AALTO UNIVERSITY SCHOOL OF SCIENCE AND TECHNOLOGY
Faculty of Electronics, Communications, and Automation

Reijo Asikainen

Quality Analysis of a Printed Natural Reference Image

Master's thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Engineering

Espoo, May 3, 2010

Supervisor: Professor Pirkko Oittinen

Instructor: Raisa Halonen, M.Sc. (Tech.)

AALTO UNIVERSITY **ABSTRACT OF**
SCHOOL OF SCIENCE AND TECHNOLOGY **MASTER'S THESIS**

Author:	Reijo Asikainen
Title of the thesis:	Quality Analysis of a Printed Natural Reference Image
Date:	May 3, 2010 **Number of pages**: 87 + 21
Department:	Department of Electrical and Communications Engineering
Professorship:	AS-75 Media Technology
Supervisor:	Professor Pirkko Oittinen
Instructor:	Raisa Halonen, M.Sc. (Tech.)

This thesis was contributed to study the image quality properties of printing papers. The main goal was to produce an automatic, objective software system for predicting human opinion on the print quality of papers. To reach this goal, the project was divided into four phases: the development of a reference image for image quality evaluation, the assessment of subjective print quality from the reference image, the programming of quality analysis software for quality attributes, and the construction of a single grade for print quality, visual quality index (VQI). Four low-level quality attributes were studied: colorfulness, contrast, sharpness, and noise. Only inkjet printing technology was covered.

In the first phase, a natural reference image was developed for subjective and objective image quality testing. Focus was placed not only on quality aspects, but also on the high-level properties of the image, i.e. naturalness, balance, and aesthetical expression. Furthermore, presenting a unique feature for a reference image of this kind, seven GretagMacbeth test colors were implemented into natural objects in the image. During later phases, subjective tests were arranged to gather the subjective reference data of print quality for software development with Matlab. Finally, the computed quality attribute scores were combined with statistical regression analysis into a single grade for the print quality of papers, VQI, accompanied with individual regression models for the quality attributes.

The outcome of the software development was three functional and statistically accurate Matlab implementations, i.e. for colorfulness, contrast, and noise, complemented with a color difference method. The implementation of the VQI was successful as well, showing remarkably strong goodness measures. However, the generalization of the regression models was compromised by the strong cross-attribute similarity of the subjective reference data, eventually preventing the feasibility of the models in real world applications. Other issues requiring attention included handling the high colorfulness of the reference image and optimizing the software to the print context.

Keywords: reference image, image quality, print quality, inkjet, test colors, subjective quality, objective quality, photo aesthetics, computed quality attributes, visual quality index

AALTO-YLIOPISTON TEKNILLINEN KORKEAKOULU

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Reijo Asikainen
Työn nimi:	Painetun luonnollisen referenssikuvan laatuanalyysi
Päivämäärä:	3.5.2010 **Sivumäärä**: 87 + 21
Osasto:	Sähkö- ja tietoliikennetekniikan osasto
Professuuri:	AS-75 Viestintätekniikka
Työn valvoja:	Professori Pirkko Oittinen
Työn ohjaaja:	DI Raisa Halonen

Tämän diplomityön tarkoituksena oli tutkia paperin vaikutusta koettuun kuvanlaatuun. Päätavoitteeksi asetettiin automaattisen, objektiivisen ohjelmistojärjestelmän kehittäminen ennustamaan ihmisen arviota paperin kuvanlaatuominaisuuksista. Tutkimusprojekti koostui neljästä vaiheesta: testikuvan suunnittelusta kuvanlaadun tutkimukseen, subjektiivisen kokonaislaadun ja laatuattribuuttien arvioinnista testikuvasta, ohjelmiston kehittämisestä ennustamaan laatuattribuutteja sekä visuaalisen laatumallin muodostamisesta ilmaisemaan kokonaislaatua yhdellä laatuarvosanalla. Tutkimuksessa käsiteltiin neljää laatuattribuuttia: värikkyyttä, kontrastia, terävyyttä ja kohinaa. Painatusmenetelmänä käytettiin mustesuihkutulostusta.

Ensimmäisessä vaiheessa luotiin luonnollinen referenssikuva kuvanlaadun subjektiivista ja objektiivista arviointia varten. Suunnittelussa painotettiin laatuominaisuuksien lisäksi korkean tason ominaisuuksia, kuten luonnollisuutta, tasapainoa, ja esteettistä vaikutelmaa. Erityispiirteenä kuvaan lisättiin seitsemän GretagMacbeth testiväriä, jotka sisällytettiin kuvassa sijaitseviin luonnollisiin esineisiin. Seuraavassa vaiheessa suoritettiin subjektiivinen testaus ihmisen visuaalisen laatuarvion mittaamiseksi, josta saatuja laatuattribuuttien referenssiarvoja käytettiin objektiivisten laatumittojen suunnittelussa Matlab-ohjelmistolle. Lopuksi kehitetyt laatumitat yhdistettiin tilastollisen regressioanalyysin avulla yhdeksi arvosanaksi paperin kokonaislaadusta, ns. visuaaliseksi laatumalliksi. Myös laatuattribuuteille muodostettiin regressiomallit.

Tutkimuksen tuloksena luotiin toimivat ja tilastollisesti tarkat objektiiviset mitat kolmelle laatuattribuuteille: värikkyydelle, kontrastille ja kohinalle. Lisäksi kehitettiin mitta värivirheen laskentaan. Myös visuaalisen laatumallin toteutuksessa onnistuttiin hyvin, ja kaikkien regressiomallien selitysasteet olivat tilastollisesti korkeita. Subjektiivisten arvosanojen samankaltaisuus laadun ja laatuattribuuttien välillä johti kuitenkin ongelmiin regressiomallien yleistämisessä, mistä johtuen mallien käyttöä ei voitu suositella reaalimaailman sovelluksissa. Erityistä paneutumista vaativat myös testikuvan suuri värikkyys sekä ohjelmallisten laatumittojen optimointi paperi- ja painatusympäristöön.

Avainsanat: referenssikuva, kuvanlaatu, painatuslaatu, mustesuihku, testivärit, subjektiivinen laatu, objektiivinen laatu, valokuvan estetiikka, lasketut laatuattribuutit, visuaalinen laatumalli

Preface

With this master's thesis, I conclude my studies toward the degree of Master of Science in communications engineering. The research for the thesis was done in the Department of Media Technology at Helsinki University of Technology as a part of the DigiQ project.

Firstly, I want to express my gratitude to my supervisor Professor Pirkko Oittinen for offering me an interesting and challenging thesis topic to work with and providing me useful tips and valuable insight during the research process. Additional thanks go to my instructor Raisa Halonen for sharing her knowledge and expertise throughout the thesis. I also want to thank all my fellow workmates in the Department of Media Technology.

Further thanks are addressed to my friend Inka Tiainen for modeling in the reference image.

Finally, I wish to thank my parents and the whole family for the endless support during my studies and throughout the graduation.

Helsinki, May 3, 2010

Reijo Asikainen

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Background

This thesis is conducted as a part of the DigiQ research project, Fusion of Digital and Visual Print Quality (TKK Media Technology n.d.). The main aim of the DigiQ project is to construct an automatic system to grade printing papers in terms of visual image quality. The automatic system is based on software algorithms, placing an obvious requirement for the system: an analog print needs to be digitized before processing. To reach the goal of automatic print quality evaluation, the system requires following process phases: the printing of a test image to a sample paper, the digitizing of the test image, analyzing paper quality properties from the digitized image with computational methods, and constructing a visual quality index (VQI) to present a visual quality appearance of the print.

The following design goals are placed on the paper evaluation system in the DigiQ project. For printing, the system accepts any image with unrestricted image content, printed with any digital printing method, to any paper grade. After digitizing a print with a predefined scanning system, the digitized print is analyzed from two viewpoints: image quality attributes and image content features. Since the system needs to function with any image available, no-reference methods are required for objective calculations. Lastly, a VQI model is developed to combine the analyzed image quality characteristics into a single grade for print quality.

This thesis concentrates mainly on the last two parts: designing algorithms to analyze image quality attributes, and constructing a VQI model from the designed algorithms. Image content analysis with objective methods is not covered. Furthermore, as a difference to the original DigiQ project goals, the algorithms are optimized only for one image, a specific reference image designed for image quality evaluation. Earlier in the DigiQ project, the covered aspects are studied mainly on two projects: the reference image development in Salmi (2008), and the objective analysis of papers in Halonen (2008a).

1.2 Challenges

Expanding the concept of image quality from digital images to digitized prints presents numerous new challenges and problem areas for quality assessment. The areas of concern are mainly related to objective, software based quality testing, while the procedure for subjective, human driven evaluation is relatively similar for both the cases. One challenge to address is image scanning, which, when done manually, is prone to user errors. A variation on paper placement between scans, for instance, may cause unexpected difficulties for objective methods operating in image pixel level (Eerola et al. 2009). Furthermore, a scanned image is not only influenced by the quality of the paper and the test image, but also by the printing technology. In the DigiQ research by Halonen (2008a), major differences were observed in the functionality of objective quality measures between two digital printing technologies: inkjet, and

electrophotography. This may lead to a need for individual methods for each printing technology, placing excessive stress for software development. Further issues include the handling of printing raster and scanning moiré, not to forget errors caused by printer and scanner (Gatter 2004, pp. 111-116). However, these issues are already considered in Halonen (2008a), where a complete digitizing system with optimized printing and scanning settings was introduced.

In addition to the concerns in scanning and printing, a paper itself presents a major challenge for quality evaluation. Most of the objective quality algorithms available till the date are designed to evaluate digital images, and the correct functionality in a print context cannot be assured without further testing. One of the problem areas is paper noise or graininess, showing quite different noise characteristics than e.g. digital noise generated by digital cameras (Halonen 2008a). Thus, when choosing objective quality algorithms for print paper quality, extensive testing and adjusting of algorithms may be needed to achieve accurate results.

Another area of concern is the variation of subjective quality estimations with changing image content. Considering the fact that visual quality must be based on subjective quality evaluations, objective ratings should be as close the subjective counterparts as possible. However, based on the findings of multiple DigiQ researches, subjective quality observations are not stable when images with variable image content are evaluated, presenting a problematical situation for objective analysis (Eerola et al. 2008a, Halonen 2008a, Leisti et al. 2008). Furthermore, the objective analysis of images with diverse themes requires no-reference methods, placing even more pressure to software development.

If the requirement of multiple images is abandoned, the situation simplifies considerably. With a single reference image, less complicated reduced-reference or full-reference methods can be developed. Furthermore, the subjective and objective estimations may be evaluated from the same image, thus increasing result comparability. This option was studied in a DigiQ study by Salmi (2008), where a single reference image was developed for the use of both the subjective and objective evaluation, with further development goals on image naturalness and balance.

1.3 Objectives and restrictions

In this thesis, the main objective is to

- design a natural, balanced, and aesthetically pleasing high quality reference image suitable for subjective and objective image quality assessment,
- arrange subjective tests to evaluate the image quality of papers from the printed reference image,
- develop software to calculate low-level image quality attribute scores from the digitized reference image prints, and
- expand software to compute a single grade for the print quality of papers, visual quality index, from the calculated quality attribute scores.

Furthermore, to address the research challenges in the scope of one Master's Thesis, the following restrictions are set to the original DigiQ research goals in this study. Firstly, the functionality of objective methods is restricted to only one reference image, the reference image developed in this study. Secondly, electrophotography is excluded and only inkjet prints are studied. Thirdly, no-reference methods are not required in objective development.

The first research objective is placed to develop a second reference image for the DigiQ project. The research findings by Salmi (2008) are considered during development, seeking for improvements in image naturalness, balance, and quality evaluation aspects. Furthermore, the new reference image includes pre-defined test colors to allow the objective analysis of color accuracy. The second objective is required to provide subjective reference data for objective software calculations. The subjective tests cover overall quality and four quality attributes: colorfulness, contrast, sharpness, and noise. As the third objective, objective measures are developed for the four quality attributes. Lastly in the fourth objective, calculated quality attribute results are combined into one grade for print quality, the VQI.

The four quality attributes of colorfulness, contrast, sharpness, and noise, are selected mainly because of their extensive use in earlier DigiQ projects. In addition to previous usage in Salmi (2008) and Halonen (2008a), Leisti et al. (2008) found the quality attributes closely related to sharpness, colorfulness, and contrast to be especially suitable for describing image quality of natural images in subjective evaluation. In objective evaluation, Eerola et al. (2008a) considered sharpness and contrast as one of the most important objective image quality measures.

In this study, the selected attributes fulfill three important aspects of quality evaluation. Firstly, the attributes are simple, common and easily explained to test subjects in a subjective test situation. Secondly, the attributes can be calculated with objective methods and a large number of algorithm implementations are available for digital images. Thirdly, the attributes provide different perspectives on paper quality, an important aspect when constructing a single quality grade based on multiple attributes.

1.4 Structure of thesis

The thesis consists of two parts: the literature review in Chapter 2, and the experimental part in Chapters 3-6. Chapter 2 gives a brief overview on the aspects underlying image quality in a print context with a slight emphasis on color aspects, and presents the descriptions of the terms and methods used in this thesis. Detailed descriptions, however, are not provided due to the large scope of the thesis. Further information can be found e.g. in Salmi (2008) for subjective evaluation, Halonen (2008a) for objective measures, and Kivinen (2009) for color evaluation.

Starting the experimental part, Chapter 3 presents the methodology and the procedure of the study, describing how everything was done during the research. The results are listed in Chapter 4 with further discussion in Chapter 5. Lastly, Chapter 6 presents conclusions by summing up the research procedure and the accomplishments of the thesis.

2 ASPECTS OF IMAGE QUALITY

2.1 Image quality

2.1.1 Definitions

The concept of image quality can be approached by first defining good quality. According to Johansson et al. (2007, pp. 150-151), good quality has always two sides: subjective opinions and an image itself. Subjective opinions are affected by complex physical and psychological parameters, while image goodness is simpler to define. For a good quality image, properties like optimal photography, technical excellence, and natural color reproduction are required. Furthermore, when assessing many images at the same time, the balanced and equal output of the images is essential for good quality experience.

Janssen (2001, pp. 19-25) expands the idea of a good image by presenting two requirements for visual information in the image: precision and reliability. For meeting the requirements, Janssen (2001) emphasizes the importance of successful image interpretation by an observer, leading thoughts to a secured observation situation and the suitable high-level characteristics of the observed image. For general image quality, Janssen (2001) lists two defining measures: image usefulness and naturalness, where usefulness refers to the precision of the internal representation of an image and naturalness to the correspondence between the internal precision and observer's knowledge of reality.

Fairly similar allocation is presented by de Ridder and Endrikhovski (2002), who divide the quality of a reproduced image into three categories: fidelity, usefulness, and naturalness. With the fidelity category, de Ridder and Endrikhovski (2002) refers to the reproduction accuracy of an observed image in comparison to the original, perfect image. The second category, usefulness, indicates image suitability for the designed task, while naturalness category is defined similarly as in Janssen (2001), the match between a reproduced image and the mental impression of an observer, affected e.g. by memory traces (de Ridder & Endrikhovski 2002).

When discussing of visual quality, it is important to understand the distinction between image quality and image fidelity. According to Silverstein and Farrell (2004), image fidelity refers to the ability to discriminate two images from each other. An image with high fidelity is reproduced accurately, without any visible information loss or distortion. Image quality, on the other hand, explains an ability to form preferences between images (Silverstein & Farrell 2004). While subjective conditions affect both the cases, image quality observation is based on the own preferences and impressions of an observer. For instance, adding a vignetting effect on a photograph decreases image fidelity, but may be regarded as a positive aspect on image quality. Thus, because of the complexity of the concept, the characterizing or quantifying image quality presents a major challenge (Silverstein & Farrell 2004).

2.1.2 Subjective and objective quality

Fundamentally, image quality is always an outcome from human sensation. As discussed in the previous chapter, human observers make the final decisions about quality based on their own visual preferences that, naturally, are not only affected by the psychophysical aspects of the observer, but also by e.g. the fidelity of the image and the observation situation. For evaluating image quality, testing with human observers, i.e. subjective evaluation, is often considered the most reliable way to estimate the quality of images (Wang & Bovik 2006, pp. 1-3). From subjective evaluation measures, the mean opinion score (MOS) is the most widely used. MOS is often regarded as the most reliable image quality measure, but since it requires numerous human observations and a specific test arrangement, MOS is also slow and expensive method in real world situations (Wang & Bovik 2006; Pappas et al. 2005).

Another option for quality evaluation, objective assessment, relies on computational models that can predict the image quality observations of humans. According to Wang and Bovik (2006, pp. 1-3), an accurate objective image quality model predicts the image quality sensation of an average human observer. In other words, strong correlations to subjective observations are essential when defining a good objective quality model. Since image quality is strongly based on subjective observations, traditional objective models such as the mean-squared-error (MSE) rarely work accurately on a quality context (Seshadrinathan et al. 2005; Wang & Bovik 2006). A good objective model may, for instance, exploit the knowledge of human visual system (HVS) in calculations.

In designing objective quality methods, three types of knowledge can be exploited in development: the knowledge of the original image, the knowledge of the process causing image distortion, or the knowledge of HVS and the subjective processes affecting quality observation (Wang & Bovik 2006, pp. 11-15). Based on the knowledge reference, measures can be described with three commonly used categories: full-reference, reduced-reference, and no-reference measures. Full-reference methods present the most popular category, describing methods where evaluation is based on the comparison of an original, perfect image and a distorted sample image. The most image quality methods in literature are full-reference methods. In many practical applications, however, an original reference is not available, and no-reference methods must be applied. These “blind” no-reference methods are very complex to develop with software, but for a human, who already has the knowledge what the image should look like, quality observation without a reference is a rather uncomplicated task. In the category between full-reference and no-reference methods are reduced-reference methods. The principle of a reduced-reference based quality evaluation system is illustrated in Figure 1. (Wang & Bovik 2006)

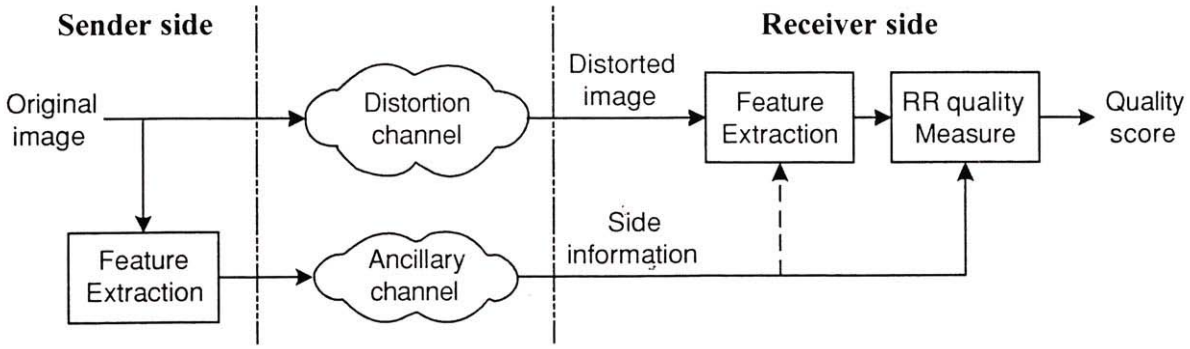


Figure 1. *A reduced-reference image quality evaluation system (Wang & Bovik 2006, p. 13)*

With reduced-reference methods, the original image is not fully available (Wang & Bovik 2006, pp. 11-15). The evaluation system is constructed in the way that only certain features of the original image are gathered and exploited in quality calculations. The extracted information is used to guide the functionality of the method, helping to identify the reasons for image distortion. Figure 1 illustrates the idea by presenting one possible implementation of a reduced-reference system, i.e. a video transfer system, where extracted image information is transferred through a secondary channel to a measurement system and analyzed with a distorted image. However, a complex evaluation system as shown in Figure 1 is not required for reduced-reference measures. A reduced-reference method can also rely on simple side information extracted from the original image, e.g. on metadata or structural image characteristics (Wang et al. 2004).

2.1.3 Image quality attributes

To describe image quality, specific image quality attributes are needed. Leisti et al. (2009) divide quality attributes to two categories: low-level, concrete attributes; and high-level, abstract attributes. For low-level attributes, e.g. colorfulness and sharpness, there exists a physical, measurable counterpart in quality evaluation. High-level attributes, e.g. naturalness and usefulness, are strongly connected to observer's visual perception and experience from image content, making them harder to quantify. There lies, however, an interesting connection between high and low-level attributes. According to Leisti et al. (2009), high-level attributes guide an observer in quality sensation by clarifying the meaning of low-level attributes for general quality. Therefore, if asked to describe a reason for a certain low-level quality attribute rating, people tend to use high-level attributes. For example, an observer may prefer a less colorful image because it is more natural. (Leisti et al. 2009)

In this thesis, four low-level quality attributes, i.e. colorfulness, contrast, sharpness, and noise, are studied in subjective and objective print context. Brief descriptions of the listed attributes are presented in this chapter, supplemented with example objective methods strongly related to the software implementations developed in this study. As examples for high-level attributes,

literature usually lists attributes like naturalness, clearness, realism, usefulness, and depth. This thesis studies high-level quality attributes only in subjective assessment, concentrating on image naturalness and, in some extent, on usefulness.

Starting from the low-level attributes of this study, Fairchild (2005, p. 87) defines colorfulness as the visual sensation of the chomatic appearance of an image, explaining the intensity of color hue for a given color stimulus. Similarly, Maureen (2003, pp. 23-24) connects colorfulness to image chroma by presenting Munsell color-order system visualized in Figure 2. In the illustration, image chroma with two color axis, red–green and blue–yellow, is defined as colorfulness. Other components of color perception are lightness, i.e. value, and hue, i.e. the placement of color in the color spectrum.

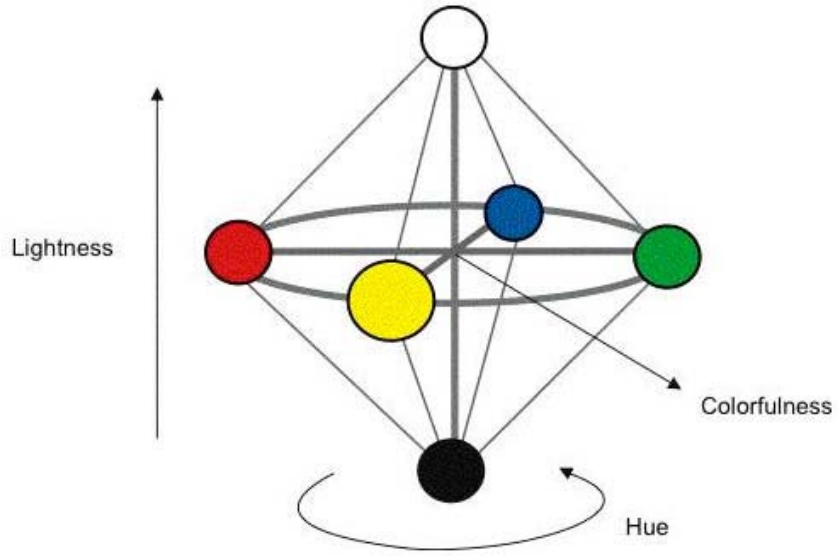


Figure 2. *Munsell color-order system, showing the perceptual structure of color (Maureen 2003)*

Hasler and Süsstrunk (2003) studied the colorfulness appearance of natural digital images with objective methods. Based on the testing of numerous colorfulness methods with diverse weighting coefficients, two metrics for images in CIELAB color space were introduced for objective colorfulness calculations:

$$M^{(1)} = \sigma_{ab} + 0.37 \times \mu_{ab} \quad \text{and} \quad M^{(2)} = \sigma_{ab} + 0.94 \times \mu_{C}\,, \qquad (1\ \&\ 2)$$

where σ_{ab} is the trigometric length of the standard deviation in a*b* space,

μ_{ab} is the distance of the center of gravity in a*b* space to the neutral axis, and

μ_{C} is the mean of chroma (Hasler & Süsstrunk 2003).

To attain a more comprehensive estimate on color quality, Hasler and Süsstrunk (2003) recommend supplementing the colorfulness metric with a color cast estimation method.

In digital processing, contrast is universally defined as an intensity difference between the highest and the lowest intensity levels in an image (Gonzalez and Woods 2008, p. 58). From the perspective of a photographer, Burian (2004, p. 250) defines contrast as the ratio of the lightest and the darkest elements in an image. For the measurement of contrast, Peli (1990) states that uniquely defined methods for contrast evaluation are not available in literature. As an example for objective contrast evaluation, Peli (1990) presents a common method for comparing contrast between two images, the root-mean-square (rms) contrast. Basically, the rms contrast is the standard deviation of image luminance values, defined as

$$rms_{contrast} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\,, \tag{3}$$

where x_i is the normalized grey level value in the range of [0, 1], and

$\bar{x}$ is the mean normalized grey level (Peli 1990).

In a print context, Kipphan (2001, p. 1122) defines sharpness as the reproduction quality and detail rendering of contrast gradients, particularly lines and picture contours, while the reference sharpness is defined by the output resolution of an original and printing equipment. In the domain of digital images, sharpness is often indicated by image blur. Marziliano et al. (2002) define blur as the attenuation of high spatial frequency values in an image. In objective blur analysis, the frequency characteristics can be studied with image filtering, e.g. by applying Sobel filters to reveal the edge structure of an image. Edges are indeed important in blur analysis, as Wang et al. (2008) regard image edges as the most important image source for no-reference measures. Gonzales and Woods (2008, pp.165-168) list two Sobel filters, a horizontal and a vertical, for finding sharp edges in the corresponding dimension. The entire Sobel approximation and the edge approximations G_x and G_y are calculated in the spatial domain as

$$G = |G_x| + |G_y|, \tag{4}$$

$$\text{where } G_x = I(x,y) * \begin{vmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{vmatrix} \quad \text{and} \quad G_y = I(x,y) * \begin{vmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{vmatrix}. \tag{5 \& 6}$$

In a no-reference blur metric by Marziliano et al. (2002), edge detection filters such as Sobel filters were exploited to analyze the spread of edges in an image. The execution of the method starts by vertical Sobel edge filtering, marking the edge locations in the image. Only luminance data is needed. Then the original image is scanned one row at the time, marking the closest local maximum and minimum value for each edge location point. The spread of an edge is defined as the difference of these local extremes. After the entire image is analyzed,

the overall blur measure is obtained by averaging the local blur values over all edge locations. (Marziliano et al. 2002)

Image noise has many characteristics, which strongly depend on the used electrical equipment or reproduction media (Johansson et al. 2007, p. 159). With digital images, Burian (2004, p.257) regards noise as random, unintended pixel values distributed across an image, a common byproduct of e.g. the amplification of digital signals. In a print context, noise can be caused by paper grain, showing as a gritty look on a print (Johansson et el. 2007). To estimate noise of digital images with objective methods, some sort of pixel masking is usually required to reveal the variance characteristics of pixels. Immerkær (1996) proposed a no-reference noise metric for estimating the variance of additive zero mean Gaussian noise in digital images. In the metric, a zero mean noise estimation operator *N* is constructed from two Laplacian masks as

$$N = 2(L_2 - L_1) = \begin{vmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{vmatrix}, \tag{7}$$

$$\text{where } L_1 = \begin{vmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{vmatrix} \quad \text{and} \quad L_2 = \frac{1}{2} \times \begin{vmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{vmatrix}. \tag{8 \& 9}$$

The variance of noise can be then calculated by applying the resulting mask *N* in the image *I* using convolution in the spatial domain with the equation

$$\sigma_n^2 = \frac{1}{36(W-2)(H-2)} \sum_{\text{image } I} (I(x,y) * N)^2, \tag{10}$$

where *W* and *H* are the width and height of the image,

I(x,y) is the image *I* at position (x,y), and

N is the applied estimation operator (Immerkær 1996).

Regarding the high-level attributes of this study, image usefulness is typically a well recognized term, and can be described as the suitability of a reproduced image to the intended task (Yendrikhovskij 2002, pp. 397-399). Image naturalness, another important high-level attribute, is commonly regarded as a fundamental aspect in high quality images (Fernandez & Fairchild 2002). Indeed, naturalness has been listed as one of the most important high-level attributes in image quality research (Leisti et al. 2009; Kim et al. 2006). Yendrikhovskij (2002, pp. 397-399) explains image naturalness as the extent of visual match between the reproduced image and the internal references of an observer, e.g. memory prototypes related to image content. Furthermore, Yendrikhovskij (2002) states that naturalness becomes important in subjective quality judgements especially when an external reference, e.g. an original image, is not available. This situation is very common in everyday life, where pictures are usually seen in their

reproduced form without any reference to compare with. Reading a magazine, for instance, or browsing through photographs in the Internet are situations without real life references.

2.2 Print quality

2.2.1 Quality of digital prints

Even though the term image quality is often used in a digital print context, a printing process itself mostly affects image fidelity, i.e. how accurately a print is reproduced in terms of the original image (Pappas et al. 2005). The affecting factors on reproduction include e.g. a printing process, technology, and material (Kipphan 2001, p. 99). Starting from the main segmentation, Johansson et al. (2007, pp. 325-336) list two main techniques for printing: the printing press and digital printing. The former offers capacity for large print-runs with high quality, but requires the preparation of a printing plate for each print. The latter is more common in non-professional use, as it relies on digital source data and allows quick printing when needed.

Kipphan (2001, p. 99) approaches print quality from a process perspective by introducing four main factors that affect the reproduction quality of prints: prepress, print, postpress and material. In digital printing, the most important factors are the prepress, print and material. The prepress stage includes image scanning, screening, system calibration, and color management. In the print stage, the printing technology and ink-to-material transportation have an effect on quality, as well as printing materials, such as printing ink and paper. (Kipphan 2001)

For digital printing, three main technologies are available: electrophotography, inkjet and dye sublimation (Johansson et al. 2007, pp. 325-336). The first two, presenting the most common consumer printing techniques, are discussed here to give an insight on differences between printing methods. Described by Johansson et al. (2007), electrophotography, also called as xerography, is a technology extensively used in laser printers, copiers and digital printing presses where copy counts of few hundred prints are needed. Inkjet, while being slower and more expensive to use, matches better for color photo printing and thanks to low purchase prices, is a very common choice for home use.

Quality wise, the main differences between electrophotography and inkjet techniques are caused by different ink characteristics and how the ink is transported to a paper. Electrophotographic printers use non-liquid ink consisting of very small particles, called as toner. The ink is pressed on the paper by a rotating drum and burned onto the paper afterwards. Inkjet printers, on the other hand, use liquid ink and spray it straight on the paper. The liquid ink soaks into the printing paper, in contrast to electrophotography, where the toner is left on the top of the paper surface. The effect can be easily observed by touching the print surface. An electrophotographic print feels rougher than the corresponding inkjet print. (Johansson et al. 2007, pp. 328-336; Kipphan 2001, pp. 60-65)

As listed by Kipphan (2001), printing material such as ink and paper should be also considered in print quality determination. For printing ink, there rarely exists many options of choice in

digital printing. The selection of papers, however, is large in commercial printing. Levlin and Söderhjelm (1999 cited in Eerola et al. 2008a) list seven most important and standard physical properties of papers: grammage, whiteness, brightness, opacity, PPS roughness, Bendtsen roughness and gloss. The descriptions of the terms are presented in Eerola et al. 2008b. Furthermore, after experimenting with the listed paper properties, Eerola et al. 2008a concluded in three most important physical paper measures: gloss, whiteness, and brightness.

In addition to the reproduction aspects like printing technology and paper properties, many complex subjective processes affect the sensation of image quality. An abstract level based approach on print quality by Eerola et al. (2008a) presents the information flow from physically measured quantities to subjective visual perception, ending to the human visual quality experience. Figure 3 shows the levels of the quality experience.

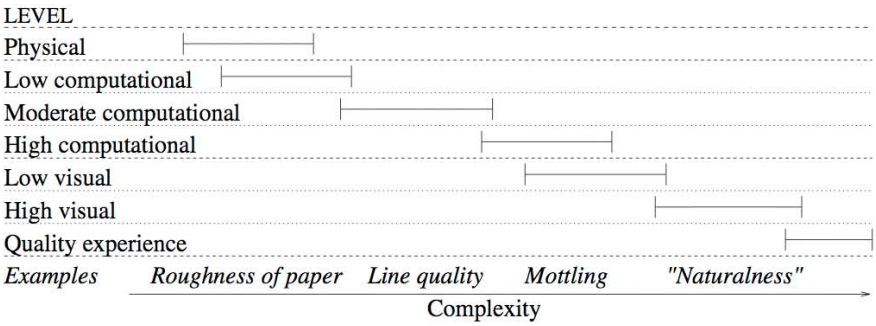


Figure 3. *The levels of print quality evaluation (Eerola et al. 2008)*

In the model by Eerola et al. (2008a), the process of image quality perception is divided in three main levels: physical, computational, and visual. The physical level includes physically measurable quantities, such as paper properties and printing technology aspects. The visual level presents a subjective point of view, containing measures for human visual observation. Between the two lies the computational level, where the physical and visual levels are combined with computational, objective methods. Above all the listed levels exists the quality experience, the final perception of print quality. (Eerola et al. 2007 & 2008a)

2.2.2 Objective print quality evaluation

Print quality evaluation differs quite strongly from the quality measurements of digital images. The main difference lies in the format: print quality is observed from a printed paper, while digital images can be evaluated from a computer monitor or any digital display. In subjective quality assessment the situation is quite similar, but in objective evaluation the process is different. While digital images are often measured automatically with computer software, print quality is usually observed manually from specific test targets using external devices, such as densitometer for e.g. color saturation measurements or spectrometer for spectral color accuracy measurements (Kipphan 2001, pp. 100-108). However, instrumental measurements usually require human attention and are thus relatively slow to use. To speed up the measurement

process, print quality can be assessed automatically by using sophisticated computer software as is the case with normal digital images. This requires the digitizing of an analog print with a scanner.

Chapter 3.2 describes one possible solution for a scanning process, which is also used in the experimental part of this thesis. After scanning, a digital print can be measured with the same automatic objective methods and software algorithms than any digital image. However, there lies a major problem. Paper characteristics and a printing technology have a major impact on a scanned image. In a recent research conducted by Halonen (2008a), less than a half of the objective quality evaluation methods worked as expected with scanned digital prints. The main cause for inaccuracies was supposedly the variable noise characteristics of scans compared to normal digital images, which negatively affected the functioning of edge detection algorithms in sharpness evaluation, overall noise measures and colorfulness estimations. Similarly, Halonen (2008a) detected major variation in algorithm functioning between printing methods, namely inkjet and electrophotographic prints. Only the simplest of the algorithms, a standard deviation based contrast algorithm, showed expected results with both the printing methods.

In addition to paper noise, the possible movement of a paper during scanning may cause major additional workload on objective method development and if unattended, serious errors to end results. This is especially the case with many objective full-reference measurement methods where a direct, pixel-level comparison to the original image is required (Eerola et al. 2009). Furthermore, because of the halftoning and screening technologies needed in printing, the effect of raster must be dealt with during a digitizing process. This can be accomplished by e.g. using relatively low resolution for scanning (Halonen 2008a, Halonen 2008b).

