Image quality metrics for the evaluation of printing workflows

Doctoral Dissertation by

Marius Pedersen

Submitted to the Faculty of Mathematics and Natural Sciences at the University of Oslo in partial fulfillment of the requirements for the degree Philosophiae Doctor (PhD) in Color Imaging
Abstract

The aim of our research is to assess image quality of prints without the involvement of human observers. The printing industry is continuously moving forward as new products and technologies are introduced to the market. The need to assess the quality is increased with this growth, for example, to verify that these technology advancements lead to higher quality prints.

In this thesis we describe the work carried out to use image quality metrics for the evaluation of printed images. The intended behavior of such metrics is to measure or predict image quality as human observers would perceive it.

Following an introduction and background on quality assessment, we introduce the concept of image quality metrics. Existing image quality metrics are classified and a survey of them is given, to show how they are constructed and their differences. Following the survey, a new image quality metric, the Spatial Hue Angle MEtric (SHAME) is proposed, which accounts for two key aspects of image quality, namely region of interest and the human visual system. The evaluation of image quality metrics against the percept is a key aspect for ensuring that the metrics can substitute or assist human observers in the assessment of quality. Therefore, existing evaluation methods are presented and analyzed, revealing the need for a new method to assess the overall performance of image quality metrics. For that reason, a new method to evaluate overall performance is proposed, based on the rank order method. Using existing methods and the new evaluation method, a set of commonly used metrics are evaluated with a set of public databases. These databases contain digital images, with a range of different distortions and quality issues, and have quality ratings from human observers.

The knowledge gathered from the evaluation of image quality metrics on the existing databases was then applied to the assessment of printed images. Since the metrics require digital images as input, a framework to digitize printed images is proposed. Using this framework a set of metrics is evaluated against human observers, which shows that none of the existing metrics predict overall image quality adequately. These findings lead us to break overall image quality into parts, more precisely quality attributes. Based on existing quality attributes a manageable set of six color printing quality attributes is proposed. The final set included: sharpness, color, lightness, contrast, artifacts, and physical. Through two experimental validation procedures these attributes are found to be a good foundation for the evaluation of color prints. The image quality metrics are then used to describe each of the proposed quality attributes separately to find the most appropriate metrics to measure the quality of each attribute. Two experiments with human observers were carried out, which acted as the basis for the evaluation and selection of metrics. The results show that for some attributes, such as sharpness, suitable metrics can be found, but additional work is needed to find metrics that correlate well with the percept for all of the attributes.

An area that may be improved with the use of image quality metrics is the reduction of quality values to a more manageable number (pooling), usually a single quality value. We have investigated the impact of pooling strategies on the performance of image quality metrics. This investigation shows that the performance is linked to the metric, and that the parameters for the pooling strategies are very important. Even with the effort spent on pooling strategies none of the evaluated metrics performed well for the color quality attribute. This lead to a proposal for a new image quality metric designed for the color attribute, Total Variation of Difference (TVD) metric, which applies a spatial filtering to simulate the human visual system before quality is calculated. A comparison against the state of the art metrics shows an increased performance for the new metric.
Lastly, we gather the work carried out in this thesis to develop a practical tool for the printing industry to assess the quality of prints using image quality metrics, named the Quality Assistant. The Quality Assistant consists of all functions needed to evaluate quality, including: a test image suite, the framework for digitizing the prints, a set of image quality metrics, and visualization tools. Through the work carried out in this thesis we have shown the applicability of image quality metrics for the evaluation of printing workflows.
Acknowledgements

There are many people who deserve to be acknowledged, without their support and help it would not have been possible to complete this work. First of all, I would like to thank my main supervisor, professor Jon Yngve Hardeberg from Gjøvik University College, who introduced me to the field of color imaging. Many thanks goes to my second supervisor Professor Fritz Albregtsen from from the Department of Informatics at the University of Oslo, and Nicolas Bonnier, my industry supervisor from Océ Print Logic Technologies. Their help and assistance has been invaluable. My deepest gratitude to all of you for your guidance and support during this work.

I would also like to thank the members of the Norwegian Color Research Laboratory at Gjøvik University College: Peter Nussbaum, Aditya Sole, Ivar Farup, Gabriele Simone, Arne Magnus Bakke, Dibakar Raj Pant, Zhaohui Wang, Raju Shresta, and others.

Many of the employees at Océ have helped during the last years, and in particular I would like to thank Christophe Leynadier, Jacques Perville, Arlette Del-Aguila, Kristyn Falkenstern, Medhi Felhi, Albrecht Lindner, Maria-Valezzka Ortiz-Segovia, Reneka Even, Christine Hurtret, and Marine Lauga. Many thanks to Océ for the pleasant work atmosphere in Créteil. Thanks to Frans Gaykema of Océ Venlo for his support and discussions.

Thanks also goes out to Zofia Baranczuk, Timothée Royer, Sébastien Akli Ajagamelle, Ali Amirshahi, Eriko Bando, Faouzi Alaya Cheikh, Alessandro Rizzi, Guanqun Cao, and many others for their co-operation, help, and assistance.

My deepest appreciation is due to my family and friends, their support has been invaluable. Finally, I would like to express my gratitude to Lene, who has borne my increasing absence with good will.

Thank you all.

Gjøvik, Norway
August 2011
Marius Pedersen
CONTENTS

1 INTRODUCTION
1.1 Motivation ................................................................. 3
1.2 Aims ........................................................................... 4
1.3 Research methods ...................................................... 4
1.4 Publications ................................................................. 5
1.5 Thesis outline ............................................................. 8

PART I BACKGROUND ................................................... 11

2 WHAT IS QUALITY? ......................................................... 13
2.1 Quality ........................................................................ 13
2.2 Image quality ............................................................. 13
2.3 Print quality ............................................................... 14
2.4 Definition used in this work .......................................... 14

3 EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY 15
3.1 Objective quality assessment ........................................ 15
3.1.1 Measuring quality using instruments ......................... 15
3.1.1.1 Densitometer ..................................................... 15
3.1.1.2 Colorimeter ..................................................... 16
3.1.1.3 Spectroradiometer ........................................... 18
3.1.1.4 Spectrophotometer .......................................... 18
3.1.1.5 Glossmeter ..................................................... 18
3.1.2 Quality evaluation in the industry .............................. 18
3.1.2.1 Process-standard offset .................................... 19
3.1.3 Standards for printing .............................................. 20
3.1.3.1 ISO 12647 ....................................................... 20
3.1.3.2 USA ............................................................. 20
Specifications for Web Offset Publications ....................... 20
General Requirements and Applications for Commercial Off-
set Lithography ............................................................. 20
3.1.3.3 Germany ........................................................ 20
Der Bundesverband Druck und Medien e.V. ....................... 20
3.1.3.4 Switzerland .................................................... 21
System Brunner ............................................................. 21
3.1.4 Test images ............................................................. 21
3.1.4.1 Altona test suite .............................................. 21
3.1.4.2 Roman16 bvdm reference images ..................... 21
10 What quality attributes are the most important? 141

10.1 State of the art .......................................................... 141
10.1.1 Image quality attributes ........................................... 142
10.1.2 Image quality models ............................................... 142

10.2 Investigation and selection of important quality attributes ................. ......... 144
10.2.1 Color ........................................................ 146
10.2.2 Lightness ...................................................... 147
10.2.3 Contrast ........................................................ 147
10.2.4 Sharpness ................................................. 148
10.2.5 Artifacts .................................................... 148
10.2.6 Physical ...................................................... 148
10.2.7 Relations between quality attributes ........................................ 149

10.3 Validation of the quality attributes - experiment I ....................... ......... 149
10.3.1 Experimental setup .................................................. 150
10.3.1.1 Test images .............................................. 150
10.3.1.2 Color workflow ........................................ 151
10.3.1.3 Viewing conditions ...................................... 151
10.3.1.4 Instructions given to the observers .................... ......... 153

10.3.2 Perceptual results........................................................ 153
10.3.3 Fitting the quality attributes to the color printing quality attributes 154
10.3.4 Relations between the attributes .................................. 154

10.4 Validation of the quality attributes - experiment II ................................ 157
10.4.1 How to validate quality attributes? .................................. 158
10.4.2 Experimental setup .................................................. 159
10.4.2.1 Images .................................................. 159
10.4.2.2 Color workflow ........................................ 160
10.4.2.3 Viewing conditions ...................................... 160
10.4.2.4 Instructions ................................................ 160

10.4.3 Fitting the quality attribute data to the color printing quality attributes . 160
10.4.3.1 Discussion on the fitting of quality attributes .................. 163
    Overlapping QAs .................................................. 163
    Independence ....................................................... 163
    Global and local issues .......................................... 164
    One child with several own children ......................... 164
    Skewness ......................................................... 165
    Dimensionality .................................................... 165

10.4.4 Observations on the color printing quality attributes ....................... 165
10.4.5 Dependence between the attributes .................................... 167
10.4.6 Validation summary ............................................... ......... 168

10.5 Summary ........................................................... 169
11 IMAGE QUALITY METRICS TO MEASURE QUALITY ATTRIBUTES 171
  11.1 Selection of image quality metrics for the color printing quality attributes 171
    11.1.1 Sharpness ................................................................. 171
    11.1.2 Color ........................................................................ 172
    11.1.3 Lightness ................................................................. 172
    11.1.4 Contrast .................................................................... 172
    11.1.5 Artifacts ................................................................. 172
  11.2 Experiment I ................................................................. 173
    11.2.1 Selected image quality metrics ........................................ 173
    11.2.2 Evaluation of the selected image quality metrics ............... 173
      11.2.2.1 Experimental setup ......................... .......................... 174
      11.2.2.2 Evaluation of image quality metrics ............. 174
        Phase 1: na"ive observers ............................................. 174
        Phase 2: expert observers ........................................... 177
    11.2.3 Investigation of image characteristics ............... 180
      11.2.3.1 Selected image characteristics .................... 180
      11.2.3.2 Experiment ............................................. 181
      11.2.3.3 Dominant color .................. .................................. 181
      11.2.3.4 Colorfulness ................................................ 183
      11.2.3.5 Lightness .................................................. 183
      11.2.3.6 Details .......................................................... 185
      11.2.3.7 Overall observations .......................... 186
  11.3 Experiment II ................................................................. 187
    11.3.1 Selected image quality metrics ........................................ 187
    11.3.2 Evaluation of the selected image quality metrics ............... 187
      11.3.2.1 Experimental setup ........................................... 187
      11.3.2.2 Experimental results ........................................ 189
        Preparation of the printed images ....................... 189
        Evaluation method ............................................ 190
        Evaluation results .................................................. 190
    11.4 Summary ................................................................. 192

12 IMPROVING PERFORMANCE OF METRICS BY POOLING 193
  12.1 State of the art ................................................................. 195
    12.1.1 General formulation ................................................. 195
    12.1.2 Quality based pooling .............................................. 196
      12.1.2.1 Minkowski pooling ........................................... 196
      12.1.2.2 Local quality/distortion-weighted pooling ........... 197
        Monotonic function pooling ...................................... 197
        Percentile pooling .................................................. 197
    12.1.3 Content based pooling .............................................. 198
      12.1.3.1 Information content-weighted pooling ............... 198
14 QUALITY ASSISTANT - TOOL FOR EVALUATING PRINT QUALITY

14.1 Pre-metric operations

14.1.1 Padding the images

14.1.2 Scanning

14.1.3 Registration

14.1.4 Registration validation

14.2 Measuring attributes using image quality metrics

14.2.1 Sharpness

14.2.1.1 Perceptual sharpness

14.2.1.2 Computational sharpness

14.2.2 Artifacts

14.2.2.1 Noise

14.2.2.2 Banding

14.2.3 Color

14.3 Visualization of results

14.3.1 Spider and bar plots

14.3.2 Quality maps

14.3.2.1 Box plots

14.3.2.2 3 and 4 dimensional plots

14.3.2.3 Color histograms

14.3.3 Scale differences

14.4 Overview of the quality assistant

14.4.1 Show test images

14.4.2 Pad images

14.4.3 Scan profile evaluation

14.4.4 Image registration

14.4.5 Image registration validation

14.4.6 Calculate results

14.4.7 Show results

14.5 Evaluation of the Quality Assistant

14.5.1 Experimental setup
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.5.1.1 Printers and print settings</td>
<td>250</td>
</tr>
<tr>
<td>14.5.1.2 Observers and instructions</td>
<td>251</td>
</tr>
<tr>
<td>14.5.1.3 Viewing conditions</td>
<td>251</td>
</tr>
<tr>
<td>14.5.1.4 Scanning and registration</td>
<td>251</td>
</tr>
<tr>
<td>14.5.2 Results and discussion</td>
<td>252</td>
</tr>
<tr>
<td>14.5.2.1 Sharpness</td>
<td>252</td>
</tr>
<tr>
<td>14.5.2.2 Noise</td>
<td>253</td>
</tr>
<tr>
<td>14.5.2.3 Banding</td>
<td>253</td>
</tr>
<tr>
<td>14.5.2.4 Color</td>
<td>255</td>
</tr>
<tr>
<td>14.5.3 Overall observations</td>
<td>255</td>
</tr>
<tr>
<td>14.6 Summary</td>
<td>256</td>
</tr>
<tr>
<td>15 Conclusion</td>
<td>257</td>
</tr>
<tr>
<td>16 Future work</td>
<td>261</td>
</tr>
<tr>
<td>16.1 Quality Assistant</td>
<td>261</td>
</tr>
<tr>
<td>16.2 Quality attributes</td>
<td>262</td>
</tr>
<tr>
<td>16.3 Image quality metrics</td>
<td>262</td>
</tr>
<tr>
<td>16.4 Image characteristics</td>
<td>262</td>
</tr>
<tr>
<td>Nomenclature</td>
<td>263</td>
</tr>
<tr>
<td>References</td>
<td>280</td>
</tr>
<tr>
<td>Part IV Appendices</td>
<td>319</td>
</tr>
<tr>
<td>A Overview of Image Quality Metrics</td>
<td>321</td>
</tr>
<tr>
<td>B Measuring Perceptual Contrast in a Multi-Level Framework</td>
<td>333</td>
</tr>
<tr>
<td>B.1 Background</td>
<td>333</td>
</tr>
<tr>
<td>B.1.1 Tadmor and Tolhurst</td>
<td>333</td>
</tr>
<tr>
<td>B.1.2 Rizzi et al.</td>
<td>334</td>
</tr>
<tr>
<td>B.1.3 Retinal-like subsampling contrast</td>
<td>335</td>
</tr>
<tr>
<td>B.2 The weighted-level framework</td>
<td>336</td>
</tr>
<tr>
<td>C Saliency Models as Gamut-Mapping Artifact Detectors</td>
<td>339</td>
</tr>
<tr>
<td>C.1 Saliency map</td>
<td>339</td>
</tr>
<tr>
<td>C.2 Applicability of saliency maps to gamut mapping artifacts</td>
<td>340</td>
</tr>
<tr>
<td>C.3 Experimental framework: saliency models as artifact detectors</td>
<td>340</td>
</tr>
<tr>
<td>C.3.1 Global strategy</td>
<td>342</td>
</tr>
<tr>
<td>C.3.2 Local strategy</td>
<td>343</td>
</tr>
<tr>
<td>D Detection of Worms in Error Diffusion Halftoning</td>
<td>347</td>
</tr>
<tr>
<td>D.1 Introduction</td>
<td>347</td>
</tr>
<tr>
<td>D.2 Proposed error diffusion worm measure</td>
<td>349</td>
</tr>
<tr>
<td>D.3 Evaluation of the error diffusion worm measure</td>
<td>351</td>
</tr>
<tr>
<td>D.3.1 Experimental setup</td>
<td>351</td>
</tr>
<tr>
<td>D.3.2 Results</td>
<td>352</td>
</tr>
<tr>
<td>D.3.2.1 Overall scores</td>
<td>352</td>
</tr>
</tbody>
</table>
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.3.3 Gradient image</td>
<td>353</td>
</tr>
<tr>
<td>D.3.4 Artistic image</td>
<td>355</td>
</tr>
<tr>
<td>D.3.5 Highlight image</td>
<td>357</td>
</tr>
<tr>
<td>D.3.6 Shadow image</td>
<td>358</td>
</tr>
<tr>
<td>D.3.7 Comparison to other metrics</td>
<td>358</td>
</tr>
<tr>
<td>D.3.7.1 Overall performance</td>
<td>360</td>
</tr>
<tr>
<td>D.4 Summary</td>
<td>362</td>
</tr>
</tbody>
</table>

### EF FROM CONTRAST TO IMAGE DIFFERENCE: TWO NEW METRICS

#### F SPECIFICATIONS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>F.1 Monitors</td>
<td>365</td>
</tr>
<tr>
<td>F.1.1 Eizo CG241W</td>
<td>365</td>
</tr>
<tr>
<td>F.1.2 Eizo CG211</td>
<td>367</td>
</tr>
<tr>
<td>F.2 Scanners</td>
<td>368</td>
</tr>
<tr>
<td>F.2.1 Microtek ScanMaker 9800XL</td>
<td>368</td>
</tr>
<tr>
<td>F.2.2 Epson Expression 10000XL</td>
<td>369</td>
</tr>
<tr>
<td>F.2.3 HP ScanJet G4050</td>
<td>370</td>
</tr>
<tr>
<td>F.3 Printers</td>
<td>371</td>
</tr>
<tr>
<td>F.3.1 Océ Colorwave 600</td>
<td>371</td>
</tr>
<tr>
<td>F.3.2 HP Designjet 10ps</td>
<td>372</td>
</tr>
<tr>
<td>F.3.3 HP Color LaserJet 4600dn</td>
<td>373</td>
</tr>
<tr>
<td>F.4 Paper</td>
<td>374</td>
</tr>
<tr>
<td>F.4.1 Océ LFM 050 Red Label</td>
<td>374</td>
</tr>
<tr>
<td>F.4.2 Océ LFM 054 Red Label</td>
<td>374</td>
</tr>
<tr>
<td>F.4.3 Stora Enso MultiCopy Original</td>
<td>375</td>
</tr>
<tr>
<td>F.4.4 HP office paper</td>
<td>375</td>
</tr>
<tr>
<td>F.5 Viewing cabinet</td>
<td>376</td>
</tr>
<tr>
<td>F.5.1 Verivide DTP</td>
<td>376</td>
</tr>
</tbody>
</table>

### G LEGAL

377
1 INTRODUCTION

For centuries attempts to achieve accurate reproductions of pictorial images in an efficient way have been tried on various materials, and it is still an active field. Interest in image reproduction spreads across several occupations, such as engineers, photographers, and scientists, and industries, such as printing, painting, photography, display, and film. Even in our digital age most people’s preference is still towards paper rather than displays [184].

