4.5.

The conjoint analysis and the determination of hit rates was performed for the Gaussian and the logistic distribution function. The hit rates turned out to be very similar with a slight advantage for the Gaussian distribution for *Study 1* and an advantage for the logistic distribution for *Study 2*.

Even if there is evidence from the Mosteller test, that the logistic distribution can better explain frequencies at large psycho-visual distances, the logistic distribution does not clearly increase hit

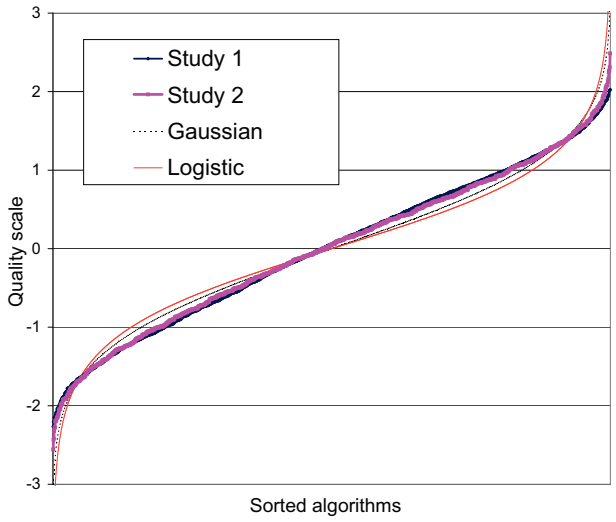


Fig. 4.4: Sorted scale values for the algorithms for *Study 1* and *Study 2* compared to model graphs using Gaussian and logistic distributions.

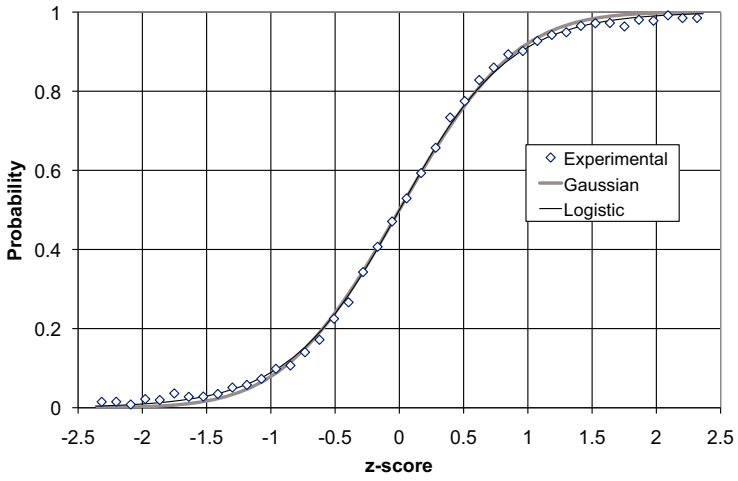


Fig. 4.5: Cumulative distribution function. Logistic versus Gaussian CDF compared to experimental data.

Distribution	*Study1*	*Study 2*
Probabilistic	81.5%	85.6%
Logistic	81.4%	85.8%

Table 4.7: Hit rates using probabilistic and logistic distributions for the two studies.

rates. The Gaussian distribution may be more appropriate at shorter distances. In view of the very similar results for the part-worths values for both distributions we did not further investigate finding a better distribution function, which could be a convolution of a Gaussian function with a logistic function[1]. The influence of the choice of distribution functions has been already discussed in earlier works [34; 36]. There was no significant difference in appropriateness of either of models [1].

Error analysis.

In Figure 4.6, we show the comparison of the theoretical error with three types of experimental errors. Due to the larger data set in *Study 1* compared to *Study 2* the estimated error is smaller.

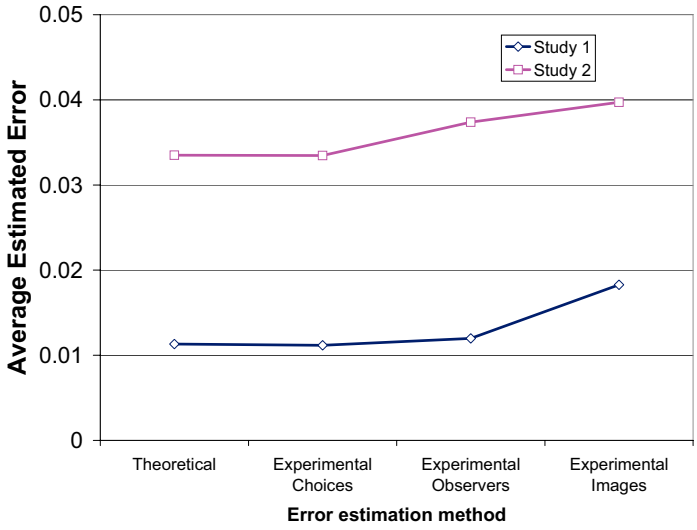


Fig. 4.6: Error estimation. Average error of parameter levels for theoretical error estimation and three types of experimental error estimations.