2.3 Color in images

2.3.1 Color fundamentals

To understand color reproduction, we must first assimilate the general aspects of color. Basically, color is just a sensation, a product of the human mind. It is formed from visible light perceived by the eyes of an observer and processed to visual information by the brain and the human visual system. What comes to light, it is one type of electromagnetic radiation, similar to radio waves or X-rays for instance. The eyes are able to detect the specific wavelengths and frequencies of electromagnetic radiation, i.e. the wavelengths between approximately 385 to 705nm. These wavelengths form the sensation of colors from deep violet to blue and through green and yellow to deep red. The wavelengths outside the red spectrum are called as infrared waves, which are sensed as heat energy. Above the violet end are ultraviolet rays, which e.g. have energy to tan human skin. Both of these names refer to the frequencies of equivalent colors. (Fairchild 1998, p. 64; Fraser et al. 2005, pp. 4-9; Johansson et al. 2007, pp. 69-70; Kipphan 2001, p. 68)

Clayton and Hashimoto (2009, pp. 235-236) divide colors in three categories: the primary, secondary, and tertiary colors. The primary colors, i.e. red, yellow, and blue, are the three main colors from which all the other colors are formed. The secondary colors are mixtures of the primary colors. Orange is made of red and yellow, purple from red and blue, and green from blue and yellow. The six tertiary colors are mixed from the primary and secondary colors located next to them on the color wheel. The color wheel showing all the 12 hues of colors is shown in Figure 4. In addition to hue, every color has a value, indicating the darkness and lightness of the hue. (Clayton & Hashimoto 2009)

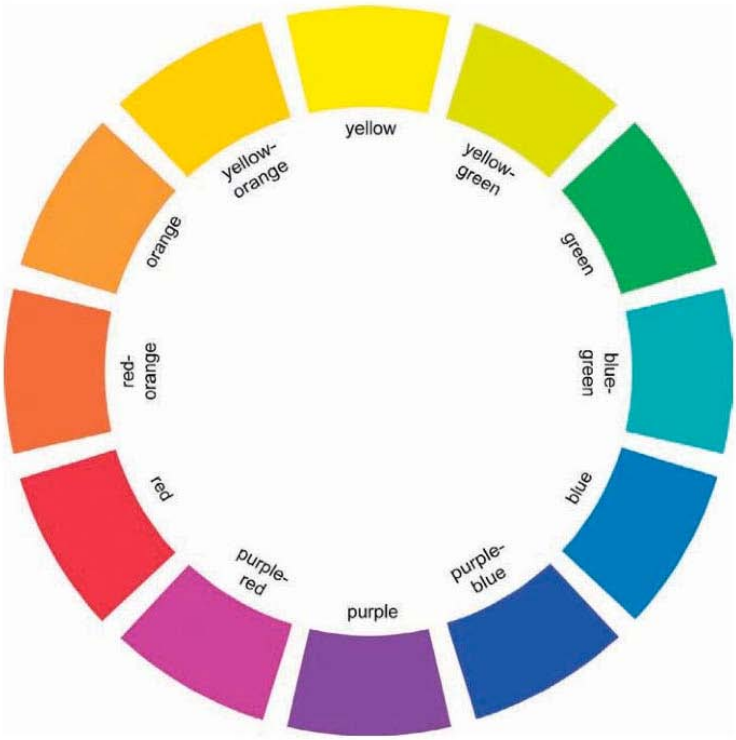


Figure 4. *The 12-hue color wheel to indicate primary, secondary, and tertiary colors. The complementary colors are places on the opposite sides of the wheel. (Clayton and Hashimoto 2009)*

The color wheel introduces an interesting aspect of colors, complementary colors. The complement color pairs are presented in the opposite sides of the wheel, e.g. red – green and yellow – purple. By adding the complementary color, the intensity of the main color decreases. For instance, mixing equal amounts of red and green results in muddy grey. Another feature linked to complementary colors is related to visual perception. Complementary color pairs attract attention on humans, especially when the colors are presented at full intensity. The complementary pairs give a lively and active feeling to a color shame, amplifying the sense of excitement in the image. (Clayton & Hashimoto 2009, pp. 237-240)

Field (1999, p. 39) defines color as a complex visual sensation that is influenced by the physical properties of a light source and a reflecting sample as well as the physiological and psychological characteristics of an observer. Fairchild (1998, pp. 64-66) expands the definition by presenting the connections between these components in the triangle of color, which is visualized in Figure 5.

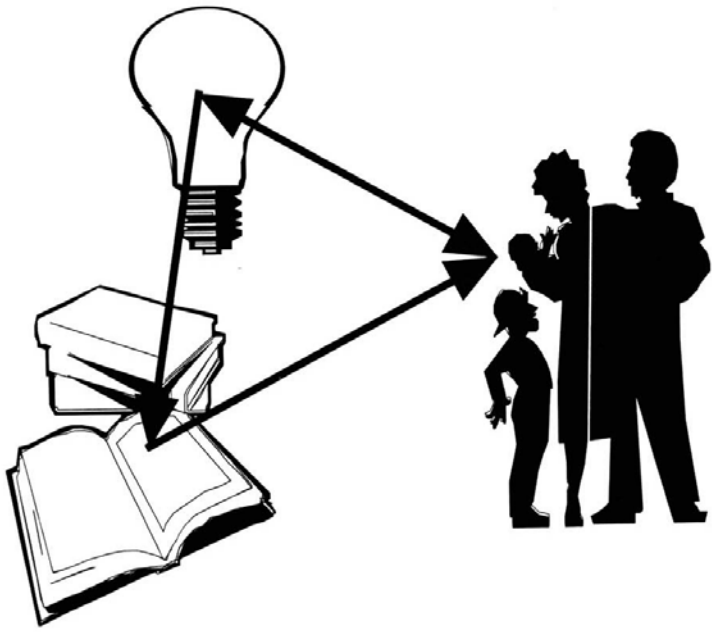


Figure 5. *The triangle of color, presenting three components needed for a color sensation: light sources, objects, and observers (Fairchild 1998, pp. 66).*

To produce color, three components are needed: a light source, a reflecting surface and an observer. Basically, in the color sensation process, the surface reflects the source light to the observer who perceives the reflection as color. The perceived color is affected by the spectral frequency characteristics of the light source as well as the physical characteristics of the reflecting area. The surface absorbs some frequencies of light and reflects others. The photoreceptors of the human eye catch the reflected light spectrum, and the human visual system processes the information to form a sensation of color. However, it is important to note that the reflected light is not the only light having an influence on color. The light source itself also affects color perception directly through numerous physical mechanisms of the human visual system, for example by adapting to dominant lighting through chromatic adaptation. (Fairchild 1998, pp. 64-66)

Ultimately, color evaluation has always two sides. It is not only affected by the physical properties of illumination and a scene, but also by human perception. While physical properties can be measured, e.g. with a colorimeter, human sensations can only be estimated (Kipphan 2001, pp. 68-69). Like in the case of image quality evaluation, this, undoubtedly, presents a major challenge for color assessment.

2.3.2 Color reproduction

In a print context, color reproduction quality is one of the most important print quality aspects (Kipphan 2001, p. 99). Field (1999, p. 1) describes color reproduction as “a process of producing a color image from some form of original scene or object”. Maureen (2003, p. 88) adds the main goal of reproduction: to create an output that looks like the original. Indeed, the perfect reproduction is not usually essential or technically achievable, since every reproduction medium has its own characteristics, and no medium can perfectly match the visual perception of the human eye (Johansson et al. 2007, p. 69-71; Maureen 2003). Thus, a certain level of compromises cannot be avoided in color reproduction (Maureen 2003).

To achieve good color reproduction, Hunt (1987, pp. 43-44) lists four requirements for successful color reproduction in the order of importance: the correctness of hue, the correctness of lightness, colorfulness proportional to original, and colorfulness and brightness similar to original. From the listed requirements, the first two are the most relevant. The first requirement emphasizes the importance of correct color balance in color reproduction. The color balance errors are easily visible in e.g. pale Caucasian skin or in neutral grey areas. With the second requirement, Hunt (1987) refers to tone reproduction and contrast levels. The correct contrast level looks natural, while high contrast results in a gaudy and exaggerated appearance and low contrast in a hazy and misty rendition. The third recommendation is connected to a special occasion with blue sky and the high ultra-violet sensitivity of certain photographic materials. The fourth recommendation is related to viewing conditions, explaining a situation where a high contrast image is observed in low light, resulting in a limited visual brightness and colorfulness sensation. (Hunt 1987)

Maureen (2003, pp. 93-94) presents a modern approach on good color reproduction by listing five principles of color image reproduction: correct mapping of memory colors, correct reproduction of neutral and white colors, control of the tone reproduction, control of the overall colorfulness, as well as control of sharpness, texture, and other visual artifacts. Similarly to Hunt (1987), Maureen (2003) raises the reproduction of memory colors as one of the most important aspects. Contrast and brightness are also mentioned in terms of the tone control. Furthermore, Maureen (2003) recommends a civilized approach on colorfulness management and warns about color casts in neutral colored image areas. The last principle refers to possible ill-effects in image reproduction, e.g. in printing, where the halftoning patterns may affect detail rendering.

Kipphan (2001, pp. 79-80) discusses color reproduction in a print context, and presents two reproduction categories: faithful reproduction, and appeal-optimizing reproduction. In faithful reproduction, the goal is to maximize the color similarity between the original and the reproduced sample. With faithful systems, automatic processes can be developed to maintain reproduction quality. With appeal-optimizing systems, however, color reproduction is context-dependent, and reproduction adjustments must be done manually. In both the cases, the ob-

serving situation, e.g. lighting conditions, must be taken into account in reproduction quality assessment. (Kipphan 2001)

As in general image quality evaluation, the assessment of color reproduction quality can be based on subjective or objective perspective. Subjective evaluation, however, is harder to execute, especially when an original reference is not available (Kipphan 2001, p. 69). In these no-reference cases, a human observer has to rely on memory traces and impressions how the colors should look like, a situation where the presence of memory colors is helpful (Hunt 1987, pp. 36-39). Memory colors and their relation to subjective color reproduction assessment is discussed in the next chapter.

For objective assessment, a common method for calculating color reproduction accuracy is the CIELAB ΔE_{ab} color difference metric recommended by CIE in 1976 (Janssen 2001, pp. 16-17). Basically, the metric is just a normal MSE metric fitted in the CIELAB (L*, a*, b*) color space. An implementation for the CIELUV (L*, u*, v*) was also introduced. ΔE_{ab} can be calculated with the equation

$$\Delta E_{ab} = \sqrt{(\Delta E^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2}, \tag{11}$$

where ΔE^*, Δa^*, and Δb^* indicate the difference between the original and the reproduced CIELAB coordinates (Janssen 2001, Maureen 2003, p. 27).

In the CIELAB color space, coordinate L* indicates the lightness of the color, whereas a* and b* are the color-opponent dimensions for green to red and blue to yellow respectively (Johansson et al. 2007, pp. 78-80). Numerous more advanced color difference methods are available, e.g. CIEDE2000 or iCAM, but they are not discussed here. A comprehensive selection of references and further information can be found in Kivinen (2009).

2.3.3 Memory colors

Humans have a tendency to consider the image with familiar objects more natural than the image without a well-known content (Yendrikhovskij et al. 1999). These familiar objects are often associated with a certain color in human mind, namely a memory color. By remembering what the colors of familiar objects should be, people are able to make rough judgements of the color reproduction of an image without seeing the original scene (Hunt 1987, pp. 36-39). However, normally the term memory color is used to denote objects that have very strong memory traces.

Maybe the most well-known memory color is human skin. Other common memory colors are related to nature, e.g. blue sky, green leafs, or other natural objects, like fruits and vegetables (Fairchild 1998, p. 154; Hunt 1987, p. 39; Olkkonen et al. 2008; Perez-Carpinell et al. 1998). Due to the nature of these objects and concepts, there exists much variation in some memory colors. The color of green leafs, for instance, changes from dark green to light green and even the base hue alters at different seasons. Skin color is even more variable (Hunt 1987, p. 39).

Nevertheless, this variation seldom presents any real problem in color evaluation, as humans naturally use their knowledge of the environment and the context to guide their perception. In other words, even without the original image to compare with, humans can make observations of preferred color reproduction based on some psychological concept of an idealized image (Fernandez & Fairchild 2002). Therefore, as a research by Olkkonen et al (2008) indicates, the naturalness of color stimuli in is highly preferable in color testing and actually increases the effects of memory colors.

As diverse as the variation in memory colors may seem, the dominant wavelengths of many common memory colors fall to a rather narrow area. Hendley and Hecht (1949) examined the distribution of natural colors in landscape images and found three distinct color groups: water and sky with dominant wavelengths in a blue region (wavelengths 459-486 nm), green plants and foliage in a yellow-green region (550-575 nm), and earth and dried vegetation in a yellow to orange-red region (576-589 nm). For human complexions, e.g. human skin, Buck and Froelich (1948 cited in Yendrikhovskij et al. 1999) state the dominant wavelength of approximately 590 nm.

Generally, two properties are often associated with memory colors. Firstly, the objects with strong memory colors are often perceived more saturated than other objects, even if all the objects have the same saturation level (Fairchild 1998, pp. 28-30). According to Perez-Carpinell et al. (1999), this is especially true with highly saturated objects. Secondly, a change in illumination has lesser effect on the observed hues of memory colors than on other colors, making them more resistant to lighting variation (Olkkonen et al. 2008). Although the strength of changes varies among different lighting types, the effect can be easily observed under many different kinds of illuminations (Olkkonen et al. 2008; Perez-Carpinell et al. 1999).

A phenomenon closely related to this property is color constancy, which Chalupa and Werner (2004) explain as a human ability to recognize object colors under different lighting conditions. In natural viewing, human observers actively use their knowledge of a scene in trying to keep colors constant and counteract the effects of changing illumination for sensed colors. It must be stressed, however, that since colors are always relative to changing conditions, such as illumination changes, or human abilities, such as chromatic adaptation and simultaneous contrast, colors are never really constant for human observers, and there exists no such a term as absolute color constancy (Fairchild 1998, pp. 156-157).

2.4 Reference images

2.4.1 Technical considerations

Proper reference images are essential for obtaining reliable and comparable test results in image quality assessment. However, selecting or creating a high quality reference is not simple. Since each test situation is unique and different number of test images is needed, it is hard to present detailed instructions for reference image selection without knowing the exact test con-

text. The situation simplifies dramatically if the test context is forgotten and only technical aspects of reference images are considered.

In image quality assessment, high quality references are generally preferred. Johansson et al. (2007, pp. 150-152) list technical requirements for selecting or creating a high quality digital reference image. Firstly, the basic technical specifications, such as resolution and color depth, must be sufficient. Generally speaking, the resolution should be as high as possible and the color depth 8 or 16 bits per channel. When storing the original, it is recommended to choose uncompressed or lossless compressed format to prevent information loss. RGB and CIELAB are the suggested color spaces for storing. However, it should be noted that using these settings may result in expanded file sizes, and therefore limit the easy distribution of files. (Johansson et al. 2007)

Field (1999, pp. 234-239) approaches the situation by presenting technical aspects especially important for color reproduction originals. A high quality reference should be ultra sharp, not exceedingly contrasty, free from graininess, and properly exposed to ensure saturated colors. To further develop color quality, special attention must be taken to evenly illuminate the whole scene. Constant lighting provides similar color rendering conditions for each part of the image and is thus essential for successful color evaluation. In practice, all of listed requirements are not always met, which can happen e.g. when searching images from the Internet. For these cases, Field (1999) lists two minimum requirements for a color original: high sharpness and the absence of colorcasts. (Field 1999)

2.4.2 Previous work

When the concept of reference image design is extended to include image content, the situation becomes more complicated. Based on the findings from various recent studies on image quality, image content has a major effect not only on subjective evaluation, but also on objective measurements (Eerola et al. 2008a, Halonen 2008a, Leisti et al. 2008). For example, in the research of Leisti et al. (2008) the subjective opinions of preferred printing papers varied greatly between the landscape and the human image. With the landscape, the paper with sharper and more saturated rendering was preferred, while with the human image the paper with more neutral and natural colors combined to less sharp details was favored. Further, the objective methods studied in Halonen (2008a) were strongly affected by image content. With the contrast method for example, significant Pearson correlations with reference data were obtained for the man and studio images, but insignificant correlations for the cactus and lake images.

Previous research on reference images have concentrated on creating image collections consisting of many different reference originals, from which suitable reference images can be chosen for each assignment individually. Figures 6 and 7 present some examples of typical picture collections used in image quality evaluation.

Figure 6. *An extract from Roman 16 reference image collection (Bundesverband Druck und Medien n.d.)*

Figure 7. *Sony sRGB Standard Images 1999 (CIE Division 08 n.d.)*

As can be observed, reference images either present some particular theme, such as a posing human, a landscape or an arrangement of objects, or attempt to combine all the essential image quality elements into one image. In the latter case, a color test chart is often embedded into the image to provide an effortless option for objective color reproduction measurements. While very practical, the resulting image with a test chart rarely looks natural for human observation.

Salmi (2008), as a part of DigiQ research project on natural reference images, approached the concept of reference images from another perspective. Instead of using many references, only one reference image was created. The objective was to design as natural and balanced reference picture as possible, while still maintaining usability on both subjective and objective testing. To preserve naturalness, no test charts of any kind were implemented into the image. The resulting reference image, named as Reference Image I in this study, is presented in Figure 8.

Figure 8. *A natural and balanced reference image by Salmi (2008), named as Reference Image I in this thesis.*

Reference Image I presents a breakfast scene with one human subject and a table full of objects. During the design process, Salmi (2008) concentrated especially on the following aspects in addition to overall balance and naturalness: wide color distribution, the presence of memory colors, diverse surface materials, the amount of details, a neutral theme, copyright issues, and perceptual and aesthetical aspects. Furthermore, the image was optimized for the measurements of following image quality attributes: sharpness, noise, contrast, colorfulness and paper gloss.

Salmi (2008) studied the useful areas and objects in Reference Image I with an eye-tracker and usefulness maps filled by test subjects. In eye-tracking tests simulating a free-view situation, the importance of a human model as an attention catcher was evident. The test subjects started the viewing from the human face, moving their gaze through the hands to the teacup and the other objects on the table, firstly concentrating on the second life-like subject, the dog picture (Salmi 2008, p. 91). The same concentration on the human model was visible when the usefulness of different image areas was asked directly from the test subjects (Salmi 2008, p. 78). However, the dog picture was not considered very useful, regardless of it presenting a natural real world object. The statistical analysis of the usefulness results is presented in Figure 9 (Salmi et al. 2009).

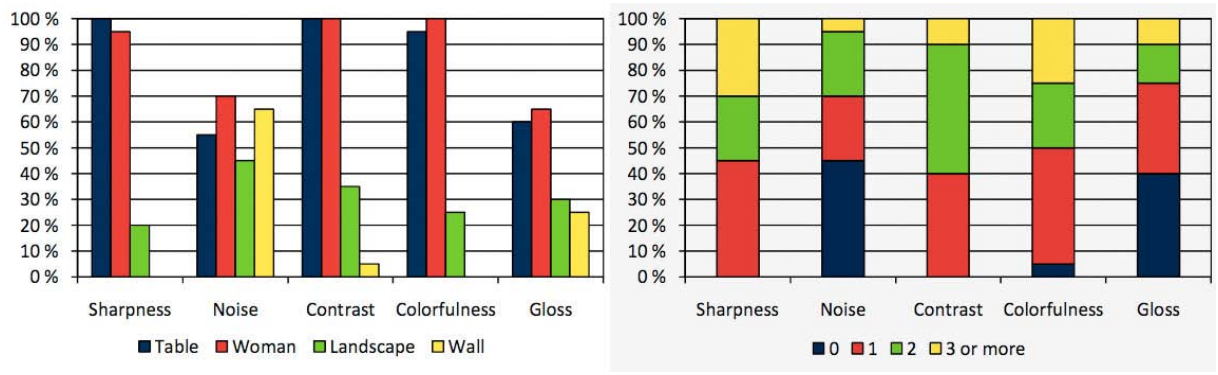


Figure 9. *The perceived usefulness of various image areas in subjective evaluation for Reference Image I (left) and the number of usable objects in quality assessment (right). (Salmi et al. 2009)*

Inspecting the left chart in Figure 9, over 90% of the test subjects found the human model useful in sharpness, contrast and colorfulness tasks. When considering individual attributes, in noise assessment the observers evaluated the whole image, while in the case of sharpness, contrast and colorfulness, the attention was more concentrated on few areas only. Indeed, in most of the cases only few objects were needed for evaluation. As presented in the right chart of Figure 9, the majority of the test subjects used only two objects maximum to reach their conclusions.

2.5 Photo aesthetics

2.5.1 Aesthetic photograph

Oxford Advanced Learner's Dictionary defines aesthetics as (1) "concerned with beauty and art and the understanding of beautiful things", and (2) "made in an artistic way and beautiful to look at" (Hornby 2006). Similarly, Field (1999, pp. 375-376) connects aesthetics to creative concepts like beauty, proportion, balance, and form, all of which are important for the workers of art, e.g. photographers, graphic designers, or artists. Regarding aesthetics in photographs, Datta et al. (2006) define aesthetics as the way for observers to characterize the beauty of photographs.

When making judgments about the aesthetical aspects of images, Field (1999, pp. 375-376) emphasizes the importance of image usage context, i.e. the function the image was designed for. Fashion images, for instance, may offer aesthetically pleasing presentation with much more colors and less realism than documentary pictures. Wright (1999, pp. 36-50) takes a less practical approach by introducing three aesthetic intentions which guide the actions and main intends of photographers in a photography situation: realism, formalism, and expressionism. The three terms describe the basic aesthetics concerns in photography, starting from recording the subject (realist) with the means of photography (formalist) by a photographer (expressionist). In other words, realism prioritizes the faithful reproduction of the scene or object being

photographed, formalism concentrates on the technical aspects of a photographic medium or capturing a scene under the restrictions of photographic equipment, and expressionism explains the viewpoint to artistic expression by the photographer. All the intentions are related, and with the careful weighting of each intention to suit the main photographic intend in hand, a photograph with high aesthetical value can be captured (Wright 1999).

Hedgecoe (2006, pp. 35-37) approaches the concept of good, aesthetical photograph from the viewpoint of a photographer by discussing about the most essential elements in a photographic image. Hedgecoe (2006) lists aspects like an interesting main subject, attractive forms or shapes, an effective composition, the intelligent use of color, the expression of feelings, and the inclusion of small details. However, while photo goodness is the product of the skills and expressionism of a photographer, an observer is always the final judge on photo goodness. (Hedgecoe 2006)

Axelsson (2007) studied photographic images with subjective methods to find the most important attributes to describe the aesthetic appeal of photographs. Based on the subjective tests arranged with a sample pool of 564 images with variable themes and image content, the following three dimensions were found to explain aesthetic appeal: Hedonic tone - familiarity, the absense of color, and expressiveness - dynamics. The first dimension regards to the familiarity of a photograph subject or objects. The second dimension emphasizes the influence of color. Black and white pictures, for example, are often considered as aesthetically pleasing. The third dimension, expression-dynamics, shows increased aesthetic appeal on dynamic scenes versus static presentations with an emphasis on expressive storytelling. In short, aesthetic appeal is strongly influenced by the familiarity, color appearance, and dynamic presentation of photographs. (Axelsson 2007)

2.5.2 Aesthetic primitives of vision

In general, an aesthetic experience is a product of the human brain and the human visual system. Peters (2007) studied the aesthetical appeal of images inspired by the principles of visual arts and insights from cognitive neuroscience. Peters (2007) concentrated on the modular construction of the human visual system and underlying cognitive research on visual perception, and formed a conclusion that an aesthetic experience can be directly derived from these modular parts of the human brain. As a result, six groups, named as the aesthetic primitives of vision, were selected to describe the way humans perceive the aesthetical cues of images: color, form, spatial organization, motion, depth, and the human body. Further, Peters (2007) listed various properties for each primitive to explain the aesthetic function of the primitives. Some example images of the primitives are shown in Figure 10.

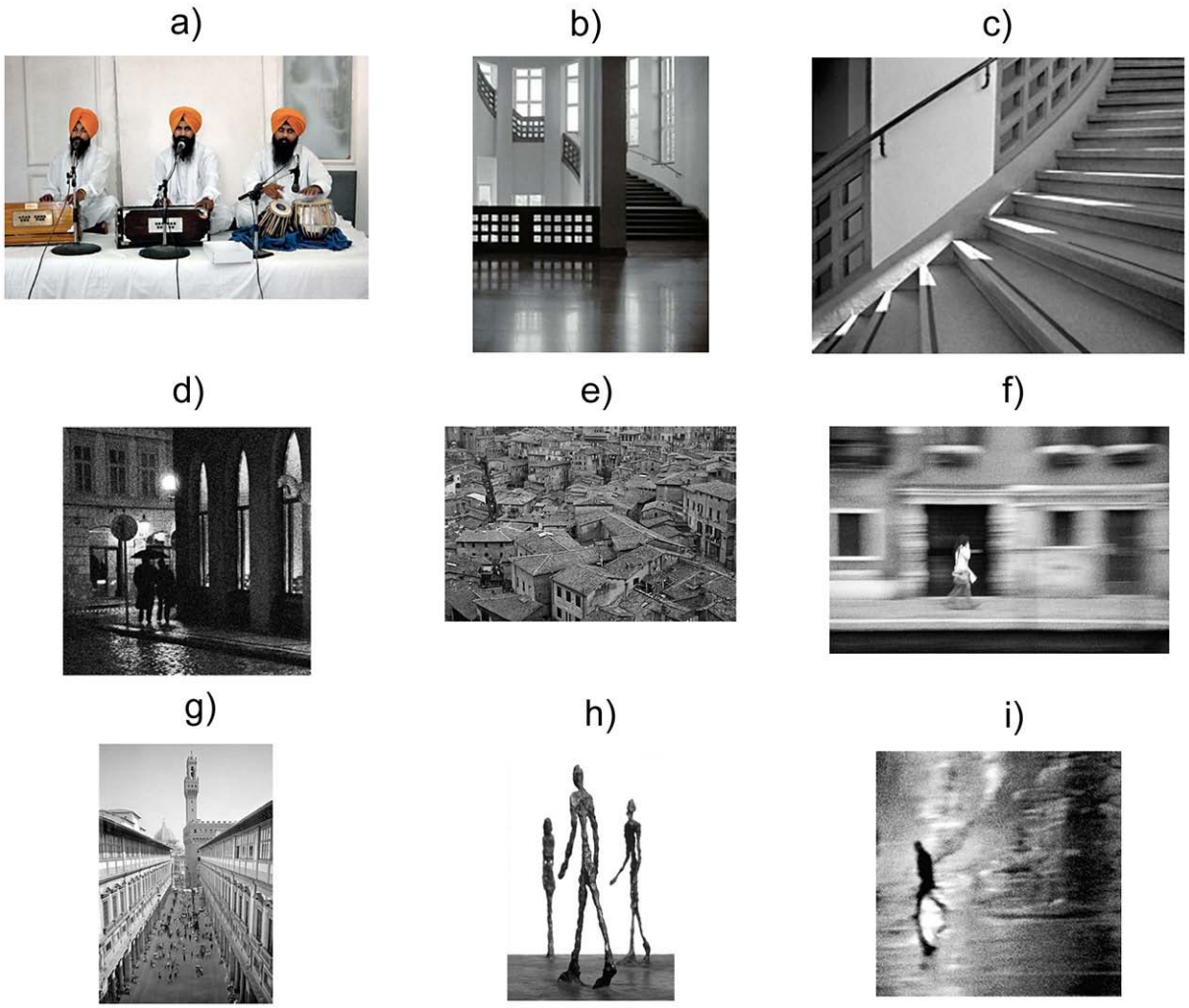


Figure 10. *Examples of the aesthetic primitives from Peters (2007), expressed in the form: id) primitive - property. a) color – only a few strong colors, b) color – dynamic range, c) form – clarity, d) form – silhouette, e) spatial organization – texture, f) motion – motion blur, g) depth – linear perspective, h) human body – principal axis, and i) human body – silhouette*

Peters (2007) considers color, the first primitive, as an important means for forming aesthetic experiences. The first property increasing the aesthetical appearance is the number of colors, especially when only a few strong colors are included. According to Peters (2007), only two or three strong colors are needed at the maximum for a pleasing aesthetic experience. The exception to this rule is monochomaticity, e.g. black and white images, which can form quite strong aesthetical cues. Another strong color property is a complementary contrast, i.e. the presence of complementary color pairs in an image. The most beautiful images are often formed when complementary colors are the only strong colors in an image. Further, Peters (2007) lists the exploitation of dynamic range as an important aesthetic cue. Generally, aesthetic appearance is

increased when the full dynamic range is exploited and complemented with smooth transitions between tonal values. (Peters 2007)

The second primitive, form, can be clarified as the shape of objects. Aesthetical sensation is increased when a form consists of strong, easily identified lines. Curved lines are especially aesthecal. Silhouettes are also considered as efficient aesthetic cues, especially when human forms are regarded. The third primitive, spatial organization, expands the form primitive into the spatial field. Considering the number of objects, less is usually better than more. Furthermore, for aesthetical image, compositional aspects such as the use of golden mean should be emphasized. Well definied textures also affect positively on aesthetics, in addition to the rhythm, repeation, and variation of textures. The fourth primitive, motion, refers to life and action in an image. However, indicating movement in a still image is not trivial. Motion blur tends to increase the feeling of movement, affecting aesthetical experience. Another technique to illustrate movement is to implement multiple distinct motion images into one image. With the fifth primitive, depth, Peters (2007) regards the illusion of three dimensions in a two-dimensional image. To increase this aesthetical cue, Peters (2007) lists techniques such as linear perspective, selective focus, and the synenergy between light and shadow areas. Finally, the sixth primitive, human body, expresses the importance of a human form in aesthetics. Peters (2007) raises up the concept of principal axes, i.e. the structural form of a human body. In stick figures, for instance, human form is easily recognizable even if no indication to body texture is presented. This simple presentation is also aesthetically beautiful. (Peters 2007)

2.5.3 Gestalt grouping

When humans look images, certain forms and patterns attract more attention and are visually more pleasing than others. Psychological studies in the field of human perception of visual elements give insight on the psychophysical processed underlying the human observation of images. The most fundamentally accepted study in this regard is the Gestalt theory of visual perception (Clayton & Hashimoto 2009, p. 28). Generally, the Gestalt theory states that humans inherently look for certain relationships and components in images, and have a tendency to form them into larger, greater image schemes (Clayton & Hashimoto 2009). Furthermore, in visual perception, the simplest, most regular, and balanced organization gains our attention and is commonly regarded as a “good” Gestalt (Spillmann & Ehrenstein 2004). The good Gestalt has e.g. a positive effect on image harmony, an important aspect of the aesthetical appeal of images (Fedorovskaya et al. 2008).

A central issue in the Gestalt theory is the figure-ground segregation, i.e. the difference between figure and ground. In general, figure is a delineated, enclosed area of certain shape, usually symmetrical, which appears to lie slightly above ground. Background is denoted by ground, which appears farther away, is usually larger and open without enclosed forms, and is partially obstructed by figure. For short, figure can be defined to have the “Ding (thing) character” and ground the “Stoff (stuff, material) character”. (Spillmann & Ehrenstein 2004)

To explain how humans process a complex stimuli field into figure and ground, the Gestalt psychology describes seven grouping factors: symmetry, parallelism, smooth continuation, closure, proximity, similarity, and common fate (Figure 11). From all the groups, similarity and common faith are especially powerful for forming surfaces in our perception. The pattern can consist e.g. of similar brightness, color, texture, or depth. (Spillmann & Ehrenstein 2004)

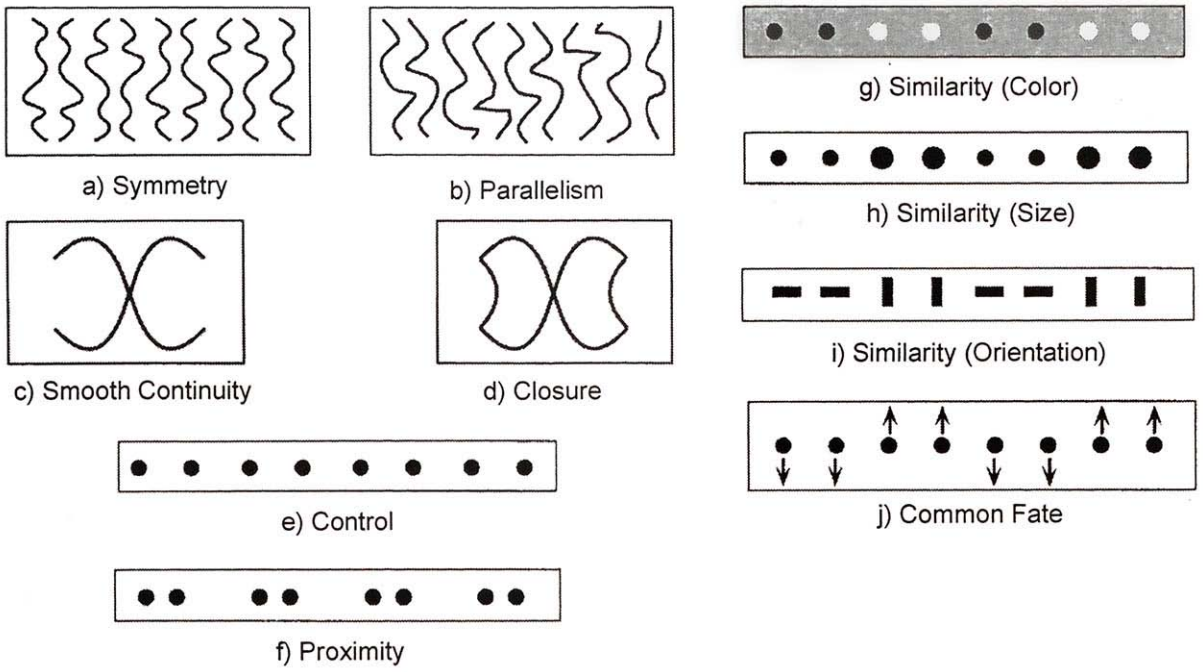


Figure 11. *Schematic examples of the Gestalt groups (Spillmann & Ehrenstein 2004)*

The first group, symmetry, explains a situation where human perception follows a symmetrical presentation over an asymmetrical. For instance, wavy, mirror-symmetric shapes are perceived as figure and asymmetrical areas as ground. Similar situation is with parallelism, where parallel lines with fixed width are regarded as figure and unparallel lines as ground. The third group of figures, smooth continuation, refers to shapes with smooth, collinear parts that tend to be linked together. Shapes with sharp edges or transitions, on the other hand, are seen as ground. Similar discrimination happens with closure, a situation where shapes with closed form are organized as figures and open ones as ground. The fifth group, proximity, describes a condition where closely adjacent stimuli combine as a figure more likely than stimuli farther apart. Similarity, the next group, refers to the uniformity of stimuli, e.g. the similarity of shape, brightness, color, size, orientation, texture, or depth. The final group, common fate, describes forms that are moving simultaneously at the same speed and direction, even when widely distributed over a visual field. Common fate is regarded as the strongest of all the Gestalt groups, overriding the other highly perceived groups: proximity and similarity. For instance, a group of shapes that are normally regarded as ground will appear as figure as soon as the group starts moving in an organized order and speed. (Spillmann & Ehrenstein 2004)

3 METHODOLOGY

3.1 Outline of research plan

This study was executed as a part of the DigiQ project to research the quality characteristics of printing papers with objective methods. The main goal was to develop software to automatically grade papers for image quality and several quality attributes. The main difference to the most previous print quality studies, however, was that objective grading was based on subjective evaluation results from printed papers. In other words, the software was developed to simulate human opinion of paper quality, and to rate papers as humans would have rated them. To achieve these goals, the project was divided in four separate phases: reference image development to create a reference image for subjective and objective testing, subjective tests to obtain subjective paper quality data from the printed reference image, software development for calculating quality attributes from scanned prints, and VQI (visual quality index) development to form one grade for print quality and more accurate quality attribute grades with regression analysis.

In the first phase, a single reference image was constructed for subjective and objective testing. The reference image development was based on a previous DigiQ research by Salmi (2008), where a natural reference image was created for subjective quality evaluation. During the development, the research findings by Salmi (2008) were closely studied to construct an improved reference image version for the DigiQ project. The following thesis document will discuss both the reference image versions, and to simplify the naming for a reader, the reference image by Salmi (2008) is named as Reference Image I, and the reference image of this thesis as Reference Image II.

Reference Image II, similarly as the previous version, presents a breakfast situation where a human is sitting by a table filled with objects related to breakfast. In addition to extremely high technical quality of the image, especial attention was placed on image naturalness, balance, and aesthetical aspects. As an extra development goal, certain test colors, i.e. red, green, blue, cyan, magenta, yellow, orange, and three neutral greys, were implemented into natural objects in the image to allow an option for advanced color balance calculations. The reference image was photographed indoors with a high quality digital SLR camera and lightened with two flash units.

After Reference Image II was successfully photographed and post processed, subjective tests were arranged for both the reference image versions at the same time. The two images were printed on 21 sample papers with an inkjet printer Epson Stylus Pro 3800 and a print size of 15 x 10 cm, following the printing process described in Halonen (2008a). Five quality categories were assessed, consisting of quality and four quality attributes, i.e. colorfulness, contrast, sharpness, and noise, with the emphasis on the first impression on evaluation. 30 test subjects, consisting mainly of media technology students and research staff from Helsinki University,

graded the sample papers using 7-point Likert scale, with anchor images setting the limits of the scale.