Many problems are encountered while working with the reproduction of images on paper, such as the limited color gamut that a reproduction device can reproduce, paper characteristics, and registration issues. Managing these problems has been a key issue for color reproduction, and has been motivation for extensive research. Since reproduction devices are not able to create a reproduction that is identical to the original, differences between the original and reproduction will occur. These differences impact the impression of the quality of the reproduction. Due to immense research, technology advancements are rapid, and new and more refined ways to deal with the limitations of a reproduction system are proposed continuously in order to achieve high quality images. One of the key factors customers consider when buying or upgrading a color reproduction system is Image Quality (IQ) [118]. Quality assessment is therefore needed to show if technology advances increase the quality of the produced image (Figure 1.1).

1.1 Motivation

There are essentially two types of IQ assessment: subjectively or objectively. Subjective evaluation is carried out by human observers, and is therefore influenced by the Human Visual

Figure 1.1: When printing a digital image, the physical printed image might present differences from the original. These differences contribute to the impression of quality. Extensive research has been carried out to improve the quality, but in order to know if this research improves quality, some kind of quality assessment is needed.
INTRODUCTION

System (HVS). Objective evaluation of IQ can be carried out in many ways, a typical way is to use measurement devices gathering numerical values. Another way is to use algorithms, commonly known as IQ metrics, in an attempt to quantify quality.

Subjective evaluation is commonly used since it is a precise way to quantify IQ [116]. However, it is both resource and time demanding, and in addition observers can be inconsistent. Because of these limitations, objective evaluation methods have been proposed. They have lower cost, are less time and resource demanding, they produce consistent results, and they require less competence of the user. One of these objective methods is IQ metrics, which are based on different assumptions and incorporate different characteristics of the HVS. An impressive number of IQ metrics have been proposed in the literature. Unfortunately, no IQ metric fully correlated with perceived quality has yet been proposed. Additionally, use of IQ metrics in the assessment of quality for printed images has not been extensively researched. Using IQ metrics in the printing process has a great potential, in terms of determining the most suitable processing of an image, to objectively assess the IQ during the printing process and of the final printed image, or to indicate where a loss of quality occurs in a print workflow.

1.2 Aims

IQ metrics have been proposed for different applications, such as video and images, and distortions, for example compression and noise. As they are becoming more correlated with perceived IQ, application of IQ metrics are becoming more popular. They can be used to monitor, evaluate, or improve quality.

The goal of this research is to adapt, use, and evaluate IQ metrics for measuring perceived IQ. Adapt refers to modifying or developing methods to apply IQ metrics to printed images, it also refers to improvement of existing IQ metrics or development of new IQ metrics, which are better correlated with subjective data than existing metrics. Use refers to the application of IQ metrics in a printing workflow, for example to measure the quality of a printed image. Evaluate refers to the evaluation of IQ metrics against perceptual data.

We will mainly focus on full-reference IQ metrics (where an original is known and it is used in the assessment of quality), and the applied field is color printing. We limit our work to halftoned images, but the methods are applicable to continuous tone images.

1.3 Research methods

Since we will investigate the relationship between IQ metrics and perceived quality, we will be carrying out psychophysical experiments which require human observers. For these experiments we will use both qualitative and quantitative methods. Qualitative methods will be used to obtain a more complete and detailed description, while quantitative method will be used for numerical analysis of data and to get a more broaden understanding.

With the perceptual data obtained from experiments, analysis will be carried out anchored in existing methods. This requires a good overview and understanding of the existing literature, before commencing the task at hand. The work carried out in this thesis contains is both theoretical, such as literature surveys, but also practical, such as experiments.
1.4 Publications

The current study has led to the publication of several papers in international peer-reviewed conferences and international journals. The relations between the publications and the thesis can be seen in Figure 1.2. Listed below are the main publications, and then the supported publications in reverse chronological order.

Main publications:


**Supporting publications:**


1.5 Thesis outline

This thesis is intended to provide the reader with the understanding required to use IQ metrics to evaluate print quality. The thesis is structured into three parts, each with several chapters, to guide the reader through the different topics of the thesis.

- Part I is divided into two chapters that introduces important background knowledge to the reader.
- Part II gives an in-depth introduction to IQ metrics in four different chapters, giving a presentation of IQ metrics, how to evaluate them, introduction of a new IQ metric, and evaluation of a set of IQ metrics.
- Part III investigates the use of IQ metrics in a printing workflow, and consists of seven chapters. The chapters address how to use metrics with printed images, which quality attributes to use, selection of IQ metrics for different attributes, improving metrics in the pooling strategy, proposal of a new metric, and at last the Quality Assistant.

Then the conclusion is given, together with proposals of future work. Appendices are given at last. An overview of the outline is shown in Figure 1.2 on the next page, linking chapters and the publications above in Section 1.4.
Figure 1.2: Overview of thesis and how they are related to the publications.
PART I

BACKGROUND
2 WHAT IS QUALITY?

Quality is a word we encounter in our daily life, and it refers to many different areas. To carry out research on the quality of prints a definition needs to be given.

2.1 Quality

A common definition of quality, regardless of field, is that quality is the conformance to requirements [90]. This definition is general, and has been adapted by many. Related definitions are given by the International Organization for Standardization (ISO), who defines quality as the totality of characteristics of an entity that bear on its ability to satisfy stated or implied needs [194] or as the ability of a set of inherent characteristics of a product, system or process to fulfill requirements of customers and other interested parties [196]. All of these definitions relate quality to some sort of requirements.

2.2 Image quality

The earliest known mentioning of IQ dates back to the invention of optical instruments, between the years 1600 and 1620 [115]. The term became more and more common with the introduction of photography and television. In the recent literature we find several definitions, such as:

- Janssen [217] defined quality in the context of visuo-cognitive systems to be the degree to which the image is both useful and natural. The usefulness of an image is the precision of the internal representation of the image and the naturalness of an image to be the degree of correspondence between the internal representation of the image and knowledge of reality as stored in memory.

- Yendrikhovskij [488] defined quality as a compromise between color realism (naturalness constraint) and color discrimination (colorfulness constraint).

- Jacobson [216] defined IQ as the subjective impression found in the mind of the observer relating to the degree of excellence exhibited by an image.

- Engeldrum [116] defined IQ as the integrated set of perceptions of the overall degree of excellence of the image.

- ISO [202] and Keelan [230] defined IQ as the impression of the overall merit or excellence of an image, as perceived by an observer neither associated with the act of photography, nor closely involved with the subject matter depicted.
WHAT IS QUALITY?

- Fairchild [122] defined color IQ as the reductions in IQ corresponding to perceptible visual differences from some ideal and the magnitude of such differences.

- The International Imaging Industry Association (I3A) defined IQ as the perceptually weighted combination of all significant attributes of an image when considered in its marketplace or application [188].

2.3 Print quality

Print quality has also been defined, and the ISO 12647 standards [193] deal with quality parameters in terms of process control for the graphic industry, and in these standards quality is related to a set of quality parameters and tolerances. This follows the common definition by Crosby [90] where quality is the conformance to requirements.

Southworth and Southworth [421] relate quality level of prints with an arbitrary or subjective judgment as to the degree of “goodness” or the absence of defects.

Field [143] adapted the framework proposed by Garvin [151] and applied it to color printing. The framework is divided into four quality facets:

- Conformance to specifications concerns the actual appearance of the image, and is the conformance to a set of tolerances.

- Excellence refers to how closely the reproduction approaches optimum or perfect appearance. Judgment of excellence is usually carried out at the proofing stage.

- Aesthetics concerns creative aspects of the image.

- Permanence relates to durability, i.e. the ability to resist influence of light, chemicals, and moisture.

In addition to these, five non-appearance factors relating to the production and economy are given; delivery urgency, rapid change, creative interface, extras, and value.

2.4 Definition used in this work

The definitions above differ in terms of assumptions and restrictions. Because of this it is important to have a definition adapted to the setting in which it is to be used. The definition of quality adopted in this work is the one by ISO [202] and Keelan [230]. This definition is a narrower definition than that of Jacobson [216] since it excludes personal attributes. However, it includes artifactual, preferential, and aesthetic attributes. This definition also allows the image to be of better quality than an original and it accounts for image enhancements, in contrast to the definition by Fairchild [122]. In practice the definition chosen corresponds to the one by Engeldrum [116], since in most relevant cases the user or customer has not taken the photograph himself.
3 EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

With the steady technology advancements, the need to verify if the new technology produces higher quality prints than the current technology increases. There are two main classes of methods to assess IQ, objective and subjective. Objective assessment can involve the use of measurement devices to obtain numerical values, alternatively IQ metrics can be used. Subjective assessment is carried out by human observers, while objective assessment does not involve observers.

3.1 Objective quality assessment

Objective quality assessment has several advantages over subjective assessment: first it is less resource demanding, in most cases more economic, faster, allows for real-time monitoring of imaging systems, easy to use for optimization of image processing systems, requires little or no knowledge to use, and usually provides more consistent results than observers. This is why it is the most common why to evaluate quality in the industry.

3.1.1 Measuring quality using instruments

Using physical instruments to measure quality has been and still is a popular way to assess quality. We will briefly mention a couple of different types of instruments useful to quantify quality.

3.1.1.1 Densitometer

A densitometer measures the level of darkness (optical density) of a print. It measures the amount of light received from a stimulus and the result is calculated using a logarithmic output. A densitometer is commonly used to measure dot gain, the phenomenon causing printed material to look darker than intended. This instrument is important in order to have correct density steps.

There are three main types of densitometers:

1. A reflection densitometer is used to measure the reflected light from a surface. A reflection densitometer has a stable light source, optics to focus the light on a pre-defined spot on a surface, a set of filters for the spectral response, and a detector to monitor
the reflected light as seen in Figure 3.1 [237, 438]. A sample is usually illuminated from above, giving an illumination angle of 90° to the surface, and viewed at 45°, following the International Commission on Illumination (Commission Internationale de l’Éclairage, CIE) annotation this is written as (0°:45°x) where x indicates the azimuthal direction of the reference plane [84]. Several other recommendations for geometry are given by CIE [84].

2. A transmission densitometer is used to measure the amount of light transmitted through a transparent material. Transmission densitometers follow the same principles as reflection densitometers, but the sensor is placed beneath the material in order to measure the transmittance (Figure 3.2).

3. A combination of the two above, measuring both reflected and transmitted light.

![Figure 3.1: Components of a reflection densitometer. A stable lightsource is used together with an infrared filter, and an aperture to focus the light, which further passes through optics before it reaches the sample surface. The reflected light from the surface passes through a color filter before it reaches a sensor, which measures the reflected light. Results are converted to a logarithmic scale before they are displayed. Figure inspired by Kipphan [237].](image)

### 3.1.1.2 Colorimeter

The tristimulus colorimeter is the simplest form of measurement for color [281], which measure tristimulus values. It has red, green, and blue photoreceptors, as the eye. A light illu-
Figure 3.2: Components of a transmission densitometer. A stable light source illuminates the film, optics to focus the light on a pre-defined spot on a surface, a set of filters for the spectral response, and a detector to monitor the reflected light are the main components.
minimates the stimuli from 45° to the normal, light reflected along the normal is collected and passed to a detector. The detector consists of three different filters with light sensitive diodes, which are combined to approximate the spectral response of the eye. A colorimeter is effective for measuring color differences, but it has some limitations. The accuracy is limited, the colorimetric values are only valid for the light source (illuminant) in the instrument, and metamerism cannot be indicated (the spectral response can be different but the tristimulus values are similar).

3.1.1.3 Spectroradiometer

A spectroradiometer measures the absolute spectral power distribution of a light source. The light is collimated by a lens into a dispersive element, and further decomposed into its spectrum [395]. The spectrum is then sampled and recorded with a set of detectors. Spectroradiometers can be used to measure both self-luminous and reflective objects. In the case of reflective objects a spectrally smooth light source and a spectrally referenced sample is required. Since these types of instruments are calibrated against a standard, the measurement uncertainty depends heavily on the reference standard [326].

3.1.1.4 Spectrophotometer

A spectrophotometer measures the ratio of reflected to incident light from a sample at many points across the visible spectrum [281]. In this instrument a light source illuminates a stimulus sample, and reflected light is passed to a spectral analyzer where the light is split into spectral components. The spectral analyzer has an advantage over the colorimeter. Unlike spectroradiometers, a spectrophotometer cannot measure self-luminous objects [395]. Therefore, spectrophotometers are useful for the calibration of printers and scanners, but not displays. A spectrophotometer has a built in light source, and for color imaging purposes the light source is usually at a 45° to the sensor. Geometry conditions are defined by CIE [84]. A schematic cross section of a spectrophotometer is shown in Figure 3.3. The spectrophotometer is the most commonly used color measurement instrument in the graphic arts and printing industry today [324].

3.1.1.5 Glossmeter

Gloss is an optical property of a surface, which is based on the interaction of light with physical characteristics of a surface. A glossmeter gives the amount of reflected light from a sample. The measurements are dependent on the material and on the angle of the illumination. Depending on the application different illumination and illumination angles are used. Gloss is measured on a scale from 0 to 100, so called Gloss Units (GU).

3.1.2 Quality evaluation in the industry

When considering printing presses there are two different approaches to ensure a stable printing process; optimized or standardized press behavior [324]. In optimized behavior one maximizes the properties of the press without consideration to external specifications or standards, which can result in unique printing conditions and usually requires custom International Color Consortium (ICC) profiles. The latter approach is to make the printing presses to conform to
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.3: Schematic cross section of a spectrophotometer. A light source illuminates the sample, the reflected (or transmitted) light is split into its spectral components by gratings (or by prisms or interference filters) before they reach the detector. Reproduction inspired by Sharma [395].

There are also commercial softwares available in the market. Quality Engineering Associates (QEA) provides test equipment for the printing industry. Their products include different measurement instruments solutions, but also scanner or camera based solutions. Their Image Quality Analysis® (IAS) system follows the ISO-13660 [197] standard, and gives the possibility to quantify print quality attributes as density, graininess, mottle, and banding. Imatest also provides software for automatic quality evaluation for digital imaging systems and sensors. In a printing setting their Gamutvision™ software is the most relevant, which visualizes gamuts and rendering intents. It also evaluates ICC profiles and includes a performance test for printers. ImageXpert® has products relevant for print evaluation as well, such as scanner based solution for automatic quality evaluation attributes as mottle, banding, text quality, and line quality.

3.1.2.1 Process-standard offset

PSO is the description of an industrially orientated and standardized procedure for the creation of print products. PSO was developed by Fogra Graphic Technology Research Association (Fogra) in co-operation with the German Printing And Media Industries Federation. PSO is in conformance with ISO 12647 [193], and should therefore ensure the quality of the production of a print product, from data creation to the finished printing product. PSO certification by Association for the Promotion of Research in the Graphic Arts Industry (UGRA)¹ contains several functions; documentation, data reception, data creation, display, digital proofing, printing plate production, printing, and illumination. In some functions, as the digital proofing

¹Visisted 02/07/11: http://www.ugra.ch/psos-certification.phtml
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

ing, both subjective (visual inspection) and objective (measurements) evaluation is carried out. In Norway PSO certification is carried out by the Norwegian Color Research Laboratory (Colorlab) following the same procedure as UGRA. In Germany PSO certification is carried out by Fogra in collaboration with der Bundesverband Druck und Medien e.V. (bvdm).

3.1.3 Standards for printing

There are international standards, such as the ISO standards, and national standard for printing. We are giving a short introduction to the most relevant ones.

3.1.3.1 ISO 12647

ISO 12647 provides standard process control aims and tolerances for various printing methods and processes. ISO 12647-1 [200] defines vocabulary, parameters, and measurement methods. ISO 12647-2 to 7 are targeted against different printing technologies, 12647-2 [201] for offset printing, 12647-3 [206] for newspaper printing, 12647-4 [207] for gravure printing, 12647-5 [198] for silk screen printing, 12647-6 [208] for flexographic printing, and 12647-7 [210] for digital proofing. These standards, such as the one for offset printing, describes control values, different paper types, tone values, solid tones, and tolerances.

3.1.3.2 USA

Specifications for Web Offset Publications Specifications for Web Offset Publications (SWOP) is an organization and the name of a set of specifications that it produces, and it aims at improving consistency and the quality of printed products. SWOP concerns only the United States, however the specifications are often referred to outside the United States [360].