For both studies we did not notice a significant difference between the experimental error computed by randomly dividing the paired comparisons into two groups and the error calculated by linear regression. However, the experimental error computed by randomly dividing the images into two groups is significantly larger in both studies. For *Study 2*, also the experimental error computed by randomly dividing the observers into two groups is significantly larger. It means, that the differences in scale values of algorithms are higher between images than between random chosen sets of choices. This suggests it is worth an effort to develop gamut mapping algorithms based on individual image properties and even personalize gamut mapping algorithms for user groups. Therefore, we discuss the issue of individualization in gamut mapping in the next chapter.

[1] Note that the re-scaling in the conjoint analysis was derived assuming normal distributions, thus the rescaling for logistic distribution is only approximative

Table 4.8: Test images

Chapter 5

Image-individualized gamut mapping algorithms

Variety is the soul of pleasure.
Aphra Behn

In the previous chapter, we concluded from the error analysis, that preferences of algorithms are image dependent. In this chapter, we first verify whether we can observe such differences also in other psycho-visual tests concerning gamut mapping. Especially, we want to know how the number of comparisons per image influences possible improvements of the model by using individual data. Further, we focus on two aspects of individualization for GMAs. The first one is individualized evaluation. We describe how to model observers preferences for individual images. To assess the accuracy of non-individualized and individualized models we use hit rates. The models are based either on psycho-visual data or on image quality measures. We also compare the accuracy of these models with maximal possible hit rates for the given data sets. Another aspect of individualization is designing an image-individualized meta-algorithm which chooses appropriate algorithms for individual images.

5.1 Error analysis – differences between images

We use data from different previous studies concerning gamut mapping algorithms. Details of the algorithms considered in these tests are not important at this point, as we are focusing here on image dependency in a psycho-visual test. Hence, here we only summarize these tests.

Study 1: Basic Study (BS)

This study [6] is a traditional benchmark study comparing some newer image dependent gamut mapping algorithms to known reference algorithms. In addition to the reference algorithms HP-minDE, SGCK [12], the following algorithms using image gamut or spatial gamut mapping have been considered: the algorithm NOptStar that is using the image gamut as described in by Giesen et. al. [27], the Kolas algorithm [39], the Zolliker algorithm [65] applied to the SGCK and NOpt-Star algorithms, and the Caluori algorithm [6]. For this study, 97 images were used, each mapped with all seven algorithms. Each possible comparison was tested at least once. We will refer to this study as *Basic study* or simply *BS*.

Study 2: Image Gamut (IG)

The topic of this study was the use of image gamut descriptions for gamut mapping [27]. The considered algorithms have used a linear or sigmoidal mapping, each of them had three possible source gamuts, namely the device gamut (sRGB) and two types of image gamut description. The six possible combinations were compared to HPminDE and SGCK, resulting alltogether in eight algorithms. 75 images have been used. Each possible comparison was made approximately twice. We will refer to this study as *Image Gamut* study or simply *IG*.

Study 3: Local Contrast (LC)

In this study, the influence of detail enhancement applied to a set of gamut mapping algorithms was investigated [65]. The study comprised the HPminDE, SGCK, SGDA [64] algorithms and a linear compression algorithm. All algorithms were compared with and without detail enhancement. 77 images were used, and 5376 comparisons have been performed. Each possible comparison was made approximately 2.5 times. We will refer to this study as *Local Contrast* study or simply *LC*.

Study 4: Individual Study (IS)

In this study [16; 17], algorithms proposed by Gatta [23], Kolas [39] and an algorithm using detail reconstruction proposed by Zolliker [65] applied to the HPMinDE algorithm were compared with the reference algorithms HPminDE and SGCK. 20 images, presented in Figure 5.4, have been used. Each possible comparison has been performed 40 times. We will refer to this study as *Individual Study* or simply as *IS*.

We summarize the number of images, comparisons and algorithms per test in Table 5.1

Study	Number of images	Number of comparisons	Number of algorithms
BS	97	2086	7
LC	77	5376	8
IG	75	4360	8
IS	20	8000	5

Table 5.1: Number of images, comparisons and algorithms in the considered studies.

The aim now is to discuss two estimation procedures for the error: experimental with choice divisions and experimental with image divisions (compare Section 2.5.5). Plots of these errors are presented in Figure 5.1.

For the BS and IG tests we cannot see much difference between experimental errors obtained with choice divisions and image divisions. There are a few possible reasons for that. In those tests, there were less comparisons for each image, so the results were less significant. There were also clearer winners in those tests: The best algorithm was optimal for most of the images, so choosing it for individual images did not change much. On the other hand, the differences between the experimental error with choice division and experimental errors with images divisions were large in tests for Local Contrast (LC) and especially for Individual Study (IS). This finding correlates with number of comparisons per number of algorithms and images.