In the third phase, quality analysis software was developed based on the subjective test data. Only Reference Image II was analyzed. Since software analysis required digital data, the printed samples were digitized based on the digitizing process described in Halonen (2008a). Four methods were successfully developed with Matlab 7.1, including methods for computing colorfulness, contrast, noise, and color balance. The design goal was placed on maximizing correlations between software calculated scores and subjective reference data. For color balance method, reference data was obtained from the original image. All the applied methods were reduced-reference methods, excluding the no-reference noise method implementation by Malanin (2006).

The accuracy of the objective method results was further improved with statistical methods in the last, fourth phase. With regression analysis and SPSS 14 software, new quality attribute grades for colorfulness, contrast, and noise were calculated. Compared to the objective scores obtained in the phase three for the same quality attributes, the result accuracy was improved by using the scores of more than one objective method for calculating single grades. Furthermore, as the final goal of the study, the objective results were combined into one grade for print paper quality, visual quality index (VQI). As was the case with the subjective tests, all the grades were given with 7-point scale, now accompanied with written explanations.

3.2 Printing and digitizing process

In the experimental part of this thesis, two kinds of samples were required: printed reference image samples for subjective tests, and digitized reference image prints for objective analysis. The samples were prepared based on a printing and digitizing process by Halonen (2008a). Figure 12 presents the full process map for generating digitized samples for objective methods. This chapter discusses the procedure briefly, while the detailed explanation of the process is available in Halonen (2008a).

The process map by Halonen (2008a, pp. 44-45) was divided in two main phases: sample printing, and print digitizing with scanning. In the first phase, both the reference images were printed on the 21 sample papers with a professional quality inkjet printer, Epson Stylus Pro 3800 (Epson 2006). For each sample paper, a paper specific ICC profile was created from a printer profiling target. With the approximate print size of 15 x 10 cm and the image resolution of 360 pixels/inch, Reference Image I offered pixel dimensions of 2126 x 1417 pixels and the print version of Reference Image II a resolution of 2134 x 1417.

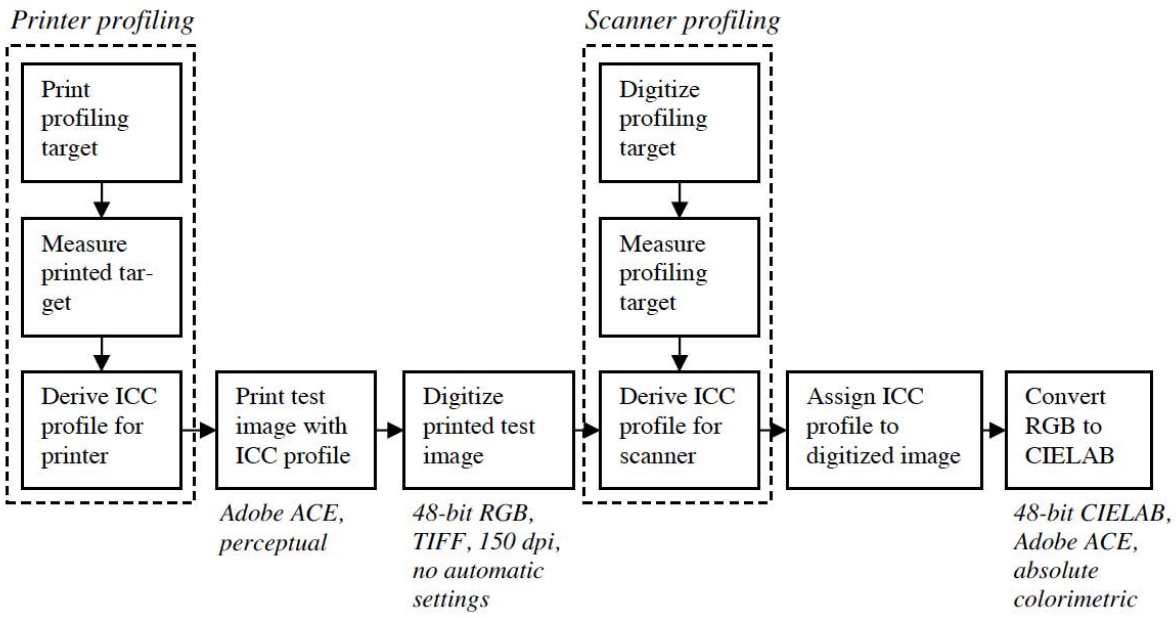


Figure 12. *The process map of the printing and digitizing process (Halonen 2008a)*

After printing, the sample was digitized with a high quality scanner, Epson Perfection V750 Pro (Epson 2007). Automatic adjustments were disabled during scanning, such as driver level sharpening and color balance adjustments. Similarly to the printing phase, the scanner was profiled for every sample paper with ICC profiles constructed from printed and digitized scanner profiling targets. The digitizing proceeded by scanning the samples with a 150 dpi sampling frequency and saving the files in a 16-bit TIFF format, resulting in images with a resolution of 876 x 584 pixels. Further, the ICC profiles were applied and the files were converted to a CIELAB colorspace. A relatively low sampling frequency of 150 dpi was chosen based on a project study by Halonen (2008b), where the accuracy of computational blur method was studied in terms of scan resolution.

3.3 Reference image

3.3.1 Outline of procedure

Reference image development was one of the main goals of this study. The plan was to create one natural, high quality reference image suitable for two different approaches on image quality evaluation: human driven subjective image quality assessment and computer software based objective quality calculation. With the goal of increasing result comparability between the approaches, only a single image was created for both the purposes. As previous research indicates, image content has a major impact on image quality evaluation results, and using different images in comparison would therefore lower result reliability (Eerola et al. 2008a, Halonen 2008a, Leisti et al. 2008).

The reference image created by Salmi (2008), named as Reference Image I in the following chapters, was selected as the base setting for the reference image of this study, Reference Image II. The setting presents a breakfast situation, where a human model is sitting by a breakfast table. The table is full of items related to breakfast, such as dishes, food, a newspaper, plants, and other ordinary dinner table items.

During the design process, special attention was placed to create the image with balanced and natural appearance, with a secondary emphasis on visual pleasantness and other aesthetical aspects. High quality was assured with advanced digital camera equipment and advanced post processing. The adequate lighting was provided with two flash units and two. Furthermore, predefined test colors were included in the image to provide option for objective color difference and color gamut measurements. To maintain image naturalness, test colors were implemented into natural objects inside the image and no test charts of any kind was required. In addition to test colors, numerous objects with memory colors were included as well, e.g. fruits, vegetables, green foliage, and blue sky.

3.3.2 General requirements

The following requirements were considered in Reference Image II creation. On subjective side, the reference image needs to be natural, aesthetically pleasing, and well balanced without distracting objects. Furthermore, memory colors are needed for subjective color quality evaluation. On objective side, the image should provide properties for accurate quality calculations such as high sharpness, strong contrast, saturated colors, and pre-defined test colors for objective color measures. Combining these two approaches into one, complete image was the major challenge in the reference image development process.

For subjective evaluation, the image should be pleasing to look at and as natural as possible. With natural representation, the image itself has less emotional impact on a test subject, thus decreasing result variance caused by emotional changes. On the other hand, naturalness is regarded as one of the most important high-level image quality attributes when evaluating general image quality (Kim et al, 2008). However, designing an image to be natural is not a simple task. According to Yendrikhovskij et al. (1999), the most critical object in the image determines the naturalness of the whole image. Therefore, the careful planning of every image object is required to create a naturally pleasant image.

Careful planning is also required when image balance is concerned. Various image objects must be intelligently placed in the picture to prevent any object from gaining too much attention from observers. According to previous research, a human subject is likely to gather the most attention, and is therefore one of the main concerns of the planning process (Salmi, 2008).

To provide a pleasing experience for test subjects, image pleasantness and aesthetical aspects should be considered. The aesthetical aspects include e.g. image composition, object grouping, color harmony, and other photographical properties better explained in Chapter 2.5. These as-

pects not only increase the appeal of the image, but also give it a visual touch, a sparkle that makes the image visually more attractive and interesting.

On subjective evaluation, two additional aspects are important to take into account as well: image suitability for different cultures and the possible presence of advertising in the image. From the cultural standpoint, the image should not contain any culturally disturbing elements, e.g. related to cultural habits or religion, which may result in unwanted misunderstandings or resentment among observers. This allows the image to be used more universally around the world. From the advertising point of view, the objects in the image should not have any direct indication to the brand of the maker or advertisement of any kind, unless otherwise agreed with the advertising party. Unintended advertising may lead to problems if the image gets more widespread attention.

As a reference image is concerned, the image must naturally offer very high general image quality, namely high image contrast, saturation, sharpness, and low noise. In other words, the original reference should be as excellent as possible to ascertain suitability for the accurate qualitative and quantitative evaluation of quality attributes. This is extremely important, since the original image defines the upper limits for image quality attributes in quality testing.

To create a reference image with reference level quality, uniform and adequate lightning is required to evenly illuminate the entire scene and to provide natural appearance for the image. Furthermore, photographic gear, such as a SLR camera body and a lens, should be chosen carefully to correspond with the high standards of image quality required in reference image shooting. Camera settings have to be adjusted as well to correspond with the photography situation in hand. After shooting, post processing with sophisticated image editing software is needed to maximize image quality.

3.3.3 Development considerations

Similarly to Reference Image I, a breakfast situation with one human model was chosen as a base setting for Reference Image II as well. After considering the research findings by Salmi (2008) and the requisitions for this work listed in the previous chapter, the following development considerations were placed on the reference image creation to complement the requirements. Some of these considerations were assessed to correct some unwanted properties discovered in the later stages of the research by Salmi (2008), while the other considerations were chosen to make the reference image more complete and usable in different image quality evaluation situations, especially in objective evaluation. A major difference compared to the previous work, however, was including test colors in the image, resulting in a highly colorful image appearance. Therefore, maintaining image naturalness and balance in this vividly colored image contributed a major challenge for this study. Table 1 compares the development considerations between the new reference image and Reference Image I by listing issues and concerns related to reference image development and proposing solutions for each issue addressed. While selecting the concerns, the research findings by Salmi (2008) were considered.

Table 1. *Reference Image II development considerations and proposed solutions in comparison to the previous DigiQ reference image (Reference Image I) constructed by Salmi (2008).*

Issue	Concern	Solution and remarks
Human model	Gathers too much attention from observers, thus decreasing the effect of the other objects in the image	Directing the gaze of the model away from the camera and image observers Newspaper reading: more attention to the objects on the table, as the gaze line guides attention to the newspaper and the surrounding objects
Table	Filled with objects	Decreasing the number of objects on the table Only relevant objects for a breakfast situation are needed
Painting on the wall	The image is small with barely visible details	Decreasing the size of the borders Choosing a simpler theme for the image Memory colors must be preserved, i.e. green foliage and blue sky or water
Background wall	Occupies a very large proportion of the image, presenting a dominant color	Choosing an appropriate wall color to suit the image style and color balance
Test colors	Objective color accuracy evaluation is demanding without proper reference colors	Implementing pre-defined test colors in the image for objective color accuracy measures Using natural objects to preserve natural appearance The print context must be considered when choosing the test colors
Color groups	The test colors and memory colors increase the number of bright saturated colors in the image, leading to unnatural and unpleasant appearance	Grouping colors on appropriate color groups to make the groups equally distracting Groups of large objects with opposite colors should be avoided
Perspective and composition	Numerous objects in a small area, all needed to be properly visible	Slightly angled perspective to show more table Avoiding wide angle distortion by using a lens with a natural focal length The shapes of large objects should not be broken, and visible Gestalt shapes should be preserved to increase image pleasantness and distribute attention evenly through the entire image

3.3.4 Test colors

Predefined test colors were implemented into Reference Image II to allow an option for objective color accuracy and balance calculations. After considering the earlier color related research conducted by Koivisto (2009) and Nurminen (2005), seven test colors were chosen to be included in the image, i.e. red, green, blue, cyan, magenta, yellow, orange. In addition, three neutral greys were implemented as well.

Choosing the first six colors turned out to be trivial, since they present the main colors of well-known additive and subtractive color models, and are universally used in numerous common display and printing technologies (Johansson et al. 2007, pp. 71-74). Furthermore, based on the research by Nurminen (2005), humans can easily name and identify the majority of these colors. According to the research, test subjects succeeded in almost perfect identification of red, green, blue, yellow, black, white, and orange. Magenta was identified with 80% accuracy and cyan with 13% accuracy. The accurate identification of test colors can be considered as a desired property for a reference image, since it simplifies the subjective evaluation of color balance and potential color problems. For these reasons, orange, being amongst the perfectly identified colors in Nurminen (2005), was added as a seventh test color for this study.

The exact number of test colors was considered as well. In the research by Koivisto (2009), increasing the number of test colors had a major positive impact on the result accuracy with various objective color accuracy calculation methods. On the other hand, test colors have a tendency to be bright and colorful, which naturally has a strong influence on the overall appearance of the image. Consequently, to preserve image balance and naturalness, the number of test colors was not increased any further.

A ColorChecker chart manufactured by GretagMacbeth company, nowadays owned by X-Rite, was chosen as a reference for the test color RGB values (X-Rite 2010). Figure 13 presents the placement of the test colors in the ColorChecker. However, the official ColorChecker values were not utilized in calculations, since the values have small variations between different charts (Fraser et al. 2005, p. 151; Pascale 2006). Instead, the measured averages of ten seperate ColorChecker charts formed the reference values of this study. These averages, named as BabelColor averages, were obtained from the research conducted by Pascale (2006). The BabelColor values for the test colors are listed in Appendix 1 for Adobe RGB (1998), CIE L*a*b*, and sRGB color spaces.

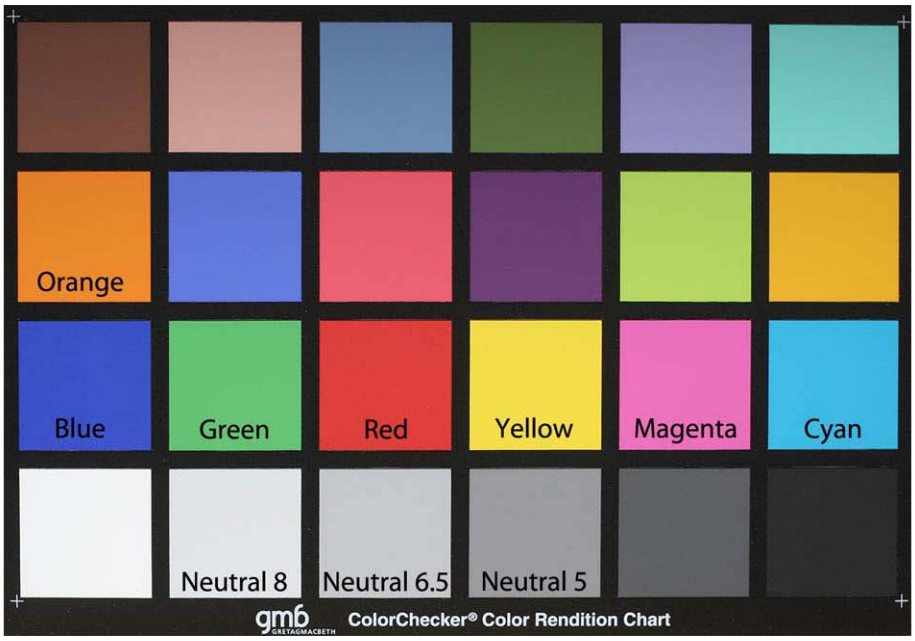


Figure 13. *A visualization of GretagMacbeth ColorChecker color chart. The test colors in Reference Image II are named on the top of the corresponding color.*

Since the reference image required high naturalness, the test colors were implemented in natural objects, and no test charts were included. Firstly, objects with appropriate colors were searched for to match the test colors. After choosing the objects with correct colors and photographing the image, the colors of the objects were further adjusted with Adobe Photoshop CS4 to equal the reference test colors. The color optimizations were commensed on full, carefully cropped objects with Photoshop color adjustment tools. Software optimization was essential, because the colors of the natural test objects varied with lighting, and it was plainly impossible to find natural objects with perfectly accurate test colors.

What comes to test color implementation, the average RGB values of pre-defined square areas in the test objects were adjusted to match the BabelColor average values of the reference colors. In other words, each test color had two placement properties, a center point and the width of the square area surrounding it (Appendix 2). The average values of these square areas were adjusted to match the corresponding reference colors.

3.3.5 Implementation

Reference Image II was photographed in the Department of Media Technology at Helsinki University of Technology, Espoo. Over the planning process, one studio room was reserved almost exclusively for the reference image development for four weeks. This arrangement proved out to be very efficient, as the development phase required thorough testing and turned out to be very time consuming. Furthermore, taking the image indoors helped in standardizing the reference picture photography situation, thus easing the possible repetition of image taking in future.

During the planning phase, dozens of different objects related to the breakfast theme were acquired and tested for suitability for the project. Since the image needed to include some objects with pre-defined test colors and some other objects with memory colors, heavy emphasis was place on acquiring the suitable, correctly colored items for the image. Furthermore, a great deal of attention was also placed on an aesthetically pleasing placement of the objects on the table, as well as grouping the colorful objects in visually attractive color groups to balance the color output of the image. For example, choosing an appropriate color for a tablecloth turned out to be a time consuming task. Because of the large surface area it occupied, the color of the tablecloth was very dominant and had a major impact on the overall look and the atmosphere of the image.

As discussed in Chapter 2.4.2, a natural reference image requires a presence of a human model. Reference Image II included a woman model with moderately light skin and dark hair sitting by a breakfast table and reading a newspaper with her gaze aimed at the newspaper on the table. This arrangement helped to reduce the predominant effect of the human as a part of the breakfast scene, thus directing more attention to the objects on the table. Furthermore, the gender of the model was considered as well. Mostly because of the colorful appearance of the image, a woman model was considered suitable for the atmosphere of the image.

The reference image was photographed with a professional digital SLR camera, Nikon D300, supplemented with a Sigma 18-50mm f/2.8 EX DC Macro lens. The camera was placed on a sturdy Manfrotto tripod for a solid, shake-free support. Two Nikon SB-600 flash units provided adequate lighting for the scene. The image was recorded in 12 megapixel resolution with 4288 x 2848 effective pixels and was taken in 14-bit raw format, i.e. Nikon NEF, using Adobe RGB (1998) color space. No cropping was required after the shooting. The following exposure settings were used: ISO 200, shutter speed 1/250 s, aperture f/7.1, and focal length 31 mm. In the 35 mm format, the focal length equaled to 46.5 mm after the 1.5x crop factor of the Nikon APS-C sized image sensor.

All the exposure settings were chosen carefully to achieve the highest quality attainable for this particular photographic situation. Firstly, the ISO value of 200, presenting the base ISO setting for Nikon D300, provided the lowest noise levels and the widest achievable dynamic range for this camera. Secondly, the 1/250 s shutter speed provided the fastest available flash sync speed for the camera, decreasing the effect of ambient lighting in the image. Thirdly, the aperture of f/7.1 was wide enough to allow ample light for a proper exposure, while providing a sufficient depth of field of 4.41 m for attaining high sharpness over the whole image. Furthermore, light diffraction was not visibly decreasing image sharpness with this moderate aperture. Lastly, the focal length 31 mm, equivalent to 46.5 mm in the 35 mm format, was chosen to ensure the natural image appearance without distracting perspective distortions. According to Westlake (2009), the focal length of 43 mm is often considered as a standard vocal length for photography with full frame cameras, providing a natural perspective for human observers.

With the exposure settings listed above, the two flashes were basically the only light sources illuminating the scene, effectively canceling the lighting effect of ambient background lights. This phenomenon was easily observed when the flash units were turned off and a photograph was taken in ambient lighting only. The resulting photograph was almost black. Thus, the reference image was photographed in decent ambient lighting, making the whole process much more pleasing for both the photographer and the human model.

Considering lighting, the two flash units were accompanied with two reflectors and one diffuser plate. Figure 14 presents the placement of the flashes and the reflectors in the final reference image.

Figure 14. *The lighting setup of Reference Image II, illustrating the placement of two flash units and two reflector plates.*

Lots of efforts were placed on achieving uniform lighting with stable white balance over the whole image. Light of the left flash was bounced from the ceiling, providing a large area of soft light over the entire scene. The right flash was shot through a diffuser with lower power, resulting in a relatively small area of harder light. The harder light effectively softened the shadows caused by the left flash, and provided extra light for improved illumination of the human model. In addition, the side flash effectively increased the three dimensional appearance and the surface texture visibility of the objects on the table, both of which are well-

known properties of side light (Hedgecoe 2005, pp.46-49). For additional lighting, two reflectors were constructed from aluminum foil. The left reflector provided extra light for the human model, while the right reflector softened the shadows on the right corner of the tablecloth. However, in the case of the entire image, some faint shadows were intentionally left on the image to increase image naturalness.

3.3.6 Software adjustments

After Reference Image II was taken, extensive post processing was commenced to maximize the image quality of the reference. The image was edited with two different image editing software, Nikon Capture NX and Adobe Photoshop CS4. Eizo ColorEdge CG242W, a high quality 24.1" LCD display designed for image processing professionals, was selected as a primary display for editing (Eizo 2009). The specifications include 1920x1200 pixel resolution, internal 16-bit color processing, 1100:1 contrast ratio with 270 cd/m^2 brightness, DVI input, and hardware calibration. Before the editing process was initiated, the display was calibrated with GretagMacbeth Eye-One 1 Monitor calibrator unit to ensure the most accurate color response attainable for the display. The calibration was performed with ColorNavigator, software included with the Eizo display, using the default settings with 2.2 gamma.

The image was processed in two phases. During the first phase, basic adjustments for saturation, contrast, curves, and sharpness were applied with Nikon Capture NX software. For natural color response, white balance was adjusted by first photographing a ColorChecker chart in the same lighting as the reference image, and later correcting the white balance with Nikon Capture NX by selecting a neutral grey patch from the photographed chart as a white balance gray point. All the adjustments were done on 14-bit raw image.

After the corrections for the entire image were completed, the image was converted to 16-bit TIFF format in Adobe RGB (1998) color space and further processed in Adobe Photoshop CS4. The second phase adjustments included cloning and correcting specific areas in the image, as well as adjusting the colors of the objects with test colors. To provide an option for future changes, Photoshop layers were exploited in all the edit steps.

Thanks to the careful planning of the reference image, extensive cloning was not required after photographing. Only three areas were corrected with cloning, including a disturbing blue object on the wall, excessive reflections on the painting, as well as some distracting hair wisps on the head of the model. However, test color adjustments required much more attention. As explained in Chapter 3.3.4, the average color of a certain area in each test color object was adjusted to match the corresponding reference test color. To retain the natural appearance of the adjusted objects, color modifications were performed on entire objects, not only on the test color area. Each test color object was carefully cropped from the image background and saved as an own layer on the top of the whole image. This practice not only preserved the natural look of the image, but also offered an option to edit the color balance of the main image without affecting the test color values.

Reference Image II was presented in two versions, one intended for display viewing and the other for the printing purposes of this study. The display version is recommended for normal usage, such as viewing from display devices or distribution in the Internet. The best quality and color balance is obtained with the reference display Eizo ColorEdge CG242W, which was used for the adjustments (Eizo 2009). The image was tested with other lower quality displays as well to ensure balanced image reproduction with a variety of display devices. The second version, the print version, was developed primarily for the use of the project and contained optimizations for the printing process described in Chapter 3.2. In comparison to the display version, the print version included minor corrections to image brightness, saturation, curves, and sharpness. The test colors remained unchanged. The corrections were applied to maximize the output quality of the test printer, Epson Stylus Pro 3800 (Epson 2006). Based on the research by Halonen (2008a), few high quality papers were chosen from the test paper pool for test prints and quality optimization.

3.4 Subjective tests

3.4.1 Outline of procedure

After Reference Image II was completed, subjective tests were conducted to provide reference data for objective quality calculation methods. The emphasis was to evaluate the effect of various quantitative image quality attributes on general image quality. Only the influence of printing papers was studied.

Subjective tests were executed by printing the print version of Reference Image II on 21 different printing papers with an inkjet printer without any modifications to the reference image between prints, and requesting test subjects to evaluate pre-defined image quality properties from the printed images. Consequently, the printer, the printing method, and the image remained the same in every sample, only the paper changed. Furthermore, to provide valuable research data of different reference images for the DigiQ project, the tests were conducted on Reference Image I at the same time.

The following quality attributes were assessed: overall image quality, sharpness, contrast, colorfulness, and noise. In the case of overall quality, test subjects were requested to evaluate the general quality of the printed image, and they were allowed to define the concept of quality according to their own personal preferences. In the case of the other four attributes, test subjects were asked to assess only the amount of the attribute concerned, not their subjective opinion about the quality of that attribute. Therefore, these four attributes were quantitative, and personal preferences were not intended to affect the results.

3.4.2 Samples

For comparison purposes in the DigiQ project, two different reference images were chosen as main images in subjective tests, i.e. the print version of the Reference Image II and Reference

Image I by Salmi (2008). Both the images were printed on 21 printing papers using an inkjet printer Epson Stylus Pro 3800 (Epson 2006). The printing process followed the procedure described in Halonen (2008a), and is also briefly discussed in Chapter 3.2. The printed samples were presented in a size of 15 x 10 cm. Because the emphasis of this study was on print quality differences arising from paper, the reference images remained unmodified between adjacent prints, and only the printing papers changed during the sample creation process.

The printing papers were selected to match the earlier DigiQ research conducted by Halonen (2008a, pp. 27-28). The paper selection included 15 inkjet papers and 6 multipurpose papers, resulting in a diverse selection of papers with varied quality characteristics. The range of variation in terms of paper characteristics is presented in Table 2.

Table 2. *The variation of paper characteristics in the sample paper selection (Halonen 2008a)*

Paper	Grammage (g/m^2)	Whiteness (%)	Brightness (%)	Fluorescence (%-units)	Permeability Bendtsen (ml/min)	Roughness PPS (μm)	Roughness Bendtsen (ml/min)	Gloss (GU)
Multipurpose	77-84	79-157	89-100	0-68	392-706	4.5-7.2	91-247	4.0-7.4
Ink-jet	98-280	91-150	86-98	7-59	0-193	0.6-5.7	0-270	2.2-96

After printing, each of the 42 samples was glued on a sheet of neutral grey cardboard. To ease sample handling and to offer a neutral background for the samples, the samples with a size of 15 x 10 cm were glued on larger cardboards with an approximate size of 21 x 15 cm, resulting in reasonably large grey borders. Furthermore, the ID number of the paper sample was written on the opposite side of each cardboard. The order of the ID numbers was randomized. An example picture of the resulting samples in a test situation is shown in Figure 16 in Chapter 3.4.5.

3.4.3 Test subjects

30 test subjects attended to the subjective tests. The most of them, 22, were media technology students from Helsinki University of Technology. The students were participating on an entry level course of media technology, Basics of Media Technology, where the subjective tests of this study were listed as a mandatory part for completing the course. The rest eight participants were students and researchers from varying fields of expertise at Helsinki University of Technology. As a result, all the test subjects had a technical background, but on average their knowledge of image quality research was not advanced.

Ages of the participants varied from 21 to 45 years, with the average of 25.6 years. The test subjects, 25 men and 5 women, completed the test in 1 hour 23 minutes on average, with the minimum time of 58 minutes and the maximum time of 1 hour 58 minutes. All the participants had good eyesight on short distances and one test subject had light color tone blindness. The

results from the color blind participant were, however, well in line with the other test subjects and therefore not rejected.

3.4.4 Test environment

The tests were arranged in the premises of the Department of Media Technology in Helsinki University of Technology, where one office room was reserved exclusively for the subjective test purposes for two weeks. The test area consisted of two tables, one reserved for a test subject, and the other for a test instructor. The table of the participant was an approximately 1 x 2 m empty white table with an adjustable seat for comfortable test experience. The other table included a computer setup for the test instructor and offered extra space for the items related to the test situation.

The room was illuminated with fluorescent lamps, providing a fairly constant color temperature of 3000 K over the test table. Approximate illumination levels ranged from 700 lx to 850 lx. During the tests, test subjects were allowed to move and place samples wherever they wanted for the best viewing experience. This helped to minimize the effect of unwanted reflections from the overhead lamps.

3.4.5 Procedure

The subjective tests consisted of five image evaluation phases. During every phase, a test subject assessed one image quality category at the time from 42 samples. Furthermore, every phase was divided in two stages: the anchor image selecting stage to define the evaluation scale limits, and the grading stage to evaluate the quality category in question. The five categories included quality, sharpness, noise, colorfulness, and contrast. From the categories listed, quality presented the only qualitative element, while the other four categories, sharpness, noise, colorfulness, and contrast, were considered as quantitative attributes for quality.

The sample image pool included two reference images, Reference Image I and II, the both of which were printed on 21 papers. The same samples were used in all the five phases. After each phase, the test subject was asked to fill a short questionnaire about the completed tasks. Similarly, at the end of the tests, an additional end questionnaire form was filled. The visual presentation of the test process is illustrated in Figure 15.

The samples were graded by using a 7-point Likert scale. The Likert scale denotes an interval scale where grade steps are equally divided and scale limits or all the grades are labeled with verbal notes or categories (Metsämuuronen 2006, pp. 62-64). For grading the samples, only the scale limits were labeled, e.g. for sharpness with a label “Extremely sharp” for the upper limit, and “Extremely unsharp” for the lower limit. The labels followed the same pattern with all the other quality categories but noise, where the scale limits were reversed: “Noiseless” was given for the upper limit and “Extremely noisy” for the lower limit. The same scale with seven grade levels was used on questionnaire answers as well, with corresponding verbal labels to describe the scale limits.

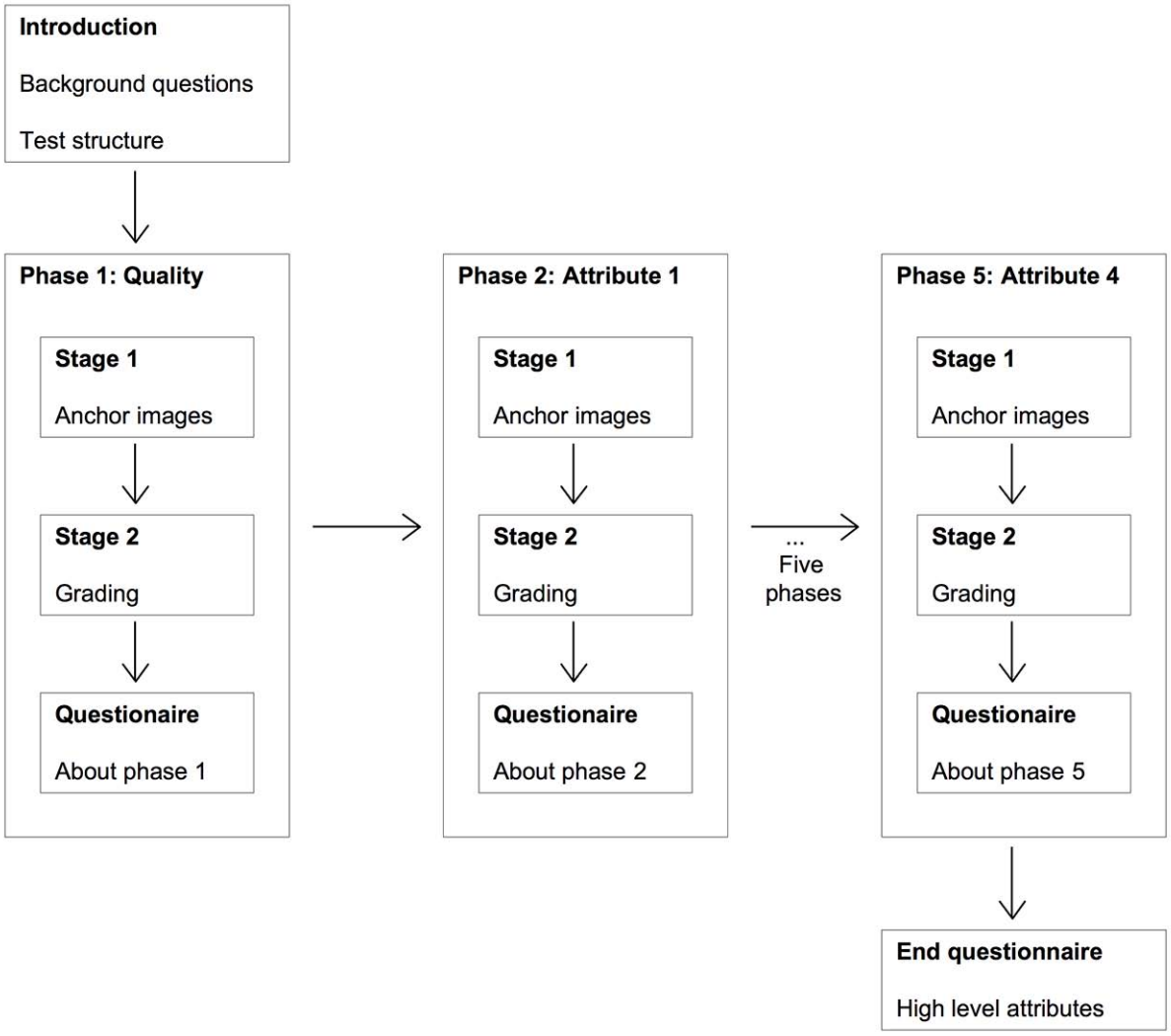


Figure 15. *The test procedure of the subjective tests*

When considering quality categories, quality was the only category where the test subjects were allowed to perform the grading based on their own personal preferences. With the other four categories, test subjects were asked to evaluate only the amount of the attribute in question, not the quality of the attribute. For example, in the case of sharpness, test persons performed grading based on how sharp the sample was, not how much they liked the sharpness of the sample. To make the evaluation simpler, quality, as the only qualitative element amongst the quality categories, was always the first category to evaluate. The other four categories with quantitative quality attributes were presented in a random order.

Before the test started, some background information about the test subject was asked, such as age, gender, eye sight, possible color blindness, and previous experience in image quality assessment. In addition, the structure and the progression of the tests were explained to the par-

ticipant. At the beginning of every new phase, an instruction paper was given to the test subject. The paper contained the instructions required for the phase, and stated the quality category under evaluation. Furthermore, to clarify the meaning of the quality category under testing, an example paper presenting a bad and a good image from that category was showed to the test subject before the first stage started. Appendix 3 presents an example page of the instructions for the quality category and two questionnaire pages in Finnish.

The first stage consisted of anchor image searching. All the samples were handed to the test subject, who searched for proper anchor samples to define the limits of the grading scale. In other words, the test subject selected one bad sample with a grade of 1 to define the lower limit of the grading scale, and one excellent sample with a grade of 7 to define the high limit. It should be noted, however, that as there were two different reference images in the sample group, the two anchor samples were not required to present the same reference image.

The second stage started by placing the selected anchor images on the table in front of the test subject. The test arrangement is illustrated in Figure 16. During grading, the test subject reviewed a sample briefly, told a grade and a sample number to the test instructor and proceeded to the next sample. After the grade was given, the test instructor recorded the information on a computer with Microsoft Excel software. The anchor samples were visible all the time for reference. In order to prevent the test subjects from learning sample numbers during successive evaluation phases, a sample number was written on the opposite side of each sample and was read only after a grade was given.

Figure 16. *An example test setting to visualize the test arrangement*

For all the five evaluation phases, the procedure remained the same. After each stage and phase, the order of the samples was randomized. During grading, the test subjects were instructed to grade samples based on their first impression. Furthermore, the participants were

not required to exploit the full scale of grades for both the reference images. For example, if Reference Image II was considered more colorful than Reference Image I, it was allowed to use a limited grading scale for either or both the images. However, from all the 42 samples including both the reference images and 21 papers, at least one sample was required to get a grade of 1 and one sample a grade of 7.

For each quality attribute, test subjects filled a question form for both the reference images about the speed and certainty of their grading decisions. In addition, test participants were asked to mark the areas they found useful in grading on a printed example image with a black marker pen. The number of areas was not limited. After all the five phases were finished, the final questionnaire was filled where the high-level aspects of the reference images were assessed with three attributes: image naturalness, balance, and pleasantness. Table 3 lists the explanations of the attributes translated in English. Finnish versions are available in Appendix 3.

Table 3. *The English translations of high-level attribute descriptions*

Naturalness	Naturalness means how real the image is. If the image is extremely natural, the observer can imagine viewing the scene of the image with his / her own eyes, like there is no image present. In the moderately natural image, the content can be considered authentic but the observer realizes he / she is only looking an image. On the other hand, the image with very low naturalness presents the content in a distorted, unreal way.
Balance	Balance expresses the balance between different objects in the image, how various items attract attention and are there objects that stand out strongly from the others.
Pleasantness	Pleasantness indicates the general visual pleasantness of the image and photographic properties, such as aesthetic expression and image composition.