General Requirements and Applications for Commercial Offset Lithography General Requirements and Applications for Commercial Offset Lithography (GRACoL) is a document containing general guidelines and recommendations that could be used as a reference source across the industry for quality color printing developed by the Graphic Communications Association (GCA), now called IDEAlliance. The different guidelines published are "GRACoL x" (GRACoL, or G, followed by the version number), 7 being the last version published in 2006.

In 2005 SWOP and IDEAlliance joined forces, and today SWOP and GRACoL are under the same organization [360].

3.1.3.3 Germany

Der Bundesverband Druck und Medien e.V. Der Bundesverband Druck und Medien e.V. (bdvm) is the German Printing and Media Industry Federation. They have for a number of years published the standard "ProzessStandard Offsetdruck" (PSO), which is a standard in co-operation with Fogra. It builds on the requirements from ISO 12647-2:2004 [201]. This work thoroughly describes the implementation of the standard with all its parameters.
3.1.3.4 Switzerland

System Brunner  System Brunner is a private swiss company that since 1975 has published their own standard "Eurostandard/Globalstandard", which sets requirements for density, dot gain, and gray balance [478]. Thorough implementation of this standard achieves the same results as bdvm and it meets the requirements of the international standard ISO 12647-2 [387].

3.1.4 Test images

There are a range of test images, both pictorial and targets, in use in the industry. We will present the most commonly used images in the industry. We limit this overview to the images mainly used for measurement instruments, images for subjective assessment will be found later in Section 3.2.4.7.

3.1.4.1 Altona test suite

bdvm is responsible for the Altona test suite, which is the "toolbox" in conformance with ISO 12647-2:2004 [201]. The test suite contains a set of offset printed A3 sheets (reference sheets), a set of ICC profiles, a set of PDF/X-3 files, and a manual. In 2005 an updated version was published as a collaboration between bdvm, European Color Initiative (ECI), European Rotogravure Association (ERA), UGRA, and Fogra. The Altona Test Suite files includes a measure sheet (Altona measure as seen in Figure 3.4(a)), a visual control image (Altona visual as seen in Figure 3.4(b)) to evaluate overprints and spot colors, and a technical file (Altona Technical as seen in Figure 3.4(c)) to evaluate overprinting and fonts.

![Altona test suite](image-url)

Figure 3.4: Altona test suite. Figures reproduced from [www.andigun.de](http://www.andigun.de), visited 28/06/11.

3.1.4.2 Roman16 bdvm reference images

The Roman16 bdvm reference images are published by the bdvm, and the suite contains 16 reference images. The reference images are specially created test motifs for visual assessment, processing and output in premedia and printing. In the images all basic colors are covered, such as CMYK and RGB. One of the 16 reference images are shown in Figure 3.5.
3.1.4.3 Visual print reference

Verband der Schweizer Druckindustrie (VSD, Association of the Swiss Printing Industry) established together with Ugra and others a set of test pages to perform quality control of the entire printing process. This set of reference prints are known as the visual print reference (Figure 3.6). The goal of the test pages is to control the printing production stage from start to end according to ISO 12467-2. The set contains eight pages with different images, and each page has an integrated chart for measurements (the Ugra/Fogra Media Wedge CMYK V3.0 and the ECI TVI10 strip).

Figure 3.5: One of the Roman16 bvdm reference images. Image reproduced from http://www.cleverprinting.de/frame_newsletter0708_1.html, visited 30/06/11.

Figure 3.6: Visual print reference. Reproduced from http://www.ugra.ch/visual-print-reference-12.phtml, visited 30/06/11.
3.1.4.4 Common test charts

A test chart is an arrangement of standardized color samples used for color comparisons and measurements. They are most commonly used to calibrate and to profile graphic devices, but also for quality control.

**MacBeth ColorChecker color rendition chart** The MacBeth ColorChecker color rendition chart was first produced in 1976. The test chart consists of 24 squares of painted samples based on Munsell colors. Color charts, such as the ColorChecker, are used to calibrate and to profile devices. The ColorChecker was commercialized by X-Rite.

![MacBeth ColorChecker color rendition chart](http://www.pbase.com/elliot/image/84099006/, visited 28/06/11)

**IT8 test charts** IT8 is a set of American National Standards Institute (ANSI) standards for color communications and control specifications. For printing the IT8.7/3 specification are used for characterization of 4-color process printing. The IT8.7/3, as seen in Figure 3.8, defines two data sets; the basic ink value data set consists of 182 patches and the extended data set is 928 patches. The first 182 patches can be used as inputs to Neugebauer equations, masking equations, color mixing model calibration, black generation, and grey balance [161]. The most common use of the IT8.7/3 test chart is for calibrating and profiling printers.

The IT8.7/2 is a color reflection target used for scanner calibration, while IT8.7/1 is a transmission target for scanners.

**ECI 2002** The ECI 2002 target (Figure 3.9) has been developed by ECI [34]. The ECI 2002 chart is a superset of the IT8.7/3 target, where all 928 patches of the IT8/7.3 are contained in the 1485 patches of the ECI 2002 chart. The chart is available in visual (structured) and random layout, the random layout is recommended for characterization of a press [109].

**Ugra/Fogra Media Wedge** The Ugra/Fogra Media Wedge CMYK is a standard tool for controlling the color transformation from data to digital proof or printing. The wedge consists of 72 patches, which are defined with area coverages of the process colors C (Cyan), M (Magenta), Y (Yellow) and K (Black). There is also a correspondence between the wedge and the most important patches of the ECI 2002 color chart. The wedge is also in correspondence with ISO 12647.
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.8: IT8.7/3 test chart for characterization of 4-color process printing. Reproduced from http://www.cgan.net/science/print/preprint/images/news43_06_0001.jpg, visited 30/06/11.

Figure 3.9: ECI 2002 test chart in visual layout. Downloaded from http://www.eci.org, visited 02/07/11.

Figure 3.10: Ugra/Fogra Media Wedge. Reproduced from http://www.ugra.ch/media-wedge-data.phtml, visited 30/06/11.
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

**TC3.5 CMYK** The GretagMacbeth TC3.5 CMYK consists of 432 patches specified in CMYK (Figure 3.11). A variant of this target is the TC3.5 CMYK+Calibration target, which has 88 additional patches placed around the target. These targets are commonly used to create CMYK profiles.

![Figure 3.11: TC3.5 CMYK test target.](image)

**Contrast resolution test target** The contrast resolution test target shown in Figure 3.12 consists of a two dimensional array of patches used to evaluate resolution. Each patch contains concentric circles of varying size and contrast. These circles represent a sampling of image detail. This target is usually only used for the black printer due to easy evaluation, and since it is likely that the printing engine for black is no different from the other colors [405].

**Addressability test target** Sigg [405] proposed an addressability test target (Figure 3.13), which will form the basis of the ISO 29112 standard for objective measurement of monochrome printer resolution [491]. This test target determines the native addressability of a digital printing system, where addressability is defined by Sigg [405] as ”the maximum number of individually controllable spots per linear inch that a printer can lay down parallel or perpendicular to the direction of substrate motion through the printing system process”. The evaluation of the test chart is subjective. The chart consists of diverging ray patterns that will show a discontinuity at the native addressability of the printing system. The authors mention the possibility of using scanner-based instruments instead of subjective evaluation.

**Edge characteristics targets** For edge raggedness and edge blurriness analysis a test target (Figure 3.14) is introduced in Zeise et al. [491]. It consists of three square printed solid regions orientated at 0, 8, and 24 degrees from the vertical. The evaluation of raggedness and blurriness is carried out within four regions on each side of the square.
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.12: Contrast resolution test target. Reproduced from Sigg [405].

Figure 3.13: Addressability test target proposed by Sigg [405]. Figure reproduced from Zeise et al. [491].
3.1.5 Measuring quality using algorithms

The other objective method is to quantify IQ using algorithms, so called IQ metrics. These metrics are models designed to automatically predict quality, either computational or perceptual quality. Computational quality relates to the calculation of quality without modeling the HVS, while perceptual quality is dependent on the viewing conditions and the observers. The advantage of metrics compared to instruments is that they do not require special hardware and usually not special software. Most algorithms require little interaction from users, minimizing the risk of human errors. These advantages over instruments and those of objective assessment methods over subjective methods give IQ metrics great potential to measure quality. A full introduction to algorithms for measuring IQ is found in Part II.

3.2 Subjective quality assessment

Subjective assessment involves psychophysical experiments, where human observers are asked to grade a set of images according to a given criterion. There are several existing methods for carrying out these types of experiments, which are introduced below. The use of subjective evaluation is important in this work, since we require subjective data for the evaluation of the objective methods. Validation against subjective data is the only method to ensure that the objective methods corresponds with perceived quality. Therefore a thorough introduction to different methods for subjective quality assessment is given.

3.2.1 Psychophysical thresholds

The first psychophysical experiments were carried out by Ernst Weber in the 1830’s. In these experiments it was found that the Just Noticeable Difference (JND) (difference threshold) is proportional to the stimulus magnitude, which today is referred to as Weber’s law. In the same time period Gustav Fechner carried out similar experiments, and in the 1860’s he published ”Elemente der Psychophysik” (Elements of psychophysics) [133], which was a study of the relationship between stimulus intensity and subjective experience. In psychophysics one can find three traditional methods for measuring thresholds; method of adjustment, method of limits, and method of constant stimuli.
3.2.1.1 Method of adjustment

In the method of adjustment the observer would adjust a parameter until a visual match with a reference stimulus is obtained. This is carried out for a given set of stimuli with varying characteristics, for example patches with varying intensity. The setting where a visual match is found is recorded. The process is repeated for all stimuli in the test set. Further, statistical analysis can be carried out to find the Point of Subjective Equality (PSE) (the average setting that gives a visual match), and the JND (the smallest difference between stimuli that can be reliably discriminated 50% of the time). Figure 3.15 shows an illustration of a set of statistics commonly calculated.

![Graph showing probability of detection vs test intensity with labels for PSE, JND, LL, and UL.]

Figure 3.15: Example showing the Point of Subjective Equality (PSE), Just Noticeable Difference (JND), Lower Limit (LL), Upper Limit (UL), and Interval of Uncertainty (IU). The PSE is located at the 50% probability of detection. The UI is the area between the LL and the UL. One JND equals the difference between the PSE and the upper or lower threshold.

Example method of adjustment: If a gray standard patch is given with an intensity of 50, the observer would be given the standard patch and another patch with higher or lower intensity, and then be instructed to adjust the intensity to match the standard (Figure 3.16). For the first trial the observer is shown the reference patch and one of the samples from the test set, in this case 30. The observer adjusts the "knob" ascending and the final intensity is 49, which is the intensity for which the observer perceives both patches as identical in intensity. Further, the same procedure is carried out for the other samples in the test set, in the second trial the observers are shown a test patch with an intensity of 70, and the final intensity is set to 52. For the third trial a patch of 40 is shown, and adjusted by the observer to 51. This is then carried out until every patch in the test set is evaluated. Even though the example here is with intensity patches, the setup can also be used to evaluate natural images.

The data gathered from the method of adjustment (Table 3.1) can be statistically analyzed to obtain information about the PSE and JND. Table 3.2 shows the results after a basic statistical analysis of the recorded data. The PSE is the grand mean, in this case 50.6. Another commonly used statistics is the JND. This can also be calculated from the gathered data. JND is usually defined as the point in the distribution that is equal to or greater than 75% of the...
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Test set

Reference

Trial 1

Trial 2

Trial 1

Ascending

Descending

Ascending

Start intensity: 30
Final intensity: 49

Start intensity: 70
Final intensity: 52

Start intensity: 40
Final intensity: 51

Figure 3.16: Example method of adjustment. The observer is given a pair of patches, the reference and one from the test set. The observer is instructed to adjust the intensity of the test patch until it has the same intensity as the reference. The intensity that the observer indicates as identical to the reference is recorded. Figure inspired by Ferwerda [141].
values in the dataset. Given that the data is normally distributed we can find the point that is equal to or greater than 75% of the data (75th percentile), this is where the z-scores (or a standard score indicates how many standard deviations an observation is above or below the mean) are approximately 0.67. The JND can be found, with some assumptions, by taking the STandard Deviation (STD) of the gathered data and multiplying it with 0.67. The JND is then \(0.67 \times \text{STD}\), and with the data from Table 3.2 we get a JND of: \(0.67 \times 2.17 = 1.45\). Based on the PSE and JND the UL and LL thresholds can be derived as \(PSE \pm JND\), giving the UL threshold to be 52.05 and the LL threshold to be 49.15. By subtracting the Lower Limit threshold from the Upper Limit threshold the Interval of Uncertainty (IU) can be found, in this case 2.9. These results can be shown in a summary sheet (Table 3.3).

**Table 3.1:** Recorded data from an experiment carried out based on the method of adjustment, where A indicates ascending, and D descending trials. See Figure 3.16 for a visual representation of the data for the first three trials.

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>Start intensity</td>
<td>30</td>
<td>70</td>
<td>40</td>
<td>60</td>
<td>35</td>
<td>65</td>
<td>45</td>
<td>55</td>
<td>42</td>
<td>62</td>
</tr>
<tr>
<td>Final intensity</td>
<td>49</td>
<td>52</td>
<td>51</td>
<td>50</td>
<td>48</td>
<td>53</td>
<td>51</td>
<td>51</td>
<td>47</td>
<td>54</td>
</tr>
</tbody>
</table>

**Table 3.2:** Statistical analysis for the method of adjustment, where the mean and STD of the ascending, descending, and grand mean of the series is calculated. The grand mean is equal to the PSE.

<table>
<thead>
<tr>
<th>Series</th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascending</td>
<td>49.20</td>
<td>1.79</td>
</tr>
<tr>
<td>Descending</td>
<td>52.00</td>
<td>1.58</td>
</tr>
<tr>
<td>Grand</td>
<td>50.60</td>
<td>2.17</td>
</tr>
</tbody>
</table>

**Table 3.3:** Statistical analysis for the method of adjustment. A summary sheet shows the relevant statistical results, such as Point of Subjective Equality (PSE), Just Noticeable Difference (JND), Upper Limit (UL), Lower Limit (LL), and Interval of Uncertainty (IU).

<table>
<thead>
<tr>
<th>Summary sheet</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PSE</td>
<td>50.60</td>
</tr>
<tr>
<td>JND</td>
<td>1.45</td>
</tr>
<tr>
<td>UL</td>
<td>52.05</td>
</tr>
<tr>
<td>LL</td>
<td>49.15</td>
</tr>
<tr>
<td>IU</td>
<td>2.90</td>
</tr>
</tbody>
</table>

### 3.2.1.2 Method of limits

The second method is the method of limits. Given the same setup as above, with a standard patch of a given intensity, this patch is compared against a test set of patches where the observer is given one patch at a time in an ascending or descending order. For each patch the observer is instructed to decide whether or not the standard patch is brighter or dimmer than
the test patch. Usually each observer is shown both ascending and descending trials to minimize adaptation and expectation errors. Similar statistical analysis to the one for the method of adjustments can be carried out to find the PSE, the JND, and other statistical measures. This method provides more precise thresholds than the method of adjustment, but at the cost of a slight increase in complexity [395]. The method of limits only provides information about the "transitions", where the observer indicates a change in one stimulus compared to another, all other stimuli do not provide any information. Because of this, the method of limits is often used to get a rough idea of where the transition occurs, then further refinement of the experiment is performed.

**Example method of limits** Following the same example as for the method of adjustment with a reference patch with an intensity of 50, and five test patches with a range of 30 to 70 in intensity are used (Figure 3.17). In the first trial the observer is shown the reference patch and the patch with an intensity of 70. The observer perceives the 70 patch as brighter than the reference. Next the patch with a lower intensity (60) is shown and perceived as brighter. Next the patch with an intensity of 50 is shown and the observer perceives this as dimmer. We then record the midpoint between 50 and 60 as the cross point, in this case 55. The same is done for an ascending trial, and then alternating order for the remaining trials. Table 3.4 shows an example of data recorded in an experiment carried out based on the method of limits. This data can further be processed to obtain the same statistical measures as for the method of adjustment; PSE, JND, STD, LL and UL thresholds, and IU. Table 3.5 shows the analysis required to obtain these measures. The PSE is the same as the mean, in this case 50. JND equals to 3.53, the UL threshold 53.53, the LL threshold 46.47, the IU 7.06 (Table 3.6).

**Table 3.4:** Recorded data from an experiment carried out based on the method of limits. **AS** indicates ascending order, **DS** descending. **Crosspoint** is the point where the observer perceives a change compared to the reference, **B** indicates that the patch was brighter than the reference, and **D** indicates a dimmer patch with lower intensity.

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series</td>
<td>DS AS DS AS DS AS DS AS DS AS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test intensity</td>
<td>70 B B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>60 B B B B B B B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 D B B B D D D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40 D B B D D D D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30 D D D D D D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crosspoint</td>
<td>55 55 45 45 45 45 55 55 55 55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.5:** Statistical analysis method of limits, where the mean and STD of the ascending, descending, and grand mean of the series is calculated.

<table>
<thead>
<tr>
<th>Series</th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascending</td>
<td>49</td>
<td>5.48</td>
</tr>
<tr>
<td>Descending</td>
<td>51</td>
<td>5.48</td>
</tr>
<tr>
<td>Grand</td>
<td>50</td>
<td>5.27</td>
</tr>
</tbody>
</table>
**EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY**

**Figure 3.17:** Example method of limits. The observers are given a pair, a reference patch and a patch from the test set. They are instructed to indicate whether the test patch is brighter or dimmer than the reference. For the first trial the reference and a test patch with an intensity of 70 are shown, and the observer judges the test patch to be brighter than the reference. Then the next patch with lower intensity is shown, and this is continued until the answer changes to dimmer. The recorded value from each trial is the average between the two values where the observer changed from one answer to the other, for the first trial this would be \(((60+50)/2) = 55\). Figure inspired by Ferwerda [141].