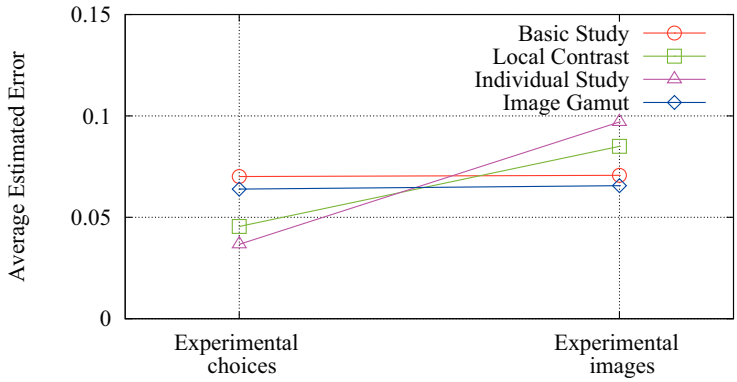


Fig. 5.1: Error estimation. Average error of attribute levels for two types of experimental error estimations.

5.2 Image-individualized evaluation of gamut mapping algorithms

We present in this section hit rates for different degrees of individualization for the considered studies. Then we evaluate image quality measures for the considered studies and compare them with data-based models.

5.2.1 Evaluating Thurstone's method

As described in the Section 2.4, we can build individual models using Thurstone's method. In Figure 5.2 we present hit rates for Thurstone's method for the different studies and different degrees of individualization. Hit rates on training sets are higher than those on test sets. Individualization always improves hit rates on training set data, however it does not always improve hit rates on the test sets. The higher hit rate on the training set is due to the overfitting of the model.

On the *BS* data set individualization does not increase the hit rate on the test sets. These test sets included only about one repetition for each comparison, so individual results are probably not stable enough to contribute to the model's accuracy.

For the *IG* study, the optimal hit rate is obtained for a linear combination of the global scale values and individualized ones. In this test, each comparison was repeated twice, which is enough to individualize the scale values but not enough to get a significantly better hit rate for these scale values than for the global scale values. The best hit rate needs a combination of global and individualized scale values.

In the *LC* study, there are about 2.5 repetitions for each comparison. As for the *IG* study, the optimal hit rate is achieved at a combination of individual and global scale values. But in the *LC* study using only the individual scale values provides a hit rate almost as high as the optimal combination of global and individualized scale values.

The largest number of repetitions for the individual comparisons between algorithms, namely 40, is present in the *IS* study. Here, we get the highest hit rate using just the individual scales values.

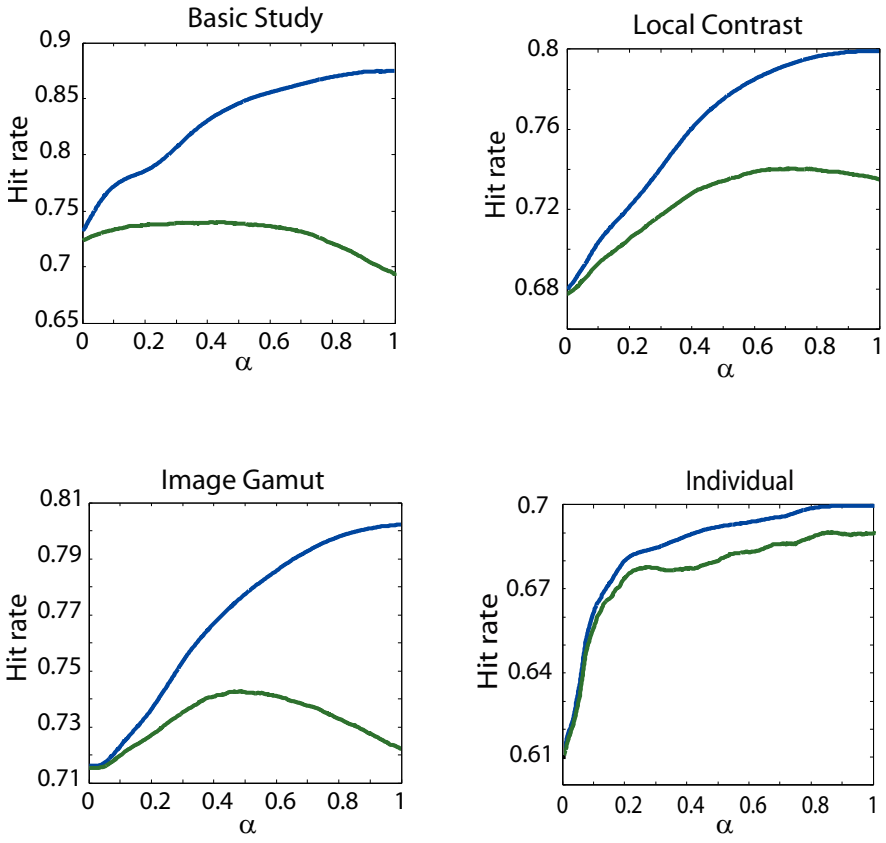


Fig. 5.2: Hit rates using Thurstone's method for: (a) Basic Study, (b) Local Contrast Study, (c) Image Gamut Study, and (d) Individual Study. The blue (higher) line shows the hit rate on the training set, the green (lower) line shows the hit rate on the test set. Scale values (sv) are computed as a convex combination of scale values for the whole population of images (sv_{gen}) and scale values for individual images (sv_{ind}), i.e., $sv = \alpha \cdot sv_{ind} + (1 - \alpha) \cdot sv_{gen}$ with $0 \leq \alpha \leq 1$.