3.5 Software development

3.5.1 Outline of procedure

After the subjective data was obtained, software was developed to calculate numeric scores for image quality attributes from printed Reference Image II samples. The goal of this phase was to simulate human opinion of image quality with objective methods. To reach this goal, the subjective test results were selected as the main reference data for software development.

The software was programmed with Matlab 7.1, a numerical computing environment designed for complex scientific calculations. During the design and testing phase, methods were created to compute numerical scores for image contrast, colorfulness, and noise. In addition, a method for color difference calculation was programmed. Sharpness method was under development as well, but was later excluded from the final software because of inaccurate results it produced. The finished implementations were reduced-reference measures, with the exception on the no-reference noise measure implementation by Malanin (2006).

The program was designed specifically for Reference Image II. In addition, the software was optimized for images printed with a specific inkjet printer, Epson Stylus Pro 3800, not for images adjusted with software. This practice required image digitizing, which restricted the accurate software operation only to the reference image printed and digitized with predefined hardware equipment and conversion settings, as the constant output was naturally required. The detailed description of the digitizing process is explained in Chapter 3.2.

Previous research on objective quality measures was extensively studied and various methods were tested to find the most suitable calculation method for every image quality attribute. Unfortunately, many tested methods failed to work correctly with digitized prints, which notably increased the difficulty level of the programming process.

3.5.2 Implementation

Software was programmed with Matlab 7.1. All the methods required a printed, digitized Reference Image II sample in L*a*b* format, constructed by following the digitizing process explained in the previous chapter. The program was divided in six parts: a main program to execute other subprograms, four quality attribute subprograms, and a subprogram for visual quality index calculation and final grading. The implementations of attribute subprograms are discussed in Chapters 3.5.3 – 3.5.7, whereas visual quality index is explained in Chapter 3.6. The Matlab code listings can be found in Appendices 4-8, with the exception on noise method, which is available in Malanin (2006).

The results of the subjective tests were used as reference data for the four main quality attribute methods, i.e. colorfulness, contrast, sharpness, and noise. However, the sharpness method was excluded in the later phases of the research. To acquire the reference data, the results from 30 test subjects were averaged to obtain a reference grade for each sample paper and quality attribute. For the fifth method, color difference method, reference data was extracted directly from the original, digital reference image.

During the testing and development phase, the main goal was to maximize Pearson correlations between reference grades and software calculated scores. For each attribute, various methods were tested to find the most accurate method available. If a previously known and tested method was obtainable, it was emphasized during the choosing and testing process. To ease the testing, related research from the field of print quality evaluation was extensively searched for during the development. However, as the availability of print quality research turned out to be extremely limited, the strongest concentration was placed on previous DigiQ studies on print quality and the other related research from the Department of Media Technology, such as Hiden (2008), Halonen (2008a), Malanin (2006), Nurminen (2005), and Salmi (2008). The main research under inspection was conducted by Halonen (2008a), since it shared many similarities to this study, e.g. the same digitizing process and the same partial goal: to objectively compute image quality parameters from printed natural images.

The main differences between Halonen (2008a) and this study concentrated the selection of test images and the exploited reference data. As opposed to the single reference image used in this study, Halonen (2008a) studied four natural images with diverse themes, i.e. images named as cactus, landscape, man, and studio. Furthermore, Halonen (2008a) used objective, instrumentally measured data from printed images as references. In this study, on the other hand, the reference data of the main quality attribute methods was based on the subjective quality evaluation results.

Table 4 lists the most common properties of the selected, final objective methods. Chapters 3.5.3 – 3.5.7 discuss the methods in detail. Furthermore, Chapter 2.1.3 presents brief descriptions of quality attributes in general, accompanied with introduction to the objective methods listed here. For sharpness and noise methods, the detailed descriptions are available in the corresponding method references or in Malanin (2006).

Table 4. *The main properties of the selected objective methods*

	Colorfulness	**Contrast**	**Sharpness**	**Noise**	**Color dif.**
Method	Own chroma	rms-contrast (eg. in Peli 1990)	None Tested: Marziliano et al. (2002) Crete et al. (2007)	Im-merkær (1996)	Chroma difference ΔC_{ab}
Implementation	$\mu_C + \sigma_C$ with masking	rms-contrast with masking	None	Malanin (2006)	ΔC_{ab} with normalization
La*b* component(s)	a* b*	L	L	L	a* b*
Component scale	[-128, 127]	[0, 1]	[0, 1]	[0, 255]	[-128, 127]
Matlab code listings	Appendix 5	Appendix 6	Malanin (2006) for Marziliano (2002)	Malanin (2006)	Appendix 7

3.5.3 Colorfulness

Halonen (2008a) studied colorfulness in the print context with chroma based algorithm $M^{(1)}$ by Hasler and Süsstrunk (2003), using four natural images of different themes and exploiting the same digitizing process as in this study. However, the obtained results with inkjet prints were unexpected. According to Halonen (2008a), the negative results reflected strong dependency on image content. In this study, similar approach for colorfulness calculation was taken but with one exception: the method was optimized for a single image only.

The colorfulness method was based on chroma *C(x,y)*, which was calculated from the a* and b* channels of digitized two dimensional image for each image pixel with the equation

$$C(x,y) = \sqrt{a(x,y)^2 + b(x,y)^2}, \quad (12)$$

where *a(x,y)* and *b(x,y)* are the a* and b* components of the La*b* image (Lindbloom 2003).

Furthermore, the chroma method was improved by masking out certain pixel level values from the resulting two-dimensional chroma image. In the masked image, all of the chroma values below pre-defined threshold value, 48, were set to 0 with the theoretical maximum at approximately 181. The goal of the masking was to enhance the effect of more saturated areas in the image by completely removing the values with lower saturation from calculations. The use of masking was based on an assumption that the quality differences in printing papers are subjectively the most visible in the reproduction of maximum quality measure values, e.g. in the most saturated areas for colorfulness and in the white and black areas for contrast. Thus, the effect of other areas was limited in calculations, and the exact value boundaries were set by trial and error to maximize Pearson correlations with the subjective reference data.

To illustrate the effect of masking with an example, Figure 17 presents the visualized chroma levels, complemented with calculated colorfulness scores for two paper samples, D20 (left) and D27 (right). In the subjective evaluation, D20 was in the group of less colorful samples and D27 was considered as one of the most colorful papers. The images on the upper row are unmasked and on the lower row masked.

In visual observation, the effect of masking was evident for the example image pair presented in Figure 17. Considering the masked images, the areas with high chroma levels were pronounced, e.g. the flower pot and the fruit bowl, with a major difference in the size of tablecloth. The unmasked images showed only minor visual differences, and the computed scores were nearly equal. The scores of the masked images, on the other hand, were correctly biased toward the subjectively more colorful image on the right side.

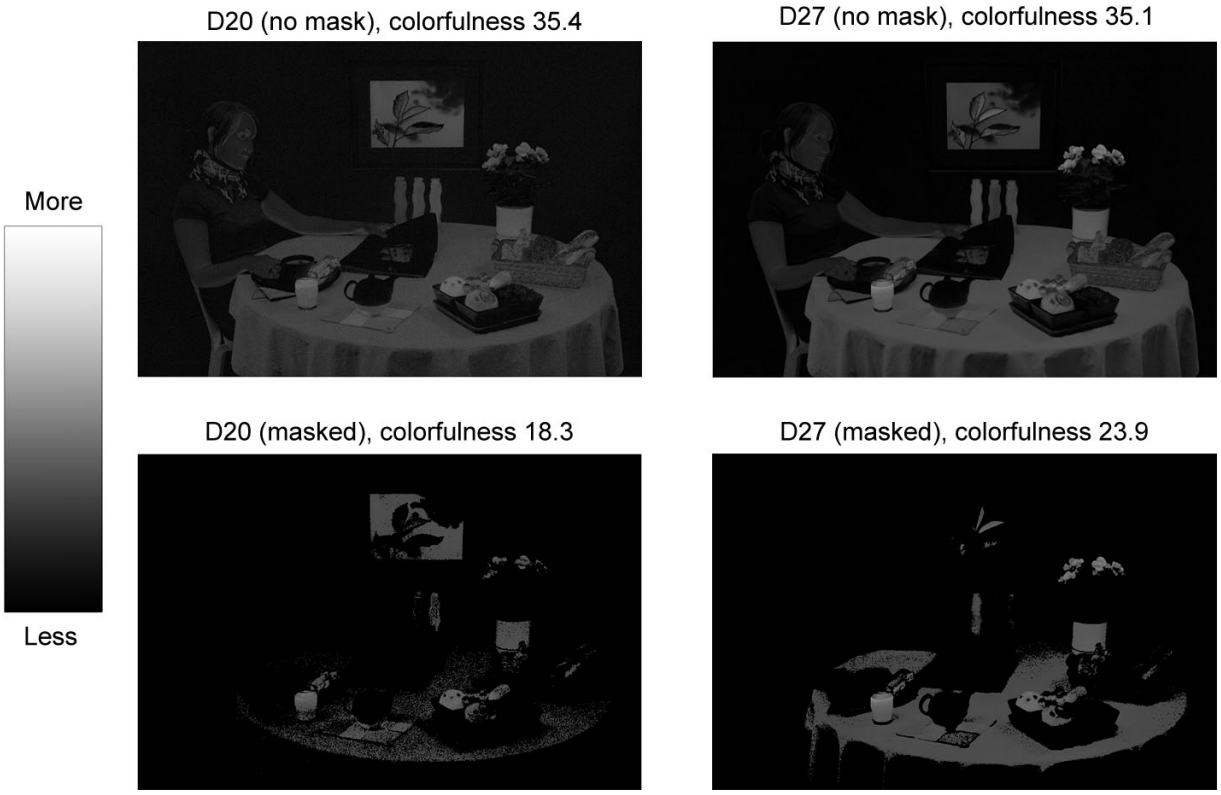


Figure 17. *The chroma level comparison of a masked and an unmasked image for two sample papers, subjectively less colorful D20 (left) and more colorful D27 (right). Whiter point indicates increased chroma level.*

After the masking was applied, a colorfulness score was computed with the equation

$$Colorfulness = \mu_C + \sigma_C, \tag{13}$$

where μ_C is the mean of chroma, and

σ_C is the standard deviation of chroma.

Before choosing Equation 13, numerous colorfulness equations listed in Hasler and Süsstrunk (2003) were tested with various weighting coefficients. However, the form of the equation turned out to have fairly minor effect on Pearson correlations, and therefore a simple equation with no weighting coefficients was chosen. In addition to chroma equations, a complex colorfulness algorithm based on a color gamut volume method by Saito and Kotera (2000) was tested with an implementation of a fellow co-worker, Mikko Nuutinen. Unfortunately, this measure showed results that seemed completely random with scanned prints and was therefore rejected.

3.5.4 Contrast

The implementation process of the contrast method was quite similar to colorfulness. In the research by Halonen (2008a), the root-mean-square (RMS) contrast method showed fairly expected correlations with inkjet samples, and was therefore chosen as the main contrast method for this study (Peli, 1990). The computation was performed on a luminance channel L to prevent color information from distracting the results.

As was the case with colorfulness, accuracy was further improved by masking. The goal was to remove the luminance variation in medium luminance levels, thus increasing the effect of higher and lower luminance levels in contrast calculation. However, only moderate masking was required. The mask was formed by setting values in a range of [0.5, 0.77] to 0.5. The exact threshold values were chosen to maximize the correlation between calculated and reference values. The equation for the RMS contrast is presented in Chapter 2.1.3 (Equation 3).

3.5.5 Sharpness

Two main methods for sharpness were tested: a blur method by Marziliano et al. (2002) and a blur method by Crete et al. (2007). The matlab code implementations for both the methods were obtained from previous DigiQ studies, namely from Malanin (2006) for Marziliano et al. (2002) and from an unreported DigiQ study for Crete et al. (2007).

In the research by Halonen (2008a), Marziliano et al. (2002) showed strong correlation with the reference data, but to the opposite direction than expected. Similar results were obtained in this study as well. Correlation was strong with both the methods, but to the unexpected direction. In other words, the sharpest samples were computed as the unsharpest and vice versa. This interesting phenomenon is further examined in the Figure 18, which presents 100% crops from two scanned example images in 100 dpi print resolution. The left sample paper D26 was subjectively evaluated in the group of sharpest papers with both the reference images, and the right sample D16 in the group of unsharpest. It should be emphasized, however, that the original subjective evaluation was conducted from printed samples, not from the scanned and digitized images as presented in Figure 18.

Observing the two images, there exists a sharpness difference between the samples, but the major difference in noise levels makes the evaluation harder. At the first glance, the right image raises the impression of higher sharpness despite of the stronger noise. The higher sharpness is especially visible in the areas of high contrast, e.g. the newspaper title, the tea cup, and the tea pan. On the other hand, the cloth under the tea pot shows more details in the left image, though the details are slightly blurred. With the right image, the details are lost under high noise. Observations remained similar with other paper samples with comparable characteristics.

Figure 18. *The sharpness comparison of two digitized paper samples, a subjectively sharp paper D26 (left) and a subjectively unsharp paper D16 (right), illustrating the unexpected visual sharpness of digitized samples.*

It can be argued if the sharpness difference was caused by the different amount of sharpening applied in the images. Even though the sharpening was turned off during the scanning, some sharpening may have been added at the driver level without informing the user, reflecting an undocumented behavior of scanner drivers. The unexpected performance may also be caused by the different noise characteristics, affecting the functionality of the scanner drivers or the sharpness methods.

To conclude, since the risk of inaccurate sharpness information of the samples definitely existed, the sharpness calculation was excluded from this study. A few simple methods for sharpness evaluation were tested as well in addition to the two main methods, e.g. a Sobel filter implementation. However, the results of the other methods followed the same pattern of incorrectly directed correlation.

3.5.6 Noise

The testing process of the noise method was the most trivial from all the objective methods studied. Similarly to the previous methods, the tested algorithm was chosen based on the research findings by Halonen (2008a), where the no-reference noise variance estimation measure by Immerkær (1996) showed expected correlation results with inkjet samples. This positive trend retained in this study as well. Therefore, excessive testing of various noise algorithms was not required, and the Immerkær (1996) Matlab code implementation by Malanin (2006) was chosen as the main objective noise evaluation measure for the ongoing research project.

3.5.7 Color difference

The development of color difference method had two major differences when compared to the other methods. Firstly, objective color difference calculation was not previously studied within the DigiQ project in the similar print context as in this thesis. Secondly, instead of obtaining reference data from subjective tests, the reference values were computed directly from the original digital reference image.

As the time for color difference method development was fairly limited, different options for color calculations were not extensively studied. Consequently, an uncomplicated chroma difference ΔC_{ab} method was chosen as the basis of color difference calculation (Imatest n.d.). Luminance data was excluded from calculations, since the focus of this method was to examine the chromatic accuracy of printing papers, not the brightness variation.

In addition, the effect of saturation variance was minimized with chroma normalization. In other words, the saturation of tested sample colors was adjusted to match the saturation of corresponding reference colors. This removed the effect of overall chroma variation from color difference calculations, the property which was already calculated with the colorfulness method. For each test color computed from a sample image, a* b* channels were normalized with the equations

$$a_{norm} = \frac{C_{ref}}{C_{samp}} \times a_{samp} \quad and \quad b_{norm} = \frac{C_{ref}}{C_{samp}} \times b_{samp}\,, \tag{14 \& 15}$$

where a_{samp} and b_{samp} are the averaged a* and b* components, and

C_{ref} and C_{samp} are the chroma values of the reference and the sample (Imatest n.d.).

After chroma normalizing, chroma difference ΔC_{ab} was calculated for each test color with the equation

$$\Delta C_{ab} = \sqrt{(a_{ref} - a_{norm})^2 + (b_{ref} - b_{norm})^2}\,, \tag{16}$$

where a_{ref} and b_{ref} are the a* and b* components of the reference color, and

a_{norm} and b_{norm} are chroma normalized a* b* values for the sample image (Imatest n.d.).

The same saturated test colors as introduced in Reference Image II were used in color difference calculations, i.e. red, green, blue, cyan, magenta, yellow, and orange (Chapter 3.3.4). Three neutral greys were excluded. All the values, including the reference values and the sample values, were averages computed from pre-defined two-dimensional square areas in the image. The reference color values were calculated from the full sized reference image with 4288 x 2848 pixel resolution, with a square size of 15 x 15 pixels for yellow and 25 x 25 pixels for the other colors. The sample values to be compared with, on the other hand, were averaged from smaller images with 876 x 584 pixels, representing the normal resolution of printed and

scanned images in this study. Therefore, to obtain comparable results, only 1/5 part of the original square dimensions was needed for sample value calculations, resulting in the area size of 3 x 3 pixels for yellow and 5 x 5 pixels for the other colors. The center points for the test colors, however, remained the same as with the original test colors (Appendix 2).

Compared to the original test colors of Reference Image II, the selected area dimensions were only a half from the originals. The main reason for smaller areas was the problem with paper movement during the scanning process, leading to inaccurate results caused by overlapping color areas. This unfortunate phenomenon is illustrated in Figure 19, which presents the placement of square areas in a double 11 x 11 pixel size for the original, downsampled reference image (left) and for an incorrectly aligned sample (right). Both the images share the same center point coordinates. In the right image, the square area clearly overlaps the borders of the magenta color patch. With a tighter area size of 5 x 5 pixels, the squares were adjusted properly inside the color borders.

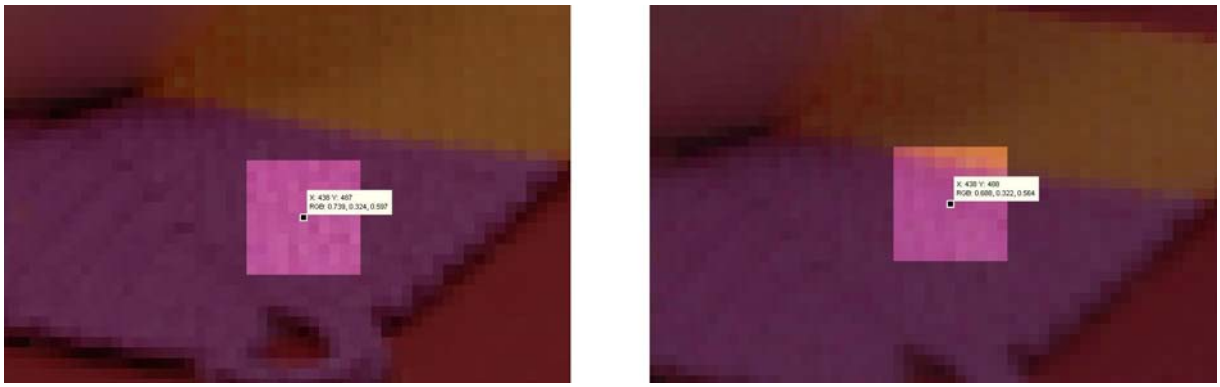


Figure 19. *The effect of paper movement during the scanning process with oversized sample areas. The left image presents a crop from the downsampled original image, and the right image shows the worst case of the sample paper set. The center coordinates are the same for both the images.*

The accuracy of the method was affected not only by the paper movements, but also by the inaccuracies caused by paper noise. Despite the fact that the pixel averaging decreased calculation errors caused by noise, the accuracy of calculations was definitely compromised especially with the square area of 3 x 3 pixels. For these reasons, the implementation of more accurate color difference methods, such as CIEDE 2000, was not considered worth the effort.

3.6 Visual quality index

3.6.1 Outline of procedure

In the final phase of this study, the software was expanded to combine the calculated quality scores from the objective Matlab methods into one grade for visual print quality, visual quality

index (VQI). This print quality grade was designed to estimate subjective image quality of prints as closely as possible, thus providing means for predicting human opinion on print quality with software. A statistical estimation method, regression analysis, was utilized with SPSS 14 software to provide statistical equations for VQI calculations. Furthermore, in the hope for more accurate results, additional regression models were generated for colorfulness, contrast, and noise. Grading was performed with 7-point interval scale, and complemented with corresponding verbal grades.

Like was the case with the software development, the reference data consisted of the subjective test results, presenting the regression dependents in regression analysis. The regression predictors, on the other hand, included the four Matlab methods, i.e. colorfulness, contrast, noise, and color difference. The VQI code shared the same restrictions on software usage as the code programmed in previous development phase, and was only usable with digitized Reference Image II prints.

3.6.2 Regression analysis models

Traditionally, regression analysis is defined as a statistical analysis method to describe a single, normally continuous variable (dependent) with multiple variables (predictors) (Metsämuuronen 2006, p. 677). In this study, predictors were gathered from the quality attribute results obtained from the objective Matlab methods, basically colorfulness, contrast, noise, and color difference. The described variables, dependents, consisted of the subjective test results for the quality categories, i.e. quality, colorfulness, contrast, and noise. The models to describe the dependents were based on the equation

$$Y = A + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \varepsilon, \tag{17}$$

where A is a constant calculated during the regression analysis,

X is a predictor and β is a weighting coefficient for the predictor,

and ε is an estimation error for the model (Metsämuuronen 2006, p.680).

While building the models with SPSS 14 software, the recommended regression settings listed in Field (2009, pp. 225-233) were used. Furthermore, the following methods of regression were utilized. For quality, the regression analysis was executed with a stepwise method with all the Matlab predictors included. For colorfulness, contrast, and noise, a hierarchical method with the enter and stepwise blocks was exploited, forcing the predictor based on the dominating quality attribute to be included in the first regression block, and adding all the rest predictors with stepwise method in the second block. The methods were chosen based on the recommendations by Field (2009, pp. 212, 225), who described a hierarchical method with forcing the first block for the most important predictor as a general starting point for model building. Furthermore, every model was carefully analyzed for goodness and reliability according

to the statistical analysis guidelines explained in the next chapter. If required, predictors were dropped out from the final model, which was indeed the case with the colorfulness model.

3.6.3 Model analysis

Each of the four models was carefully analyzed for goodness and reliability. After considering the recommendations by Metsämuuronen (2006) and Field (2009), the model analysis was based on the following guidelines listed in Table 5. In addition, the presence of outliers was considered in the data. However, as the regression dependents consisted of the calculated averages of 30 test subjects and the predictors were based on the same data, there were no single sample outliers in the data.

	R^2	**Sig F Change**	**Sig**	**Durbin-Watson**	**VIF**	**Condition Index**
SPSS section	Model Summary	Model Summary	ANOVA	Model Summary	Coefficients	Collinearity Diagnostics
Guideline	1 the best	< 0.05	< 0.05	[1, 3]	< 10	< 30

Table 5. *The guidelines for regression model analysis with SPSS, based on the recommendations by Field (2009) and Metsämuuronen (2006).*

As a base analysis for models, Field (2009, p. 237) lists four general recommendations to be followed when analyzing regression analysis models generated with SPSS. Firstly, the value of R Square (R^2), presented in the Model Summary section of SPSS result print, should be as near the value one as possible. R^2 is the coefficient of determination of the model, describing how many percent from the variance of the dependant the model explains, and is therefore the main indicator for model goodness (Metsämuuronen 2006, p. 686). Secondly, Sig F Change in the Model Summary section should be less than 0.05 for each added predictor, indicating a statistically significant change in the model caused by applying the predictor. Further, the significance of the added predictor can be tested with ANOVA analysis as well, where the column Sig on results should contain values under 0.05 to indicate a statistically significant change. Finally, Durbin-Watson statistic in Model Summary section should present a value between one and three, ideally two. Durbin-Watson examines the realization of the regression assumption of independent regression errors, where the value two indicates totally independent errors, an ideal situation for regression analysis (Field 2009, p. 236).

Another important aspect in model analysis is the multicollinearity characteristics of predictors. Multicollinearity is caused by strong correlations between predictors, reducing model reliability by increasing the standard deviation of a weighting coefficient β and limiting the size of R, thus presenting an unwanted property for regression variables (Metsämuuronen 2006, p. 689; Field 2009, pp. 223-224). To analyse multicollinearity, two diagnostic measures were examined for each model: VIF (Variance Inflation Factor) in Coefficients section of SPSS re-

sults, and Condition Index in Collinearity Diagnostics section. In the case of VIF, Field (2009, pp. 223-224) states that no hard and commonly accepted rules are available for recommended values. However, Myers (1990 cited in Field, 2009) suggests VIF values higher than 10 as a real concern for multicollinearity. For Condition Index, Belsleyn et al. (1980 cited in Metsämuuronen, 2006, p. 700) lists two boundary values. With Condition Index values higher than 15, there might be some problems with multicollinearity. With values over 30, the problems are major.

3.6.4 Model generalization

The next step in model analysis was to examine whether or not the model can be generalized beyond the selected sample to the whole population. In the case of this study, a generalized model would give reliable results with all inkjet papers, not only with the selection of 21 papers used in the subjective tests. To generalize a model, three regression assumptions must be met: the linearity of residuals (regression errors), the homoscedasticity of residuals, and the normal distribution of residuals (Field 2009, pp. 247-251). If any of the assumptions is violated, the model cannot be generalized to the whole population. However, even with violated assumptions, the model can be perfectly good for describing the selected sample (Field 2009, p. 251).

The linearity and the homoscedasticity of residuals were examined from residual charts based on the recommendations by Field (2009, pp. 247-251). Two kinds of residual charts were created: a scatter plot for full model residuals and partial regression plots for the residuals of the regression predictors. In SPSS, a scatter plot with the following axis was generated for each model: on Y-axis, the standardized residuals; and on X-axis, the standardized predicted values of the dependent variable. Similar charts are presented in Figure 20, which also shows the general guidelines for checking the assumptions of homoscedasticity and linearity. Furthermore, partial plots for each regression predictor were created, presenting the scatter plot of dependent residuals and all the predictors when the both variables are regressed separately on the remaining predictors.

To analyze scatter plots, the following properties listed by Field (2009) were examined. If the chart contains dots randomly spread over the chart, the result is good with no linearity and heteroscedasticity in the residuals. If the dots are more spread out in some points of the chart than in others, forming a linear tunnel from the one side to the other, an assumption of homoscedasticity is probably violated. The nonlinear shape of dots estimates the violation of non-linearity assumption. If both of the described dot forms are combined into one chart, both the assumptions are violated. For partial plots, the analysis recommendations remained the same. However, the shape of partial plots is generally ascending or descending, indicating the relationship between the predictor and the dependent. The assumptions must be met with every chart, otherwise the model generalization may be compromised. (Field 2009, pp. 247-251)

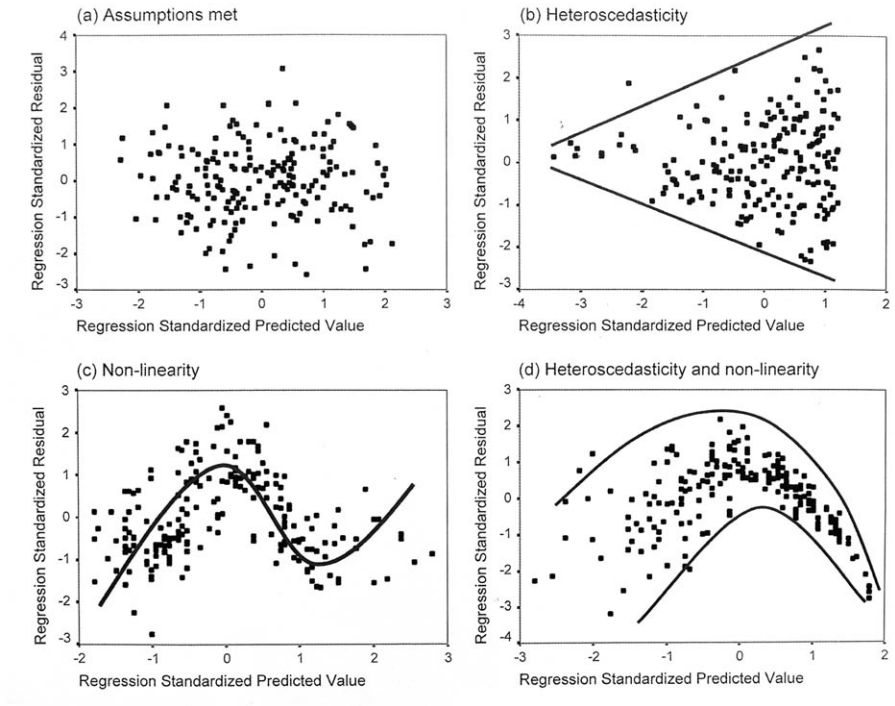


Figure 20. *Recommendations for checking the assumptions of homoscedasticity and linearity on regression analysis models (Field 2009, p. 248)*

The normality assumption was checked by examining the histograms of standardized residuals and the normal propability plots (P-P plots) generated in SPSS. According to Field (2009, pp. 247-251), the histograms of standardized residuals should form a continuous, normally distributed shape. Furthermore, the P-P plots should present a straight, diagonal line of dots. If the following requirements are not met, the regression assumption of normally distributed residuals is probably violated.

3.6.5 Model implementation and grading

Similarly to the software development phase, the regression model equations were implemented and computed in Matlab 7.1. The calculations were added as a part of the previously development code, thus allowing an easy execution of all the required Matlab calculations at the same time. A code listening for the regression analysis calculations and grading is presented in Appendix 8 and the main code to execute all the Matlab calculations is listed in Appendix 4.

What comes to grading, the model equations generated grades with the same numeric scale as in the reference data. In other words, calculated numeric grades were distributed between 1 and 7, similarly to the 7–point Likert scale used in the subjective tests, and no further modifications or changes to scaling were required. To simplify the grades further, verbal grades were

also given for each paper sample. The verbal grading was based on the recommendation ITU-R BT.500-7, 1995, widely applied in the imaging industry, presenting five category labels for quality evaluation: excellent, good, fair, poor, and bad (Engeldrum 2000, pp. 124-126). Furthermore, two additional labels, very good and satisfactory, were included to obtain the corresponding verbal labels for all the seven numeric grade levels. The equivalence of the verbal and numeric grades was as follows: Excellent >=6.5, Very good [5.5, 6.5[, Good [4.5-5.5[, Fair [3.5, 4.5[, Satisfactory [2.5, 3.5[, Poor [1.5, 2.5[, and Bad <1.5.

4 RESULTS

4.1 Reference image

4.1.1 Quality aspects

The goal on the reference image development was to create a natural, balanced, and technically high quality reference image suitable for subjective and objective testing. Two versions of the image were prepared, a display version intended for display viewing with common display devices, and a print version optimized for the print quality testing. The display version of the final reference image, named as Reference Image II in this thesis, is presented in Figure 21. The larger, entire page version of the image is available in Appendix 9. For further details, Figure 22 shows the RGB histogram of the reference image and a 100% 200 dpi crop displaying the smaller details captured in the image. Figure 23 presents the comparison of the unedited and edited image.

Figure 21. *The display version of Reference Image II*

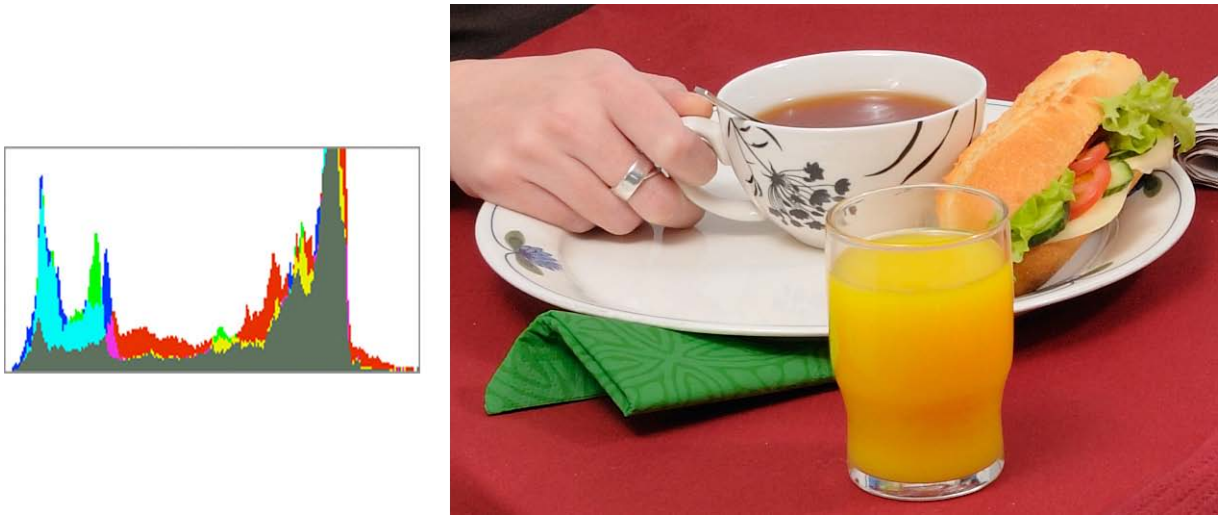


Figure 22. *The RGB histogram of Reference Image II and a 100% 200 dpi crop to reveal the finer details available in the reference image.*

Figure 23. *The comparison of the unedited, straight out of the camera version of Reference Image II (left) and the final, fully edited image (right)*

Reference Image II presents a breakfast situation with a human model sitting in front of a table. The model is reading a newspaper, directing attention from the model to the center of the table. The objects on the table were chosen to correspond with the breakfast theme. When selecting objects, five main properties were considered: quality aspects, the implementation of test colors, the implementation of memory colors, color balance, and aesthetical aspects.

Due to the use of flashes and a relatively large f/7.1 aperture resulting in minimal light diffraction, extremely high sharpness over the whole image was attained. All the objects were well inside the high sharpness area, with only the front side of the table showing slight signs of sharpness reduction, mainly on the apple and on the front tip of the tea cloth. However, that was fixed by applying light extra sharpening on the affected areas and was only visible under critical display viewing on 100% pixel per pixel level. What comes to lighting, the two Nikon SB-600 flashes provided sufficient light to properly expose the image, resulting in constant

lighting without distracting, strong shadows. As proven with the RGB histogram in Figure 22, the image was well exposed without lost highlights in any RGB channel. The only visible area with lost details, the specular highlight on the top of the tea pot caused by a flash reflection, caused no concerns because of its small size.

When considering quality evaluation, the reference image was required to provide objects and areas to effectively assess image sharpness, contrast, noise, and colorfulness. In the case of sharpness, the most detailed objects were placed on the middle of the table, including the breadbasket, the newspaper, and the human model. The camera was focused on the breadbasket, providing the highest sharpness in the nearby areas. For contrast evaluation, the frame of the wall picture was chosen to offer high contrast between bright and dark areas. As for subjective noise, the darker areas on the red tablecloth and the dark shirt of the model were considered as good starting points for evaluation. What comes to colorfulness, the vast variety of saturated objects offered numerous options for colorfulness evaluation.

4.1.2 Test colors and memory colors

Seven test colors and three neutral grey areas were implemented in the image to enable an option for objective color accuracy measures. Figure 24 shows the color objects with test colors on the top of black and white background. The test color values and locations in a pixel level are listed in Appendix 2. When comparing the test values to the reference color values obtained from BabelColor averages, the following error levels were attained: Adobe RGB ±1 value, sRGB ±2 values, and L*a*b* ±2 values (Pascale, 2006). Two exceptions to the error levels were listed for sRGB values: for yellow test color, B channel with the error of -4; and for orange test color, B channel with the error of -3.

In addition to test colors, memory colors were considered in object selection as well. The following areas and objects were the main sources for memory colors: a picture on the wall with blue sky and green leaves, a plant with red flowers and additional hues for green foliage, a fruit bowl with various well-known fruits, a breakfast bread filled with vegetables, and a human model with caucasian skin color. A breadbasket included three familiar breads with less saturated colors.