**Table 3.6:** Statistical analysis method of limits. The summary sheet shows the relevant statistical results, such as Point of Subjective Equality (PSE), Just Noticeable Difference (JND), Upper Limit (UL), Lower Limit (LL), and Interval of Uncertainty (IU).

<table>
<thead>
<tr>
<th>Summary sheet</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PSE</td>
<td>50.00</td>
</tr>
<tr>
<td>JND</td>
<td>3.53</td>
</tr>
<tr>
<td>UL</td>
<td>53.53</td>
</tr>
<tr>
<td>LL</td>
<td>46.47</td>
</tr>
<tr>
<td>IU</td>
<td>7.06</td>
</tr>
</tbody>
</table>
Improvements of the method of limits  Improvements to the methods of limits have been suggested as well, such as staircase methods, Parameter Estimation by Sequential Testing (PEST) [434], and QUEST [470]. The first improvement, staircase, usually starts with a high intensity stimulus, and reduces the intensity until the observer reaches the point where the stimulus is perceived as dimmer, then the staircase is reversed and the same is done in the opposite direction until the observer perceives the stimulus as brighter (Figure 3.18). This process continues until the results converge. This method is also known as ”1-down-1-up”. Statistical methods to analyze the results from the staircase method are available; see for example Feeny et al. [136] who suggest different methods to calculate the PSE.

![Figure 3.18: Example of the staircase method. Blue diamond points indicate reduction in intensity, and red squares increasing intensity.](image)

PEST [434] uses changes in step size to focus the adaptive track, stopping the track when the estimate is adequately defined (Figure 3.19). The starting level is usually significantly different from the expected final value, and a number of samples are shown. After each sample is shown at a fixed level a statistical test is applied to indicate if the performance at that level is significantly better or poorer than the targeted performance (for example 75% correct detections). If above the targeted performance, the level is changed to the opposite side of the expected final value, and the procedure is repeated. Then the step size is halved, and the process is repeated until the answers converge. The final estimate is simply the final value determined by the trial placement procedure.

QUEST [470] can be considered as an improvement of PEST, where prior information is used along with the data to guide the placement of trials (Figure 3.20). It is a method where the test intensity is chosen so the data gives the best fit to what is called a psychometric function. An advantage of QUEST is that it provides the maximum information gain per trial about the threshold value, and doing so increased efficiency, allowing the threshold to be found in the smallest number of trials.

Leek [264] gives a good introduction, comparison, and discussion of the staircase method, PEST, and QUEST.
Figure 3.19: PEST example. The first trials are usually far from an expected final value. When the answers are higher than a targeted performance, the test intensity changes. This is done until the answers converge. Blue diamond points are correct answers, while red squares are wrong answers.

Figure 3.20: Example QUEST. The first trials are usually far from an expected final value. When the answers are higher than a targeted performance, the test intensity changes. This is done until the answers converge. Blue diamond points are correct answers, while red squares are wrong answers.
3.2.1.3 Method of constant stimuli

The last method is the method of constant stimuli, where the property of a stimulus is not related from one trial to the next, but the stimulus is presented randomly. The advantage of a random order is that the correct answer cannot be predicted by the observer, minimizing expectation errors. However, the disadvantage is that it requires a large number of trials and is therefore time demanding. Statistical analysis can also be carried out on the data from this method to obtain the PSE and JND. This method has been shown to be inefficient compared to adaptive methods [469], such as QUEST.

Given a stimulus with a reference value, a set of test stimuli ($\phi$) are created surrounding the intensity range of the reference. Each stimulus is shown to the observer repeatedly in a random order, and the observer is asked to answer whether the stimulus is above or below the reference in intensity. The number of yes and no responses for the test stimuli are recorded, and the proportion ($p$) of yes responses is calculated. Based on the proportion of responses a psychometric function can be constructed. Then a table with the proportion of detections and corresponding $z$-scores for the intensity values can be constructed using the normal distribution table. Further, to obtain the PSE and other statistical measures additional analysis is required. The linear regression between the $z$-scores and test intensities is calculated, resulting in the slope and intercept of the fitting. These can further be used to obtain the PSE defined as $-(\text{intercept}/\text{slope})$ and the STD defined as $1/\text{slope}$. UL, LL and the IU are similar to the ones above.

![Example method of constant stimuli](image)

**Figure 3.21:** Example of the method of constant stimuli, with a reference test intensity patch of 50 and 5 patches in the test set. The reference is compared against one of the patches from the test set. The observer is asked to judge whether the test patch is brighter or dimmer than the reference. Figure inspired by Ferwerda [141].

**Example method of constant stimuli** An experiment is set up with 20 trials per test intensity. For each test intensity the number of patches brighter and dimmer than the reference is recorded (Table 3.7). Based on this, the percentage of patches judged as brighter than the reference can be calculated, and plotted as a psychometric function (Figure 3.22), and further the $z$-scores can be derived (for example in Microsoft Excel using the normsinv function or norminv in Matlab, which calculates the inverse of the cumulative distribution function). The $z$-scores can then be plotted against the test intensities as seen in Figure 3.23. The linear regression between the $z$-scores and the test intensities will reveal the slope and intercept, which can be used to find the PSE. In this case the linear regression is $0.07x - 3.79$, where the
slope is 0.07 and the intercept −3.79. The PSE is then the negative of the intercept divided by the slope −(−3.79)/0.07 = 48.88. The STD is calculated as 1/0.07, the JND then becomes 0.67 × STD = 0.67 × (1/0.07) = 8.67. This can be visualized as seen in Figure 3.24, where the PSE is the point where 50% of the test patches are perceived as brighter and the remaining 50% as dimmer. The IU is the distance between the UL and LL, surrounding the PSE. The area between the PSE and one of the thresholds is equal to one JND.

Table 3.7: Example constant stimuli. Each test has 20 trials, and based on the number of trials judged as dimmer the percentage of trials being dimmer can be calculated. This information can further be used to obtain z-scores.

<table>
<thead>
<tr>
<th>Test intensity</th>
<th>Brighter</th>
<th>Dimmer</th>
<th>Percentage brighter</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>2</td>
<td>18</td>
<td>0.1</td>
<td>-1.28</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>16</td>
<td>0.2</td>
<td>-0.84</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>11</td>
<td>0.45</td>
<td>-0.13</td>
</tr>
<tr>
<td>60</td>
<td>17</td>
<td>3</td>
<td>0.85</td>
<td>1.04</td>
</tr>
<tr>
<td>70</td>
<td>19</td>
<td>1</td>
<td>0.95</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Figure 3.22: Example constant stimuli. The test intensities are plotted against the percentage of trials perceived as being brighter.

3.2.2 Psychophysical scaling

Psychophysical threshold methods are useful to determine color tolerances, compression limits, and more. However, the goal is often to determine the scale of perception compared to a single threshold. In order to achieve this, scaling methods have been proposed to obtain the relationship between physical changes and perceived changes. There are three common methods for conducting psychophysical scaling experiments: category judgment, pair comparison, and rank order [116].
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.23: Example constant stimuli. Linear regression is carried out between the test intensities and the z-scores, which reveals the slope (0.07) and intercept (-3.79). This information is used to obtain the statistical measures.

Figure 3.24: Example constant stimuli. The PSE is located at the 50% probability of detection. The interval of uncertainty (IU) is the area between the lower limit threshold (LL) and the upper limit threshold (UL). One JND equals the difference between the PSE and the upper limit or lower limit thresholds.
3.2.2.1 Pair comparison

In pair comparison experiments observers judge quality based on a comparison of image pairs, and the observer is asked which image in the pair is the best according to a given criterion, for example which has the highest quality or is the least different from an original (Figure 3.25). These experiments can be either forced-choice, where the observer needs to give an answer, or the observer is not forced to make a decision and may judge the two reproductions as equals (tie). In the case of pair comparison experiments no information on the distance between the images is recorded, making it less precise than category judgment, but less complex. A Matlab script for the calculation of scale values from pair comparison data is found in Green and MacDonald [162].

In these experiments the observer evaluates \( n \) reproductions for \( m \) reference images, resulting in \( \frac{n(n-1) \times m}{2} \) comparisons. Usually in pair comparison experiments each pair of reproductions is shown twice, changing the position of the right and left reproductions to avoid bias, giving \( mn(n-1) \) comparisons. With an increasing number of reproductions the number of trials increases very rapidly, which makes it unsuitable for experiments involving many reproductions.

Using Thurstone’s Law of Comparative Judgment [437], data collected from pair comparison experiments can be transformed into interval scale data. The results from this transformation represent the distance of a given stimulus from the mean score of the set being evaluated. When calculating scaled values several assumptions must be satisfied [408]:

- Each sample has a single value that can describe its quality.
- Each observer estimates the quality of this sample with a value from a normal distribution around this actual quality.
- Each sample has the same perceptual variance.
- Each comparison is independent.

However, all of these assumptions are not always valid in IQ experiments.

Pair comparison is the most popular method to evaluate e.g. gamut mapping [80], and is often preferred due to its simplicity, requiring little knowledge by the user.

**Example pair comparison** The observers are shown a pair of images, and instructed to select one of them according to the instructions given. Given three different reproductions (A, B, and C), all combinations of them are shown. For a given pair the observer selects one, and the result is recorded in an \( n \times n \) raw data matrix, in this case \( 3 \times 3 \). In this matrix the value one is given in column \( i \) and row \( j \) for reproduction \( i \) as compared with reproduction \( j \). An example of these matrices for five different observers is shown in Table 3.8.

Then an \( n \times n \) summed frequency matrix (Table 3.9) can be computed by summing all raw matrices (Table 3.8). From the summed frequency matrix a percentage matrix (Table 3.10) can be found, by dividing the frequency matrix by the number of observations, in this case 5. This matrix shows the percentage of when a reproduction was preferred over another. From the summed frequency matrix (Table 3.9) a Logistic Function Matrix (LFM) (Table 3.11) can be found by using the formula proposed by Bartleson [29]:

\[
LFM = \ln \left( \frac{f + c}{N - f + c} \right),
\]

(3.1)
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.25: Example pair comparison experiment. For the first trial the observer judged the left patch to be closer to the reference, the same with the second trial, and in the third trial the right. The observer judges all combinations of pairs.

Table 3.8: Frequency matrices for five observers each judging three different reproductions (A, B, and C). A 1 indicates that the reproduction was preferred over another, while 0.5 indicates an equal preference.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

(a) Observer 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

(b) Observer 2

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

(c) Observer 3

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

(d) Observer 4

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

(e) Observer 5

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.9: Summed frequency tables based on the frequency tables for each observer (Table 3.8).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>B</td>
<td>4.5</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>3.5</td>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.10: Percentage matrix based on the frequency table (Table 3.9).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>B</td>
<td>0.9</td>
<td>-</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.7</td>
<td>0.8</td>
<td>-</td>
</tr>
</tbody>
</table>
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

$y = 0.7903x^{1.5}$

Figure 3.26: LFM values from Table 3.11 plotted against z-scores computed using the inverse of the standard normal cumulative distribution based on the percentages from Table 3.10. The slope of the linear regression gives the coefficient to calculate the final z-scores, in this example 0.79.

where $f$ is the value from the frequency matrix, $N$ is the number of observations, and $c$ is an arbitrary additive constant. Bartleson [29] proposed 0.5, which is the most commonly used constant [162, 302].

Table 3.11: Logistic Function Matrix (LFM).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-1.61</td>
<td>-0.69</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1.61</td>
<td>-</td>
<td>-1.10</td>
</tr>
<tr>
<td>C</td>
<td>0.69</td>
<td>1.10</td>
<td>-</td>
</tr>
</tbody>
</table>

Values in the LFM can be transformed into z-scores using a simple scaling coefficient. This coefficient is found by looking at the relationship between the inverse of the standard normal cumulative distribution for the percentage matrix (converts percentages into z-scores) and the LFM values, which is done using linear regression (Figure 3.26). The scaling coefficient for this example is the slope of the regression line. The z-scores are in the same way as in the method of constant stimuli: in Microsoft Excel using the normsinv function or in Matlab with norminv. The coefficient (slope) for this example is 0.79, which is further used to obtain the z-score matrix (Table 3.12) by multiplying the LFM matrix with the coefficient.

Table 3.12: Z-score matrix.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>-1.27</td>
<td>-0.55</td>
</tr>
<tr>
<td>B</td>
<td>1.27</td>
<td>-</td>
<td>-0.87</td>
</tr>
<tr>
<td>C</td>
<td>0.55</td>
<td>0.87</td>
<td>-</td>
</tr>
</tbody>
</table>

The mean z-score for each reproduction is found by taking the average of each column.
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Table 3.13. The 95% Confidence Interval (CI) is found by using the number of observations:

\[ CI = 1.96 \times \frac{\sigma}{\sqrt{N}}, \]  

(3.2)

where \( \sigma \) is the standard deviation and \( N \) the number of observations. Since the z-scores have a scale with units equal to \( \sigma \sqrt{2} \) the standard deviation can be set to 1, giving the confidence interval of the mean to be \( 1.96 \times (1/\sqrt{N}) \). For the given example: \( 1.96 \times (1/\sqrt{5}) = 0.62 \). The 95% confidence interval is then the mean z-scores \( \pm CI \). The most common way to visualize z-scores is by an error bar plot, as shown in Figure 3.27. The mean z-score value is indicated by the center square, and the whiskers on each line show the 95% CIs. If two CIs overlap the two reproductions are not considered to be significantly different with a 95% confidence (as seen between reproduction B and C), if they do not overlap the difference between the two CIs they are statistically significant with 95% confidence (as seen between reproduction A and C).

It has been shown that each sample does not have the same perceptual variance for IQ experiments [107]. Because of this Montag [294] proposed a new way to calculate confidence intervals:

\[ CI = \pm 1.96 \sigma_{obs}, \]  

(3.3)

where

\[ \sigma_{obs} = b_1(n - b_2)^{b_3}(N - b_4)^{b_5}, \]  

(3.4)

where \( b_1 = 1.76, b_2 = 3.08, b_3 = 0.613, b_4 = 2.55, b_5 = 0.491 \), \( n \) is the number of stimuli, and \( N \) the number of observations. This method takes into consideration the number of stimuli, while the method in Equation 3.2 does not. Equation 3.2 is the most commonly used method, and unless stated otherwise all CIs calculated in this work has been made using Equation 3.2.

Table 3.13: Mean z-scores. Higher values are better. CI = 1.96 \times (1/\sqrt{5}) = 0.62.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.91</td>
<td>-0.20</td>
<td>-0.71</td>
</tr>
</tbody>
</table>

3.2.2.2 Category judgment

In category judgment the observer is instructed to judge an image according to a criterion, and the image is assigned to a category (Figure 3.28). Five or seven categories are commonly used, with or without a description of the categories. This method is an extension of Thurstone’s law of comparative judgment, and it is based on the law of categorical judgment [442]. One advantage of category judgment is that information on the distance between images is recorded, but the task is more complex than pair comparison for the observers. Additionally, since the judgments are more subjective the results are more observer dependent. Category judgment experiments are often faster than pair comparison, with fewer comparisons necessary. Because of this it is better for large experiments compared to pair comparison. When designing this type of experiment, one can consider showing all reproduction of the same scene at once, or they can be shown separately. The observers also have the ability to assign more than one reproduction to the same category. A Matlab script for the calculation of scale values from category judgment data is found in Green and MacDonald [162].
Existing methods for the Evaluation of Print Quality

Figure 3.27: Error bar plot of z-scores from a pair comparison experiment. The mean z-scores for three different reproductions are plotted with a 95% CI. Reproduction A can be considered statistically significant from reproduction C with 95% confidence, while A is not statistically significantly different from B.

Figure 3.28: Category judgment experiment. The observer is shown the entire image set, and all stimuli are judged according to a criteria based on a seven step category scale with 1 being the best. In this example the stimuli with 50 in intensity has been judged as best and assigned to category 1.
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Example category judgment  As for pair comparison an $n \times n$ frequency matrix is calculated from the raw experimental data (Table 3.14), where each column contains the frequency that one reproduction is judged to belong to a category. These are summed to obtain a frequency matrix (Table 3.15). Based on the frequency matrix a cumulative frequency matrix, containing $n \times (n - 1)$ values, is calculated (Table 3.16), and this is converted into a cumulative percentage matrix (Table 3.17).

This matrix is then converted into a z-score matrix following the same steps as for pair comparison, first obtaining an LFM (Table 3.18), which is converted into z-scores (Table 3.19) based on a scaling coefficient found by linear regression between the LFM values and the z-scores of the cumulative percentage matrix. The coefficient for this example is 0.78.

Table 3.14: Category judgement example. Five observers judging each reproduction twice, results in 10 observations.

<table>
<thead>
<tr>
<th>(a) Observer 1</th>
<th>(b) Observer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 6 5 4 3 2 1</td>
<td>7 6 5 4 3 2 1</td>
</tr>
<tr>
<td>A 0 0 0 0 2 0 0</td>
<td>A 0 0 0 0 2 0 0</td>
</tr>
<tr>
<td>B 0 0 1 1 0 0 0</td>
<td>B 0 0 0 1 1 0 0</td>
</tr>
<tr>
<td>C 1 1 0 0 0 0 0</td>
<td>C 0 0 2 0 0 0 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) Observer 3</th>
<th>(d) Observer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 6 5 4 3 2 1</td>
<td>7 6 5 4 3 2 1</td>
</tr>
<tr>
<td>A 0 0 0 0 2 0 0</td>
<td>A 0 0 0 0 1 1 0</td>
</tr>
<tr>
<td>B 0 0 1 1 0 0 0</td>
<td>B 0 0 0 1 1 0 0</td>
</tr>
<tr>
<td>C 0 2 0 0 0 0 0</td>
<td>C 0 0 1 1 0 0 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(e) Observer 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 6 5 4 3 2 1</td>
</tr>
<tr>
<td>A 0 0 0 0 2 0 0</td>
</tr>
<tr>
<td>B 0 0 0 2 0 0 0</td>
</tr>
<tr>
<td>C 0 2 0 0 0 0 0</td>
</tr>
</tbody>
</table>

Table 3.15: Category judgment example: frequency table.