The results correlate well with the results from an error analysis (see Section 5.1). The experimental error increase between choice sampling and image sampling is the highest in the IS test. A high error by images sampling means, that results from one image are not a good prediction for other images, as there is no algorithm in the given test, that is best on all images. This is the test, where individualization improves hit rates the most. Also, for the LC study we could notice a significant difference between errors based on choice sampling and image sampling. In the IG test, where there was no significant difference between these two types of errors, mixtures of general and individual results are only slightly better than for the non-individualized results. In the BS, where there is also no difference between these types of errors, individualization does not improve the results at all.

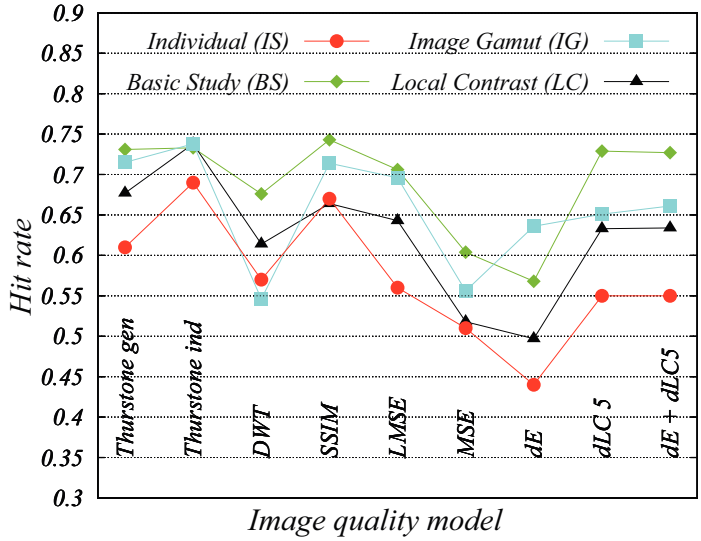


Fig. 5.3: Hit rates obtained by different methods for four studies. Here *Thurstone_gen* refers to general (non-individualized) Thurstone's method and *Thurstone_ind* refers to individualized Thurstone's method.

5.2.2 Evaluating the image quality measures

Individualization can increase hit rates, if we have enough data for single images and there is no clear "winner" algorithm for all images. Now that we know this, we look for an image quality measure that predicts observers' preferences good enough to use it for individualization, i.e. better, than non-individualized Thurstone's method. Below we compare the accuracies obtained by image quality measures and Thurstone's method for the discussed tests. The results are presented in Figure 5.3.

On all the data sets the *S*tructural *SIM*ilarity index measure (SSIM) proved to be the best performing image quality measure, i.e., it provided the highest hit rates. On the *BS* data set, the results obtained with SSIM are even better than those coming from the individualized Thurstone's method. On the other studies, the individualized Thurstone's method gives better results. As discussed before, a likely reason for this behavior is the size of the *BS* study as the performance of Thurstone's method improves with increasing number of comparisons. It is worth noting that the hit rates for SSIM are comparable to the hit rates obtained for the general Thurstone's method, or, in case of the IS test, even much higher.

The two pointwise image quality measures that we considered, namely $Q_{\Delta E}$ and the mean square error Q_{MSE}, scored lower than their competitors, often showing hit rates close to random choice, i.e., 50%. The likely reason is that all gamut mapping algorithms tested in these studies already optimize color preservation in some way, and thus observers' choices are more affected by detail preservation. In particular, clipping algorithms, for example the HPminDE algorithm, are optimizing the mapped image against the pointwise distance measures, but ignores detail preservation.

The quality measures LMSE and LC, which embody detail preservation differences, perform better than pointwise measures, but still not as good as the SSIM measure.

5.2.3 Theoretical limit of hit rates

Let us notice, that the theoretical limit hit rate of 1.0 is almost never achieved, because observers usually differ in their choices and even the decisions of a single observer are typically inconsistent, i.e., the same person, under the same conditions makes sometimes a different choice on the same images in repeated paired comparison.

If we have choice data with many repetitions for each choice, then we can estimate a better (choice data dependent) limit for the hit rate than the ideal 1.0, namely, the *maximally achievable hit rate* as follows: let f_{ij} be the frequency that algorithm i has been preferred over algorithm j in a comparison, i.e., the number of times an image mapped using algorithm i has been preferred over the same image mapped by algorithm j divided by the total number that i and j have been compared. If we have same number of repetitions for each comparison (which was the case in IS test), we can define the maximal achievable hit rate as follows:

$$\mathrm{HR}_{max} = \frac{\sum_{i<j} max(f_{ij}, f_{ji})}{\text{number of pairs of algorithms}} \tag{5.1}$$

Computing a *maximally achievable hit rates* requires having meaningful f_{ij} for single images. This was the case only in IS test, where for each pair of algorithms for each image we had $N = 40$ comparisons. In other tests this number was much smaller. Hence we use the IS data set to check how close the best performing quality measures SSIM and Thurstone's method come to the maximally achievable hit rate. All the images considered in this test are presented in Figure 5.4 The hit rates computed for the different images in the *IS* test set are shown in Figure 5.5.