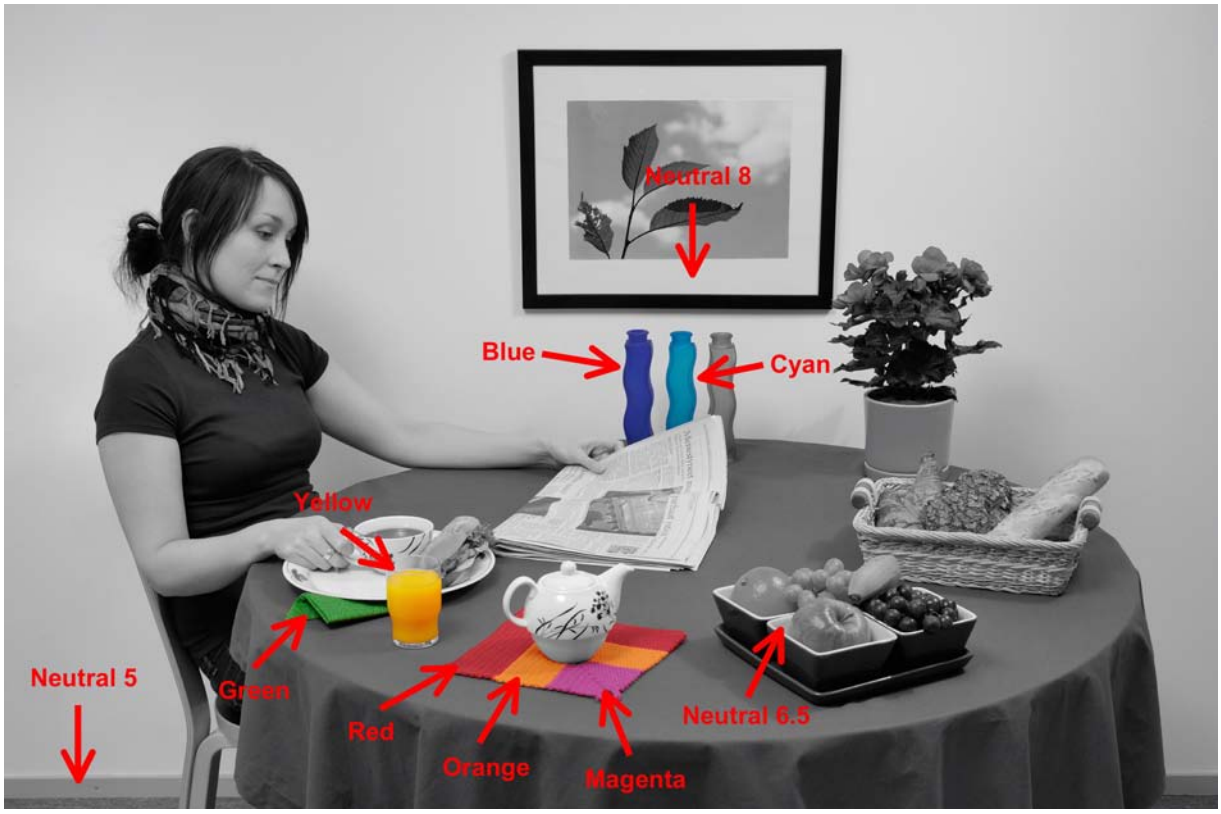


Figure 24. *The natural objects with test colors in Reference Image II*

4.1.3 Color balance and aesthetical aspects

With all the colorful objects required, attaining aesthetically pleasing color balance presented a major challenge for the development. From the objects selected, the tablecloth was certainly the most visible one. The red color was chosen after many trials and errors to match the colorful appearance of the image. With less colorful tablecloths, objects with strong colors gained too much attention and distracted the overall color balance. The tested tablecloth colors included brown, bright orange, dark blue, dark green, and light yellow. From the listed options, the red tablecloth with no surface texture was considered the most balanced choice in terms of overall colorfulness and the contrast between the objects and the tablecloth.

From the aesthetical point of view, image composition and object placements were carefully planned to provide a visually pleasing image. The base structure of the image was formed with three shapes: the rectangular painting, the round table, and the sitting human model. Three flower vases with wavy forms, interrupted by the inclined newspaper corner, complemented these simple shapes delicately. The hands of the model presented another visible shape in the image, forming a natural line from the human body to the objects on the table. Furthermore, to enhance the symmetrical appearance of the image, the objects on the table were arranged on a round form with the newspaper on the center. The round form also simplified object place-

ment, increasing the visibility of individual objects without other items blocking the line of view.

Many of the choices were made keeping in mind the Gestalt grouping suggestions of the attractive forms and shapes for a human observer (Chapter 2.5.3). From the Gestalt groups, especially closure and similarity gained attention during the development. The major shapes in the image, e.g. the table, the painting, the human model, and the object groups on the table, were presented in the way that full, closed shapes were preserved without major interruptions caused by the other objects. Similarity was also considered, with the most obvious areas of similarity visible in the three wavy flower vases and the round assortment of the objects on the table. The Gestalt forms were used to orient more attention to the table, since the human model was known to gather the most attention (Salmi 2008). Another important reason for preserving the Gestalt forms was just to increase the attractiveness of the image.

In addition to Gestalt grouping, color groups were also considered in object placement. Since the image was filled with colorful objects, the colors were distributed over the image evenly to prevent any area from gaining too much attention over the other. Furthermore, the theory of the opposite colors was regarded when forming the color groups by avoiding the presence of complement colors in nearby objects. This, however, turned out to be quite demanding, as there existed numerous other aspects to consider as well, such as the size of the objects and other compositional properties. In the final reference image, the most obvious complement color pair was formed with the green napkin and the red tablecloth. However, the strongly saturated, yellow juice glass brought balance to this colorful setting.

4.2 Subjective tests

4.2.1 Grades for quality categories

Subjective tests were carried out to obtain subjective quality data for 21 printing papers. With the effort of 30 test subjects, 42 samples were graded in five quality categories: quality, colorfulness, contrast, sharpness, and noise. In addition to Reference Image II, the evaluation was performed on Reference Image I by Salmi (2008) as well. However, as the scope of this study was limited, the result analysis concentrated mainly on the individual results of Reference Image II and the differences between the two reference images.

Figure 25 presents the visualization of averaged subjective grades as a function of sample papers for Reference Image II. The standard deviations of the grades are shown in Figure 26. For further details, the numeric results for both the reference images are listed in Appendices 10 and 11.

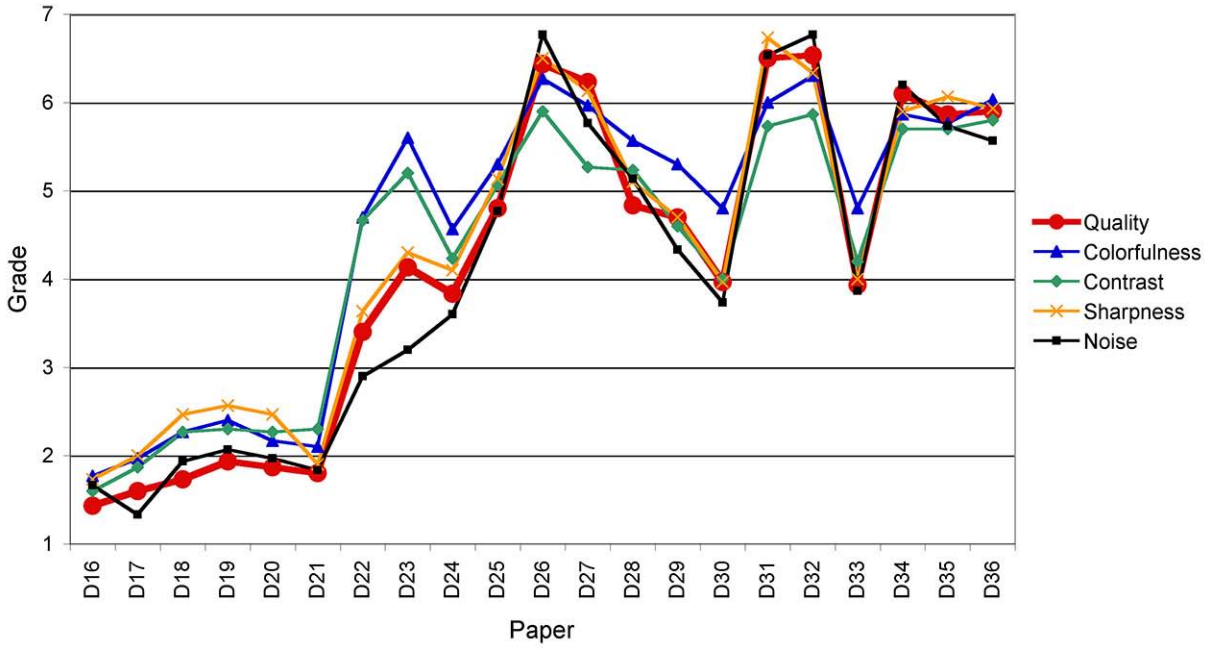


Figure 25. *The grades of quality attributes for Reference Image II printed on 21 sample papers, averaged from the results of 30 test subjects*

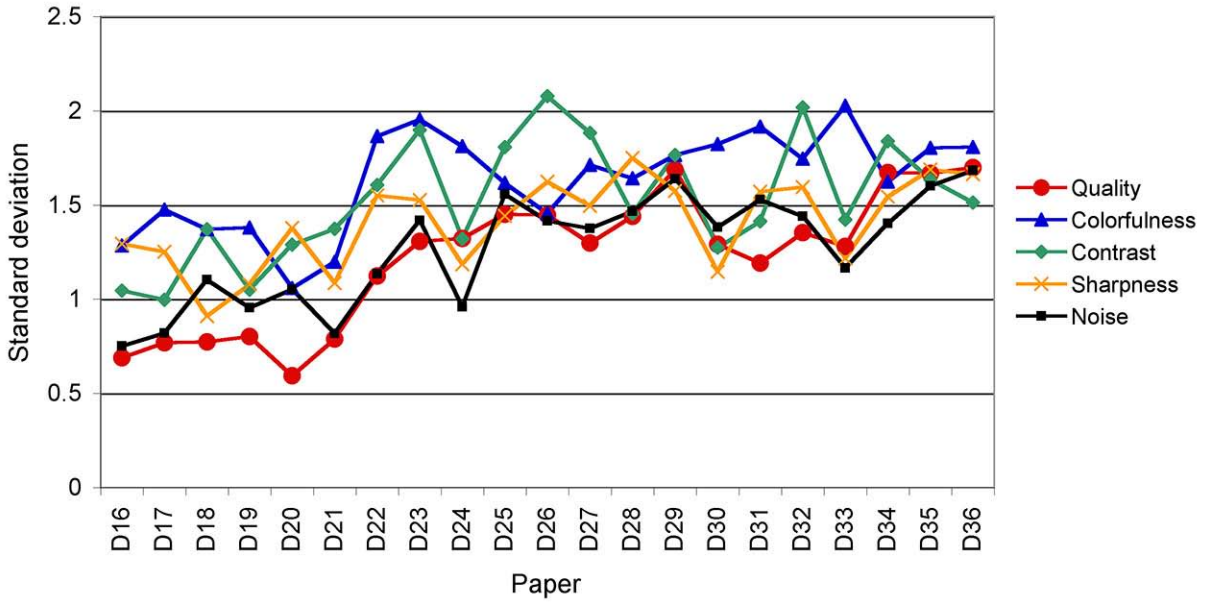


Figure 26. *The standard deviation of the grades*

As Figure 25 shows, the grades were well adjusted between the anchor limits of 1 and 7. In visual observation, the most distinctive property was the significantly lower grades for the first six sample papers (D16 – D21) on all the five quality categories. This property was expected, since the listed papers presented the only multipurpose papers in the sample set and the other

15 papers were intended especially for inkjet printing. However, this dissimilarity caused an empty gap of grades between the two paper types in the area of low to medium grades, ranging from approximately 1 grade to noise, and 1.5 grades for quality and sharpness, to even 2.5 grades for colorfulness and contrast. Considering the higher grades, the best three papers (D26, D31, D32) stood out quite clearly from the sample set. Again, this was expected, since the first two papers were used for optimizing the print version of the reference image, and the research by Halonen (2008a) reflected similar results.

What was more interesting, though, was the high similarity between the quality attribute grades and the quality grades. As Figure 25 indicates, with the exception of two papers (D22, D23), the range of grade variation between the five quality categories was one full grade at the maximum, which indicated strong cross-category dependency in the results. The similarity in grades was observed in correlation analysis as well.

Table 6 lists correlations between the quality grades and the quality attributes with two correlation methods, the Pearson product-moment correlation and the Spearman's rank-order correlation, also known as Spearman's rho. All the correlations were statistically significant at 0.001 level. Between quality and the quality attributes, sharpness and noise correlated more strongly than colorfulness and contrast. Nevertheless, with all correlations exceeding 0.95, the correlations were much higher than anticipated, indicating a strong dependency between quality and other attributes in subjective analysis for the sample paper selection used.

Table 6. *The correlations between the overall subjective quality and the four quality attributes for Reference Image II*

	Pearson	**Spearman**
Colorfulness	0.962 ($p<0.001$)	0.976 ($p<0.001$)
Contrast	0.962 ($p<0.001$)	0.954 ($p<0.001$)
Sharpness	0.994 ($p<0.001$)	0.980 ($p<0.001$)
Noise	0.985 ($p<0.001$)	0.983 ($p<0.001$)

4.2.2 Evaluation speed and certainty

In addition to quality attribute grading, the test subjects estimated the speed and certainty of their grading decisions. Figures 27 and 28 list the averaged results, as well as the standard deviation of the results for both the reference images tested.

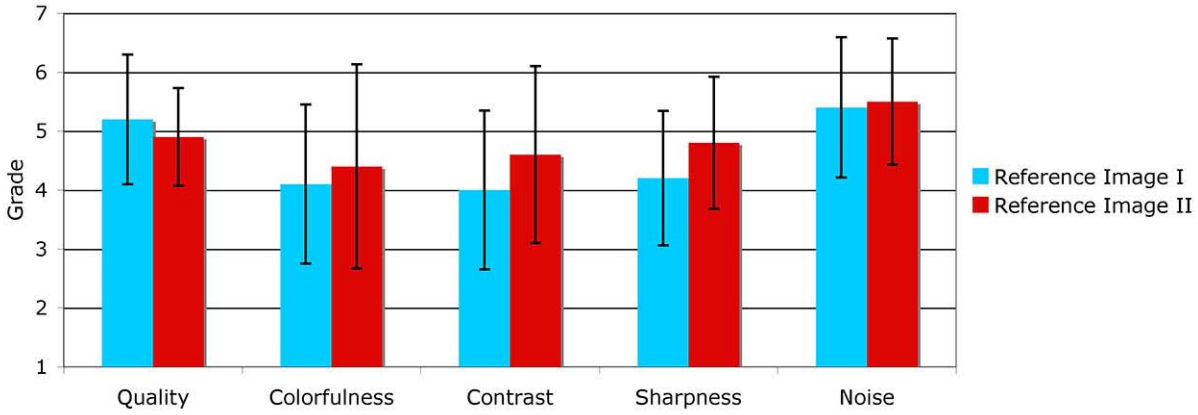


Figure 27. *The subjective speed of assessment*

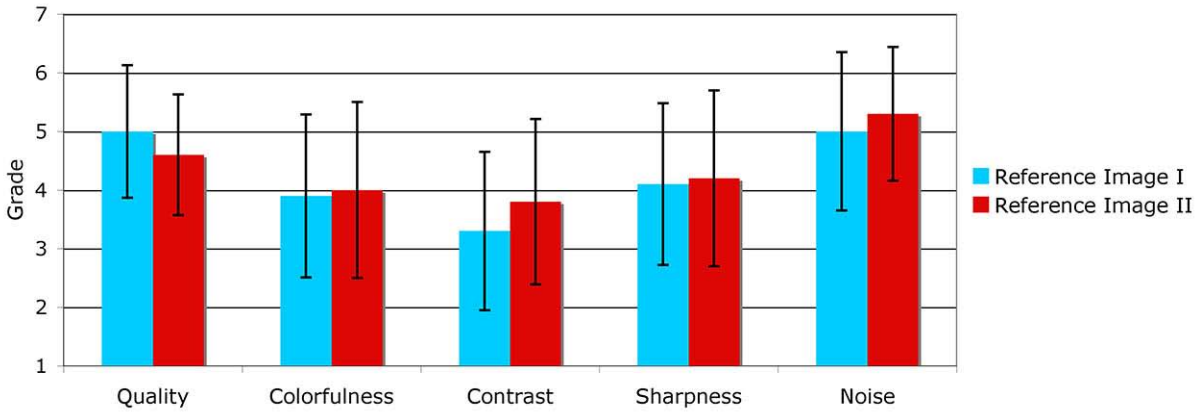


Figure 28. *The subjective certainty of assessment*

Considering all the grades and both the images, each category except the certainty grade of contrast got the average grade of 4 or higher, which was regarded as an acceptable result for a reference image. In the case of the quality categories, quality and noise were graded faster and with more certainty than the other three categories. For quality, standard deviations were lower than average as well. Furthermore, Figure 26 showed the lowest standard deviations of Rederence Image II for quality and noise. Standard deviation behaved similarly in the case of Reference Image I as well. These findings suggest that in the terms of evaluation speed, certainty, and reliability, the quality and noise categories were the strongest for both the reference images.

When comparing the grades between the two images, differences were fairly subtle. While Reference Image I had a slight edge on the speed and certainty of quality, Reference Image II had a minor advantage in the estimation speed of colorfulness and a moderate advantage of approximately 0.5 grades on the speed of sharpness and contrast, as well as in the certainty of contrast. Thus, in the light of these results, contrast and sharpness evaluation showed an advantage for Reference Image II, albeit the differences were minor.

4.2.3 High-level attributes

The aesthetical perspective of the reference images was studied with three high-level attributes: image naturalness, balance and pleasantness (Figure 29). While Reference Image II obtained fair grades of 4 in every category, Reference Image I graded higher in balance and pleasantness by one full grade, which was considered as a notable difference in favor for Reference Image I.

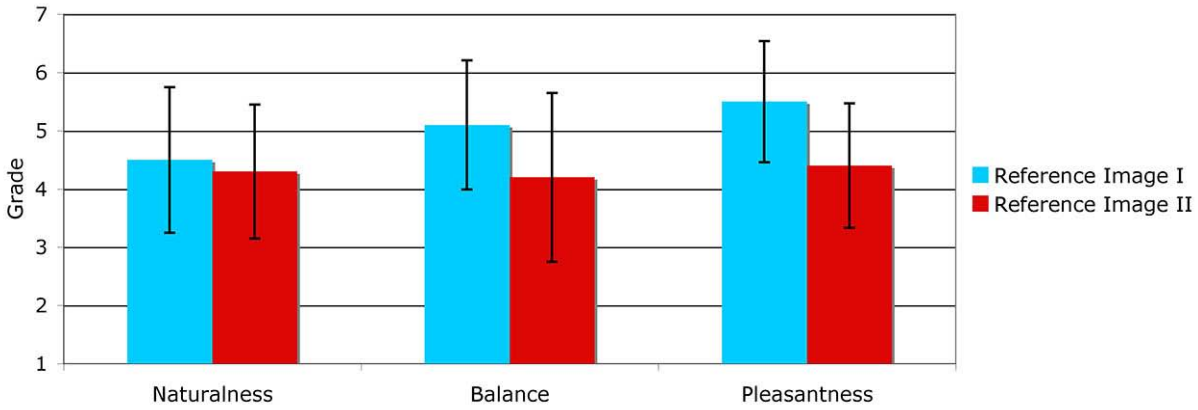


Figure 29. *The high-level aspects of the reference images*

The issue was examined further by assessing the grades of individual test subjects. Figure 30 presents the subtraction of high-level grades between the two reference images for each of the 30 test subjects. The grade difference was calculated by subtracting the grade of Reference Image I from the grade of Reference Image II. This way, the bar on the left hand side intends stronger grade for Reference Image I, and the right side for Reference Image II respectively. Furthermore, Table 7 lists how many times the reference images got higher grades in relation to each other. The cases with no difference were also counted.

Table 7. *The number of cases where the corresponding reference image got a higher grade when compared to the other reference image*

	Ref. Image I	**No difference**	**Ref. Image II**
Naturalness	15	3	12
Balance	18	5	7
Pleasantness	21	5	4

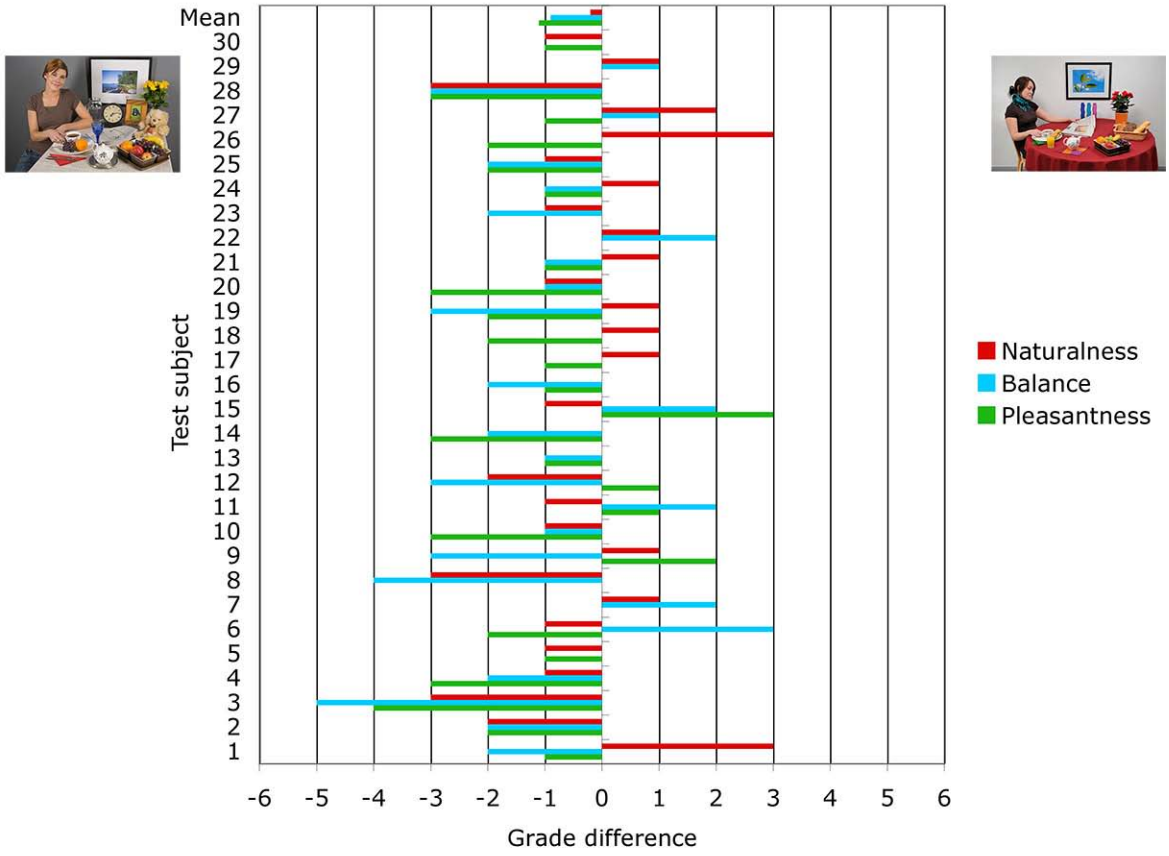


Figure 30. *The subtraction between the high-level grades of the two reference images for each test subject. The left hand side denotes better grade for Reference Image I, and the right side for Reference Image II.*

Inspecting Figure 30, the bars are apparently concentrated to the left hand side, denoting the better grades for Reference Image I. While with naturalness the difference between the images was fairly subtle, with balance and pleasantness the effect was pronounced. This phenomenon was easily observed in Table 7 as well, where the count numbers are considerably higher for Reference Image I with balance and pleasantness. Thus, considering the average values presented in Figure 29, the listed observations supported the notified one grade difference in average grades with balance and pleasantness.

4.2.4 Useful areas in evaluation

Similarly to the research by Salmi (2008), information about the useful image areas and objects for quality evaluation was gathered during the subjective tests. Figure 31 shows the areas of interest from 30 test subjects for Reference Image II, combined from hand made pen markings to a single image for each quality category with Adobe Photoshop CS4. The usefulness of a region increases with the redness level of the region. However, it should be noted that as the

method for producing the image was fairly inaccurate, any extensive conclusions should not be made based on Figure 31 only.

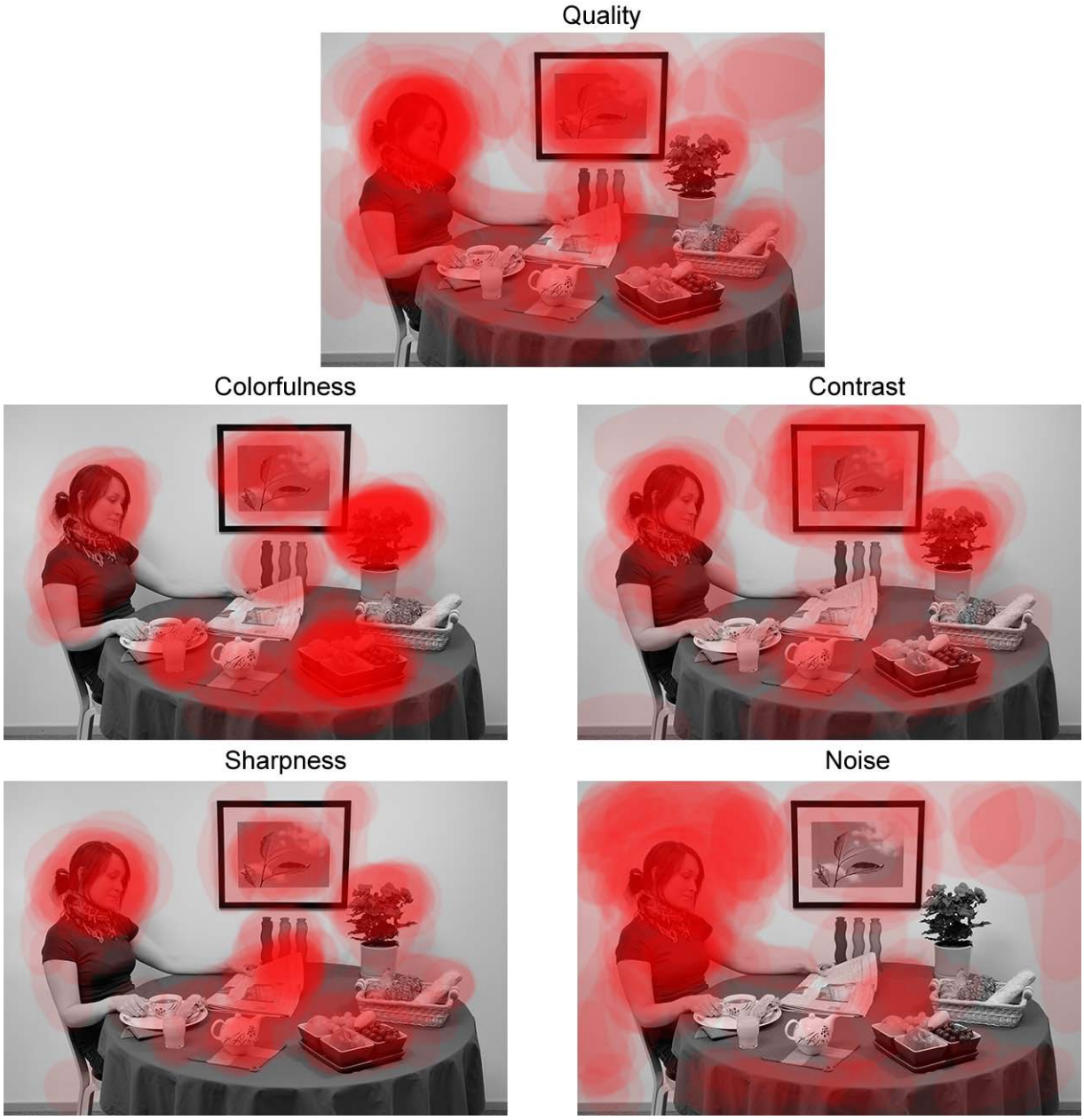


Figure 31. *Useful regions in subjective evaluation for Reference Image II. The redder the area is, the more useful it was considered by the test subjects.*

Based on the visual presentation of Figure 31, Table 8 lists the most useful regions and objects for each quality category divided in two usefulness levels: highly useful and fairly useful. The levels were defined by calculating the pixel level subtraction from the red channel and the black and white background image. For highly useful level, the objects with greater than 150

pixel difference were counted, while fairly useful level was formed from objects with >50 pixel difference. The maximum value for pixels was 255.

Table 8. *Useful objects and areas in subjective evaluation, based on Figure 31*

	Quality	**Colorfulness**	**Contrast**	**Sharpness**	**Noise**
Highly useful	Human (face) Painting Human (arm) Newspaper	Flower Fruits Orange juice Painting	Human (face) Painting (border) Painting Flower	Newspaper Human (face)	Human (face) Background Human (body)
Fairly useful	Human (body) Flower Vases Fruits Breadbasket Tea pan&cloth Tea set Background	Human (face) Human (arm) Vases Tea pan&cloth Table cloth	Human (body) Fruits Tea pan&cloth	Tea pan&cloth Flower Painting Bread basket Fruits Vases	Human (arm) Table cloth Painting (border)

As was expected from the findings by Salmi (2008), the human model, especially her face, turned out to be the most important area for subjective image quality evaluation. In four categories of five, the face was amongst the most useful regions of observation. Furthermore, in every quality category at least one additional area was considered highly useful. In the case of less useful regions, the attention was scattered more evenly over various objects, resulting in three to eight distinct areas of interest per quality category. With quality, all the areas excluding the tablecloth were considered at least fairly useful.

4.3 Software development

The main intention on software development was to provide means for automatic quality analysis of inkjet printing papers. For this purpose, four different calculation methods were developed for quality estimation with Matlab 7.1 software. Three of the methods concerning quality attributes, i.e. colorfulness, contrast, and noise, were based on the subjective reference data obtained from the subjective tests. The fourth method, color difference, used the original Reference Image II as a reference. A sharpness method was under development as well, but was not finished because of the unexpected difficulties faced during the testing.

For the three quality attribute methods, the development was based on maximizing correlations between calculated attributes and the subjective reference data. The correlations for final methods are presented in Table 9. A visual presentation of method results, in addition to numeric scores, is available in Appendix 12.

Table 9. *The correlations between calculated attributes and corresponding subjective attributes*

	Pearson	Spearman
Colorfulness	0.903 (p<0.001)	0.672 (p=0.001)
Contrast	0.946 (p<0.001)	0.905 (p<0.001)
Noise	-0.946 (p<0.001)	-0.952 (p<0.001)

All the correlations were statistically significant at 0.001 level in expected directions except the Spearman's rho of colorfulness (at 0.01 level). With all the Pearson correlation coefficients exceeding or equaling 0.9, the correlation results were higher than expected at the beginning of the development process. Comparing all the three methods, colorfulness presented slightly lower Pearson correlation and distinctly lower Spearman coefficient than the other methods, but the correlations were considered impressive nevertheless.

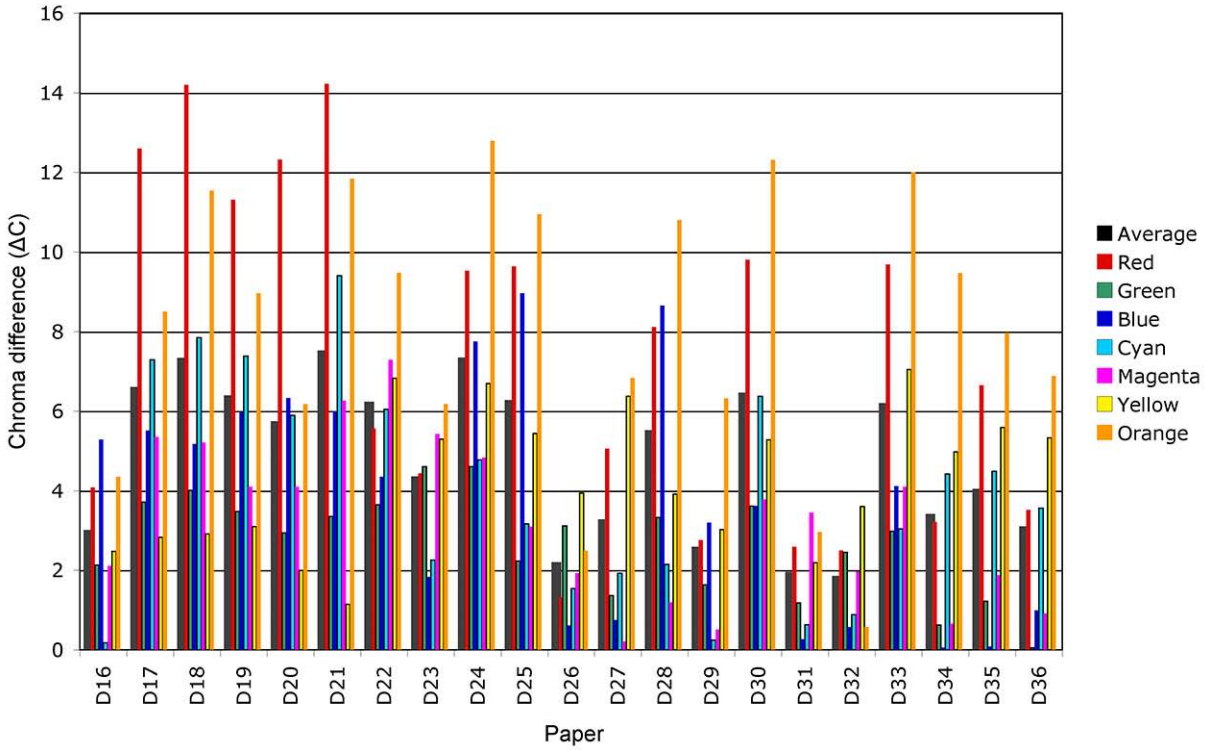


Figure 32. *Color difference for the test colors, presented as chroma difference in relation to the original Reference Image II. Longer bar indicates stronger defect in color accuracy.*

Figure 32 presents the results from color difference method. Color difference is indicated as the chroma difference (ΔC_{ab}) between the original, unmodified Reference Image II and the scanned, digitized print of each sample paper. Inspecting the results, ΔC_{ab} was the highest with red and orange on average, and the lowest with green. Furthermore, based on the subjective

test results of paper quality (Figure 25), the best quality papers (D26, D31, D32) presented the lowest color difference. On the other hand, the group of multipurpose papers with the lowest quality (D16-D21) contained the highest color differences, with the exception of D16.

4.4 Visual quality index

4.4.1 Model goodness and grading

In the final phase of the research, SPSS software and statistical regression analysis were utilized to combine the objectively calculated scores from the Matlab methods into one grade for print quality, Visual Quality Index (VQI). In addition, more accurate models were generated for colorfulness, contrast, and noise. The papers were graded for every quality category with numerical 7-point interval scale accompanied with verbal grades.

Figure 33 presents the variables used in the final regression analysis models for each of the four quality categories. Predictors in regression analysis, consisting of the software calculated scores from the Matlab methods, are presented on the left side. The right side includes the regression analysis dependents derived from the subjective test results, presenting the variables to be estimated with the predictors on the left.

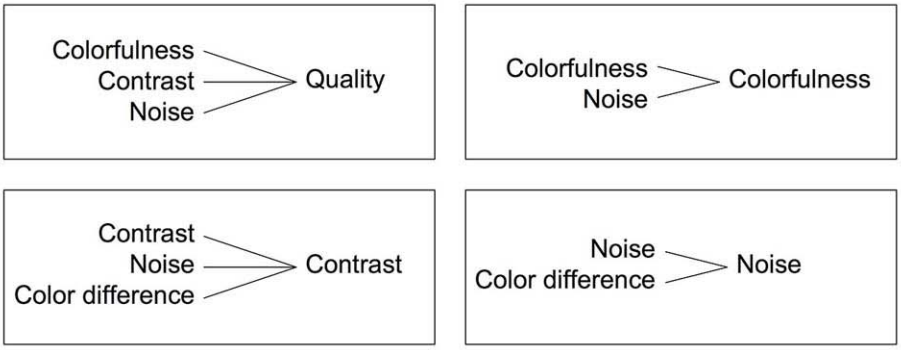


Figure 33. *Variables and their connections in the four regression analysis models. On the left side, predictors obtained from the objective Matlab methods. On the right side, dependents acquired from the subjective tests.*

Quality, contrast, and noise models were generated with the methods described in Chapter 3.6.2, i.e. with a stepwise method for quality and with a hierarchical enter + stepwise method for colorfulness, contrast, and noise. In the case of colorfulness, the contrast predictor was removed from the stepwise block to increase model reliability. With the other three models, no manual adjustments on regression blocks were required. The resulting models are listed in Table 10, presenting the coefficients for the regression equation (Equation 17). Diagnostic measures for statistical analysis of model goodness and reliability are presented in Table 11.