<table>
<thead>
<tr>
<th>7 6 5 4 3 2 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 0 0 0 0 7 3 0</td>
</tr>
<tr>
<td>B 0 0 1 6 3 0 0</td>
</tr>
<tr>
<td>C 1 5 3 1 0 0 0</td>
</tr>
</tbody>
</table>

Table 3.16: Category judgement example: cumulative frequency table.

<table>
<thead>
<tr>
<th>7 6 5 4 3 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 0 0 0 0 7 10</td>
</tr>
<tr>
<td>B 0 0 1 7 10 10</td>
</tr>
<tr>
<td>C 1 6 9 10 10 10</td>
</tr>
</tbody>
</table>

The next step consists of calculating an $n \times (n - 2)$ difference matrix (Table 3.20), which is found by taking the z-score values and subtracting the adjacent z-score value. The mean
Table 3.17: Category judgement example: Cumulative percentage matrix.

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
<td>0.6</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.18: Category judgement example: LFM matrix.

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-3.04</td>
<td>-3.04</td>
<td>-3.04</td>
<td>-3.04</td>
<td>0.76</td>
<td>3.04</td>
</tr>
<tr>
<td>B</td>
<td>-3.04</td>
<td>-3.04</td>
<td>-1.85</td>
<td>0.76</td>
<td>3.04</td>
<td>3.04</td>
</tr>
<tr>
<td>C</td>
<td>-1.85</td>
<td>0.37</td>
<td>1.85</td>
<td>3.04</td>
<td>3.04</td>
<td>3.04</td>
</tr>
</tbody>
</table>

Table 3.19: Category judgement example: z-score matrix.

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-2.36</td>
<td>-2.36</td>
<td>-2.36</td>
<td>-2.36</td>
<td>0.59</td>
<td>2.36</td>
</tr>
<tr>
<td>B</td>
<td>-2.36</td>
<td>-2.36</td>
<td>-1.43</td>
<td>0.59</td>
<td>2.36</td>
<td>2.36</td>
</tr>
<tr>
<td>C</td>
<td>-1.43</td>
<td>0.29</td>
<td>1.43</td>
<td>2.36</td>
<td>2.36</td>
<td>2.36</td>
</tr>
</tbody>
</table>

Table 3.20: Category judgement example: difference matrix.

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.95</td>
<td>1.77</td>
</tr>
<tr>
<td>B</td>
<td>0.00</td>
<td>0.93</td>
<td>2.02</td>
<td>1.77</td>
<td>0.00</td>
</tr>
<tr>
<td>C</td>
<td>1.72</td>
<td>1.15</td>
<td>0.93</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Average</td>
<td>0.57</td>
<td>0.69</td>
<td>0.98</td>
<td>1.58</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 3.21: Category judgement example: boundary estimates.

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundary</td>
<td>0</td>
<td>0.57</td>
<td>1.27</td>
<td>2.25</td>
<td>3.83</td>
<td>4.42</td>
</tr>
</tbody>
</table>

Table 3.22: Category judgement example: scale values.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.36</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2.36</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>1.43</td>
<td>3</td>
</tr>
</tbody>
</table>
of the column from this difference matrix is used to calculate the category boundaries (Table 3.21), which is done by setting the origin (between category 6 and 7) to zero and adding the adjacent mean values.

Final z-scores, or scale values, are found by subtracting the z-scores from the boundary estimates (Table 3.22). Confidence intervals are calculated similarly to pair comparison, based on the number of observations.

### 3.2.2.3 Rank order

For rank order experiments the observer is presented with a number of images, who is asked to rank them based on a given criterion. Rank order can be compared to doing a pair comparison of all images simultaneously. If the number of images is high, the task quickly becomes challenging to the observer. However, it is a fast way of judging many images and a simple type of experiment to implement. A Matlab script for the calculation of scale values from rank order data is found in Green and MacDonald [162].

![Figure 3.29: Rank order example. The observer ranks the reproductions from best to worst according to a given criteria.](image)

#### Example rank order

For each stimulus the rank is collected. This data can be transformed into pair comparison data based on comparative paired comparison modeling [91]. The simplest way of doing this is to compute the rank of each stimulus. Given that stimulus \( i \) is ranked \( j \) for \( k_j \) times, its rank score (\( RS_j \)) is found by:

\[
RS_j = \frac{\sum_{j=1}^{n} (n-j)k_j}{N(n-a)},
\]

where \( n - j \) is a weighting factor, resulting in the lowest in the rank to have a factor of 0. These rank scores can further be converted into z-scores. To do so, one needs to examine the difference between the rank order method and the pair comparison methods. For pair comparison a stimulus \( i \) gives a scaled value \( s_i \):

\[
s_i = \frac{1}{n-1} \sum_{j=1}^{n} z \left( \frac{\sum_{k=1}^{n} np_{jki}}{n} \right),
\]
where $z()$ indicates an operator to transform the proportion of choices to z-scores, $p_{kj}$ is a binary value representing whether a stimulus $i$ is greater than stimulus $j$. In order to convert rank order data into a pair comparison matrix we have the following equation:

$$RS_i = \frac{1}{(n-1)n} \sum_{k=1}^{n} n \sum_{j=1}^{n} np_{kij} = \frac{1}{n-1} \sum_{j=1}^{n} \left[ \frac{\sum_{k=1}^{n} np_{kij}}{n} \right].$$ (3.7)

Scaled values are obtained by applying the operator to Equation 3.7:

$$s_i = z \left( \frac{1}{n-1} \sum_{j=1}^{n} \left[ \frac{\sum_{k=1}^{n} p_{kij}}{n} \right] \right).$$ (3.8)

Given three reproductions (A, B, and C) and five observers we can obtain the ranking for each observer (Table 3.23).

**Table 3.23: Example rank order: the ranking of five different observers for three reproductions (A, B, and C)**

<table>
<thead>
<tr>
<th>Observers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

This raw data can be transformed into pair comparison data, and we can get a frequency matrix (Table 3.24). This matrix can be processed with the same method as for pair comparison to obtain z-scores.

**Table 3.24: Example rank order: raw rank order data transformed into a pair comparison frequency matrix.**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>3</td>
<td>-</td>
</tr>
</tbody>
</table>

### 3.2.3 Other methods

In addition to the three methods introduced, there are also other methods for image quality evaluation.

#### 3.2.3.1 Mean opinion score

Mean Opinion Score (MOS) is defined by the International Telecommunication Union (ITU) [189] for audio quality, but has also been extensively used for IQ [322, 399, 482]. Observers are
told to judge the quality from one to five where five is the best quality (similar to category judgment), it is also common to have a five point descriptive quality scale (bad, poor, fair, good, excellent) (Figure 3.30). The MOS is the arithmetic mean of the individual scores $S$ (Equation 3.9), and is usually given with a 95% confidence interval.

$$MOS = \frac{1}{n} \sum_{i=1}^{n} S_i,$$

where $S_i$ is the score from one observation, and $n$ is the total number of observations. This method assumes that the data is normally distributed. The MOS can also be used to calculate the Difference Mean Opinion Score (DMOS), where the MOS for the reference is then subtracted from the MOS for the other images. This generates a DMOS for each image. The DMOS value for a particular test represents the subjective quality of the test relative to the reference.

### 3.2.3.2 Double stimulus impairment scale

ITU [191] recommends a procedure similar to MOS and category judgment for the evaluation of television pictures. The method, Double Stimulus Impairment Scale (DSIS), consists of showing a reference picture first, then an impaired picture. The observer is instructed to judge the quality of the impaired picture. A scale from one to five is used, and adjective descriptions are recommended. Two different descriptions are suggested, the impairment scale (very annoying, annoying, slightly annoying, perceptible but not annoying, imperceptible) and the quality scale (excellent, good, fair, poor, bad) as seen in Figure 3.30. MOS can be derived from the experimental data.

### 3.2.3.3 Double stimulus continuous quality scale

The Double Stimulus Continuous Quality Scale (DSCQS) is another method proposed by ITU [191]. In this method the observers are simply asked to assess the overall quality of each presentation by inserting a mark on a vertical scale. The scales provide a continuous rating system to avoid quantizing errors, but they are divided into five equal lengths which correspond to the normal ITU five-point quality scale. Each rating is converted from mea-
measurements of length on the score sheet to normalized scores in the range 0 to 100 (Figure 3.30 left side), and then MOS can be calculated. Sheikh et al. [398] used a similar setup to the DSCQS. However, the raw scores were converted into difference scores and further into z-scores.

3.2.3.4 Triplet comparison

ISO 20462-2 (Photography: Psychophysical Experimental Methods for Estimating Image Quality - Part 2: Triplet Comparison Method) [203] presents another way of doing psychophysical experiments, where three images are judged simultaneously [231, 235]. In this method the stimuli are presorted in three or more categories, so that only those stimuli falling within certain ranges are subsequently rated against one another [231]. ISO proposes the three following categories: "favourable", "acceptable", and "unacceptable". The three images in each triplet are not ranked, but rated against a five-category scale (similar to ITU’s recommendation) or as the one ISO proposes; "Favourable", "Acceptable", "Just acceptable", "Unacceptable", and "Poor". It is also possible to use a seven point category scale in the case of many images. This method allows two or three stimuli to have the same rating within a triplet and a magnitude for the differences between stimuli can be quantified instead of a sign as in pair comparison. The complete process is shown on Figure 3.31. The number of sample combinations is fewer than for pair comparison (Figure 3.32), with the number of comparisons (N) equal to:

\[ N = \frac{n(n-1)}{6}, \]  

(3.10)

where \( n \) is the number of samples. It has been shown that the triplet comparison is almost 50% faster than pair comparison [203]. A function is also specified for the combinations of samples to be shown [203]:

\[ f(i) = 1 + \text{modulo}(i-1,n), \]  

(3.11)

where modulo indicates the remainder of the division of \((i-1)\) by \(n\). For the case of \( n = 7 \), we get the following combinations \((1, 2, 4), (2, 3, 5), (3, 4, 6), (4, 5, 7), (5, 6, 1), (6, 7, 2)\) and \((7, 1, 3)\) without duplication. Scheffe’s method [295] is applied to obtain an interval scale, as described in Annex E of ISO 20462-2 [203]. This can further be converted into JNDs, as described in Annex F of ISO 20462-2 [203].

![Figure 3.31: Steps of the triplet comparison method. First the stimuli set is divided into categories, then the number of samples are reduced, before triplet comparison is used for the evaluation according to a criteria. Finally interval scaling is used to obtain scaling.](image)
3.2.3.5 Quality ruler

The quality ruler uses a Standard Quality Scale (SQS) with a fixed numerical scale so that the stimuli are anchored to a physical standard, with one unit corresponding to one JND; it also has a zero point. A hardcopy quality ruler setup is shown in Figure 3.33, where the test stimuli is placed above the ruler. The ruler consists of several reference stimuli ordered from highest to lowest quality, spaced at approximately three JNDs, and labeled with an integer. The use of reference stimuli results in more reliable results when assessing large sets of stimuli spanning a wide range of quality. The observer can slide the ruler back and forth in order to compare the test stimuli with the reference stimuli. Since the reference stimuli are labeled 3, 6, 9, and so on, the observer can specify an integer between two reference stimuli, where one integer difference corresponds to approximately one JND. A softcopy quality ruler is also defined by ISO [204]. The quality ruler method [204, 230, 231] is more suitable for measuring larger quality differences than for example the triplet comparison method.

An average SQS score per stimulus can be easily obtained, and gives a quality value directly expressed in JNDs. Another advantage of the quality ruler method is that scores from different experiments are easily compared. However, the process of obtaining SQS scores is complex and delicate, and requires a lot of work in the implementation stage [373]. The quality ruler is recommended to be used with artifactual attributes (for example noise or banding), rather than preference attributes (for example color, contrast, and tone), since observers generally agree that a higher level of artifacts are less desirable. In the case of preferential attributes, the observers might not agree since there are large individual differences. Keelan and Urabe [231] recommends sharpness as a good reference attribute due to its strong effect on IQ and low variability among observers.
3.2.4 Experiment aspects

In addition to the experimental method there are several other aspects that need to be taken into account, such as the number of observers, their expertise, and the viewing conditions.

3.2.4.1 Number of observers

The number of observers in subjective experiments is important for the statistical analysis usually performed on the results, and the precision of the statistics [116, 117]. When a large number of observers are used, the average is more likely to be consistent with "overall IQ" [274] and the precision of the estimated values increases [116]. It is complicated to answer precisely how many observers are required, but guidelines can be found. In scale values (z-scores) the precision increases with the square root of the number of observers. A number between 10 and 30 observers are recommended for typical scaling applications [116, 117]. For the evaluation of gamut mapped images it is recommended by CIE to have at least 15 observers carrying out pair comparison, category judgment, or ranking experiments [80]. 15 observers are also recommended by the ITU [191] for the evaluation of television pictures. Keelan and Urabe [231] recommend a minimum of 10 observers for obtaining relative quality values for JND, and a minimum of 20 observers for obtaining absolute quality scores on the SQS scale [204].

There is usually a trade-off between the number of stimuli and the number of observers. Due to time restrictions it is often more desirable to have a large number of observers than a large stimuli material [395].

Recently, larger experiments have been carried out, such as for the Tampere Image Database (TID) [365] with 654 observers in different locations, or using the Internet, as done by Qiu and Kheiri [367], Simone et al. [413], or Fairchild [123], where an almost unlimited number of observers can be recruited.

\[\text{Experiment accessible on } \text{http://hdri.cs.nott.ac.uk/siq/} \text{ (14/02/2011)}\]
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

If we collect numerous observations even trivial differences are declared statistically significant. Therefore, it is also important to control the amount of observations (number of data samples). When case when two stimuli are judged to be different when they are not is commonly referred to as type II errors in statistics [32], also known as a false negative since it fails to reject a false null hypothesis. Additionally, time and resources will be wasted by collecting too many observations, often for minimal gain.

3.2.4.2 Observer type

The expertise and background of the observers will influence the results of experiments [103, 107, 108, 116, 175]. Usually observers are divided into two types (of groups): experts and naïve. Those considered to be experts usually have experience in judging or evaluating images [116]. The experts can, to a larger degree, distinguish among attributes and they often have a more precise scale than non-experts. In experiments where small differences are needed to be quantified experts are usually more suitable than non-experts [116]. It has been shown that experts have a stronger consensus in their response than non-experts [108, 357]. Experts also look at more regions of smaller and more precise size than non-experts [108], which also has been verified by eye-tracking experiments [339].

Cultural differences have also been found to influence IQ experiments [139, 140]. However, the differences found by Fernandez and Fairchild [139], Fernandez et al. [140], and Gast and Tse [152] were small, and are therefore usually not taken into account.

3.2.4.3 Observer characteristics

A portion of the population has color vision deficiencies, and thus have a decreased ability to perceive differences between some of the colors that others can distinguish. This will influence how they perceive images, and therefore they are usually not considered as optimal observers for IQ assessment experiments. It is therefore considered good practice to test all observers for color deficiency using pseudoisochromatic plates, for example with an Ishihara color test or the Dvorine test [31].

3.2.4.4 Experiment duration

A high number of stimuli will increase the time spent by observers, and recommendations indicate that the duration of an experiment should be limited to avoid observers fatigue [191, 252]. International Telecommunication Union [191] recommends not more than 30 minutes, while, Larabi [252] recommends that the median time over the observers should not be more than 45 minutes. In research by Van Der Linde et al. [447] observers showed on average a sustained level of concentration/effort during an eye tracking experiment lasting 1 hour.

3.2.4.5 Number of stimuli

Keelan and Urabe [231] state that a minimum of three scenes should be used in order to obtain relative quality values of JND. For the SQS scale a minimum of six stimuli is required [204]. ISO 20462-1 [202] reports that the number of test stimuli should be equal to or exceed three scenes, and preferably be equal to or exceed six scenes. CIE [80] recommends to use at least one of the obligatory test images specified by CIE, together with at least three additional
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.34: Ski image from the CIE guidelines, which is obligatory for evaluation of gamut mapping algorithms.

images, for the evaluation of gamut mapping algorithms. Field [143] indicates that between five and ten images are required to evaluate color image quality issues.

The number of stimuli used is often depending on other aspects, such as the number of observers required, the experimental method, and the precision of the results.

3.2.4.6 Type of stimuli

There are also recommendations and guidelines for the selection of test stimuli. Two different types of test stimuli are typically found in IQ experiments; pictorial images and research images (i.e. test targets) [143]. Pictorial images are the most commonly used test stimuli, since observers are confident in judging them. However, they must be chosen with care since the content might influence the results. Pictorial images are often difficult to use with instruments since consistent measurements are difficult to produce [143]. Research images are artificially created test images, often made to test a specific problem. They have the advantage over pictorial images that they are content free and often have areas that can be read by measuring instruments. One of the best known research image is the Macbeth ColorChecker Color Rendition Chart (Figure 3.7).