The hit rates obtained using Thurstone's method with individualization is always very close to the maximally achievable hit rate for all images. For many images, the two hit rates are even equal. The hit rates achieved by SSIM are lower, but generally close to the one for Thurstone's method with individualization and much higher than for Thurstone's method without individualization. Only on three images out of 20, SSIM performs worse than Thurstone's method without individualization.

5.3 Image-individualized gamut mapping algorithms

5.3.1 Using SSIM to construct an image-individualized gamut mapping algorithm

The results from the previous sections suggest that we can design a meta gamut mapping algorithm, that chooses a "best" gamut mapping for a given image from a class of mappings. Here, "best" is meant with respect to an image quality measure that proved to be well suited to predict the perceived quality of a mapping. Again, the previous sections suggest that SSIM is suitable as such a measure. This approach is also supported by previous studies [16; 50], showing that different gamut mapping algorithms perform differently on different images, i.e., one can improve the quality of mapped image by choosing the best algorithm for this image, instead of using the same algorithm for all images.

Formally, the meta-algorithm can be described as follows: for a given quality measure Q (in our case SSIM) and a given image I, let $I_1, \ldots, I_n$ be the mappings of this image using n different mapping algorithms. Choose the mapping I_k such that $Q(I_k) \geq Q(I_i)$ for all $i = 1, \ldots, n$.

Fig. 5.4: Images used in the *IS* study.

5.3.2 Using psycho-visual data to construct an image-individualized gamut mapping algorithm

We can build an individualized algorithm also based on the data obtained in a psycho-visual test. It can be done in the same way as based on an image quality measure, but using the psycho-visual data to find the scale values of algorithms for images. One must not use the same data for choosing an algorithm for images and testing. Hence we were computing scale values on 90% of the data and used 10% for validation. This process was repeated 10 times and a mean scale value for the meta-algorithm was computed.

5.3.3 Validation of the meta-algorithm

Thurstone's method can easily be adapted to compare the quality of the meta algorithm and the individual algorithms on which the meta-algorithm builds. We used the data from the *IS* study to validate two meta-algorithms, one using SSIM, another one using scale values from individualized Thurstone's method on the training set as quality measure. Remember that IS study comprised

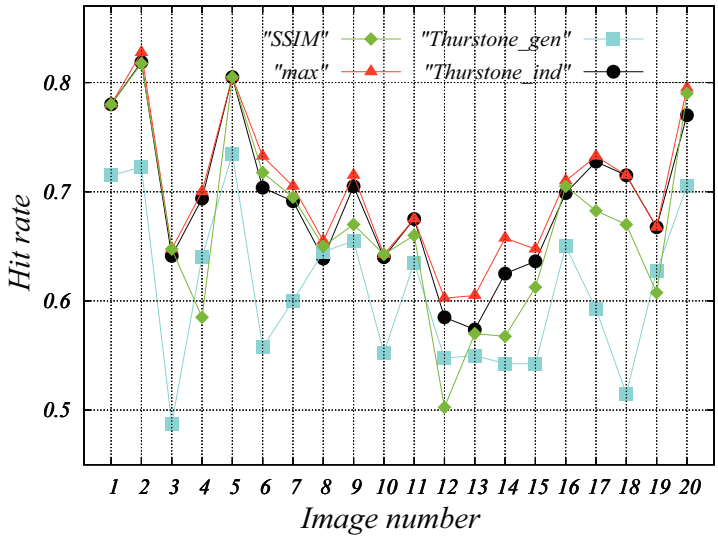


Fig. 5.5: Hit rates obtained by different methods for the different images in the *IS* test set. Again, *Thurstone gen* refers to Thurstone's method without individualization, and *Thurstone ind* refers to Thurstone's method with individualization.

twenty images, each of them mapped by five algorithms. We did not carry out an additional psycho-visual test, but adjusted the data from the original IS study. We extended the data as follows: From the original study we got the data in the form of F-matrices for each image, where f_{ij} is the number of comparisons where algorithm i was preferred over algorithm j. Now apart from five basic algorithms we consider a meta-algorithm. For each image this algorithm is the same as one of the five basic algorithms, indicated by the given model as the optimal for this image. Hence, if the optimal algorithm according to this model is for instance k, we can assume, that the meta-algorithm for this image would have the same F- values as algorithm k. Hence we put $f_{meta,i} := f_{ki}$ for $k \neq meta$. Setting $f_{meta,k} = 0.5$ is a natural assumption, as for the considered image k and $meta$ are the same algorithm. We can validate more than one meta-algorithm using this adjustment method. We compared the SSIM-individualized meta-algorithm and the Thurstone-individualized meta-algorithm with five basic algorithms. In Figure 5.6 we summarize the results of the comparison.

Both image-individualized algorithms perform better, than any single algorithm. The meta-algorithm using SSIM performs only slightly worse than the meta-algorithm based on individualized Thurstone's method. However, the algorithm based on the individualized Thurstone's method is not practical, as it requires conducting a psycho-visual test for every image that we want to map. Still, the SSIM measure has its limitations, e.g., as can be seen in Figure 5.5 the meta-algorithm predicts choices for image number 4, 12 or 19 worse than the Thurstone's individualized scale values. Note, that we applied SSIM only for the L-coordinate of the CIELAB space, and thus image quality effects based on the color coordinates have been neglected.