Table 10. *Regression coefficients for the regression models. The first column lists regression dependents i.e. models, the second shows constants for the models, and the last columns present the weighting coefficients for each regression predictor.*

Model	A	β (Colorfulness)	β (Contrast)	β (Noise)	β (Color dif.)
Quality	-0.939	-0.083	50.573	-1.085	0
Colorfulness	5.610	0.100	0	-1.051	0
Contrast	-5.168	0	50.459	-0.652	0.178
Noise	9.647	0	0	-1.370	-0.204

Table 11. *Statistical measures for the regression models, including the regression steps for each model. The columns two to five reflect the changes caused by adding more predictors in the model, while the last three columns present measures for the final model only. The relevant values for the final models are highlighted with light blue color and the values exceeding the predefined analysis limits with an asterisk (*).*

Model	Predictors added	R^2	Sig F Change	Sig (ANOVA)	Durbin - Watson	VIF	Condition Index
Quality	Noise	0.933	0.000	0.000		5.465	5.341
	Contrast	0.960	0.002	0.000		7.926	26.305
	Colorfulness	**0.974**	0.008	0.000	1.262	4.444	84.535*
Colorfulness	Colorfulness	0.815	0.000	0.000		3.019	4.812
	Noise	**0.978**	0.000	0.000	2.029	3.019	27.190
Contrast	Contrast	0.894	0.000	0.000		9.416*	5.409
	Noise	0.958	0.000	0.000		6.067	9.236
	Color dif.	**0.977**	0.001	0.000	1.777	2.720	104.129*
Noise	Noise	0.895	0.000	0.000		1.555	6.369
	Color dif.	**0.925**	0.016	0.000	1.211	1.555	8.476

The first survey to the finished models revealed remarkably high level of goodness for all the models. The coefficients of determination indicated by R^2, exceeding over 0.92 for all the models and over 0.97 for the first three models, were regarded as a clear demonstration that the inclusion of regression analysis improved the accuracy of the objective analysis. Furthermore, the objective quality method, newly introduced with the regression analysis, showed the similar high explanation levels as the quality attribute models.

The high level of correspondence between the objective models and the subjective reference data was attained in correlation analysis as well. As listed in Table 12, the correlations ranged from 0.942 to 0.989 and were statistically significant at 0.001 level, reflecting the similar high level of model goodness like with the R^2 measure. This observation was trivial in the case of

Pearson correlation, as the squared correlation is indeed the coefficient of determination R^2 (Field 2009, p. 224). Furthermore, when considering the three quality attribute models only, clear improvements were attained in correlations when compared to the correlation results of the single attribute Matlab methods and the same subjective data (Table 9), especially with Spearman correlations. From the three models, colorfulness improvement was the most pronounced, with Pearson correlation increasing from 0.903 to 0.989 and Spearman from 0.672 to 0.942. With the noise method, the gains were quite modest.

Table 12. *Correlations between the regression models and the subjective test resuls.*

	Pearson	**Spearman**
Quality	0.987 (p<0.001)	0.961 (p<0.001)
Colorfulness	0.989 (p<0.001)	0.942 (p<0.001)
Contrast	0.989 (p<0.001)	0.976 (p<0.001)
Noise	0.962 (p<0.001)	0.963 (p<0.001)

In addition to increased correlations, another benefit for using regression models was related to the scaling of calculated scores. The regression equation automatically scaled the results to correspond with the reference data, thus providing result grades with the same 7-point scale as was used in the subjective tests. This enabled an easy option for visual comparison of the model results and dependents, as can be seen in Figure 34.

In visual inspection of all the four models, the quality and contrast models followed the reference grades with the most accuracy, while with the noise model errors were pronounced. Furthermore, the noise model showed the concentration of errors on the papers with the lowest and highest grades. With the contrast model, on the other hand, the grade errors were the most visible with highly graded papers. In general level, the errors were quite modest with the maximum errors as follows: 0.97 for quality, 0.50 for colorfulness, 0.47 for contrast, and 1.11 for noise. Thus, considering error levels on maximum errors, colorfulness and contrast had an edge over quality and noise.

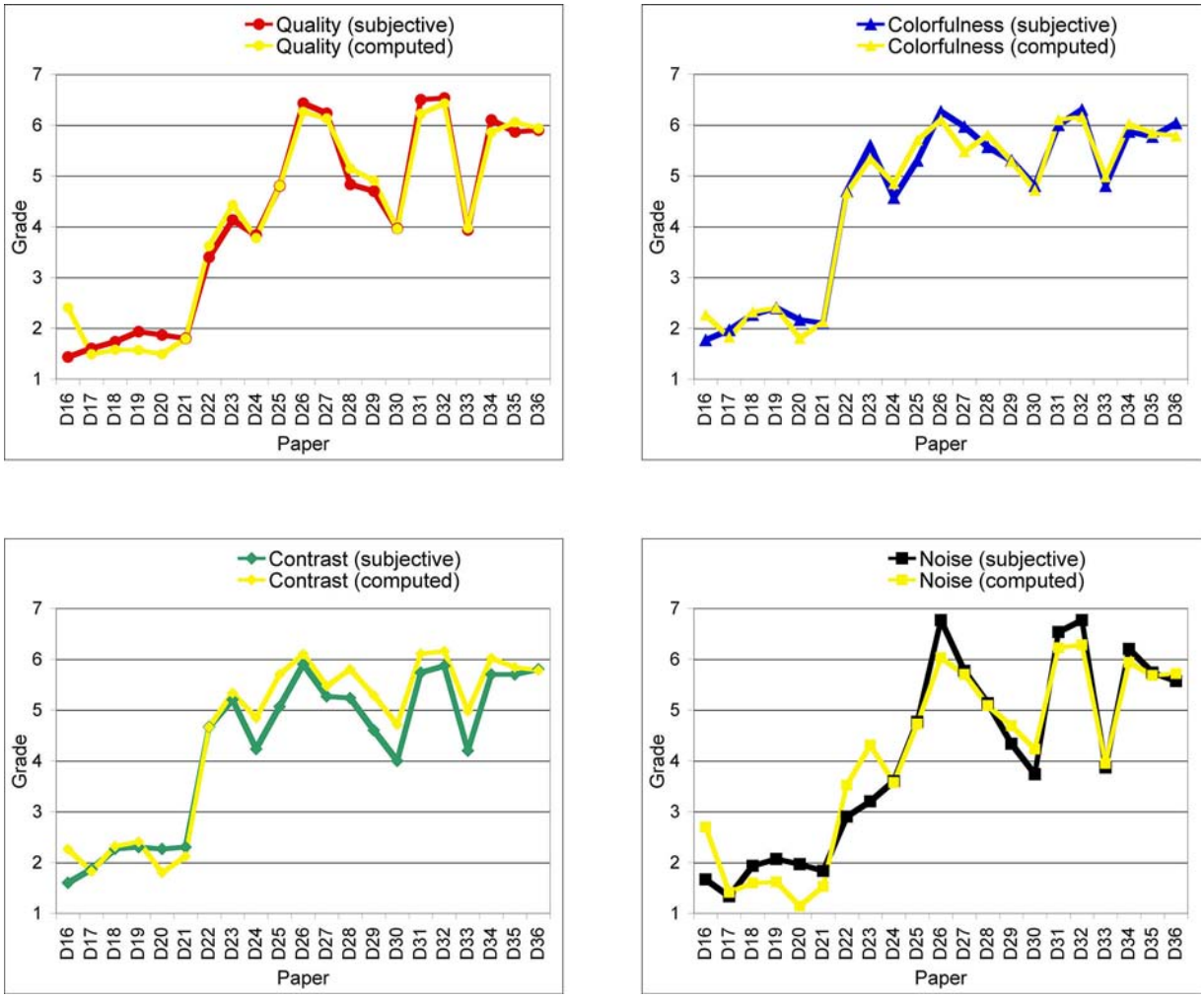


Figure 34. *Visualized results of the regression models in comparison to the subjective reference data. In charts, the title "Subjective" refers to the regression dependents, and "Computed" to the results of regression models.*

To study a question whether or not the grading errors were pronounced in the different parts of the grading scale, Figure 35 shows the regression model grades as a function of the subjective reference grades. With perfectly equal results, the dots should follow a straight line indicated by gray color in the charts.

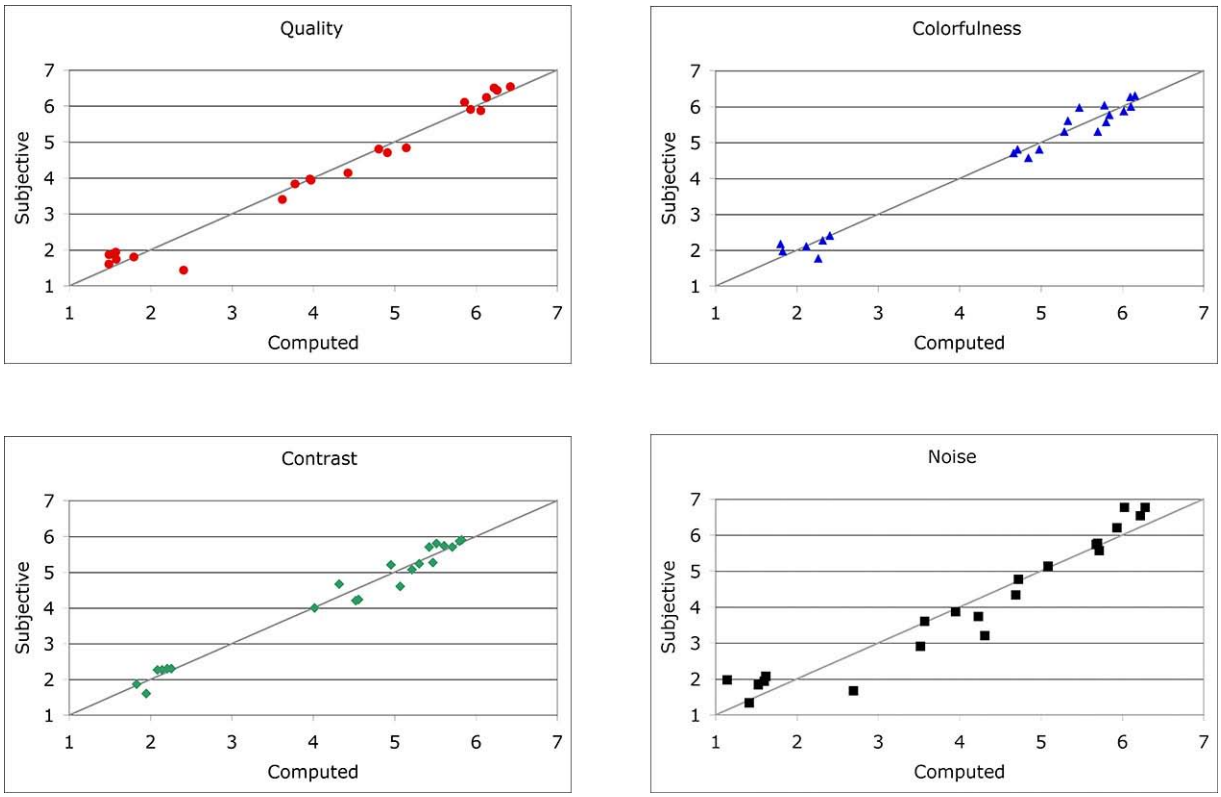


Figure 35. *Regression model grades (Computed) as a function of the reference grades (Subjective) with diagonal lines*

For the quality model, errors were distributed fairly evenly over the entire grading scale. With colorfulness and contrast, a subtle increase in error levels was observed with the higher grades ranging from 4.5 to 5.5. In the case of noise, errors were more pronounced, but fairly evenly distributed over the grading scale, with a slight increase on the grades around 4 and on the both extremes of the scale. With the other models, errors were minor on the scale limits.

To simplify the result reading, all the numeric grades were complemented with corresponding verbal grades as explained in Chapter 3.6.5. Table 13 shows the verbal key of grades for the regression models, complemented with the verbal key of subjective reference data for straightforward comparison. The written verbal grades for the regression models are available in Appendix 13, showing a screenshot from the actual execution of the main Matlab code (Appendix 4).

Table 13. *The verbal key of grades for the regression models (Computed), complemented with the corresponding grades for the subjective reference data (Subjective). The grades are described with the following verbal key: (7) Excellent, (6) Very good, (5) Good, (4) Fair, (3) Satisfactory, (2) Poor, and (1) Bad. The grade differences between the pairs are marked with light blue color.*

	Quality		Colorfulness		Contrast		Noise	
Paper	Subjective	Computed	Subjective	Computed	Subjective	Computed	Subjective	Computed
D16	1	2	2	2	2	2	2	3
D17	2	1	2	2	2	2	1	1
D18	2	2	2	2	2	2	2	2
D19	2	2	2	2	2	2	2	2
D20	2	1	2	2	2	2	2	1
D21	2	2	2	2	2	2	2	2
D22	3	4	5	5	5	4	3	4
D23	4	4	6	5	5	5	3	4
D24	4	4	5	5	4	5	4	4
D25	5	5	5	6	5	5	5	5
D26	6	6	6	6	6	6	7	6
D27	6	6	6	5	5	5	6	6
D28	5	5	6	6	5	5	5	5
D29	5	5	5	5	5	5	4	5
D30	4	4	5	5	4	4	4	4
D31	7	6	6	6	6	6	7	6
D32	7	6	6	6	6	6	7	6
D33	4	4	5	5	4	5	4	4
D34	6	6	6	6	6	5	6	6
D35	6	6	6	6	6	6	6	6
D36	6	6	6	6	6	6	6	6

Inspecting the verbal key of the regression models, numerous papers obtained the “Very good” (6) grades, but not a single one scored “Excellent” (7). With the subjective reference data, five “Excellent” (7) grades were given for three papers. Nevertheless, the maximum grade difference stayed in one grade at all the cases. When considering the grading of individual papers, 7 papers out of the total 21 papers got the correct verbal grades with all the models. However, the most important model, quality, gave correct results in 15 cases out of 21.

4.4.2 Model analysis

Looking back at Table 11, columns two to five present the progression of regression analysis with the regression steps required. As was expected from the highly correlating single attributes, R^2 was above 0.8 even after the first regression step for all the models. However, each model was improved when more predictors were added in the model. One or two extra predictors were sufficient in every case, and not a single model required the total of four predictors. Considering step significance, all the steps were statistically significant with the levels ranging from 0.000 to 0.016, clearly meeting the predefined analysis guidelines with the maximum value of 0.05 (Chapter 3.6.3). This was the case with ANOVA significance levels as well, where each step was significant at 0.000 level.

Analyzing the models in general, Durbin–Watson statistic was within the predefined limits of [1, 3] in all the cases, with good results on colorfulness and contrast. However, the quality and noise models showed rather low Durbin–Watson values with an approximate result of 1.2. This may indicate some problems with the regression assumption of independent regression errors with these two models (Field 2009, p. 236). In the case of multicollinearity analysis, problems were more apparent. Only the noise method showed acceptable results with VIF and Condition Index measures. With colorfulness, the values were within predefined limits, but the Condition Index was near the maximum value of 30. However, in the case of quality and contrast models, the problems with multicollinearity were evident. With quality, VIF was just beneath the limit of 8, but the Condition Index value of 84.535 noticeably exceeded the recommendation of 30. With contrast, the situation was even worse with both the values exceeding the limits, resulting in 9.416 for VIF and 104.129 for Condition Index.

4.4.3 Model generalization

Even if a regression model shows perfectly acceptable results regarding a coefficient of determination, the independence of residuals, and multicollinearity characteristics, it has to meet three additional regression assumptions for residuals to be generalized over the used sample to the whole population: the assumptions of linearity, homoscedasticity, and normality (Field 2009, pp. 247-251). To analyze the assumptions, recommendations explained in Chapter 3.6.4 were followed. Figure 36 shows the required charts for the quality model. The charts for the colorfulness, contrast, and noise models are listed in 14-16.

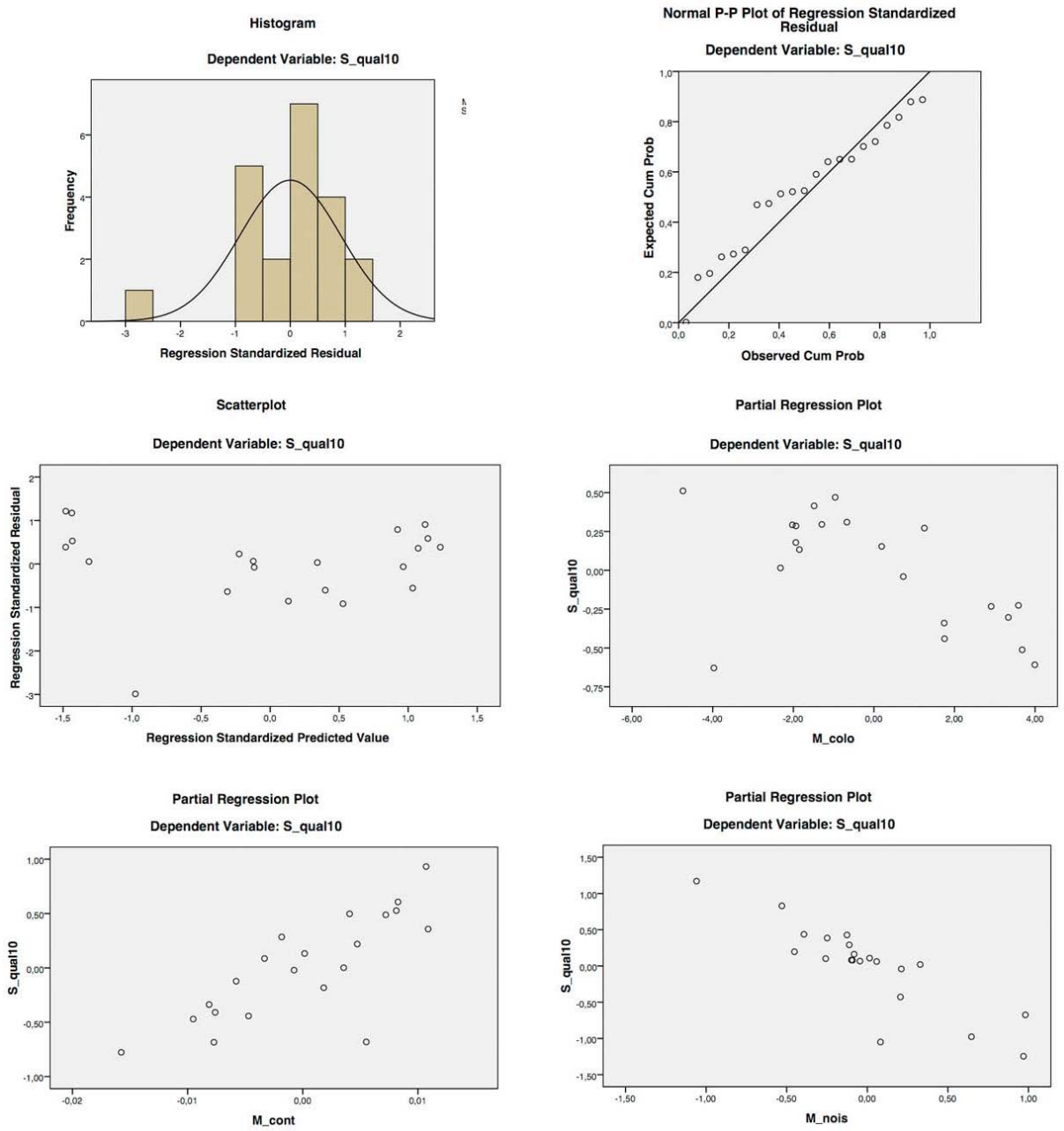


Figure 36. *The regression charts for the quality model, generated with SPSS for checking the regression assumptions of model generalization. In charts, S_qual10 refers to quality, while M_colo, M_cont, and M_nois indicate the regression predictors, i.e. colorfulness, contrast, and noise.*

Starting with the normality assumption for the quality model, the histogram chart in Figure 36 showed problems with the normal distribution of residuals. While some shape of normal distribution was visible, the empty spot on the left side of the histogram and the absence of the residuals in the histogram limits indicated problems with the normality assumption. Similarly, the normal probability plot (P-P plot) showed points of discontinuity in otherwise relatively

straight line, predicting non-normally distributed residuals. With the other models, only the colorfulness model showed a normally distributed histogram with an almost straight P-P plot line. The contrast model presented problems with the linearity of the P-P plot, and the noise model showed similar results as the quality model.

Problems were observed with the linearity assumption of the quality model as well. Rather than showing the expected form of randomly scattered dots around the center in the scatterplot, the dots formed a slightly curved line with an empty spot on the left side of the chart. With the partial regression plots, however, only minor problems were detected. Excluding few obvious outliers in the charts, the lines were fairly linear with only a slight indication of curved shape for the partial plot of the colorfulness predictor. With the other models, similar problems were observed. In scatterplots, the colorfulness and contrast models showed a moderately large area of empty space between dots in the center, while the noise model had the most curved form. However, similarly to quality, the partial plots were fairly linear with some visible outliers.

The third assumption, the homoscedasticity of residuals, showed lesser problems with the quality model. In the scatterplot, the variance in y-axis was fairly constant throughout the chart, excluding one obvious outlier in the point of (-1, -3) and the empty area on the left. With the partial plots, the distribution of dots was fairly expected, showing only slight signs of variance differences on the edges of the dot forms. In the case of the other models, more signs of heteroscedasticy were observed. With colorfulness and contrast, the empty areas in scatterplots predicted problems with the assumption. Furthermore, in the contrast model the partial plot of color difference was concentrated strongly to the right hand side. The results from the noise model were not more promising either, with problems on the scatterplot and the partial plot of color difference.

To summarize, since all the models violated at least two of the total three regression assumptions in some extent, it was fairly straightforward to conclude that the models could not be generalized over the used sample paper set. The violations were, however, quite minor in the most cases. The quality model violated the normality and linearity assumptions, with fair results on homoscedasticity. The colorfulness model showed good results on the model normality, but less so with the linearity and homoscedasticity assumption. With the contrast and noise models, all the three assumptions were more or less violated.

5 DISCUSSION AND RELIABILITY ANALYSIS

5.1 Reference image

5.1.1 Quality aspects

In reference image development, the design goals of high naturalness, balance, aesthetical appearance, and quality were undoubtly well achieved. In the light of the previous DigiQ reference image by Salmi (2008), Reference Image I, the subjective tests of Reference Image II showed similar, good results when considering the usability of the image for subjective testing. As a new feature, Reference Image II presented in-the-image implemented test colors, which had an obvious effect on overall balance of the image with increased colorfulness. This aspect arguably affected the high-level observations of the image, where especially the colorful tablecloth gathered the attention of observers. In addition to high-level differences between reference images, the subjective tests showed interestingly similar results for all the quality categories, predicting a strong relationship between overall print quality and individual image quality attributes.

From the quality point of view, the reference image development process was considered successful. With thorough planning on object selection and lighting accompanied with high quality photographing equipment and careful post processing, the resulting Reference Image II certainly matched the quality requirements placed for a reference image with high sharpness, vivid colors, and contrasty, noise-free presentation. Although the ongoing study only required prints with a small print size of 15 x 10 cm, the highly optimized 12 mpix reference image included details for much larger prints (Figure 22).

One aspect strongly affecting the overall appearance of the image was the implementation of the test colors. The selection of highly saturated objects, accompanied with natural objects containing memory colors, resulted in the image with undoubtly colorful appearance. This arguably affected the overall naturalness and balance of the image, the issue further discussed in the following chapter. The colorful appearance also led to the exclusion of large grey areas from the image, as large and dark areas were not considered suitable for the otherwise colourful presentation. Regarding the reliability of test colors, a very accurate implementation was attained in the Adobe RGB colorspace with the maximum errors of ±1 pixel value compared to the BabelColor reference values (Appendix 1). In a print context, however, paper movement during scanning may cause unexpected errors for objective methods.

Another aspect requiring lots of attention was providing constant and adequate lighting over the whole scene. With two flash units with relatively low power and rather small illumination areas, the correct flash placement was crucial for attaining desirable results. The final image, while being well lightened with natural appearance, would have gained a benefit from the presence of a larger light source on the right side of the image, providing more natural lighting

with softer shadows and allowing more options for light and object placements. For future reference image projects, the importance of lighting cannot be underestimated.

5.1.2 Feasibility for subjective testing

The arrangement of the subjective tests offered an invaluable opportunity for studying the usability aspects of Reference Image II in subjective assessment. The inclusion of Reference Image I in the same tests increased the reliability of results further, offering a viable image to compare with. The feasibility for subjective testing was considered mainly from three standpoints: the speed and certainty of the evaluation, the high-level aspects of the images, and the usefulness of the image objects and areas. In a general level, the usability of both the reference images was regarded as good for subjective testing.

Looking to the findings from the subjective tests, both the images showed good or fair results for the estimation speed and certainty with the grades ranging from 3.5 to 5.5 (Figures 27 and 28). With both the measures, differences between the five quality categories were relatively small, mostly inside the variation of one grade from the maximum of 7. Between the images, the results did not reveal a definite winner in any category, showing only a half grade edge for Reference Image II on contrast and sharpness.

While the differences between the reference images were quite small in the estimation speed and certainty, more result variation was present with the other high-level aspects, i.e. naturalness, pleasantness, and balance. On average, the results were in a good level with the grades ranging from 4 to 5.5. However, a fairly considerable one grade edge for Reference Image I in balance and pleasantness raised discussion (Figure 29). Reasons for the differences were searched for from the open comments collected during the tests.

In the verbal comments of Reference Image II, many indications of the overly pronounced red color of the tablecloth were found, affecting the estimations of all the three high-level attributes. Also the overall colorfulness with many saturated objects in the image was listed as a negative high-level aspect in few cases. On the other hand, the vivid appearance was regarded as a positive point for colorfulness evaluation. Another commented area was the wall color, which, in few cases, was listed as gloomy for Reference Image II and as natural for Reference Image I. Regarding the images in general, especially the smiling model with the gaze aimed to an observer gained lots of attention in Reference Image I, an aspect that was linked to image pleasantness in many comments. Further area of consideration was the arrangement of the objects in the table, where both the messy appearance with little free space in Reference Image I and the organized and spacious arrangement of Reference Image II gained positive comments.

Regarding the comments, the vivid appearance of Reference Image II obviously had a negative impact on the high-level gradings. This was expected, since the development of the image with seven bright test colors and numerous memory color objects was a challenge itself and the vivid appearance could not be avoided. Thus, the highly colorful tablecloth was chosen to correspond with the colorful mood of the image. The positive effect of the smiling model in

Reference Image I was expected as well. In Reference Image II, however, one of the design goals was to decrease the attention gathered by the human model, which led to moving the gaze out from an observer to the newspaper on the table.

Evaluating the issue with the red tablecloth further, in may be discussed if the poor paper quality of the questionnaire papers affected the high-level results. With the multipurpose paper used in printing the questionnaires, the red color of the tablecloth in Reference Image II was visibly exaggerated. When Reference Image II was printed on high quality papers or in evaluated in display viewing, the tablecloth was much more neutral. Furthermore, the objective color difference method showed severe problems on the reproduction of red and orange colors with poor quality papers, an issue further discussed in Chapter 5.2. However, as the test subjects were instructed to concentrate only on the effect of image content when evaluating high-level attributes, the bad paper quality should not have affected the results if the assignment was followed properly.

Changing the topic to the useful objects and areas in Reference Image II, an interesting remark was observed: the image did not contain any completely useless object (Figure 31 and Table 8). Indeed, practically all the available areas gained some usefulness markings in the quality category. In the other categories, the only area without attention was the tea set consisting of the tea cup, the tea plate, and the green napkin. However, the low usefulness was not considered as a major weakness, since the tea set was essential for the breakfast theme. Furthermore, the tea set formed a natural continuum for the hand, and the napkin presented a natural place for the bright green test color.

When comparing Figure 31 to the usefulness results presented in Salmi (2008, p.78) and Salmi et al. (2009), the similarity in the markings was easily observed. The areas of attention, with a strong focus on the human face, were fairly well scattered in both of the cases. The similarities observed, however, were not straightly comparable to Reference Image I used in this thesis, since the usefulness images in Salmi (2008) were created from development versions, not from the final image.

To summarize, the design goal of evenly scattered attention over the entire image was met fairly well. To prevent misconceptions, it should be emphasized that even though the entire image was regarded useful in Figure 31, all the marked areas were not used by all the test subjects. On the contrary, the most of the test subjects marked only two or three regions for each category. The research by Salmi (2008) reflected similar results with only one or two object groups selected on average per a test subject (Salmi et al. 2009).

5.1.3 Subjective quality grading

From a reliability standpoint, the results of subjective tests always leave room for discussion. In this thesis, challenges were met with the careful planning of the test situation and providing the proper test instructions. The all time presence of the test instructor helped the test subjects to concentrate on the task in hand, an aspect that gained positive recognition in the open, ver-

bal answers of the tests. On the other hand, the long, an average of 1 h 23 min test execution time had a negative impact on sustaining a stable attention level throughout the tests, another issue recognized in the open answers. In a general level, a fairly large number of 30 test subjects helped to diminish the effect of answer variation. Since only few definite outliers were spotted in the results, the raw data was left intact without any grade removals. Regarding grade variations, the standard deviation of grades shared homogeneous variation characteristics between all the quality categories (Figure 26). The variation range of 0.5 to 2 was expected in the case of subjective evaluation, where the individual preferences of human observers can have a considerable effect on evaluation grades.

Observing the subjective grades in general, the results showed surprisingly strong correspondence between the five quality categories. As illustrated in Figure 25, the line forms in the chart shared very similar characteristics in visual observation. Statistical analysis supported this observation, showing extremely high, over 0.95 Pearson and Spearman correlations between subjective quality and quality attributes (Table 6). These results, the high quality correlations and the similar grade figures, may suggest that for many printing papers, different image quality aspects are closely related to each other. In other words, a change in overall print quality indicates a quite similar change in image quality attributes. On the other hand, the use of subjective tests may also lead to these kinds of similarities, as the overall quality of the sample can easily affect the subjective observations of image quality attributes as well. However, this possible behaviour was taken into account in the subjective test planning, mostly by asking the test subjects to estimate the amount of attributes, not the quality of attributes, and by requiring a brief assessment for each sample.

5.2 Software development

Considering the project objectives, the outcome of the software development was satisfying: three functional objective methods for image quality attributes with surprisingly good, over 0.9 Pearson correlations to the subjective reference data and an added bonus, a color difference method to test the usability of test color areas in the reference image. The exclusion of the sharpness method was unfortunate, but as described in Methodology section, hard to correct in the scope of this study. The debate over the reasons of good correlations for quality attribute methods brought up three affecting aspects: the exclusion of electrophotography, Pearson correlations as the main goodness indicator, and the method optimization with masking in the case of the colorfulness and contrast methods. Furthermore, the high color difference levels with the red and orange test colors raised discussion.

The thesis restriction to cover only inkjet prints, not electrophotography, definitely simplified the design process. Though an option for testing the designed methods on EPG prints of Reference Image II was not available, the experiments with other digitized EPG samples showed strong indications that the EPG technology would have required different methods for accurate operation. Halonen (2008a) reflected similar observations. Furthermore, considering the high

correlations in general, the fact that the quality attribute models were fine-tuned with using Pearson correlations as the reference goodness indicator arguably had a positive effect on the achieved correlations. Spearman correlations, on the other hand, showed similar, over 0.9 correlations with the colorfulness and contrast methods, but with noise the lower correlation of 0.672 indicated the lack of optimization in that respect.

The colorfulness and contrast methods were optimized with masking, increasing the Pearson correlations from 0.794 to 0.946 (contrast) and 0.767 to 0.932 (colorfulness). It may be argued if the masking operations were necessary, since based on the testing with other images, masking values were highly dependent on image content. Nevertheless, as the development of no-reference methods was not the objective of this study, the reduced-reference methods with masking offered a feasible option for optimizations. With the noise method, optimizations were not required for high correlations and the direct method implementation by Malanin (2006) was selected.

The development of the value masks was mainly based on two assumptions: the assumption that paper noise affects mostly the pixels with medium pixel values, and the assumption that quality differences between papers are the strongest in the limits of the measures. As the print version of Reference Image II was optimized to show the best quality with the best papers in the paper selection, it was assumed that the differences between paper quality appeared the most visible on the reproduction of the maximum levels of the measure under testing. Thus, with both the models, the masks were set to diminish the effect of neutral pixel values, enhancing the influence of the maximum pixel levels in calculations.

In subjective evaluation, the pronounced red color of the tablecloth, especially with the low quality papers, raised discussion about the reasons behind the issue. Examining the matter of red color further, the objective color difference results in Figure 32 showed that for five of the total six multipurpose papers, the strongest chroma error was observed in the red test color with the orange coming as the second. Similar results were obtained when considering the averages of chroma differences over the entire paper selection, showing 8.07 average color difference for orange and 7.28 for red. The next largest chroma error, 4.28 for yellow, was almost the half from orange and red. As a further notice, the red and orange shared one converging characteristic: the average wavelengths of the colors were the longest inside the test color selection, i.e. 650 nm for red and 610 nm for orange, calculated as weighted averages over the visible light spectrum from the spectral data of the BabelColor test colors (BabelColor 2009).

These findings, combined to the related issues discovered in subjective evaluation, suggested that the colors with long wavelengths, i.e. orange and red in the area of 610 to 650 nm, are strongly influenced by paper quality in inkjet printing. Furthermore, this color error may affect the reproduction of certain memory colors in the orange–red area, e.g. human skin, showing a dominant wavelength close to 590 nm (Buck and Froelich 1948, cited in Yendrikhovskij et al., 1999). The finding was concerning, since Hunt (1987, pp. 43-44) regards human skin as the

critical area of color reproduction, with lower than normal reproduction tolerances. Thus, the use of good quality papers is suggested especially for inkjet works with human content.

5.3 Visual quality index

5.3.1 Goodness of models

Regression analysis presented a viable statistical analysis method for further improving the results obtained from the Matlab methods. In addition to more accurate grades for the quality attributes, regression analysis enabled an option to generate an objective model for the print quality, the visual quality index. Concerning the goodness of the final models, the coefficients of determination showed statistically very strong results with the values over 0.97 for all the other models but the noise model with 0.925. However, regardless of the high determination values, problems were faced when analyzing regression assumptions. The problems led to a conclusion that none of the models could be generalized over the used sample, thus limiting the correct function of the models only to the 21 papers used in the study.

From the quality attribute models created, the gains on model goodness were the smallest with the noise model. This was partly explained by the fact that the noise model required only two predictors, noise and color difference. Furthermore, as the noise predictor offered the best correlations to the subjective data to begin with, the effect of color difference predictor was quite subtle in the model calculations.

When considering the grades of individual papers, an interesting remark was observed. As shown in Figure 34, the paper D16 caused unexpected difficulties for objective evaluation, resulting in high scale errors with all the models. This was especially the case with the quality model, where D16 presented the most pronounced estimation error of all the papers. However, when D16 was excluded from the quality model results for a test purpose only, the maximum error of quality decreased from 0.97 to 0.38, showing improved model correspondence with the other good models, colorfulness and contrast. Furthermore, the average errors of the models were quite similar: 0.20 average error for quality, 0.19 for colorfulness, 0.18 for contrast, and 0.38 for noise. Indeed, from these standpoints, the first three models shared very similar goodness characteristics.

One reason for the diverging results with the paper D16 may lie in the fact that the paper material contained no fluorescent substances, thus presenting a slight yellow color cast on the paper. In Table 2 showing the variation of paper characteristics in the sample paper selection, D16 is the paper causing the fluorescence percent scale to start from zero. The paper was advertised as an ecological paper, which explains the absence of fluorescence material.

5.3.2 Realization of regression assumptions

Regarding the fulfillment of the regression assumptions, all the models showed violations on the assumptions, and model generalization could not be justified. One obvious reason for the

problems was the strong multicollinearity characteristics of the predictors and the dependents. As shown in Table 6, the correlations between the reference attributes and overall quality were extremely strong. Obviously, this high level of correlation reflected to the scores computed with Matlab methods as well. According to Metsämuuronen (2006, p.679), predictors should correlate only fairly to a dependent and not too strongly with each other. High correlation causes problems with multicollinearity, decreasing the model reliability (Field 2009, pp. 223-224).