There are several guidelines regarding the characteristics of test stimuli. They should for example include several levels (low, medium, and high) of several different characteristics (such as tonal distribution, detail level, and saturation) [143]. It is recommended to test a broad range of images to reveal different quality issues [176]. For a complete overview of characteristics of test images we refer the reader to Field [143] and CIE [80].

3.2.4.7 Standard test images

CIE recommendation for gamut mapping  The CIE guidelines for gamut mapping [80] gives one image as obligatory for the evaluation of gamut mapping; the ski image (Figure 3.34). In addition to this image ten graphics (Canon Development Americas computer graphics images) and three other pictorial images (Sony sRGB standard images) are recommended.

ISO 12640 test images  ISO 12640 defines different sets of test images for evaluation of different processes.
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.35: The eight natural images found in ISO 12640-1 [195].

Figure 3.36: The eight natural images found in ISO 12640-2 [199].

The image set in ISO 12640-1 [195] consists of 8 natural (Figure 3.35) and 10 synthetic images. The natural images include skin tones, images with detail in the highlights or shadows, neutral colors, memory colors, and colors in areas that are difficult to reproduce. The synthetic images include resolution charts and uniform vignettes. The image set was developed for comparison of color output systems such as printing, proofing, and color facsimile, and therefore they are in CMYK format.

ISO 12640-2 [199] defines XYZ/sRGB standard color image data. The image set consists of 15 color images, encoded as both 16-bit CIEXYZ and 8-bit RGB digital data, for the evaluation of quality changes. The set has eight natural images (Figure 3.36) and seven synthetic images, spanning a range of different characteristics. This set is optimized for viewing on a reference sRGB display in the reference sRGB viewing environment, and relative to CIE standard illuminant D65 (Description of viewing illumination is found in Section 3.2.4.9).

ISO 12640-3 [209] provides a test image data set with a large color gamut related to illuminant D50. The 18 test images, 8 natural (Figure 3.37) and 10 synthetic, are encoded as 16-bit CIELAB digital data. The natural images are 16 bits per channel, while the synthetic images are 8 bits per channel.

ISO 12640-4 [212] specifies a standard set of wide gamut display-referred color images. These are encoded as 16-bit Adobe RGB digital data. These images compliment the existing images of ISO 12640-2 which are based on the sRGB display gamut. These test images have a larger color gamut than sRGB, and these images will require much less aggressive color re-rendering going to print than sRGB encoded images.

Canon Development Americas computer graphics images  Canon Development Americas proposed ten computer graphics (Figure 3.38), which are recommended by CIE for the evaluation of gamut mapping algorithms.

Sony sRGB standard images  The Sony standard images [420] are three photographs provided by Sony for CIE [80] for the evaluation of gamut mapping (Figure 3.39), being a part

Figure 3.37: The eight natural images found in ISO 12640-3 [209].
of the recommended images from CIE. The set consists of two studio scenes, representing a portrait and a party image, and an outdoor image with a picnic theme.

**Kodak lossless true color image suite** Kodak proposed the Kodak lossless true color image suite [241], which is a set of 24 images (22 outdoor and two studio images) as seen in Figure 3.40. The photographs originally appeared on a Kodak Photo CD. This image suite has become a very popular suite for evaluating different aspects of imaging. However, these images were made when digital images was a new concept. Therefore, their quality is limited, and they are probably not comparable to current digital photos.

**DigiQ** Halonen et al. [167] proposed the DigiQ image suite, which consists of three test images for print quality evaluation (Figure 3.41). In designing the images, aspects taken into consideration included a recognizable theme, memory colors and shapes, surface materials, detail information and salience of objects. The images in the suite are 16-bit TIFF images in Adobe RGB format, with a resolution of 360 DPI and a print size of 100 × 150 mm.

### 3.2.4.8 Stimuli labeling

Labeling the stimuli is important for registration of results. However, if the label is visible to the observer they might ”solve” the code and respond according to the code and not to the task. Therefore, placing code visible to the observer is usually not considered good practice [116, 117]. A common practice is to put the code on the back side of the stimuli. If the code needs to be visible to the observer, it should be obscure. One option is for example to use a four digit code of random numbers (1000 possible combinations), or a two letter code of random characters (625 possible combinations).
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY

Figure 3.40: Kodak lossless true color image suite.

Figure 3.41: DigiQ image suite.
3.2.4.9 Viewing conditions

**Controlled and uncontrolled environments** Regardless of the experimental type the experimental condition can be either controlled or uncontrolled. Controlled experiments are carried out in a laboratory where the viewing conditions meet standards (such as described by CIE [80]), while uncontrolled experiments can be carried out in the field or on the web. Web-based experiments get more popular, since they are easy to carry out, free, time-saving, and the number of observers is almost unlimited. The most important reason not to carry out uncontrolled experiments has been that the environments are not standardized, however research have shown small differences between controlled and uncontrolled experiments [300, 422, 503].

**Viewing distance** The perceived quality of an image is related to the distance at which the observers view it [116]. A change in the viewing distance will change the information that our HVS can perceive or detect, and will therefore also change the quality. Thus the viewing distance should be kept constant when conducting experiments, in order to have similar conditions for all observers.

In experiments where the viewing distance is not controlled, the variability of the results might increase and the results spread. In specific experimental methods, such as the quality ruler, a headrest bar is used to fix the distance from the stimuli to the observer (Figure 3.33).

**Viewing illumination** The illumination under which the stimulus is viewed has an impact on the perceived quality. The intensity of the illumination (illuminance), which is measured in lumens per square meter or lux, influences the responses from human observers. For emissive displays, for example a Liquid Crystal Display (LCD) monitor, the amount of light that the screen produces is measured in candelas per square meter. Controlling the illuminance is important since an increase in illuminance will increase the colorfulness, better known as the Hunt effect [395], it has also been shown that the perceived hue can change with a change in luminance.

Another important aspect is the spectral power distribution of the illumination, also known as the Correlated Color Temperature (CCT). Since the color appearance changes with the spectral distribution of the illumination, it is important to include this aspect in the experimental design. This is also an important factor when comparing results from different experiments. The most common light sources are standardized by the CIE; the light sources are given a code, for example D5000 (D50 short) where the D stands for daylight and the four digit number indicates the CCT. When conducting experiments involving both monitors and printed stimuli the CIE [80] recommends using a monitor with a D65 white point while the prints should be viewed under D50.

The geometry of the illumination could have an impact on the results, for example when viewing glossy prints. It is recommended to illuminate the test stimuli from an angle of 45° from the normal to the sample, and view the sample normal to the surface [116]. This is commonly achieved by the use of a normalized viewing booth.

The viewing conditions are important to control, and the CIE [80] gives guidelines for different conditions depending on the task at hand. For printed samples and transparencies we refer the reader to ISO 3664:2009 [211] and on experiment design to ISO 20462-1:2004 [202].
3.2.4.10 Instructions

The instructions given to the observers are among of the most important aspects when designing an experiment. In order to obtain useful and meaningful results, the observers need to be given clear instructions. There are several considerations to make when writing the instructions, such as what are the observers going to judge, in what context, what are the criteria for the judgment, and are there any definitions needed for the observers to carry out their task? First the aspect of what should be judged needs to be decided upon, this could be overall IQ or a specific attribute (for example sharpness or a type of artifact). Very often it is important to give a definition together with the question in order for all observers to judge the images based on the same foundation. In some cases visual references are useful, they can be used as anchors. It is important that the question asked is precise and clear, and does not leave any doubt or inconsistencies. The context in which the experiment is carried out is also an important aspect, for example if you are to judge IQ, in the context of office documents, official letters, or a private context.

3.3 Summary

In this chapter we have presented the common methods to measure IQ, both objective and subjective. First objective methods were presented focusing on instruments, such as spectrophotometers and colorimeters. Additionally, quality assessment in the industry with relevant standards and test image were introduced. Further, the most common subjective assessment methods were explained in detail together with experimental aspects, such as number of observers, test images, and instructions.
EXISTING METHODS FOR THE EVALUATION OF PRINT QUALITY
PART II

IMAGE QUALITY METRICS
4 Existing image quality metrics

An IQ metric is an objective mathematical way to calculate quality without human observers. Most of the IQ metrics in the literature have the same structure, as shown in Figure 4.1. The images (original and reproduction, or for some metrics just the reproduction) go through color transformations from a standard color space (such as sRGB) to a more suitable color space to simulate the HVS (such as an opponent color space). In this color space characteristics of the HVS is incorporated. The next step consists of calculating the quality, usually pixel-wise that results in a quality value for every pixel in the image. Most metrics also do pooling, where the quality values from the previous step are reduced to a lower number of values (usually one quality value).

There are fundamentally three different types of IQ metrics; no-reference (also referred to as blind-reference), reduced-reference, and full-reference (Figure 4.2). For the first type only the reproduction is available, and the calculation of quality is based only on the reproduction without use of the original (Figure 4.2(a)). In reduced-reference some information of the reproduction and the original is used in the calculation of quality (Figure 4.2(b)), such as a sub-signal or the histogram. In the last group both the complete reference and the reproduction are available (Figure 4.2(c)). The two first types of IQ metrics are considered as more difficult tasks than full-reference [464].

Many IQ metrics have been proposed in the literature, a brief summary of more than 100 metrics was given by Pedersen and Hardeberg [354]. These metrics stem from different ideas, and they have been made for different purposes, such as to quantify a distortion, to benchmark, to monitor quality, to optimize a process, or to indicate problem areas. Because of this it is important to categorize and identify their scopes of use, but also to evaluate their performance. Existing surveys, such as the one by Wang and Bovik [456] mostly focused on grayscale IQ.
metrics, and the one by Avcibas et al. [14] covers simple statistical metrics. In this chapter we carry out a survey of color and grayscale IQ metrics. Our goal is to classify IQ metrics into separate groups, and to evaluate their correspondence with human observers. Such a survey can be used to select the most appropriate IQ metric for a given problem or distortion. It will also provide a better understanding of the state of the art of IQ metrics, which can be used to improve or develop new IQ metrics better correlated with perception.

We limit our survey to full-reference IQ metrics, where both the complete original and reproduction are used in the calculation of quality, since the original is often available. Additionally, more work has been carried out on full-reference IQ metrics than reduced-reference and no-reference.

In the literature many different terms have been used, such as IQ, image difference, image fidelity, and image similarity. IQ metrics refer to metrics with the goal of predicting IQ, while image difference metrics aims at predicting the perceived magnitude of differences between an original and a reproduction [221] without considering how the differences contribute to perceived IQ, image fidelity metrics predict visibility of image reproduction errors [500], and image similarity metrics measures how one image matches the other [378]. If possible we will use the general term IQ, which in this work also covers the other terms since prediction of differences are seen as one of the steps towards predicting IQ [224]. However, it does not mean that there is a relationship between difference and quality [407], since a difference in some cases can increase quality. If the original image is considered to be ideal any differences in the reproduction will lead to a loss of quality [221], making metrics based on distances suitable to measure IQ as well.

In addition, we mainly use the term metric as a general term even though not all metrics
4.1 Classification of image quality metrics

Classification of existing IQ metrics is a good starting point for selecting the best metrics for a given setting, such as to evaluate print quality. Without a classification of metrics one does not have an organized way to telling the difference between the different metrics. Such an organization illuminates the relationship between metrics, and thereby increasing the understanding of them. Classification helps in the decision making of what metric to use, but also in the development of new and improved metrics.

4.1.1 Existing classification of image quality metrics

Since IQ metrics have been proposed based on different approaches they can be divided into different groups. These groups usually reflect different aspects of the metrics, such as their intended use or construction. Several different researchers have classified metrics into groups, even though it can be difficult to find sharp boundaries between the numerous IQ metrics in the literature.

Avci̇bas et al. [14] divided IQ metrics into six groups based on the information they use:

- Pixel difference-based measures such as mean square distortion.
- Correlation-based measures, that is, correlation of pixel values, or of the vector angular directions.
- Edge-based measures, that is, measure of the displacement of edge positions or their consistency across resolution levels.
- Spectral distance-based measures, that is, the Fourier magnitude and/or phase spectral discrepancy on a block basis.
- Context-based measures are based on various functionals of the multidimensional context probability.
- HVS-based measures are measures that are either based on the HVS-weighted spectral distortion measures or (dis)similarity criteria used in image base browsing functions.

Callet and Barba [59] divided IQ metrics into two distinct groups:

- Those who use a HVS model for low level perception, such as sub-band decomposition and masking effect.
- Those who use little information about the HVS for error representation, and push the effort on prior knowledge on introduced distortion.

The authors comment that IQ metrics belonging to the last group fail to be robust since they are specialized.

Wang and Bovik [456] classified IQ metrics based on three criteria:
EXISTING IMAGE QUALITY METRICS

- Full-reference, no-reference and reduced reference metrics.
- General-purpose and application specific metrics.
- Bottom-up and top-down metrics.

These categories are based on information about the original image, knowledge about the distortion process, and knowledge about the HVS. The authors state that it is not possible to draw sharp boundaries between the different groups, but that a crisp classification of IQ metrics can be helpful.

Chandler and Hemami [70] divided IQ metrics into three different groups:

- Mathematically convenient metrics, which only operate on the intensity of the distortion.
- Metrics based on near-threshold psychophysics, these metrics take into account the visual detectability of the distortions.
- Metrics based on overarching principles such as structural or information extraction.

Thung and Raveendran [436] divided full-reference IQ metrics into three groups similarly to Chandler and Hemami [70]:

- Mathematical metrics.
- HVS-based metrics.
- Others.

Recently, Seshadrinathan and Bovik [391] divided techniques for image and video quality into three main categories:

- HVS-based approaches.
- Structural approaches.
- Information theoretical approaches.

The authors also mention that IQ assessment can be broadly classified as bottom-up and top-down approaches, where bottom-up approaches model the HVS and top-down approaches characterize specific features of the HVS.

4.1.2 Proposal for classification of image quality metrics

As seen above there are many different ways to group IQ metrics. In order to present the various approaches we have divided the IQ metrics into four groups:

- Mathematically based metrics, which operate only on the intensity of the distortions. These metrics are usually simple, such as the Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR).
- Low-level based metrics, which take into account the visibility of the distortions using for example Contrast Sensitivity Functions (CSFs), such as the Spatial-CIELAB (S-CIELAB) [499].
• High-level based metrics, which quantify quality based on the idea that our HVS is adapted to extract information or structures from the image. The Structural SIMilarity (SSIM) [458], which is based on structural content, or the Visual Image Fidelity (VIF) [396], which is based on scene statistics, are examples of metrics in this group.

• Other metrics, which are either based on other strategies or combine two or more of the above groups. One example is the Visual Signal-to-Noise Ratio (VSNR) [70], which takes into account both low- and mid-level visual properties, and the final stage incorporates a mathematically based metric.

The three first groups are similar to the classification by Chandler and Hemami [70]. However, there are metrics that do not fit directly into one or more of three first groups, requiring a fourth classification group.

4.2 Survey of existing image quality metrics

For each of these groups we will present some selected metrics. Since an impressive number of IQ metrics have been proposed [340, 354] the selection of metrics will be based on standards, existing literature, frequency of use, and novelty. For an extensive overview of existing metrics we refer the reader to Appendix A.

In addition to introducing a selection of metrics, we will also identify differences between the metrics. For this a test target has been adopted. The target (Figure 4.3) has been developed by Halonen et al. [167] to evaluate print quality, and is a part of the image suite from DigiQ (Section 3.2.4.7). In this test target ten specific color patches from the GretagMacbeth ColorChecker were incorporated as objects in the scene; blue, green, red, yellow, magenta, cyan, orange, and three neutral grays. This target should show differences between a wide variety of the metrics, since it contains different frequencies, colors and objects. Since we are applying full-reference IQ metrics, a reproduction is needed. This has been obtained by taking the original image, converting it to sRGB (which is the original used in the comparison), and further applying an ICC profile (EuroScale Uncoated V2) with the relative colorimetric rendering intent, then converting the image to sRGB, using Adobe Photoshop CS3 for all procedures. This should result in many different quality changes, which will provoke differences in the metrics.

4.2.1 Mathematically based metrics

The first group of metrics, mathematically based ones, has been very popular probably due to their easy implementation, and they are convenient to use for optimization. These metrics usually only work on the intensity of the distortion $E$:

$$E(x,y) = I_O(x,y) - I_R(x,y),$$

(4.1)

where $I_O$ is the original image, $I_R$ is the reproduction, $x$ and $y$ indicate the pixel position.
EXISTING IMAGE QUALITY METRICS

Figure 4.3: Test target from Halonen et al. [167] to differentiate between IQ metrics. The original image has been processed by applying an ICC profile with the relative colorimetric rendering intent.

4.2.1.1 Mean squared error

MSE is a mathematically based metric; it calculates the cumulative squared error between the original image and the distorted image. Most of the metrics in this group are strict metrics, that is, where (\(\rho(\text{IO}(x), \text{IR}(y))\)) is essentially an abstract distance, with the following properties: \(\rho(\text{IO}(x), \text{IR}(y)) = 0\) if \(\text{IO}(x) = \text{IR}(y)\), symmetry, triangle inequality, and non-negativity. MSE is given as:

\[
\text{MSE} = \frac{1}{MN} \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} [E(x,y)]^2,
\]

(4.2)

where \(x\) and \(y\) indicate the pixel position, \(M\) and \(N\) are the image width and height. Several others of these metrics have been proposed as well, such as the PSNR and Root Mean Square (RMS). The MSE and other metrics have been used due to its easy calculation and analytical tractability [435]. These simple mathematical models are usually not well correlated with perceived IQ [70, 456]. Still, they have been of influence to other and more advanced metrics, such as the PSNR-HVS-M [366] that takes into account contrast masking and CSFs to model the HVS.