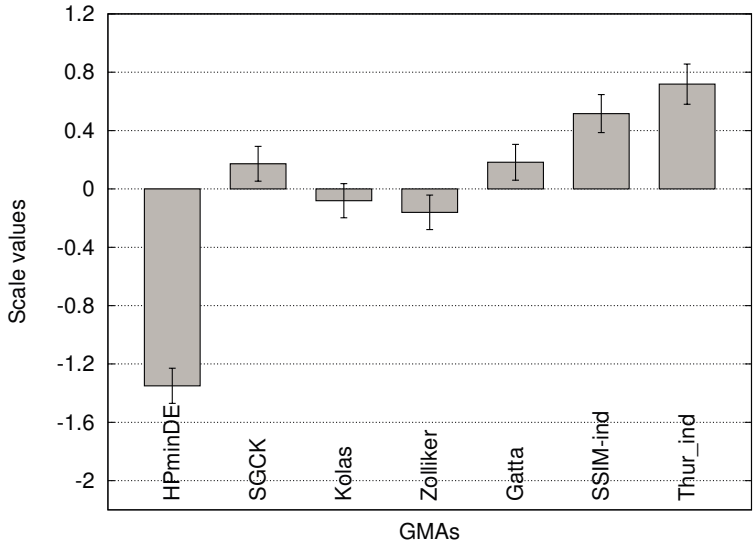


Fig. 5.6: Scale values of five algorithms considered in the *IS* study, plus the meta-algorithm based on the SSIM measure and Thurstone's individualized scale values. Error bars show one standard deviation and are computed analytically as described in Section 2.5.3.

Chapter 6

Conclusion

Finish each day and be done with it.
You have done what you could.
Ralph Waldo Emerson

This thesis investigates the main methods of modelling the quality of gamut-mapped images with the goal to optimize gamut mapping algorithms. We have introduced the concept of conjoint analysis for evaluating the performance of parametrized gamut mapping algorithms. It extends the well-established Thurstone's method by allowing to decrease the number of required comparisons to a manageable size while testing parametrized algorithms. We have presented methods to verify appropriateness of the model assumptions for evaluating given choice data. An important topic for future work on conjoint analysis is investigating parameters using the fact, that some of them are continuous. Without setting constant levels one could better model distribution of preferences. Moreover, one can improve the time-efficiency of a test by optimizing the test design. In particular, adaptive test designs are of interest.

We have introduced a new formula for error estimation, based on error propagation. It has been shown, that it performs better than previous methods. Our method does not impose a significant overhead on top of computing of the scale values themselves. This error estimation method can replace previous methods because it is more accurate for a larger range of psycho-visual scale values and number of compared algorithms. The computation of experimental errors could give further insights: It allowes verifying the appropriateness of the above methods. Additionally, different types of experimental errors (choice based, image based or observer based) allow to test for the homogeneity between observers or images.

We have discussed image quality measures as an alternative to psychovisual tests for evaluating gamut mapping algorithms. We have introduced a new measure, which takes into account color distance and details preservation, which both are important factors in gamut mapping. The use of hit rates allows to assess different models. With them we could identify the best performing image quality measure for the considered tests.

Using error analysis, we found evidence, that preferences of algorithms are not homogeneous between images. As a consequence, we have applied individualization concepts with respect to images as a technique to improve evaluation and construction of gamut mapping algorithms. We have considered two kinds of individualization. The first one: image quality measures, which by default model the quality of individual images. The second one: based on data-driven models, improved by individualization techniques. We have shown, that good individualized models perform better than non-individualized. In particular, models based on 'good' image quality measures perform better, than non-individualized data-based models (without even requiring psycho-visual test data).

An important step in optimizing gamut mapping algorithms is the design of a practical image-individualized meta-algorithm. On an example we have shown, that it can perform better than any of single algorithms used for the design of this meta-algorithm.

Many directions of improving image quality measures are still to be investigated. Two basic directions are: improving the formula describing similarity of images and defining color spaces, where distances correlate better with distances perceived by the human visual system. Along with improving image quality measures the performance of image quality based image-individualized meta-algorithms would also increase. By the current methods of individualizing an algorithm with respect to images, one has to compute mapped images for all considered algorithms.

It would be more time-efficient, if the optimal algorithm could be predicted based only on the statistics of the original image and a data base assigning for the given image statistics an optimal algorithm. However, this requires more research on preferences of different algorithms depending only on the scene present in the original.

REFERENCES

[1] J. Berkson. Maximum likelihood and minimum Chi-Square estimates of the logistic function. *Amer. Stat. Assn. Jour.*, pages 50–130, 1955.

[2] R. D. Bock and L. V. Jones. *The Measurement and Prediction of Judgment and Choice*. Holden-Day, San Francisco, 1968.

[3] N. Bonnier, F. Schmitt, H. Brettel, and S. Berche. Evaluation of spatial gamut mapping algorithms. In *14th Color Imaging Conference*, pages 56–61, Scottsdale, AR, 2006. IS&T/SID.