The problem of highly correlating regression variables was evident after the subjective tests. For future research, the inclusion of different quality attributes may be discussed. In this study, however, the choices for quality attributes were rather limited mainly for two reasons: only the paper effect on quality was evaluated, and a single reference image was used. For instance, the assessment of high-level attribute differences, e.g. naturalness, between papers with the same image in every sample would have arguably offered a major challenge for test subjects, and even more for the development of the objective method to predict the results. What comes to the exclusion of sharpness from the objective methods, sharpness was not considered essential in the regards of the VQI model, since an extra regression predictor with high correlation to the dependent would have most likely presented only a minor or non-existent increase on model goodness. Nevertheless, according to Eerola et al. 2008a and Leisti et al. 2008, sharpness presents an important quality attribute for quality evaluation, and an individual sharpness model would have arguably been useful for estimating the quality properties of printing papers.

The multicollinearity problem raised a question about the rationality of creating regression models from the image quality attributes with extreme correlations, especially when model generalization is essential. In this study, however, regression analysis was regarded as the most straight-forward way to construct a single grade for overall quality, visual quality index, from numerous objective scores. The fundamental goal in the VQI creation was to exploit different quality aspects in calculations, i.e. the four Matlab methods, thus providing more reliable quality measure. In future research, alternative quality attributes could be searched for. Furthermore, it can also be discussed if increasing the sample paper selection from 21 papers would have improved the regression results considerably. However, knowing the excessive work Halonen (2008a) conducted when selecting and characterizing the current paper selection, this may be left for future research as well.

A further remark was observed when analyzing the residual charts generated with SPSS in Figure 36 and Appendices 14-16. A large number of charts showed an empty space inside the residual plots. One possible cause for the missing residuals may lie in the reference data. In the subjective results of quality, colorfulness, and contrast, there existed an empty gap with no grades ranging from 1.5 grades for quality to 2.5 grades for colorfulness and contrast. It may be argued if these empty areas resulted in abnormal residual distributions, thus affecting the model generalization results. On the other hand, even though the subjective results of noise

showed a more even grade distribution than the other quality categories, the scatterplot chart on the regression result contained a similar empty spot with no residuals.

In literature regarding statistical analysis, the data transformation of regression variables is often offered as an option for correcting the models violating regression assumptions of homoscedasticity and normality (Metsämuuronen 2006; Field 2009). In this study, however, data transforms were not properly tested because of the complexity and the uncertainty of the issue. According to Glass et al. (1972 cited in Field, 2009, p. 155), data normalizing transforms are seldom considered worth the effort because of uncertain results. Furthermore, Field (2009, p. 251) states that data transformations not necessarily affect the residuals, compromising the main reason to use them.

6 CONCLUSIONS

This study was contributed as a part of the DigiQ project with a certain goal: to automatically assess the image quality of printing papers. The goal was met partially in this thesis with the development of visual quality index (VQI), software for grading papers based on the human observations of image quality. This achievement, however, did not come without restrictions to the original DigiQ project goals. Firstly, the analysis was done on papers with a specific reference image printed on them. Secondly, only inkjet printing technology was studied. Thirdly, the no-reference requirement for objective methods was abandoned. Furthermore, to ensure a stable image output, the printing and scanning process had to follow predefined digitizing guidelines.

Four research objectives were placed for the experimental part, which, in a shorter form, were as follows: developing a reference image for subjective and objective quality evaluation, arranging subjective tests for obtaining subjective paper quality grades, developing quality estimation software based on the subjective reference grades, and forming the VQI model. Image quality was determined with four low-level quality attributes: colorfulness, contrast, sharpness, and noise.

To address the first objective, a single reference image of high quality was developed for subjective and objective evaluation with an emphasis on image naturalness, balance, and aesthetical expression. The resulting image, Reference Image II, presented a breakfast theme with a human model sitting by a table full of breakfast related objects. As an extra feature, seven test colors equaling the GretagMacbeth references were implemented into natural objects in the image.

To meet the second objective, subjective tests were arranged to Reference Image II, accompanied with a previous DigiQ reference image by Salmi (2008), i.e. Reference Image I. The goal was to obtain subjective reference data about the visual appearance of papers for objective methods. In addition to addressing overall quality and the four quality attributes, reference images were assessed for several high-level aspects, i.e. image naturalness, balance, and pleasantness. The assessment proved both the images to be suitable for subjective testing, with good assessment ratings in evaluation speed and certainty as well as in the high-level aspects. The comparison between the images raised discussion about the high colorfulness of Reference Image II, especially when the red tablecloth was concerned. Further area of notice was the high similarity of the observation grades between overall quality and the quality attributes, an issue that affected the reliability of the VQI model.

Software development with Matlab, the third objective, resulted in functional objective measures for three quality attributes, i.e. colorfulness, contrast, and noise, as well as an additional color difference method. The exclusion of the sharpness method was related to the unexpected sharpness characteristics of the digitized paper samples. The finished quality attribute meas-

ures showed surprisingly high correlations with the subjective reference data. However, like was the case with subjective references, cross-correlations between the measures were on a high level. In the analysis of color difference results, the strongest errors were seen in the red - orange area, raising discussion of the connections to subjective test answers about the red tablecloth, the most pronounced individual color in Reference Image II. All the methods were reduced-reference measures excluding the no-reference noise method.

The final objective was addressed with statistical regression analysis. The developed regression model for quality, the VQI, showed excellent goodness figures with a high coefficient of determination, estimating the visual quality appearance of papers very accurately. Similarly, highly accurate results were obtained for the regression models of the three quality attributes. Unfortunately, the high similarity of the reference data raised problems with model generalization, leading to a conclusion that either the VQI or the three other models cannot be utilized reliably outside the test context.

To summarize, all the research objectives were met with acceptable accuracy. The developed reference image proved out to be usable in subjective and objective evaluation, combining high technical quality in a natural and aesthetically pleasing appearance. The objective print quality model, while very accurate with the papers used in testing, failed to ensure reliable operation outside the test paper selection. The issue with generalization lay in the quality attributes, which, on the other hand, turned out as good candidates for subjective quality assessment and provided statistically accurate explanation of the quality concept in objective estimation, but showed too similar results to be fully usable in regression analysis. Accompanied with other quality attributes with different characteristics, the selected quality attributes would arguably provide a solid starting point for a regression based image quality model – a definite area for future research.

REFERENCES

Axelsson, Ö. (2007) Towards a psychology of photography: dimensions underlying aesthetic appeal of photographs. *Perceptual and Motor Skills* 105. pp. 411-434.

BabelColor (2009) *ColorChecker* (online). [Accessed 2.3.2010]. Available in WWW form: <URL:http://www.babelcolor.com/main_level/ColorChecker.htm>

Bundesverband Druck und Medien (n.d.) *roman 16 bvdm Reference Images* (online). [Accessed 7.3.2010]. Available in WWW form: <URL: http://www.roman16.com/en>

Burian, P. (2004) *Mastering digital photography and imaging*. Alameda, California, USA, SYPEX Inc. 272 p.

CIE Division 8 (n.d.) *TC8-04 Sony sRGB standard images 1999* (online). [Accessed 7.3.2010]. Available in WWW form: <URL:http://www.colour.org/tc8-04/test_images/Sony>

Clayton, M. & Hashimoto, A. (2009) *Visual design fundamentals: a digital approach*, third ed. Boston, Massachusetts, USA, Course Technology. 405 p.

Crete, F., Dolmiere, T., Ladret, P. & Nicolas, M. (2007) *The blur effect: perception and estimation with a new no-reference perceptual blur metric.* SPIE Electronic Imaging Symosium, San Jose, California, USA, 28 January - 1 February 2007. 11 p.

Datta, R., Joshi, D., Li, J. & Wang, J. (2006) Studying aesthetics in photographic images using a computational approach. *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision Part III* 3953. pp. 288-301.

de Ridder, H. & Endrikhovski, S. (2002) Image quality is FUN: reflections on fidelity, usefulness and naturalness. *SID Symposium Digest of Technical papers* 33, 1. pp. 986-989.

Eerola, T., Kamarainen, J.-K., Leisti, T., Halonen, R., Lensu, L., Kälviäinen, H., Oittinen, P. & Nyman, G. (2008a) *Finding best measurable quantities for predicting human visual quality experience*. In Proceeding of the IEEE International Conference of Systems, Man and Cybernetics, Singapore, 12-15 October 2008. pp. 733-738.

Eerola, T., Kämäräinen, J.-K., Leisti, T., Halonen, R., Lensu, L., Kälviäinen, H., Nyman, G. & Oittinen, P. (2008b) *Is there Hope for Predicting Human Visual Quality Experience?* In Proceedings of the SMC'08 IEEE International Conference on Systems, Man, and Cybernetics, Singapore, 12-15 October 2008. pp. 725-732.

Eerola, T., Kämäräinen, J-K., Lensu, L. & Kälviäinen, H. (2007) *Visual print quality evaluation using computational features.* In Proceedings of the ISVC'07 International Symposium on Visual Computing, California, USA, 26-28 November 2007. pp. 403-413.

Eerola, T., Kämäräinen, J-K., Lensu, L. & Kälviäinen, H. (2009) *Framework for applying full reference digital image quality measures to printed images.* In Proceedings of the SCIA'09 Scandinavian Conference on Image Analysis, Oslo, Norway, 15-18 June 2009. pp. 99-108.

Eizo (2009) *Eizo ColorEdge CG242W brochure* (online). [Accessed 20.2.2010]. Available in WWW form:
<URL:http://www.eizo.com/data/downloads/brochures/pdf/ColorEdge_DI_0903.pdf>

Engeldrum, P. (2000) *Psychometric scaling: A toolkit for imaging systems development*. Winchester, Massachusetts, USA, Imcotek Press. 185 p.

Epson (2006) *Epson Stylus Pro 3800 brochure* (online). [Accessed 20.2.2010]. Available in WWW form:
<URL:http://www.epson.com/cmc_upload/0/000/080/385/SP3800Brochure_092806.pdf>

Epson (2007) *Epson Perfection V750-M Pro brochure* (online). [Accessed 20.2.2010]. Available in WWW form:
<URL:http://www.epson.com/cmc_upload/0/000/142/238/V700_V750_InfoSheetR1.pdf>

Fairchild, M. (1998) *Color appearance models*. Chichester, West Sussex, England, John Wiley & Sons Ltd. 417 p.

Fairchild, M. (2005) *Color appearance models*, second ed. Chichester, West Sussex, England, John Wiley & Sons Ltd. 385 p.

Fedorovskaya, E., Neustaedter, C. & Hao, Wei (2008) *Image harmony for consumer images*. IEEE ICIP'08 International Conference on Image Quality, San Diego, California, USA, 12-15 October 2008. pp. 121-124.

Fernandez, S.R. & Fairchild, M.D. (2002) *Observer preferences and cultural differences in color reproduction of scenic images*. IS&T/SID CIC'02 Color Imaging Conference: Color Science and Engineering Systems, Technologies and Applications, Scottsdale, Arizona, USA, 12-15 November 2002. pp. 66-72.

Ferzli, R. & Karam, L. J. (2007) *A no-reference objective image sharpness metric based on just-noticeable blur and probability summation*. IEEE ICIP'07 International Conference on Image Processing, San Francisco, California, USA, 16-19 October 2007. IEEE. pp. 445-448.

Field, A. (2009) *Discovering statistics using SPSS*, third ed. London, England, Sage Publications. 821 p.

Field, G. (1999) *Color and its reproduction*, second ed. Sewickley, Pennsylvania, USA, GATFPress. 475 p.

Fraser, B., Murphy, C. & Bunting, F. (2005) *Real world color management*, second ed. Berkeley, California, USA, Peachpit Press. 582 p.

Gatter, M. (2004) *Getting it right in print*. London, England, Laurence King Publishing. 172 p.

Gonzalez, R. C. & Woods, R. E. (2008) *Digital image processing*, third ed. Upper Saddle River, New Jersey, USA, Prentice Hall. 954 p.

Halonen, R. (2008a) *Determination of quality from printed natural images.* MSc Thesis. Helsinki University of Technology, Faculty of Chemistry and Materials Sciences. Espoo, Finland. 81 p.

Halonen, R. (2008b) *The influence of scan resolution and sample processing on blur computation results from digitized natural images.* Unpublished report. Helsinki University of Technology, Department of Media Technology. Espoo, Finland. 18 p.

Halonen, R., Leisti, T. & Oittinen, P. (2008) *The influence of image content and paper grade on quality attributes computed from printed natural images.* In Proceedings of the NIP24: International Conference on Digital Printing Technologies, Pittsburgh, Pennsylvania, USA, 6-11 September 2008.

Hasler, D. & Süsstrunk, S. (2003) *Measuring colourfulness in natural images.* In Proceedings of the IS&T/SPIE EI'03 Electronic Imaging, Santa Clara, California, USA, 20-24 January 2003. pp. 87-95.

Hedgecoe, J. (2006) *Valokuvaajan suuri tietokirja*, fifth edition. London, England, Dorling Kindersley. 288 p.

Hendley, H. & Hecht, S. (1949) The colors of natural objects and terrains, and their relation to visual color deficiency. *Journal of the Optical Society of America* 39, 10, pp. 870-873.

Hiden, J. (2008) *Valokuvaa sokeasti arvioivien algoritmien yhdistetty suorituskyky laatuun pohjautuvassa kuvien luokittelussa*. MSc Thesis. Helsinki University of Technology, Faculty of Electronics, Communications and Automation. Espoo, Finland. 123 p.

Hunt R., 1987, *The reproduction of colour*, fourth ed. Tolworth, England, Fountain press. 640 p.

Imatest (n.d.) *Colorcheck appendix: Algorithms and reference formulas*. (online). [Accessed 20.2.2010]. Available in WWW form:
<URL:http://www.imatest.com/docs/colorcheck_ref.html>

Immerkær, J. (1996) Fast noise variance estimation. *Computer Vision and Image Understanding* 64, 2. pp. 300-302.

Janssen, R. (2001) *Computational image quality.* Bellingham, Washington, USA, SPIE. 143 p.

Johansson, K., Lundberk, P. & Ryberg, R. (2007) *A guide to graphic print production*, second ed. Hoboken, New Jersey, USA, John Wiley & Sons. 455 p.

Kim, Y.J., Rhoder, L. P., Cheung, S., Westland, W., Choe, S. L. & Kim, C. (2006) *Image colour-quality modelling for mobile LCDs.* In Proceedings of the IS&T/SID 14th Color Imaging Conference, Scottsdale, Arizona, USA, 6-10 November 2006. pp. 159-164.

Kipphan, H. (2001) *Handbook of print media*. Berlin, Germany, Springer. 1207 p.

Kivinen, H. (2009) *Computational prediction of perceived colour difference in digital image of complex scene*. MSc Thesis. Helsinki University of Technology, Faculty of Chemistry and Materials Sciences. Espoo, Finland. 103 p.

Koivisto, L. (2009) *Digitaalikameran värintoiston mittaaminen kuvaan piilotettujen värikohteiden avulla*. MSc Thesis. Helsinki University of Technology, Department of Automation and Systems Technology. Espoo, Finland. 97 p.

Leisti, T., Halonen, R., Kokkonen, A., Weckman, H., Mettänen, M., Lensu, L., Ritala, R., Oittinen, P. & Nyman, G. (2008) *Process perspective on image quality evaluation.* IS&T/SPIE EI '08 Electronic Imaging, San Jose, California, USA, 27-31 January 2008.

Leisti, T., Radun, J., Virtanen, T., Halonen, R. & Nyman, G. (2009) *Subjective experience of image quality: attributes, definitions and decision making of subjective image quality.* IS&T/SPIE EI'09 Electronic Imaging, San Jose, California, USA, 18-22 January 2009. pp. 72420D-1-72420D-9.

Li, X. (2002*) Blind image quality assessment*. IEEE ICIP'02 International Conference on Image Processing, Rochester, New York, USA, 22-25 September 2002. pp. 449-452.

Lindbloom, B. (2003) *Useful color equations: Lab to LCH(ab)* (online). [Accessed 11.12.2009]. Available in WWW form: <URL:http://www.brucelindbloom.com/index.html?Equations.html>

Malanin, M. (2006) *Digitaalisten kuvien laadunarviointi ja kuva-agentti –pohjainen lähestymistapa painokuvien valintaan*. MSc Thesis. Helsinki University of Technology, Department of Automation and Systems Technology. Espoo, Finland. 77 p.

Marziliano, P., Dufaux, F., Winkler, S. & Ebrahimi, T. (2002) *A no-reference perceptual blur metric*. In Proceedings of the IEEE ICIP'02 International Conference on Image Processing, Rochester, New York, USA, 22-25 September 2002. pp. 57-60.

Maureen, S. (2003) *Field guide to digital color*. Natick, Massachusetts, USA, A K Peters Ltd. 326 p.

Metsämuuronen, J. (2006) *Tutkimuksen tekemisen perusteet ihmistieteissä*, second ed. Helsinki, Finland, International Methelp KY. 1324 p.

Nurminen, T. (2005) *Värien nimeäminen ja laskennallinen ennustaminen luonnollisista värikuvista*. MSc Thesis. Helsinki University of Technology, Department of Computer Science and Engineering. Espoo, Finland. 97 p.

Olkkonen, M., Hansen, T. & Gegenfurtner, K. (2008) Color appearance of familiar objects: effects of object shape, texture, and illumination changes. *Journal of Vision* 8(5), 13. pp. 1-16.

Pappas, T., Safranek, R. & Chen, J. (2005) Perceptual criteria for image quality evaluation. In: Bovik, A. (ed) *Handbook of image and video processing*. Burlington, Massachusetts, USA, Elsevier Academic Press. pp. 939-960

Pascale, D. (2006) *RGB Cooridinates of the Macbeth Color Checker* (online). Montreal, Quebec, Canada, The BabelColor Company. Updated 1 June 2006 [Accessed 31.3.2009]. Available in WWW form: <URL:http://www.babelcolor.com/download/RGB%20Coordinates%20of%20the%20Macbeth%20ColorChecker.pdf>

Peli, E. (1990) Contrast in complex images. *Journal of the Optical Society of America A* 7, 10. pp. 2032-2040.

Perez-Carpinell, J., de Fez, M., Baldovi, R. & Soriano, J. (1998) Familiar objects and memory color. *Color research and application* 23. pp. 416-427.

Peters, G. (2007) *Aesthetic primitives of images for visualization*. IEEE IV'07 International Conference on Information Visualization, Zurich, Switzerland, 2-6 July 2007. pp. 316-325.

Saito, R. & Kotera, H. (2000) *Extraction of image gamut surface and calculation of its volume*. IS&T/SID CIC'00 Color Imaging Conference, Scottsdale, Arizona, USA, 7-10 November 2000. pp. 330-334.

Salmi, H. (2008) *Constructing a balanced reference image for visual print quality evaluation*. MSc Thesis. Helsinki University of Technology, Department of Automation and Systems Technology. Espoo, Finland. 103 p.

Salmi, H., Halonen, R., Leisti, T., Oittinen, P. & Saarelma, H. (2009) *Development of a balanced test image for visual print quality*. IS&T/SPIE Electronic Imaging 2009, San Jose, California, USA, 18-22 January 2009. pp. 72420B-1-72420B-11.

Seshadrinathan, K., Sheikh, H., Wang, Z. & Bovik, A. (2005) *Structural and information theoretic approaches to image quality assessment*. In: Blum, R. & Zheng, L. (eds.) Multisensor Image Fusion and Its Applications. Florida, USA, CRC Press. pp. 473-501.

Silverstein, D. & Farrell, J. (1996) *The relationship between image fidelity and image quality*. In Preceedings of the ICIP'06: International Conference of Image Processing, Lausanne, Switzerland, 16-19 September 1996. pp. 881-884.

Spillmann, L. & Ehrenstein, W. (2004) *Gestalt factors in the visual neurosciences.* In: Chalupa, L. & Werner, J. (eds.) The visual neurosciences vol. 2. London, England, The MIT Press. 813 p.

TKK Media Technology (n.d.) *DigiQ - Fusion of Digital and Visual Print Quality* (online). [Accessed 26.2.2010]. Available in WWW form: <URL:http://owww.media.hut.fi/digiq/index.html>

Umbauch, S. (2005) *Computer imaging: Digital image analysis and processing*. Florida, USA, CRC Press. 657 p.

Wang, Z. & Bovik, A. C. (2002) *Why is image quality assessment so difficult?* IEEE IC-ASSP'02 International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida, USA, 13-17 May 2002. pp. 3313-3316.

Wang, Z. & Bovik, A.C. (2006) *Modern image quality assessment*, Vol. 2. San Rafael, California, USA, Morgan & Claypool Publishers. 156 p.

Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. (2004) Image quality assessment: from error visibility to structural similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING* 13, 4. pp. 600-612.

Wang, X., Tian, B., Liang, C. & Shi, D. (2008) *Blind image quality assessment for measuring image blur*. CISP'08 Congress on Image and Signal Processing, Sanya, Hainan, China, 27-30 May 2008. pp. 467-470.

Westlake, A. (2009) *What is a 'standard' lens?* (online). [Accessed 12.1.2010]. Available in WWW form: <URL:http://blog.dpreview.com/editorial/2009/03/what-is-a-standard-lens.html>

Wright, T. (1999) *The photography handbook*. London, England, Routledge. 198 p.

X-Rite (2010) *GretagMacbeth ColorChecker Classic* (online). [Accessed 20.2.2010]. Available in WWW form: <URL:http://www.xrite.com/product_overview.aspx?ID=1192>

Yendrikhovskij, S. (2002) *Image quality and colour categorisation*. In: Lindsay, W., MacDonald, M. & Ronnier, L. (eds.) Colour image science: exploiting digital media. Chichester, England, John Wiley & Sons Ltd. pp. 393-420.

Yendrikhovskij, S., Blommaert, F. & de Ridder, H. (1999) Color reproduction and the naturalness constraint. *Color research and application* 24. pp. 52-67.

Zakia, R. (2002) *Perception and imaging: photography – a way of seeing*. Woburn, Massachusetts, USA, Focal Press. 410 p.

APPENDICES

APPENDIX 1 Babel color reference values

		Adobe RGB			L*a*b* (CIE D50)			sRGB		
Color No.	Color name	R	G	B	L*	a*	b*	R	G	B
15	red	151	52	59	42.18	54.89	28.79	176	48	56
14	green	101	148	78	55.05	-38.09	31.62	71	149	72
13	blue	53	64	143	28.65	15.60	-50.52	43	62	147
18	cyan	65	135	164	50.55	-27.97	-28.14	0	136	166
17	magenta	165	85	147	51.82	49.79	-13.90	188	84	150
16	yellow	227	198	53	82.23	4.05	79.84	238	200	22
7	orange	197	122	54	62.56	35.13	58.05	220	123	45
20	neutral 8	199	200	199	81.01	-0.57	0.18	200	201	201
21	neutral 6.5	159	160	160	66.30	-0.43	-0.08	160	161	161
22	neutral 5	119	121	121	50.83	-0.69	-0.27	120	121	121

APPENDIX 2 Reference image test color values and coordinates

Test color values in the reference image. Adobe RGB (1998) color space was the original color space, and the L*a*b* and sRGB values were calculated directly from the image after absolute colorimetric color space conversion.

	Adobe RGB (original)			L*a*b* (absolute colorimetric)			sRGB (absolute colorimetric)		
Test color	**R**	**G**	**B**	**L***	**a***	**b***	**R**	**G**	**B**
red	152	52	59	41	52	27	176	48	57
green	101	149	78	56	-38	33	69	150	72
blue	53	63	143	29	17	-50	45	61	147
cyan	65	135	164	51	-28	-27	0	136	167
magenta	165	84	147	51	49	-16	188	83	150
yellow	227	198	52	82	3	80	238	199	18
orange	197	122	53	62	34	58	219	123	42
neutral 8	200	200	200	81	0	0	201	201	201
neutral 6.5	160	160	160	66	0	0	161	161	161
neutral 5	120	120	120	51	0	0	121	121	121

Errors in test color values.

Color space	**Error (values)**
Adobe RGB	+-1
sRGB	+-2
L*a*b*	+-2

Exceptions:

sRGB, yellow B error -4
sRGB, orange B error -3

Test color locations in the reference image, indicated by the center coordinates (center point) and the width and height (center area) of the averaged square area used to calculate the test values.

	Center point		Center area
Test color	**y**	**x**	**width**
red	2293	1743	51
green	2135	1172	51
blue	1322	2231	51
cyan	1330	2381	51
magenta	2381	2143	51
yellow	2040	1445	31
orange	2335	1861	51
neutral 8	966	2358	51
neutral 6.5	2193	2748	31
neutral 5	2771	164	31

APPENDIX 3 Subjective test instructions and questionnaires

PAINETUN KUVAN LAATU

Omakohtainen arviosi tulostetun kuvan **laadusta.**

Arvosteluasteikko

Kuvien arvostelussa käytetään seitsenportaista asteikkoa välillä 1-7.

	1	2	3	4	5	6	7	
erittäin huonolaatuinen	☐	☐	☐	☐	☐	☐	☐	erittäin laadukas

Vaihe 1.

Eteesi jaetaan 42 kuvan pino ja tehtävänäsi on löytää niiden joukosta testijoukon **yksi edustaja arvosteluasteikon kumpaankin ääripäähän,** ts. arvosanaryhmiin 1 ja 7**.** Valitsemasi kaksi kuvaa asetetaan **ankkurikuviksi** (kuviksi, joihin voit verrata muita testikuvia) vaiheen 2 arviointia varten. Kuvien etsimisen saat toteuttaa haluamallasi tavalla.

Huomioita:

- Ankkurikuvien **ei tarvitse esittää samaa kuva-aihetta** (eli olla samoja kuvia)**.**
- Pyri samalla tarkastelemaan kuvajoukkoa yleisellä tasolla, jotta saisit käsityksen laadun hajonnasta eri kuvien välillä.
- Käsittelethän kuvanäytteitä varoen :)

Vaihe 2.

Tässä vaiheessa tehtävänäsi on arvioida yksittäisten kuvien **laatua** vertaamalla niitä yksitellen ensimmäisessä vaiheessa valitsemiisi ankkurikuviin. Tarkoituksesi on antaa jokaiselle kuvalle arvosana kokonaislaadusta asteikolla 1-7. Huomaa kuitenkin, että **molempien kuvasisältöjen ei tarvitse** saada arvosanoja yhdestä seitsemään.

Toteutus:

- Sinulle jaetaan ensimmäisessä vaiheessa selailemasi kuvanippu.
- Arvioi kuvien **laatua** yksitellen aloittaen kasan päällimmäisestä.
- Arvion tehtyäsi sano näytteen laatuarvosana (1-7) ja näytteen numero (1-42, joka löytyy paperin toiselta puolelta) testin ohjaajalle, joka merkitsee sen ylös.
- Pinoa arvostelemasi näytteet pöydälle väärinpäin.
- Jatka näin, kunnes kaikki kuvat on käyty läpi.

Huomioita:

- Tarkoituksenasi on antaa **ensivaikutelma** arvosteltavan kuvan laadusta. Älä siis tarkastele testikuvia kovin kauaa.
- Arvosteluasteikon **ääripäissä (1 & 7) voi olla monta kuvaa**.
- Kuvat esitetään satunnaisessa järjestyksessä ja ne on myös numeroitu satunnaisesti.

PAINETUN KUVAN LAADUN ARVIOINTI (1/2)

TESTIKUVA 1

Alla näet kuvan toisesta juuri arvostelemastasi testikuvasta. Merkitse kuvaan mustaa tussia käyttäen ne alueet, jotka koit hyödylliseksi **laatua** arvioidessasi. Hyödyllisyydellä viitataan tässä arvostelun helppouteen ja tehokkuuteen. Voit merkitä useita alueita tarpeen mukaan.

Vastaa lisäksi seuraaviin kysymyksiin ruksaamalla oikea vaihtoehto ruudukosta.

Kuinka **nopeaa** oli arvioida laatua tästä kuvasta?

	1	2	3	4	5	6	7	
erittäin hidasta	☐	☐	☐	☐	☐	☐	☐	erittäin nopeaa

Kuinka **varma** olit arvioistasi?

	1	2	3	4	5	6	7	
erittäin epävarma	☐	☐	☐	☐	☐	☐	☐	erittäin varma

Muita kommentteja / ajatuksia laadun arviointiin liittyen tästä kyseisestä testikuvasta.

_____

_____

_____

_____

_____

LOPPUARVIOINTI (1/3)

TESTIKUVA 1

Kuvan **luonnollisuus**. Luonnollisuudella tarkoitetaan kuvan todenmukaisuuden määrää. Jos kuva on äärimmäisen luonnollinen, katsoja voi kokea katsovansa kuvan maisemaa kuin omin silmin, ikään kuin kuvaa ei olisi välissä ollenkaan. Kohtuullisen luonnollisessa kuvassa kuvan sisältö voi olla sinänsä aidon näköistä, mutta katsoja kokee katsovansa kuitenkin vain kuvaa. Kuva, jonka luonnollisuus on taas erittäin pieni, esittää asiat vääristyneellä, epätodenmukaisella tavalla tavalla.

	1	2	3	4	5	6	7	
erittäin epäluonnollinen	☐	☐	☐	☐	☐	☐	☐	erittäin luonnollinen

Kuvan **tasapainoisuus.** Tasapainoisuudella tarkoitetaan eri asioiden tasapainoa kuvassa, miten eri esineet herättävät huomiota ja erottuuko joku asia erityisen vahvasti muista.

	1	2	3	4	5	6	7	
erittäin tasapainoton	☐	☐	☐	☐	☐	☐	☐	erittäin tasapainoinen

Kuvan **miellyttävyys**. Miellyttävyydellä tarkoitetaan kuvan yleistä visuaalista miellyttävyyttä ja valokuvallisia ominaisuuksia, kuten esteettistä ilmaisua ja sommittelua.

	1	2	3	4	5	6	7	
erittäin epämiellyttävä	☐	☐	☐	☐	☐	☐	☐	erittäin miellyttävä

APPENDIX 4 Matlab code for the main program

```
% Reijo Asikainen (2009)
% DigiQ project
%
% Main function for image quality calculations.
% Programmed for the Master's thesis:
% "Objective image quality analysis of printed natural reference picture"
%
% This algorithm was programmed for the DigiQ project to estimate image quality
% properties of different inkjet papers. To ensure comparable results,
% a scanned version of the original, printed reference image is required.
%
% Algorithm functions:
% Algorithms named 'colorfulness', 'contrast', 'noise', and 'color_difference'
% are used to calculate low level scores for corresponding attributes.
% Algorithm 'quality' combines the results to four final scores (image
% quality, colorfulness, contrast and noise) using the calculation models
% obtained from regression analysis.
%
% Grading categories:
% Excellent     > 6.5
% Very good     5.5 - 6.5
% Good          4.5 - 5.5
% Fair          3.5 - 4.5
% Satisfactory  2.5 - 3.5
% Poor          1.5 - 2.5
% Bad           < 1.5
%
% Quality   Better score -> Better overall quality
% Colorf.   Better score -> More color
% Contrast  Better score -> More contrast
% Noise     Better score -> Less noise
%
% It should be noted that from these attributes, only quality indicates
% qualitative aspects of images. Other attributes specify quantitive magnitudes,
% not quality.
%
% Results are saved in two files:
% 'Results_matlab.txt' - Unadjusted numeric data from first four algorithms
% 'Results_scores.txt' - Numeric quality scores and verbal grades
%
% Input:
% Reads and evaluates all .tif files from a given directory (filepath).
% Images must be in 48bit LAB format.


function calculations_main(filepath)

warning off all

% Color reference values and positions
color_reference = xlsread('color_reference_matlab.xls');
% Regression coefficients for quality score calculations
coefficients = xlsread('regression_coefficients.xls');

% Title formatting for the result files and screen printing
text_matlab = strcat('Matlab results\n\n', 'Filename\t\t', 'Colorf\t', 'Cont\t', 'Noise\t', 'Dif(av)\t',
'Dif(R)\t', 'Dif(G)\t', 'Dif(B)\t', 'Dif(C)\t', 'Dif(M)\t', 'Dif(Y)\t', 'Dif(O)\n');
text_scores = strcat('Scores and grades\n\n', 'Filename\t\t', 'Quality\t\t', 'Colorfulness\t', 'Contrast\t',
'Noise\n');
text_view = strcat('Filename\t\t\t', 'Quality\t\t\t\t', 'Colorfulness\t\t', 'Contrast\t\t\t', 'Noise');
text_title = strcat('\nScores and grades\n\n', 'Grading categories:\n>6.5\t\tExcellent\n', '5.5-6.5\t\tVery
good\n', '4.5-5.5\t\tGood\n', '3.5-4.5\t\tFair\n', '2.5-3.5\t\tSatisfactory\n', '1.5-2.5\t\tPoor\n',
'<1.5\t\tBad\n\n', 'Quality\t\tBetter score -> Better overall quality\n', 'Colorf.\t\tBetter score -> More
color\n', 'Contrast\tBetter score -> More contrast\n', 'Noise\t\tBetter score -> Less noise\n');
disp(sprintf(text_title));
disp(sprintf(text_view));


tic
file = dir(strcat(filepath, '\*.tif'));

for i = 1:1:size(file);

    % Reads image file
    image = imread(file(i).name);
    % Converts image to double. L* channel scaled to [0, 100], a* and b* channels to [-128, 127]
    image_double = lab2double(image);
    % Only L* channel, scaled to [0 1]
    image_L = image_double(:,:,1)/100;


    % COLORFULNESS
    % Parameter: Lab image, all channels (double)
    colorfulness_value = colorfulness(image_double);

    % CONTRAST
    % Parameter: Luminance channel, scaled to [0, 1]
    contrast_value=contrast(image_L);
```

```
    % NOISE
    % Parameter: Luminance channel, scaled to [0, 255]
    noise_value = noise(255*image_L);

    % COLOR DIFFERENCE
    % Parameters: Lab image, all channels (double) & color reference table
    difference_value=color_difference(image_double, color_reference);

    % QUALITY SCORES
    % Parameters: Regression coefficient table and calculated values for all image quality attributes
    % (colorfulness, contrast, noise, color difference)
    [quality_scores quality_grades quality_grades_view] = quality(coefficients, colorfulness_value, con-
trast_value, noise_value, difference_value);


    % Formatting for the result text files and screen printing
    text_matlab = strcat(text_matlab, file(i).name, '\t', num2str(colorfulness_value), '\t',
num2str(contrast_value), '\t', num2str(noise_value), '\t', num2str(difference_value(8)), '\t',
num2str(difference_value(1)), '\t', num2str(difference_value(2)), '\t', num2str(difference_value(3)), '\t',
num2str(difference_value(4)), '\t', num2str(difference_value(5)), '\t', num2str(difference_value(6)), '\t',
num2str(difference_value(7)), '\n');
    text_scores = strcat(text_scores, file(i).name, '\t', num2str(quality_scores(1)), '\t', quality_grades{1},
'\t', num2str(quality_scores(2)), '\t', quality_grades{2}, '\t', num2str(quality_scores(3)), '\t', qual-
ity_grades{3}, '\t', num2str(quality_scores(4)), '\t', quality_grades{4}, '\n');
    text_view = strcat(file(i).name, '\t', num2str(quality_scores(1),'%.2f'), '\t', quality_grades_view{1}, '\t',
num2str(quality_scores(2),'%.2f'), '\t', quality_grades_view{2}, '\t', num2str(quality_scores(3),'%.2f'), '\t',
quality_grades_view{3}, '\t', num2str(quality_scores(4),'%.2f'), '\t', quality_grades_view{4});

    % Results to screen
    disp(sprintf(text_view));

end


% Execution time
t=toc;
disp(sprintf('\nElapsed time: %.1f seconds.\n', t))

% Results to text files
result_file = fopen('Results_matlab.txt','wt');
fprintf(result_file, text_matlab);
fclose(result_file);

result_file = fopen('Results_scores.txt','wt');
fprintf(result_file, text_scores);
fclose(result_file);
```

APPENDIX 5 Matlab code for the colorfulness method

```
% Reijo Asikainen (2009)
% DigiQ project
% Colorfulness algorithm for the Master's thesis:
% "Objective image quality analysis of printed natural reference picture"
%
% Input:
% 16 bit Lab image (double)


function out = colorfulness(image)

image_c=sqrt(image(:,:,2).^2 + image(:,:,3).^2);

% Chroma masking: values under 48 are set to 0.
% Only areas with strong chroma are included.
for i=1:1:size(image,1);
    for j=1:1:size(image,2);
        if (image_c(i,j)<48)
            image_c(i,j)=0;
        end
    end
end

colorfulness=mean2(image_c)+std2(image_c);

out = colorfulness;
```