4.2.1.2 \(\Delta E_{ab}^*\)

Metrics measuring color difference also belong to the group of mathematically based metrics. One of the most commonly used color difference formula stems from the CIELAB color space specification published by the CIE [2004], with the idea of a perceptually uniform color-space. In such a color space it is easy to calculate the distance between two colors, by using the Euclidean distance. In engineering, this is known as the RMS value, which is the square root of the MSE. The \(\Delta E_{ab}^*\) takes as input a sample color with CIELAB values \(L_r^*, a_r^*, b_r^*\) and a reference color \(L_o^*, a_o^*, b_o^*\). The distance between the two color samples is given by

\[
\Delta E_{ab}^* = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2},
\]

(4.3)
EXISTING IMAGE QUALITY METRICS

where $\Delta L^* = L_o^* - L_r^*$, $\Delta a^* = a_o^* - a_r^*$ and $\Delta b^* = b_o^* - b_r^*$.

The CIELAB formula has served as a satisfactory tool for measuring perceptual difference between uniform color patches. The HVS is not as sensitive to color differences in fine details as compared to large patches, yet the CIELAB color formula will predict the same visual difference between the two cases since it does not have a spatial variable [500]. Because of the popularity of the $\Delta E_{ab}^*$ it has also been commonly used to measure natural images, where the color difference of each pixel of the image is calculated. The mean of these differences is usually the overall indicator of the difference between the original and the reproduction:

$$\overline{\Delta E_{ab}^*} = \frac{1}{MN} \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} \Delta E_{ab}^*(x,y),$$

(4.4)

Other measures of the $\Delta E_{ab}^*$ can be the minimum value or the maximum value in the computed difference. Because of the widespread use [21, 41, 78, 169, 233, 234, 339, 350, 356, 382, 419, 462, 498, 500], its position in the industry, and because it is often used as a reference metric we consider the $\Delta E_{ab}^*$ to be included in an overview of IQ metrics. Extensions of the $\Delta E_{ab}^*$ have been proposed when it became apparent that it had problems, especially in the blue regions. First the CIE proposed the $\Delta E_{94}^*$ [83] and later the $\Delta E_{00}^*$ [275]. These are of increasing complexity compared to the $\Delta E_{ab}^*$, and therefore many still use the $\Delta E_{ab}^*$.

4.2.1.3 $\Delta E_E$

Another very interesting color difference formula was proposed by Oleari et al. [329] for small and medium color differences in the log-compressed OSA-UCS (Optical Society of America’s Committee on Uniform Color Scales) space. Because of its different origin and promising results [411, 412] we include it in this overview of IQ metrics. The formula is defined as following:

$$\Delta E_E = \sqrt{(\Delta L_E)^2 + (\Delta G_E)^2 + (\Delta J_E)^2},$$

(4.5)

where

$$L_E = \frac{1}{b_L} \ln \left[ 1 + \frac{b_L}{a_L} (10L_{OSA}) \right],$$

(4.6)

with $a_L = 2.890, b_L = 0.015$,

$$G_E = -C_E \cos(h),$$

(4.7)

$$J_E = C_E \sin(h),$$

(4.8)
EXISTING IMAGE QUALITY METRICS

with

\[ h = \arctan \left( \frac{-J}{G} \right), \]  
\[ C_E = \frac{1}{b_C} \ln \left[ 1 + \frac{b_C}{a_C} (10C_{OSA}) \right], \]

with \( a_C = 1.256, b_C = 0.050, \)

\[ C_{OSA} = \sqrt{G^2 + J^2}. \]

\( \Delta E_E \) is statistically equivalent to \( \Delta E_{00}^* \) in the prediction of many available empirical datasets. However, \( \Delta E_E \) is the simplest formula providing relationships with visual processing. These analyses hold true for CIE 1964 Supplementary Standard Observer and D65 illuminant. This formula can also be used in the same way as \( \Delta E_{ab}^* \) to measure the color difference of natural images by taking the average color differences over the entire image.

4.2.1.4 Difference between the metrics

The metrics introduced above are based on different ideas and they have different origins, because of this they will be different in the way they work. In order to understand the fundamental differences between the metrics we will investigate their similarities and differences.

Since the MSE has been shown to correlate poorly with perceived IQ we do not consider it in this part, where we will look at comparison between the \( \Delta E_{ab}^* \) and \( \Delta E_E \). The difference between these two is in the color space, since both are based on the Euclidean distance between two color values. Both color spaces are opponent color spaces, where both have a lightness, red-green, and blue-yellow axis. The OSA-UCS color space has a better blue linearity than the CIELAB color space [299]. The OSA-UCS system also has the unique advantage of equal perceptual spacing among the color samples [286].

Both the \( \Delta E_{ab}^* \) and \( \Delta E_E \) have been computed for the test image (Figure 4.3), and the results can be seen in Figure 4.4. Both results have been normalized with the maximum value, making it easier to compare the results. The first apparent difference is the difference in the blue vase, which is due to known issues in the blue regions of the CIELAB color space [50]. It is also interesting to notice the differences in the dark regions, this could be because the OSA-UCS color space, which \( \Delta E_E \) is based on, compensate for the reference background lightness [337]. There is also a difference in the skin tones, where \( \Delta E_E \) has almost no difference between the original and the reproduction, while \( \Delta E_{ab}^* \) indicates a larger difference.

It is important to state that we do not evaluate whether the the differences are correlated with perceived difference or quality, we merely show the difference between the metrics. Evaluation of the IQ metrics can be found in Chapter 7.

4.2.2 Low-level based metrics

Metrics classified as low-level based metrics simulates the low level features of the HVS, such as CSFs or masking. However, most of these metrics use a mathematically based metric, for example one of the metrics introduced above, in the final stage to calculate quality.
**EXISTING IMAGE QUALITY METRICS**

Figure 4.4: Comparison of color difference formulas, a) $\Delta E^*_{ab}$ and b) $\Delta E$. The difference is calculated between an original image (Figure 4.3(a)) and a reproduction (Figure 4.3(b)). White indicates highest difference between the original and the modified version of the test target, black a low difference.

### 4.2.2.1 S-CIELAB

When it became apparent that the $\Delta E^*_{ab}$ was not correlated with perceived image difference for natural images, Zhang and Wandell [499] proposed a spatial extension of the formula (Figure 4.5). This metric should fulfill two goals, a spatial filtering to simulate the blurring of the HVS and a consistency with the basic CIELAB calculation for large uniform areas.

![S-CIELAB workflow](image)

Figure 4.5: S-CIELAB workflow. The spatial filtering is done in the opponent color space, and the color difference is calculated using $\Delta E^*_{ab}$.

The image goes through color space transformations, first the RGB image is transformed into CIEXYZ and then further into the opponent color space ($O_1$, $O_2$, and $O_3$):

\[
O_1 = 0.279X + 0.72Y - 0.107Z,
\]
\[
O_2 = -0.449X + 0.29Y - 0.077Z,
\]
\[
O_3 = 0.086X - 0.59Y + 0.501Z.
\]

Now the image contains a channel with the luminance information ($O_1$), one with the red-green information ($O_2$), and one with blue-yellow information ($O_3$). Then a spatial filter is applied, where data in each channel is filtered by a 2-dimensional separable spatial kernel:

\[
f = k \sum_i w_i E_i,
\]

(4.12)
where

\[ E_i = k_i e^{\frac{-(x^2 + y^2)}{\sigma_i^2}}, \] (4.13)

and \( k_i \) normalize \( E_i \) such that the filter sums to one. The parameters \( w_i \) and \( \sigma_i \) are different for the color planes as seen in Table 4.1. \( k \) is a scale factor, which normalizes each color plane so that its two-dimensional kernel \( f \) sums to one.

Table 4.1: The parameters used for the spatial filtering in S-CIELAB, where \( w_i \) is the weight of the plane and \( \sigma_i \) is the spread in degrees of visual angle.

<table>
<thead>
<tr>
<th>Plane</th>
<th>Weights ( w_i )</th>
<th>Spreads ( \sigma_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminance</td>
<td>0.921</td>
<td>0.0283</td>
</tr>
<tr>
<td></td>
<td>0.105</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>-0.108</td>
<td>4.336</td>
</tr>
<tr>
<td>Red-Green</td>
<td>0.531</td>
<td>0.0392</td>
</tr>
<tr>
<td></td>
<td>0.330</td>
<td>0.494</td>
</tr>
<tr>
<td>Blue-Yellow</td>
<td>0.488</td>
<td>0.0536</td>
</tr>
<tr>
<td></td>
<td>0.371</td>
<td>0.386</td>
</tr>
</tbody>
</table>

Finally the filtered image is transformed into CIE-XYZ, and this representation is further transformed into the CIELAB color space where the \( \Delta E_{ab}^* \) is applied to calculate color differences. The final result, S-CIELAB\(^1\) representation, consists of a color difference for each pixel, these values are usually averaged to achieve one value for overall image difference.

The S-CIELAB metric was originally designed for halftoned images, but it has also been used to calculate overall IQ for a wide variety of distortions. It is often used as a reference metric due to this simple implementation and since it formed a new direction of IQ metrics, in addition it is often used by researchers [5, 6, 18, 21, 41, 45, 138, 173, 220, 233, 339, 350, 351, 356, 412, 490, 498, 500]. It has also lead to many other IQ metrics, which use the same framework as S-CIELAB, such as \( S_{DOG} \)-CIELAB [5, 6] (Appendix E) where the spatial filtering was performed with the difference-of-gaussians, and S-CIELAB\(_J \) [224] where the CSF was modified in terms of adjusting the peak sensitivity to correlate better with complex stimuli. Because of this it is natural to include S-CIELAB in an overview of IQ metrics.

### 4.2.2.2 S-DEE

\( \Delta E_E \) also has a spatial extension, Spatial-DEE (S-DEE), proposed by Simone et al. [412]. This metric follows the S-CIELAB framework, but the color difference is calculated with \( \Delta E_E \) instead of \( \Delta E_{ab}^* \). The spatial filtering with CSFs is also a bit different from the original S-CIELAB, since the S-DEE incorporates CSFs from Johnson and Fairchild [224], where the filters are implemented in the frequency domain rather than in the spatial domain as in S-CIELAB.

The luminance filter is a three parameter exponential function, based on research by Movshon and Kiorpes [308].

\[ CSF_{lum}(p) = a \cdot p^c \cdot e^{-b \cdot p}, \] (4.14)

---

\(^1\)16/02/2011: The implementation of S-CIELAB as described here is available at http://white. standford.edu/~brian/scielab/scielab.html
where \( a = 75 \), \( b = 0.22 \), \( c = 0.78 \), and \( p \) is represented as Cycles Per Degree (CPD).

The luminance CSF is normalized so that the DC modulation is set to 1.0, by doing this the luminance shift is minimized. This will also enhance any image differences where the HVS is most sensitive to them [223]. For the chrominance CSF, a sum of two Gaussian functions are used.

\[
CSF_{\text{chroma}}(p) = a_1 \cdot e^{-b_1 \cdot p^2} + a_2 \cdot e^{-b_2 \cdot p^2},
\]

(4.15)

where different parameters for \( a_1, a_2, b_1, b_2, c_1, \) and \( c_2 \) have been used as seen in Table 4.2.

**Table 4.2: The parameters used in S-DEE for the spatial filtering in the frequency domain of the chrominance channels.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Red-Green</th>
<th>Blue-Yellow</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 )</td>
<td>109.14130</td>
<td>7.032845</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>-0.00038</td>
<td>-0.000004</td>
</tr>
<tr>
<td>( c_1 )</td>
<td>3.42436</td>
<td>4.258205</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>93.59711</td>
<td>40.690950</td>
</tr>
<tr>
<td>( b_2 )</td>
<td>-0.00367</td>
<td>-0.103909</td>
</tr>
<tr>
<td>( c_2 )</td>
<td>2.16771</td>
<td>1.648658</td>
</tr>
</tbody>
</table>

The filters are applied in the \( 0_10203 \) color space as for S-CIELAB, but the final step consists of applying the \( \Delta EE \) rather than the \( \Delta E^* \) to get the color differences. The overall image difference is achieved by averaging the results over the image.

Since the \( \Delta EE \) has promising results, so has the S-DEE [412], and because it challenges the standard calculation of the CIELAB color difference it should be included in this survey of IQ metrics. A variant of the S-DEE, \( S_{\text{DOG}} \)-DEE, was proposed by Ajagamelle et al. [5, 6] (Appendix E), where the spatial filter is carried out by the difference-of-gaussians.

### 4.2.2.3 Adaptive image difference

Wang and Hardeberg [459] proposed an Adaptive Bilateral Filter (ABF) for image difference metrics. The filter blurs the image, but preserves edges, which is not the case when using CSFs. The bilateral filter is defined as:

\[
h(x) = k^{-1}(x) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\varepsilon)c(\varepsilon, x)s(f(\varepsilon), f(x))d\varepsilon,
\]

(4.16)

where

\[
k(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(\varepsilon, x)s(f(\varepsilon), f(x))d\varepsilon,
\]

(4.17)

and where the function \( c(\varepsilon, x) \) measures the geometric closeness between the neighborhood center \( x \) and a nearby point \( \varepsilon \):

\[
c(\varepsilon, x) = \exp\left(-\frac{(\varepsilon - x)^2}{2\sigma_d^2}\right).
\]

(4.18)

The function \( s(\varepsilon, x) \) measures the photometric similarity between the neighborhood center \( x \)
EXISTING IMAGE QUALITY METRICS

and a nearby point ε:
\[ s(\epsilon, x) = \exp \left( -\frac{(f(\epsilon) - f(x))^2}{2\sigma_r^2} \right). \] (4.19)

\( \sigma_d \) is the geometric spread determined by the viewing conditions in pixels per degree:
\[ \sigma_d = \frac{n/2}{180/\pi \cdot \tan^{-1}(l/(2m))}, \] (4.20)

where \( n \) is the width of the image, \( l \) is the physical length in meters, and \( m \) is the viewing distance in meter. \( \sigma_r \) is the range spread and is determined with image entropy:
\[ \sigma_r = K/E, \] (4.21)

where \( K \) is a constant to rescale the entropy into an optimized value and
\[ E = -\sum_i p_i \log(p_i), \] (4.22)

where \( p_i \) is the histogram of the pixel intensity values of an image.

The filtering is performed in the CIELAB color space, and the color difference formula used was \( \Delta E_{ab}^* \), because of this the metric can be said to follow the same overlaying framework as S-CIELAB.

Because of the new type of filtering preserving edges, which has shown to produce good results for amongst others gamut mapping [40], this metric is interesting to compare with, and it shows promising results compared to other metrics [459].

4.2.2.4 Comparison of selected metrics within the group

The difference between the metrics in this group is mainly in the filtering stage and in the quality calculation. Since the quality calculation is carried out with a color difference formula (\( \Delta E_{ab}^* \) or \( \Delta E_E \)), we will focus on the filtering. The S-CIELAB and S-DEE use CSFs, where they differ in terms of the normalization of the filters. On the other side, ABF uses a bilateral filter preserving edges.

Figure 4.6 shows the difference of an image filtered with the three different metrics. We can see that S-CIELAB (the second image from the left) is slightly darker than the original (left image), which is caused by the CSF and has been reported earlier [224]. The image filtered with the method from S-DEE does show such a lightness shift since the DC component of the CSF is normalized to one. We can also notice the blurring of the images from S-CIELAB and S-DEE, and the loss of details. The S-DEE has a CSF which enhances the frequencies where the HVS is most sensitive, and therefore it will in some cases preserve more details than S-CIELAB. The ABF on the right has a softer "look" compared to the other metrics, most noticeable is the preservation of the hard edges, which is the advantage of the bilateral filter.
EXISTING IMAGE QUALITY METRICS

Figure 4.6: Comparison between the different spatial filtering methods. From left to right we can see the original image, the image filtered with S-CIELAB, the image filtered with S-DEE, and the image filtered with ABF. All images have been computed for the same viewing conditions. In S-CIELAB we perceive a lightness shift due to the fact that the luminance filter is not normalized, in S-DEE the luminance filter is normalized which corrects this issue, the ABF has a softer look since it smooths on each side of an edge using an edge preserving bilateral filter.

4.2.3 High level based metrics

High-level based metrics quantify quality based on the idea that our HVS is adapted to extract information or structures from the image.

4.2.3.1 SSIM

The SSIM index proposed by Wang et al. [458] attempts to quantify the visible difference between a distorted image and a reference image. This index is based on the Universal Image Quality (UIQ) index [455]. The algorithm defines the structural information in an image as those attributes that represent the structure of the objects in the scene, independent of the average luminance and contrast. The index is based on a combination of luminance, contrast, and structure comparison. The comparisons are done for local windows in the image, the overall IQ is the mean of all these local windows. The SSIM is specified as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

(4.23)

where $\mu$ is the mean intensity for signals $x$ and $y$, and $\sigma$ is the standard deviation of the signals $x$ and $y$. $C$ is a constant defined as

$$C_1 = (K_1L)^2,$$

(4.24)

where $L$ is the dynamic range of the image, and $K_1 \ll 1$. $C_2$ is similar to $C_1$ and is defined as:

$$C_2 = (K_2L)^2,$$

(4.25)
where $K_2 \ll 1$. These constants are used to stabilize the division of the denominator. SSIM is then computed for the entire image as:

$$\text{MSSIM}(X,Y) = \frac{1}{W} \sum_{j=1}^{W} \text{SSIM}(x_j, y_j),$$  \hspace{1cm} (4.26)

where $X$ and $Y$ are the reference and the distorted images, $x_j$ and $y_j$ are image content in local window $j$, and $W$ indicates the total number of local windows. Figure 4.7 shows the SSIM flowchart, where the input is the original signal (image) and the distorted signal (image).