[4] N. Bonnier, F. Schmitt, M. Hull, and Leynadier C. Spatial and color adaptive gamut mapping algorithms. In *15th Color Imaging Conference*, pages 267–272, Scottsdale, AR, 2007. IS&T/SID.

[5] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs, i. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[6] U. Caluori and K. Simon. An RGB Color Management Concept based on an Improved Gamut Mapping Algorithm. In *Color Imaging XIV: Processing, Hardcopy and Applications*, volume 7241, page 724118A. SPIE, 2009.

[7] K. Chrzan and B. Orme. An overview and comparison of design strategies for choice-based conjoint analysis. *Research Paper Series*, 2000.

[8] CIE (Comité d'Etudes sur la Colorimétrie). *CIE Proceedings 1931*. Cambridge University Press, Cambridge, 1932.

[9] CIE (Comité d'Etudes sur la Colorimétrie). *Supplement No. 2 of CIE Publication 15: Recommendations on Uniform Color Spaces, Color-Difference Equations, Psychmetric Color Terms*. IS&T/SID, 1978.

[10] CIE (Comité d'Etudes sur la Colorimétrie). *CIE Publication 116: Industrial Colour-Difference Evaluation*. IS&T/SID, 1995.

[11] CIE (Comité d'Etudes sur la Colorimétrie). *CIE Publication 142: Improvement to Industrial Colour-Difference Evaluation*. IS&T/SID, 2001.

[12] CIE (Comité d'Etudes sur la Colorimétrie). *CIE Publication 156: Guidelines for the Evaluation of Gamut Mapping Algorithms*. IS&T/SID, 2004.

[13] CIE (Comité d'Etudes sur la Colorimétrie). *CIE Publication 159: A Color Appearance Model for Color Management Systems: CIECAM02*. IS&T/SID, 2004.

[14] H. A. David. *The method of paired comparison*. Hafner Press, New York, 1969.

[15] J. Dijk. *In search of an Objective Measure for the Perceptual Quality of Printed Images*. PhD thesis, Delft University of Technology, 2004.

[16] F. Dugay. Perceptual evaluation of colour gamut mapping algorithms. Master's thesis, The Norwegian Color Research Laboratory - Høgskolen i Gjøvik (Norway), 2007.

[17] F. Dugay, I. Farup, and J. Y. Hardeberg. Perceptual evaluation of color gamut mapping algorithms. *Color Research and Application*, 33(6):470–476, 2008.

[18] Fritz Ebner and Mark D. Fairchild. Developement and Testing of a Color Space (IPT) with Improved Hue Uniformity. *6th Color Imaging Conference*, pages 8–13, 1998.

[19] Peter G. Engeldrum. *Psychometric Scaling, A Toolkit for Imaging Systems Development*. Imcotek Press, Winchester MA, USA, 2000.

[20] A. M. Eskicioglu and P. S. Fisher. Image Quality Measures and Their Performance. *IEEE Transactions on Communications*, 43(12):2959–2965, December 1995.

[21] M. D. Fairchild. *Color Appearance Models*. Addison-Wesley, 1998.

[22] M. D. Fairchild and G. M. Johnson. iCAM framework for image appearance, differences, and quality. *Journal of Electronic Imaging*, 13:126–138, 2004.

[23] I. Farup, C. Gatta, and A. Rizzi. A multiscale framework for spatial gamut mapping. *IEEE Transactions on Image Processing*, 16(10):2423–2435, October 2007.

[24] I. Farup, J. Y. Hardeberg, and M. Amsrud. Enhancing the SGCK colour gamut mapping algorithm. In *Proc. Second European Conference on Color in Graphics, Imaging and Vision*, pages 520–524, Aachen, Germany, 2004.

[25] D. Gayle, H. Mahlab, Y. Ucar, and A.M. Eskicioglu. A full-reference color image quality measure in the DWT domain. In *13th European Signal Processing Conference*, September 2005.

[26] J. Giesen, K. Mueller, E. Schuberth, L. Wang, and P. Zolliker. Conjoint analysis to measure the perceived quality in volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1664–1671, November 2007.

[27] J. Giesen, E. Schuberth, K. Simon, and P. Zolliker. Image-dependent gamut mapping as optimization problem. *IEEE Transactions on Image Processing*, 16(10):2401–2410, October 2007.

[28] H. Grassmann. *Die lineare Ausdehnungslehre*. Wiegand, Leipzig, 1844, English translation, 1995, by Lloyd Kannenberg, *A new branch of mathematics*, Chicago, 1844.

[29] H. Grassmann. Zur Theorie der Farbenmischung. *Ann. Physik*, 89:69–84, 1853.

[30] J. Guild. The colorimetric properties of the spectrum. *Phil. Trans. Roy Soc. London*, A 230:149–187, 1931.

[31] H. Gulliksen. A least squares solution for paired comparisons with incomplete data. *Psychometrika*, 21(2), June 1956.

[32] J. Y. Hardeberg, E. Bando, and M. Pedersen. Evaluating colour image difference metrics for gamut mapping images. *Coloration Technology*, 124(4):243–253, July 2008.