APPENDIX 6 Matlab code for the contrast method

```
% Reijo Asikainen (2009)
% DigiQ project
% Contrast algorithm for the Master's thesis:
% "Objective image quality analysis of printed natural reference picture"
%
% Input:
% L* channel from 16 bit Lab image, scaled to [0, 1]


function out = contrast(image)

% Contrast masking: values between 0.5 to 0.77 are set to 0.5
for i=1:1:size(image,1);
    for j=1:1:size(image,2);
        if (image(i,j)<0.77 & image(i,j)>0.5)
            image(i,j)=0.5;
        end
    end
end

contrast=std2(image);

out = contrast;
```

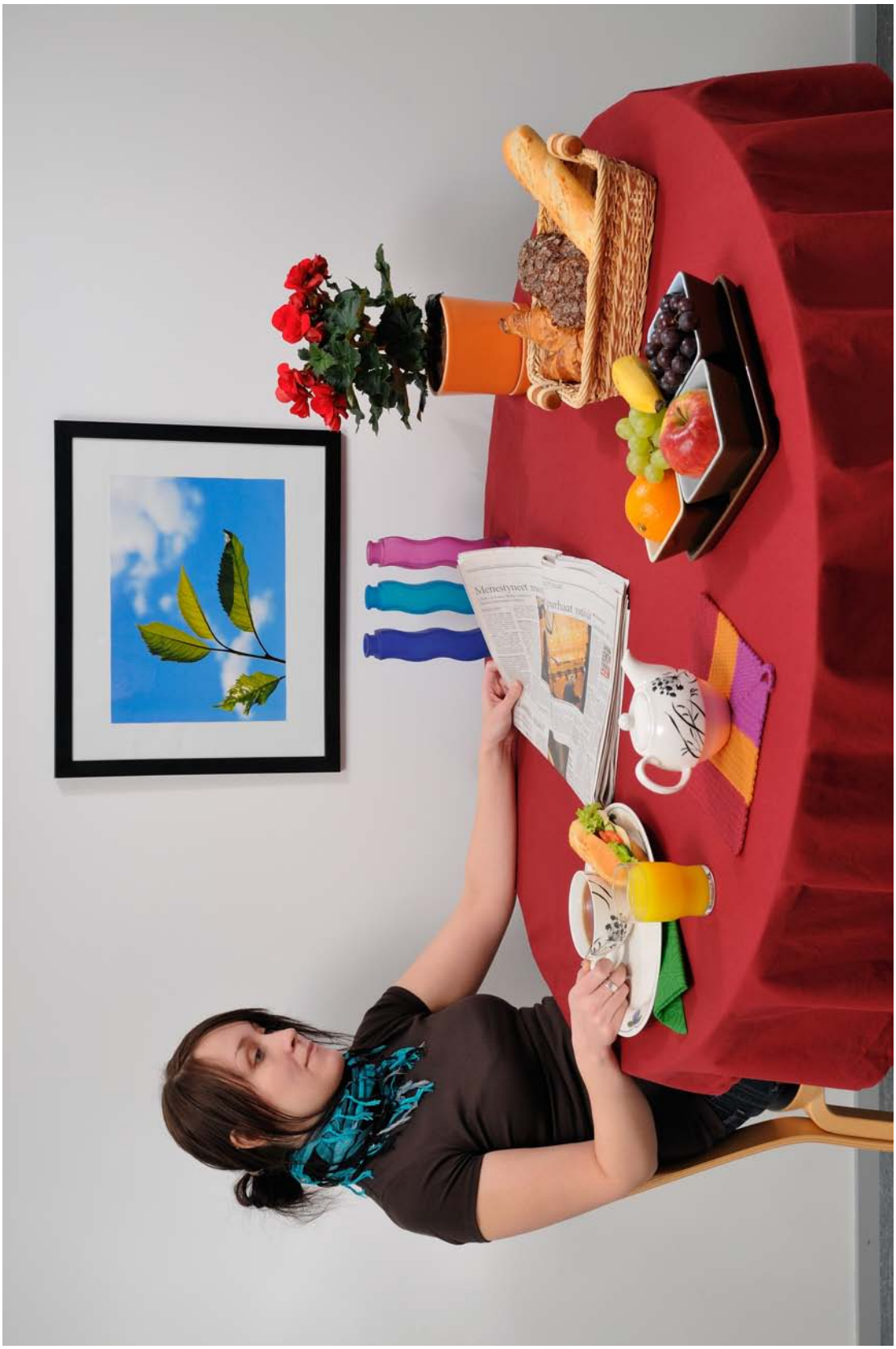
APPENDIX 7 Matlab code for the color difference method

```
% Reijo Asikainen (2009)
% DigiQ project
% Color difference algorithm for the Master's thesis:
% "Objective image quality analysis of printed natural reference picture"
%
% Input:
% 16 bit Lab image (double)
% Color reference table


function out = color_difference(image, reference_unadjusted)

% New color reference table for the correct image size

size_orig=[reference_unadjusted(1,1), reference_unadjusted(2,1), 3];
ratio = size(image)./size_orig;
reference=zeros(10,11);
% Lab values
reference(1:10,1:3)=reference_unadjusted(6:15,1:3);
% Mean area values
reference(1:10,4)=round(reference_unadjusted(6:15,10).*ratio(1));
reference(1:10,5)=round(reference_unadjusted(6:15,11).*ratio(2));
reference(1:10,6)=round(reference_unadjusted(6:15,12).*mean(ratio(1:2),2));
% Mean area coordinates & chroma
for i=1:10;
    reference(i,7)=reference(i,4)-reference(i,6)/2;
    reference(i,8)=reference(i,4)+reference(i,6)/2;
    reference(i,9)=reference(i,5)-reference(i,6)/2;
    reference(i,10)=reference(i,5)+reference(i,6)/2;
    reference(i,11)=sqrt(reference(i,2)^2+reference(i,3)^2);
end

% Calculates average Lab values from the image
results = zeros (size(reference,1)-3, 3); % no grays included (-3)
for i=1:1:size(reference,1);
    results(i,1) = mean2(image(reference(i,7):reference(i,8), refer-
ence(i,9):reference(i,10), 1)); %L
    results(i,2) = mean2(image(reference(i,7):reference(i,8), refer-
ence(i,9):reference(i,10), 2)); %a
    results(i,3) = mean2(image(reference(i,7):reference(i,8), refer-
ence(i,9):reference(i,10), 3)); %b
end

% Chroma normalizing
results_chroma = results;
for i=1:1:size(reference,1)-3;
    chroma = sqrt(results(i,2)^2 + results(i,3)^2);
    results_chroma(i,2) = reference(i,11)/chroma * results(i,2);
    results_chroma(i,3) = reference(i,11)/chroma * results(i,3);
end

% Color difference calculations
% Delta C, chroma normalized
difference_chroma = zeros (size(reference,1)-3, 1);
for i=1:1:size(reference,1)-3;
    difference_chroma(i) = sqrt( (reference(i,2)-results_chroma(i,2))^2 + (reference(i,3)-
results_chroma(i,3))^2 );
end

% Average color difference
difference_chroma(i+1) = mean(difference_chroma);

out = difference_chroma;
```

APPENDIX 8 Matlab code for the visual quality index and grading

```
% Reijo Asikainen (2009)
% DigiQ project
% Quality score and grade calculation algorithm for the Master's thesis:
% "Objective image quality analysis of printed natural reference picture"
%
% Combines image quality values to single scores of quality,
% colorfulness, contrast and noise. The scores are based on regression
% analysis models.
%
% Input:
% Regression coefficient table for score calculations
% Calculated values for all the image attributes
%
% Output:
% Scores and grades in the following order: quality, colorfulness, contrast, noise
% scores: numerical scores
% grades: verbal scores
% grades_view: verbal scores with formatting
%
% Grading categories:
% Excellent     > 6.5
% Very good     5.5 - 6.5
% Good          4.5 - 5.5
% Fair          3.5 - 4.5
% Satisfactory  2.5 - 3.5
% Poor          1.5 - 2.5
% Bad           < 1.5


function [scores grades grades_view] = quality(coefficients, colorfulness, contrast, noise,
difference)

scores=zeros(4,1);
for i=1:1:4;

    % Numerical scores
    scores(i) = coefficients(i,1) + colorfulness*coefficients(i,2) + con-
trast*coefficients(i,3) + noise*coefficients(i,4) + difference(8)*coefficients(i,5);

    % Verbal grades
    grades(i)={'No score'};
    grades_view(i)={'No score'};

    if scores(i) >=6.5
        grades(i)={'Excellent'};
        grades_view(i)={'Excellent'};
    elseif scores(i) >=5.5
        grades(i)={'Very good'};
        grades_view(i)={'Very good'};
    elseif scores(i) >=4.5
        grades(i)={'Good'};
        grades_view(i)={'Good\t'};
    elseif scores(i) >=3.5
        grades(i)={'Fair'};
        grades_view(i)={'Fair\t'};
    elseif scores(i) >=2.5
        grades(i)={'Satisfactory'};
        grades_view(i)={'Satisfactory'};
    elseif scores(i) >=1.5
        grades(i)={'Poor'};
        grades_view(i)={'Poor\t'};
    elseif scores(i) <1.5
        grades(i)={'Bad'};
        grades_view(i)={'Bad\t\t'};
    end

end
```

APPENDIX 9 The final reference image, Reference Image II



APPENDIX 10 Subjective test results for Reference Image II

Subjective grades, averages, Reference Image II

Paper	Quality	Colorfulness	Contrast	Sharpness	Noise
D16	1.43	1.77	1.60	1.73	1.67
D17	1.60	1.97	1.87	2.00	1.33
D18	1.73	2.27	2.27	2.47	1.93
D19	1.93	2.40	2.30	2.57	2.07
D20	1.87	2.17	2.27	2.47	1.97
D21	1.80	2.10	2.30	1.90	1.83
D22	3.40	4.70	4.67	3.63	2.90
D23	4.13	5.60	5.20	4.30	3.20
D24	3.83	4.57	4.23	4.10	3.60
D25	4.80	5.30	5.07	5.13	4.77
D26	6.43	6.27	5.90	6.50	6.77
D27	6.23	5.97	5.27	6.13	5.77
D28	4.83	5.57	5.23	5.10	5.13
D29	4.70	5.30	4.60	4.70	4.33
D30	3.97	4.80	4.00	3.97	3.73
D31	6.50	6.00	5.73	6.73	6.53
D32	6.53	6.30	5.87	6.33	6.77
D33	3.93	4.80	4.20	4.00	3.87
D34	6.10	5.87	5.70	5.90	6.20
D35	5.87	5.77	5.70	6.07	5.73
D36	5.90	6.03	5.80	5.93	5.57

Standard deviations of grades, Reference Image II

Paper	Quality	Colorfulness	Contrast	Sharpness	Noise
D16	0.69	1.28	1.04	1.29	0.75
D17	0.77	1.48	1.00	1.25	0.82
D18	0.77	1.37	1.37	0.91	1.10
D19	0.80	1.38	1.05	1.08	0.95
D20	0.59	1.06	1.29	1.38	1.05
D21	0.79	1.20	1.37	1.09	0.82
D22	1.12	1.87	1.61	1.55	1.13
D23	1.31	1.95	1.90	1.53	1.42
D24	1.32	1.81	1.32	1.18	0.96
D25	1.45	1.62	1.81	1.44	1.56
D26	1.45	1.45	2.08	1.62	1.41
D27	1.30	1.71	1.88	1.49	1.38
D28	1.44	1.64	1.45	1.75	1.47
D29	1.68	1.77	1.77	1.57	1.64
D30	1.29	1.82	1.27	1.14	1.38
D31	1.19	1.92	1.41	1.57	1.53
D32	1.35	1.74	2.02	1.59	1.44
D33	1.28	2.03	1.42	1.22	1.16
D34	1.67	1.62	1.84	1.54	1.40
D35	1.67	1.80	1.63	1.69	1.60
D36	1.70	1.81	1.51	1.66	1.68

APPENDIX 11 Subjective test results for Reference Image I

Subjective grades, averaged, Reference Image I

Code	Quality	Colorfulness	Contrast	Sharpness	Noise
D16	1.43	1.70	1.77	1.53	1.67
D17	1.27	1.97	2.10	2.13	1.40
D18	1.63	2.07	2.53	2.30	1.67
D19	1.63	2.43	2.73	2.57	2.07
D20	2.20	2.37	2.53	1.97	2.37
D21	1.43	2.13	2.37	1.87	1.43
D22	3.87	5.17	4.90	4.00	3.10
D23	3.27	5.37	5.00	4.00	2.27
D24	4.30	4.60	4.63	4.17	4.40
D25	5.43	5.17	5.03	5.03	5.43
D26	6.30	5.53	5.63	5.93	6.67
D27	6.33	5.50	5.20	5.57	6.23
D28	5.63	5.53	5.07	4.87	5.47
D29	5.37	5.10	4.87	4.20	4.80
D30	4.20	4.57	4.17	4.03	4.43
D31	6.13	5.50	5.27	6.13	6.57
D32	6.70	5.77	6.00	6.03	6.43
D33	4.40	4.80	4.17	3.70	4.30
D34	6.17	5.63	5.53	5.37	6.13
D35	6.33	5.67	5.57	5.73	6.10
D36	5.93	5.73	5.43	5.23	5.63

Standard deviations of grades, Reference Image I

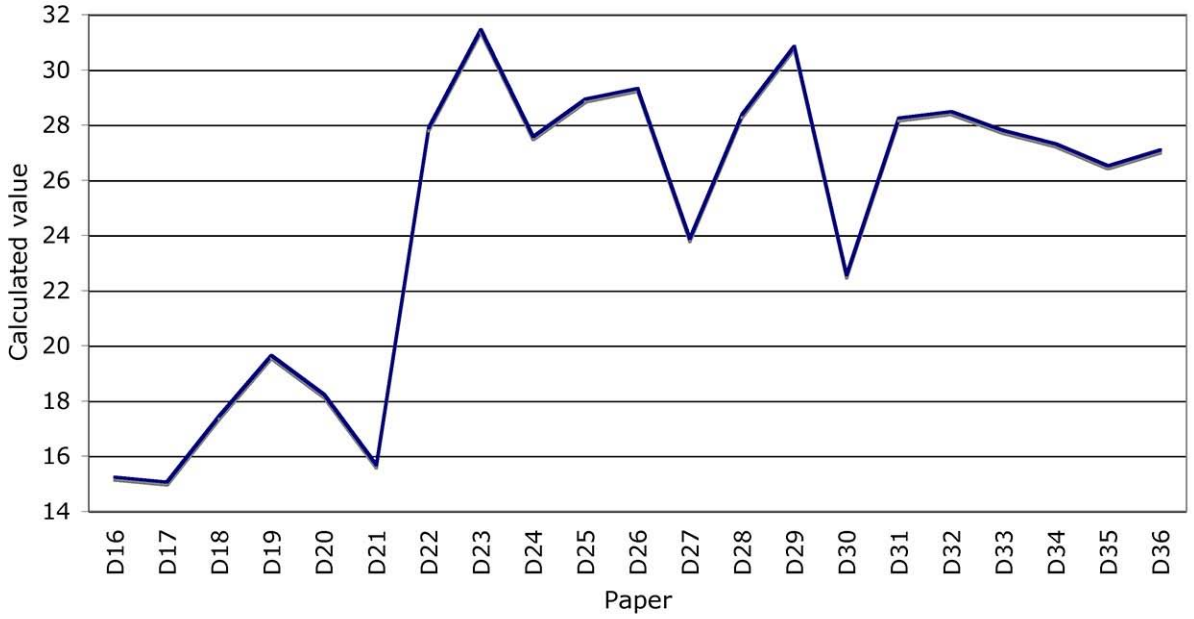
Paper	Quality	Colorfulness	Contrast	Sharpness	Noise
D16	0.91	0.87	1.10	0.87	0.70
D17	0.60	0.98	1.00	1.32	0.63
D18	0.93	1.03	1.50	1.27	0.96
D19	0.85	0.78	1.42	1.17	0.99
D20	1.03	0.99	1.24	1.09	0.95
D21	0.63	0.98	1.40	1.07	0.69
D22	1.04	1.47	1.56	1.46	1.15
D23	0.86	1.68	1.91	1.46	1.49
D24	1.30	1.38	1.66	1.20	1.27
D25	1.63	1.54	1.18	1.32	1.47
D26	1.30	1.83	1.71	1.77	1.47
D27	1.62	1.49	1.63	1.72	1.45
D28	1.39	1.39	1.06	1.38	1.49
D29	1.15	1.46	1.59	1.61	1.50
D30	1.45	1.53	1.37	1.26	1.27
D31	1.32	1.69	1.64	1.74	1.52
D32	1.25	1.43	1.32	1.46	1.29
D33	1.43	1.35	1.65	1.36	1.15
D34	1.47	1.47	1.83	1.99	1.76
D35	1.66	1.51	1.38	1.71	1.54
D36	1.77	1.51	1.32	1.52	1.49

APPENDIX 12 Numeric results and visual, unscaled charts for the Matlab methods

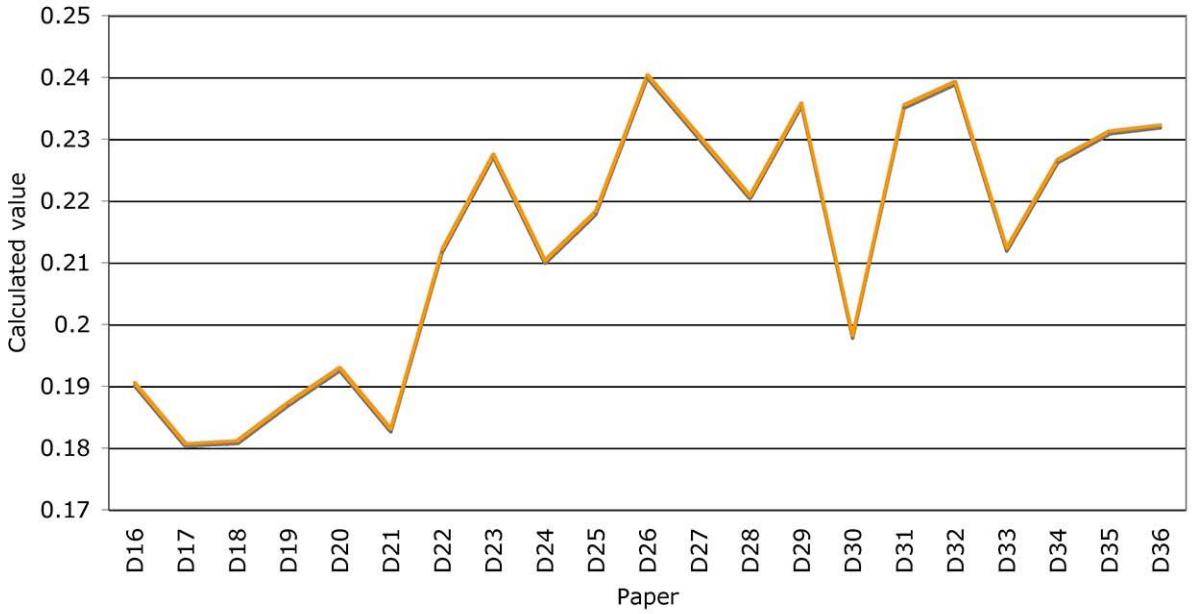
Numeric scores for colorfulness, contrast, and noise

Paper	Colorfulness	Contrast	Noise
D16	15.2	0.191	4.64
D17	15.1	0.181	5.04
D18	17.4	0.181	4.79
D19	19.6	0.187	4.92
D20	18.3	0.193	5.36
D21	15.7	0.183	4.82
D22	27.9	0.212	3.55
D23	31.5	0.228	3.26
D24	27.6	0.210	3.35
D25	28.9	0.218	2.67
D26	29.3	0.240	2.32
D27	23.9	0.231	2.40
D28	28.4	0.221	2.52
D29	30.9	0.236	3.24
D30	22.6	0.198	3.00
D31	28.3	0.236	2.22
D32	28.5	0.239	2.19
D33	27.8	0.212	3.24
D34	27.3	0.227	2.21
D35	26.5	0.231	2.30
D36	27.1	0.232	2.42

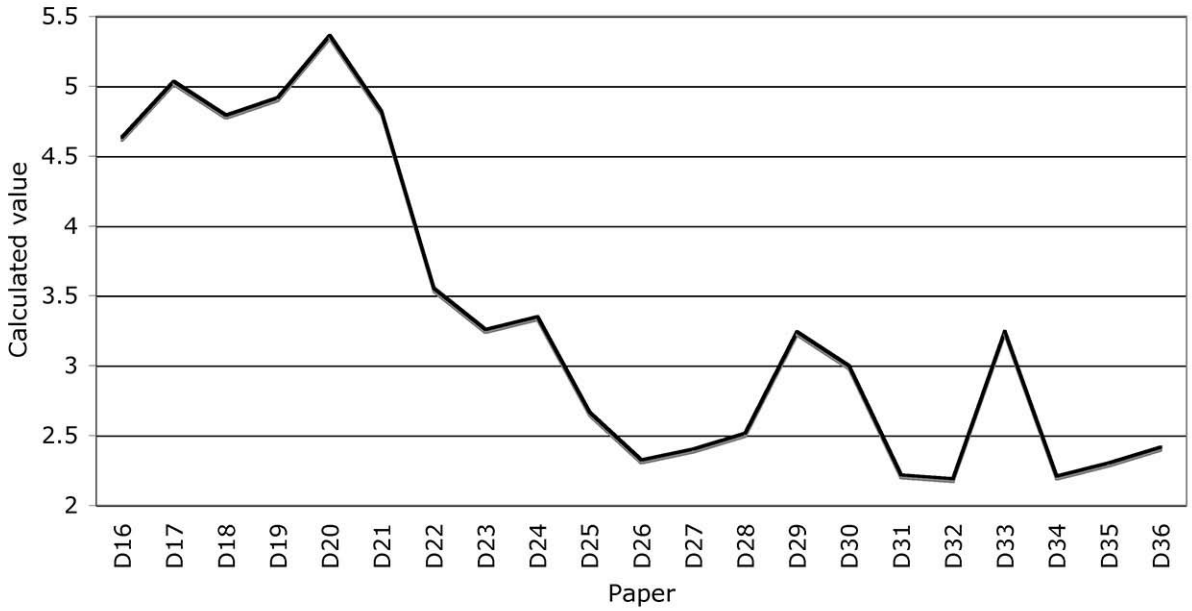
Colorfulness



Contrast



Noise



Color difference

Paper	Average	Red	Green	Blue	Cyan	Magenta	Yellow	Orange
D16	2.94	4.08	2.13	5.29	0.18	2.11	2.47	4.35
D17	6.54	12.60	3.71	5.51	7.29	5.35	2.83	8.50
D18	7.27	14.20	4.01	5.18	7.85	5.21	2.91	11.54
D19	6.33	11.31	3.48	5.97	7.38	4.10	3.10	8.96
D20	5.68	12.32	2.94	6.33	5.89	4.10	2.00	6.18
D21	7.46	14.23	3.35	5.99	9.40	6.26	1.14	11.84
D22	6.17	5.56	3.65	4.35	6.04	7.29	6.83	9.48
D23	4.29	4.44	4.60	1.83	2.26	5.43	5.29	6.18
D24	7.28	9.53	4.60	7.75	4.77	4.83	6.69	12.80
D25	6.21	9.63	2.23	8.96	3.17	3.10	5.44	10.95
D26	2.14	1.32	3.11	0.62	1.54	1.93	3.94	2.49
D27	3.22	5.06	1.37	0.75	1.93	0.21	6.37	6.84
D28	5.45	8.12	3.33	8.66	2.15	1.20	3.91	10.80
D29	2.53	2.76	1.63	3.20	0.24	0.51	3.02	6.32
D30	6.40	9.80	3.61	3.62	6.37	3.77	5.28	12.32
D31	1.90	2.59	1.18	0.27	0.63	3.45	2.19	2.96
D32	1.79	2.50	2.45	0.57	0.88	1.97	3.60	0.58
D33	6.14	9.69	2.98	4.11	3.04	4.10	7.05	12.00
D34	3.35	3.22	0.62	0.06	4.42	0.66	4.97	9.47
D35	3.98	6.65	1.22	0.08	4.48	1.88	5.59	7.97
D36	3.04	3.52	0.05	0.99	3.56	0.92	5.33	6.89

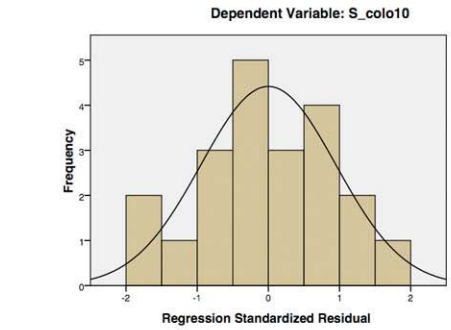
APPENDIX 13 A screenshot from the execution of the main Matlab program

```
>> calculations_main(path)

Scores and grades

Grading categories:
>6.5        Excellent
5.5-6.5     Very good
4.5-5.5     Good
3.5-4.5     Fair
2.5-3.5     Satisfactory
1.5-2.5     Poor
<1.5        Bad

Quality     Better score -> Better overall quality
Colorf.     Better score -> More color
Contrast    Better score -> More contrast
Noise       Better score -> Less noise

Filename            Quality             Colorfulness        Contrast            Noise
D16i_studio_R.tif   2.40    Poor        2.26    Poor        1.95    Poor        2.70    Satisfactory
D17i_studio_R.tif   1.48    Bad         1.82    Poor        1.83    Poor        1.41    Bad
D18i_studio_R.tif   1.57    Poor        2.32    Poor        2.14    Poor        1.60    Poor
D19i_studio_R.tif   1.57    Poor        2.40    Poor        2.20    Poor        1.62    Poor
D20i_studio_R.tif   1.49    Bad         1.80    Poor        2.08    Poor        1.14    Bad
D21i_studio_R.tif   1.79    Poor        2.12    Poor        2.26    Poor        1.53    Poor
D22_studio_R.tif    3.62    Fair        4.66    Good        4.32    Fair        3.52    Fair
D23_studio_R.tif    4.42    Fair        5.33    Good        4.95    Good        4.31    Fair
D24_studio_R.tif    3.77    Fair        4.85    Good        4.56    Good        3.57    Fair
D25_studio_R.tif    4.80    Good        5.70    Very good   5.21    Good        4.72    Good
D26_studio_R.tif    6.26    Very good   6.10    Very good   5.82    Very good   6.03    Very good
D27_studio_R.tif    6.13    Very good   5.47    Good        5.47    Good        5.70    Very good
D28_studio_R.tif    5.14    Good        5.80    Very good   5.30    Good        5.09    Good
D29_studio_R.tif    4.91    Good        5.29    Good        5.07    Good        4.69    Good
D30_studio_R.tif    3.96    Fair        4.71    Good        4.02    Fair        4.23    Fair
D31_studio_R.tif    6.22    Very good   6.11    Very good   5.61    Very good   6.22    Very good
D32_studio_R.tif    6.42    Very good   6.16    Very good   5.80    Very good   6.28    Very good
D33_studio_R.tif    3.97    Fair        4.98    Good        4.52    Good        3.95    Fair
D34_studio_R.tif    5.86    Very good   6.02    Very good   5.43    Good        5.94    Very good
D35_studio_R.tif    6.06    Very good   5.84    Very good   5.71    Very good   5.68    Very good
D36_studio_R.tif    5.93    Very good   5.78    Very good   5.51    Very good   5.72    Very good

Elapsed time: 16.3 seconds.

>>
```

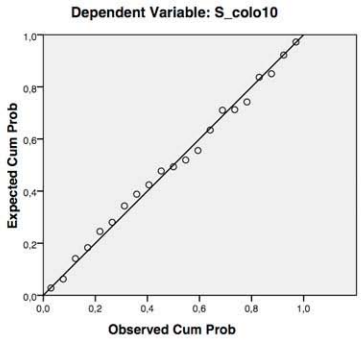
APPENDIX 14 Colorfulness model: regression charts generated by SPSS

In regression charts, S refers to subjective results, M to Matlab method, qual = quality, colo = colorfulness, cont = contrast, and cbal = color balance.

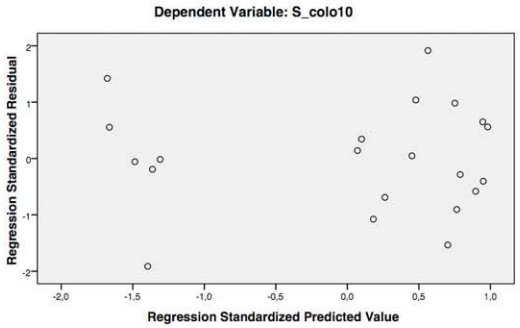
Histogram



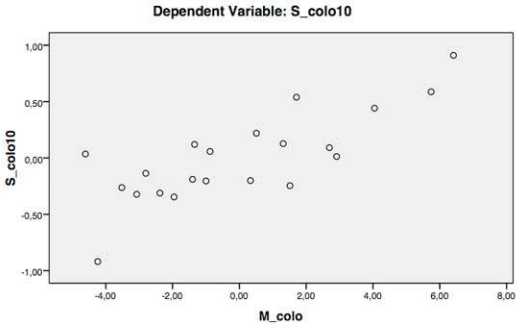
Normal P-P Plot of Regression Standardized Residual



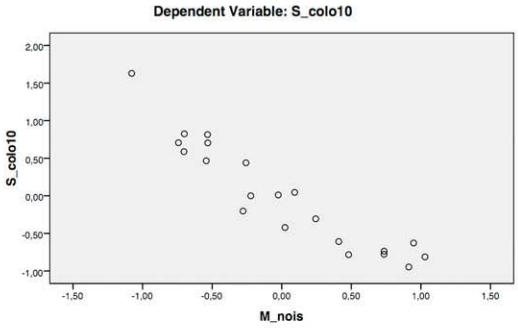
Scatterplot



Partial Regression Plot



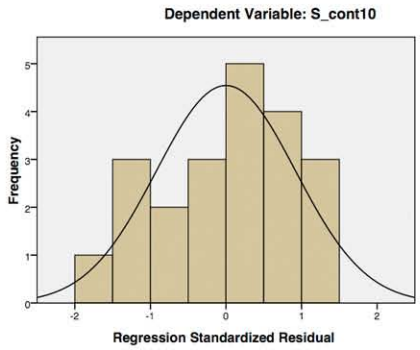
Partial Regression Plot



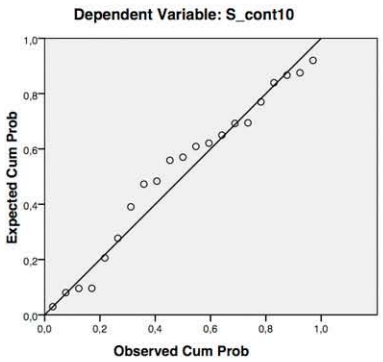
APPENDIX 15 Contrast model: regression charts generated by SPSS

In regression charts, S refers to subjective results, M to Matlab method, qual = quality, colo = colorfulness, cont = contrast, and cbal = color balance.

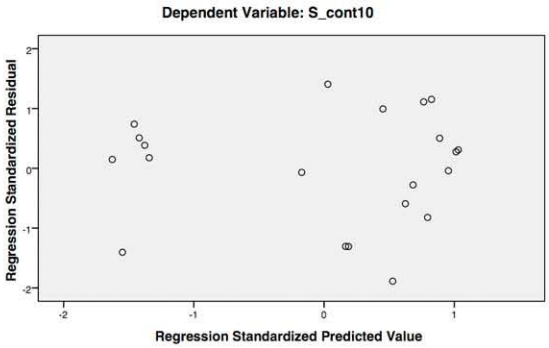
Histogram



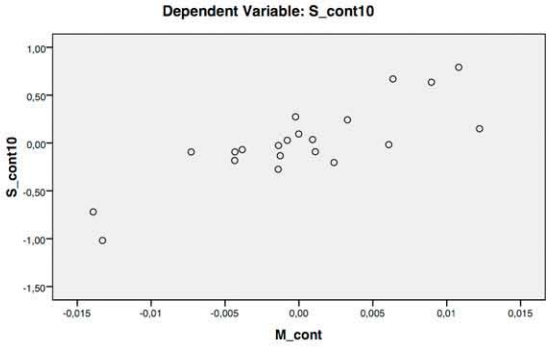
Normal P-P Plot of Regression Standardized Residual



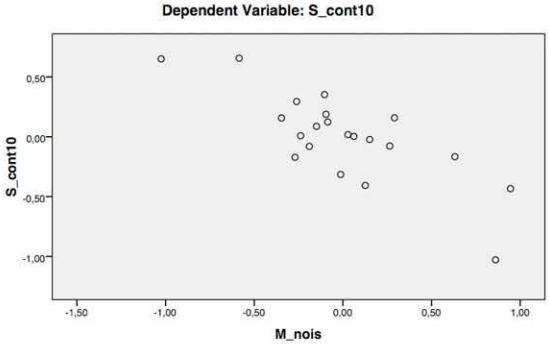
Scatterplot



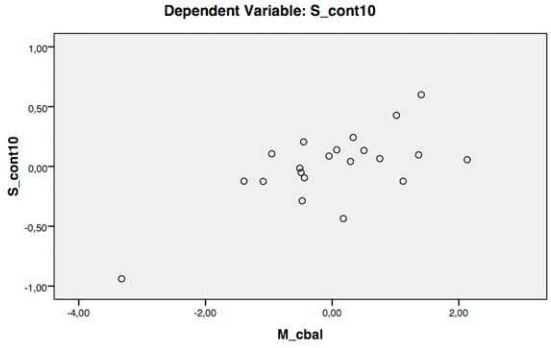
Partial Regression Plot



Partial Regression Plot



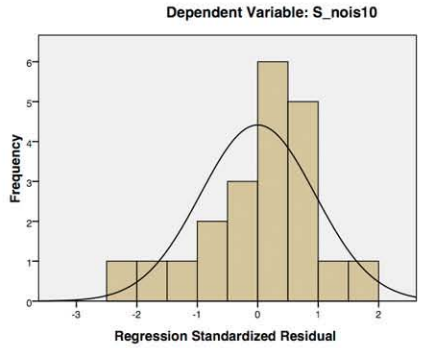
Partial Regression Plot



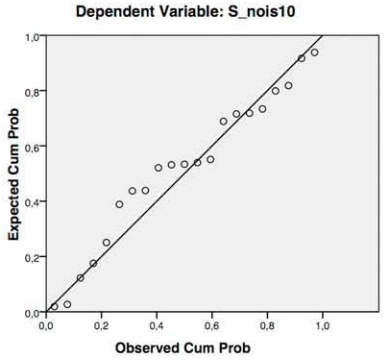
APPENDIX 16 Noise model: regression charts generated by SPSS

In regression charts, S refers to subjective results, M to Matlab method, qual = quality, colo = colorfulness, cont = contrast, and cbal = color balance.

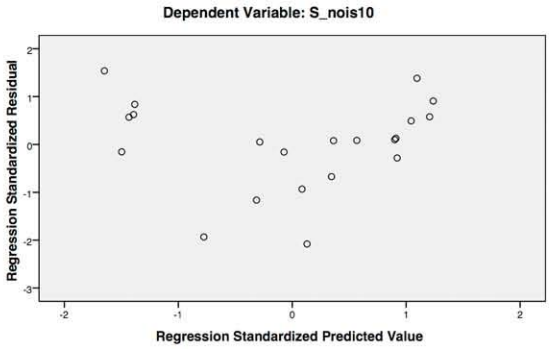
Histogram



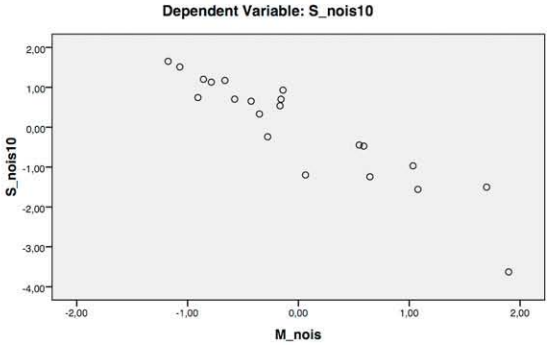
Normal P-P Plot of Regression Standardized Residual



Scatterplot



Partial Regression Plot



Partial Regression Plot