[33] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[34] R. H. Hohle. An empirical evaluation and comparison of two models for discriminability. *Journal of Mathematical Psychology*, 3:174–183, 1966.

[35] A.C. Hurlbert, K. Wolf, B.E. Rogowitz, and T.N. Pappas. The contribution of local and global cone-contrasts to colour appearance: a retinex-like model. In *Proceedings of the SPIE 2002*, volume 4662, pages 286–297, San Jose, CA, 2002.

[36] J. E. Jackson and M. Fleckenstein. An evaluation of some statistical techiniques used in the analysis of paired comparison data. *Biometrics*, 13:51–64, 1957.

[37] Byoung-Ho Kang, Maeng-Sub Cho, Jan Morovic, and Ronnier M. Luo. Gamut compression algorithm development on the basis of observer experimental data. In *8th Color Imaging Conference*, volume 8, pages 268–272. IS&T/SID, November 2000.

[38] B. Keelan. *Handbook of Image Quality*. Marcel Dekker, Inc., 2002.

[39] O. Kolas and I. Farup. Efficient hue-preserving and edge-preserving spatial color gamut mapping. In *15th Color Imaging Conference*, pages 207–209. IS&T/SID, November 2007.

[40] E. H. Land. Recent advances in the retinex theory. *Vision Research*, 26:7–21, 1986.

[41] R. D. Luce and J.W. Tukey. Simultaneous conjoint measurement. *Journal of Mathematical Psychology*, 1:1–27, 1964.

[42] J. C. Maxwell. Experiments on color, as perceived by the eye, with remarks on color blindness. *Trans. Roy. Soc. Edinburgh*, 21:275–297, 1855/57.

[43] R. McDonald and K.J. Smith. CIE94 - a new color-difference formula. *Journal of the Society of Dyers & Colourist*, 111(12):376–379, 1995.

[44] A. A. Michelson. *Studies in Optics*. U. Chicago Press, 1927.

[45] A. B. Millen, J. Bunge, and J. C. Handley. Ranked data analysis of a gamut-mapping experiment. *Journal of Electronic Imaging*, 10(2):399–408, 2001.

[46] E. D. Montag. Empirical formula for creating error bars for the method of paired comparison. *Journal of Electronic Imaging*, 15(1):010502 1–3, 2006.

[47] N. Moroney, M.D. Fairchild, R.W.G. Hunt, C. Li, M.R. Luo, and T. Newman. The CIECAM02 Color Appearance Model. In *10th Color Imaging Conference*, pages 23–27. IS&T/SID, Nov 2002.

[48] J. Morovic. *To Develop a Universal Gamut Mapping Algorithm*. PhD thesis, University of Derby, UK, 1998.

[49] J. Morovic. *Colour Gamut Mapping*. WileyBlackwell, ISBN 0470030321, 2008.

[50] J. Morovic. *Colour Gamut Mapping*, chapter 12. WileyBlackwell, ISBN 0470030321, 2008.

[51] J. Morovic and Y. Yang. Influence of test image choice on experimental results. In *11th Color Imaging Conference*, pages 143–148. IS&T/SID, 2003.

[52] J. H. Morrisey. New method for the assignement of psychometric scale values from incomplete paired comparisons. *Journal of the Optical Society of America*, 45(5):373–378, 1955.

[53] F. Mosteller. 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*, 16:203, 1951.

[54] R. Schmidt, F. Lang, and G. Thews. *Physiologie des Menschen*. Springer, 29th edition, 2005.

[55] R. Sekuler and R. Blake. *Perception*. McGraw-Hill, 4 edition, 2002.

[56] Iris Sprow, Zofia Baranczuk, Tobias Stamm, and Peter Zolliker. Web-based psychometric evaluation of image quality. In *Image Quality and System Performance VI*, page 72420A. SPIE, 2009.

[57] L Thurstone. A law of comparative judgement. In *Psychological Review*, pages 273–286, 1927.

[58] H. v. Helmholtz. Über die Zusammensetzung von Spektralfarben. *Ann. Physik*, 94:1–28, 1855.

[59] B. Wandell. *Foundations of Vision: Behavior, Neuroscience and Computation*. Sinauer Press, 1994.

[60] Z. Wang and A. C. Bovik. A Universal Image Quality Index. *IEEE Signal Processing Letters*, 9(3):81–84, March 2002.

[61] W. D. Wright. A re-determination of the trichromatic coefficients of the spectral colours. *Trans. Opt. Soc. London*, 30:144–164, 1928–29.

[62] G. Wyszecki and W. Stiles. *Color Science*. Wiley-Interscience, 1982.

[63] Th. Young. On the theory of light and colours. *Philos. Trans. Roy Soc. London*, 92:210–271, 1802.

[64] P. Zolliker and K. Simon. Continuity of gamut mapping algorithms. *Journal of Electronic Imaging*, 15(1):13004, March 2006.

[65] P. Zolliker and K. Simon. Retaining local image information in gamut mapping algorithms. *IEEE Transactions on Image Processing*, 16(3):664–672, March 2007.