![SSIM flowchart](http://www.ece.uwaterloo.ca/~z70wang/research/ssim/)

**Figure 4.7: SSIM flowchart, signal x and y goes through a luminance and contrast measurement before comparison of luminance, contrast and structure. A combination of these results in the final similarity measure. Picture from Wang et al. [458].**

The SSIM has received a lot of attention since its introduction, gone through extensive evaluation, and it has influenced a number of other metrics, such as the color version SSIM-IPT by Bonnier et al. [41] and the color version of UIQ by Toet and Lucassen [439]. Because of this and the interesting idea of using structural information to measure IQ we include SSIM in this survey.$^2$

### 4.2.3.2 Visual information fidelity

Sheikh and Bovik [396] proposed the Visual Information Fidelity (VIF) criterion, which is an extension of the Information Fidelity Criterion (IFC) by the same authors [397]. It quantifies the Shannon information present in the reproduction relative to the information present in the original. The natural scene model used is a Gaussian Scale Mixture model in the wavelet domain, and as a HVS model they use an additive white Gaussian noise model.

The reference image is modeled by a Gaussian Scale Mixture in the wavelet domain. $c$ is a collection of $M$ neighboring wavelet coefficients from a local patch in a subband. Then $c$ is modeled as $c = \sqrt{z} u$, where $u$ is a zero-mean Gaussian vector and $\sqrt{z}$ is an independent scalar random variable. The VIF assumes that the distortion of the image can be described locally

---

$^2$16/02/2011: The implementation of SSIM can be found at [http://www.ece.uwaterloo.ca/~z70wang/research/ssim/](http://www.ece.uwaterloo.ca/~z70wang/research/ssim/)
as a combination of a uniform wavelet domain energy attenuation with added independent additive noise. So that visual distortion is modeled as a stationary, zero-mean, additive white Gaussian noise process in the wavelet domain: \( e = c + n \) and \( f = d + n \), where \( e \) and \( f \) are the random coefficient vectors for the same wavelet subband in the perceived original and perceived distorted image. \( c \) and \( d \) are random vectors from the same location in the same subband for the original and distorted image. \( n \) denotes the independent white Gaussian noise with the covariance matrix \( C_n = \sigma^2_n I \), where \( \sigma^2_n \) is a HVS model parameter (variance of the internal neuron noise) and \( I \) denotes the set of spatial indices for the random fields. The VIF is calculated as the ratio of the summed mutual information in the subbands, which can be written as following:

\[
\text{VIF} = \frac{I(C; F|z)}{I(C; E|z)} = \frac{\sum_{i=1}^{N} I(c_i; f_i|z_i)}{\sum_{i=1}^{N} I(c_i; e_i|z_i)},
\]

where \( i \) is the index of local coefficients patches, including all subbands.

The VIF\(^3\) criterion has shown to perform well compared to other state of the art metrics [253, 396, 398], and combined with the use of statistical information it is interesting to compare it against more traditional metrics.

### 4.2.3.3 Comparison of selected metrics within the group

It is difficult to compare the metrics within this group since SSIM produces a map and VIF produces a single value. To be able to compare the SSIM against the other metrics, such as \( \Delta E_E \), we show the resulting map from the target (Figure 4.8). Since SSIM is a grayscale metric the images have been converted from color to grayscale\(^4\). SSIM show the biggest difference in the dark regions of the image (Figure 4.8). It is also interesting to notice the similarity between the \( \Delta E_E \) (Figure 4.4(b)) and SSIM (Figure 4.8). To verify the similarities between these two difference maps a 2-D correlation coefficient has been used:

\[
r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)}},
\]

(4.27)

where \( \overline{A} \) is the mean of one map, and \( \overline{B} \) is the mean for the other map, respectively. This method has previously been used to compare results from eye tracking maps [17, 339], and to compare an IQ metric to the regions marked by observers [61]. The correlation coefficient between these maps gives a value of -0.75 (minus since the \( \Delta E_E \) has the highest difference as max, while SSIM has the highest difference as min. \( m = 2126 \) and \( n = 1417 \), which gives \( N = 2126 \times 1417 = 3012542 \), indicating similarities between the two maps.

\(^3\)16/02/2011: The implementation of VIF is available at [http://live.ece.utexas.edu/research/quality/](http://live.ece.utexas.edu/research/quality/).

\(^4\)For all grayscale metrics we have converted the image using the rgb2gray function in Matlab unless stated otherwise, following the recommendation of Wang et al. 16/02/2011: [https://ece.uwaterloo.ca/~z70wang/research/ssim/](https://ece.uwaterloo.ca/~z70wang/research/ssim/). rgb2gray is defined as: \( 0.2989 \ast R + 0.5870 \ast G + 0.1140 \ast B \), for an RGB image. 17/04/2011: [http://www.mathworks.com/help/toolbox/images/ref/rgb2gray.html](http://www.mathworks.com/help/toolbox/images/ref/rgb2gray.html)
4.2.4 Other approaches

Others metrics considered in this group are based on other approaches or metrics combining two or more of the above groups.

4.2.4.1 Visual signal to noise ratio

Chandler and Hemami [70] proposed a metric, Visual Signal to Noise Ratio (VSNR), based on near-threshold and suprathreshold properties of the HVS, incorporating both low-level features and mid-level features. The metric consists of two stages; first contrast thresholds are used to detect visible distortions in the image, which is done in the wavelet domain by computing the Contrast Signal to Noise Ratio (CSNR). Then the contrast detection threshold is computed based on the CSNR, which is done for each octave band. The contrast is further compared to the detection threshold, and if above the distortion is supratreshold (visible). In this case a second stage is carried out, where a model of global precedence is proposed to account for the mid-level properties of the HVS. The global precedence takes into account that contrast of distortions should be proportioned across spatial frequency. The final metric is computed as the combination of perceived contrast of the distortion and disruption of global precedence.

The VSNR is an interesting metric since it is based on contrast thresholds, therefore it will only take into account the visible difference, different from the CSF based metrics where the entire image is modulated. In addition the use of both low-level features and mid-level features is compelling, and therefore we include VSNR\(^5\) in this survey.

4.2.4.2 Color image similarity measure

Lee et al. [261] proposed the Color Image Similarity Measure (CISM) with the intention for evaluating a halftoning algorithm. The CISM can be divided into two parts; one dealing with the HVS and one with structural similarity. The input images are converted to the CIELAB

\(^5\)16/02/2011: The code for VSNR can be downloaded at http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html
color space where they are filtered with CSFs from Sullivan et al. [428] and Näsänen [313] for the luminance channel and from Mullen [309] for the chrominance channels. After filtering the images are converted to RGB before each channel is used as input to the SSIM [458] framework (as introduced above). From each of the RGB channels a mean SSIM value is obtained, which is combined in order to get one value for IQ. The final value is a weighted sum of the three channel:

\[
CISM = \sum_i w_i SSIM_i, \quad (4.28)
\]

where \(i\) indicates the channel, and \(w\) is the weight (\(1/3\) suggested by Lee et al. [261]).

### 4.2.4.3 Comparison of selected metrics within the group

Since both the metrics in this group produce one quality value, and not a map it is difficult to show the differences between them. Nevertheless, they have one fundamental difference: VSNR is made for grayscale images while CISM for color images. Also the difference between the spatial filtering should be noted, where VSNR is built on contrast visibility and CISM on CSFs.

### 4.3 Summary

In this chapter we have discussed the structure of existing IQ metrics, and proposed a classification into four different groups. Further, a survey of metrics is given, where metrics from each of the classifications groups have been presented and differences and similarities between them have been discussed.
EXISTING IMAGE QUALITY METRICS
5 SHAME: A NEW IMAGE QUALITY METRIC

Many of the existing IQ metrics weight the content of the images equally, even though it is a well-known fact that observers put more emphasis on certain regions when they judge IQ [114]. It has also been shown that using information about where observers gaze can increase the performance of IQ metrics [339, 356] and increase compression without loosing quality [47]. Based on this IQ metrics can be improved by using a metric that takes into account regions of interest (i.e. does not weight all pixels equally). However, IQ metrics should also simulate the HVS, since our perception of quality is a function of the viewing conditions [98], such as the distance an image is viewed at. Combining these two key aspects of IQ could potentially result in a metric giving results that are better correlated with the results of human perception. Therefore, in this chapter, we present a new IQ metric taking these two aspects into account. The proposed metric, Spatial Hue Angle MEtric (SHAME), is based on the hue angle algorithm [177, 178], and two different spatial filtering methods are tested. We will first give an overview of the hue angle algorithm, and then the two spatial filtering methods.

5.1 The hue angle algorithm

Hong and Luo [177, 178] proposed a full-reference color image difference metric built on the CIELAB color difference formula [85]. This metric is based on the known fact that systematic errors over the entire image are quite noticeable and unacceptable. The metric is based on some conjectures; summarized from Hong and Luo [177, 178] these are:

- Pixels or areas of high significance can be identified, and suitable weights can be assigned to these.
- Pixels in larger areas of the same color should be given a higher weight than those in smaller areas.
- Larger color difference between the pixels should get higher weights.
- Hue is an important color perception for discriminating colors within the context.

The first step is to transfer each pixel in the image from $L^*$, $a^*$, $b^*$ to $L^*$, $C_{ab}^*$, $h_{ab}$. Based on the hue angle ($h_{ab}$) a histogram from the 360 hue angles is computed, and sorted in ascending order based on the number of pixels with same hue angle to an array $k$. Then weights can be
applied to four different parts (quartiles) of the histogram, and by doing this Hong and Luo corrected the drawback that the CIELAB formula weights the whole image equally. The first quartile, containing \( n \) hue angles, is weighted with 1/4 (that is, the smallest areas with the same hue angle) and saved to a new array \( hist \). The second quartile, with \( m \) hue angles, is weighted with 1/2. The third quartile, containing \( l \) hue angles, is given 1 as a weight and the last quartile with the remaining hue angles is weighted with 9/4.

\[
hist(i) = \begin{cases} 
  k(i) \cdot 1/4, & i \in \{0, ..., n\} \\
  k(i) \cdot 1/2, & i \in \{n + 1, ..., n + m\} \\
  k(i) \cdot 1, & i \in \{n + m + 1, ..., n + m + l\} \\
  k(i) \cdot 9/4, & \text{otherwise}
\end{cases}
\] (5.1)

The average color difference, computed using \( \Delta E_{ab}^* \), is calculated for all pixels having the same hue angle and stored in \( CD[hue] \). Then the overall color difference for the image, \( CD_{image} \), is calculated by multiplying the weights based on the quartiles for every pixel with the average CIELAB color difference for the hue angle

\[
CD_{image} = \sum_{0}^{359} hist[hue] \cdot CD[hue]^2 / 4.
\] (5.2)

### 5.2 Spatial filtering

We propose two different spatial filtering methods for the hue angle algorithm. The first spatial filtering is adopted from S-CIELAB [499] (Section 4.2.2.1). The advantage of using this spatial filtering is that has been extensively evaluated [18, 21, 41, 45, 138, 169, 173, 220, 233, 234, 339, 350, 351, 356, 490, 498, 500], and it has shown to produce good results for a wide variety of distortions.

The image goes through color space transformations, first the RGB image is transformed into CIEXYZ and further into the opponent color space (\( O_1, O_2 \), and \( O_3 \)) [499].

\[
O_1 = 0.279X + 0.72Y - 0.107Z,
\]

\[
O_2 = -0.449X + 0.29Y - 0.077Z,
\]

\[
O_3 = 0.086X - 0.59Y + 0.501Z.
\]

Now the image contains a channel with the luminance information (\( O_1 \)), one with the red-green information (\( O_2 \)), and one with blue-yellow information (\( O_3 \)). Then a spatial filter is applied, where data in each channel is filtered by a 2-dimensional separable spatial kernel:

\[
f = k \sum_i w_i E_i,
\] (5.3)

where

\[
E_i = k_i e^{-(x^2 + y^2)/\sigma_i^2},
\] (5.4)

and \( k_i \) normalize \( E_i \) such that the filter sums to one. The parameters \( w_i \) and \( \sigma_i \) are different for
the color planes as seen in Table 5.1. $k$ is a scale factor, which normalize each color plane so its two-dimensional kernel $f$ sums to one.

Table 5.1: The parameters used for the spatial filtering, where $w_i$ is the weight of the plane and $\sigma_i$ is the spread in degrees of visual angle.

<table>
<thead>
<tr>
<th>Plane</th>
<th>Weights $w_i$</th>
<th>Spreads $\sigma_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminance</td>
<td>0.921</td>
<td>0.0283</td>
</tr>
<tr>
<td></td>
<td>0.105</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>-0.108</td>
<td>4.336</td>
</tr>
<tr>
<td>Red-Green</td>
<td>0.531</td>
<td>0.0392</td>
</tr>
<tr>
<td></td>
<td>0.330</td>
<td>0.494</td>
</tr>
<tr>
<td>Blue-Yellow</td>
<td>0.488</td>
<td>0.0536</td>
</tr>
<tr>
<td></td>
<td>0.371</td>
<td>0.386</td>
</tr>
</tbody>
</table>

The second spatial filtering proposed is adopted from Johnson and Fairchild [223]. By specifying and implementing the spatial filters using CSFs in the frequency domain, rather than in the spatial domain as the first spatial filtering, more precise control of the filters is obtained [223], but usually at the cost of computational complexity. Results also show increased performance of metrics where the spatial filtering is performed in the frequency domain [223, 351, 412]. The luminance filter is a three parameter exponential function, based on research by Movshon and Kiorpes [308].

$$CSF_{\text{lum}}(p) = a \cdot p^c \cdot e^{-b \cdot p}, \quad (5.5)$$

where $a = 75$, $b = 0.22$, $c = 0.78$, and $p$ is represented as Cycles Per Degree (CPD).

The luminance CSF is normalized so that the DC modulation is set to 1.0, by doing this the luminance shift is minimized, compared to the first spatial filtering method. This will also enhance any image differences where the human visual system is most sensitive to them [223]. For the chrominance CSF, a sum of two Gaussian functions are used.

$$CSF_{\text{chroma}}(p) = a_1 \cdot e^{-b_1 \cdot p^c_1} + a_2 \cdot e^{-b_2 \cdot p^c_2}, \quad (5.6)$$

where different parameters for $a_1, a_2, b_1, b_2, c_1$ and $c_2$ have been used as seen in Table 5.2. A graph of the luminance and chrominance CSFs is shown in Figure 5.1.

Table 5.2: The parameters used for the spatial filtering in the frequency domain of the chrominance channels.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Red-Green</th>
<th>Blue-Yellow</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>109.14130</td>
<td>7.032845</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-0.00038</td>
<td>-0.000004</td>
</tr>
<tr>
<td>$c_1$</td>
<td>3.42436</td>
<td>4.258205</td>
</tr>
<tr>
<td>$a_2$</td>
<td>93.59711</td>
<td>40.690950</td>
</tr>
<tr>
<td>$b_2$</td>
<td>-0.00367</td>
<td>-0.103909</td>
</tr>
<tr>
<td>$c_2$</td>
<td>2.16771</td>
<td>1.648658</td>
</tr>
</tbody>
</table>
5.3 Applying spatial filtering to the hue angle algorithm

The images are spatially filtered with the previously introduced spatial filtering methods. This results in a filtered original and a filtered modified version of the original, which are used as input to the hue angle algorithm, as shown in Figure 5.2. This can be considered as an extension of the S-CIELAB (Section 4.2.2.1) flowchart on Figure 4.5.

The hue angle algorithm, filtered respectively with the first and second filter, is from now on referred to as SHAME-I and SHAME-II. The new metric will theoretically have several key features from both the S-CIELAB and the hue angle measure:

- Weight allocation: pixels in larger areas of the same color should be weighted higher.
- Simulation of the spatial properties of the human visual system.
- Undetectable distortions are ignored.
- Suitable for different kind of distortions, not only color patches.
- Generates one value for easy interpretation.
5.4 Classification

SHAME can be considered as an approach combining several different features, and would be classified together with the metrics in the group "other approaches" in Section 4.2.4.

5.5 Summary

In this chapter a new IQ metric, SHAME, was proposed. SHAME takes into account the HVS and incorporates information about region of interest. Two different spatial filtering methods are suggested, the one from Zhang and Wandell [499] and the one from Johnson and Fairchild [223]. Evaluation of the metric is found later in Chapter 7.
6 HOW TO EVALUATE IMAGE QUALITY METRICS

There are a number of algorithms involved in the reproduction of an image, for instance gamut mapping, halftoning, and compression. To identify the best reproduction among a number of variants of the same reproduction algorithm (e.g. JPEG compression), a psychophysical experiment can be carried out. This will result in a scale with the visual difference of the reproductions from the original. These psychophysical experiments are both time and resource demanding. Because of this, objective methods such as IQ metrics have been introduced to entirely or partially eliminate psychophysical experiments. However, in order to know if an IQ metric correlates with the human percept, some kind of evaluation of the metric is required. The most common evaluation of IQ metrics is by comparing the results of the metrics to the results of human observers. In this chapter we investigate the different methods to evaluate IQ metrics.

6.1 Evaluation methodology

We will give a brief introduction to the methodology for evaluation of IQ metrics. An experiment is set up following the instructions given in Section 3.2, where a set of images is evaluated by a group of observers. The results from the observers can be statistically analyzed, as suggested in Section 3.2. The mean values from the subjective evaluation are used as the basis against which the metrics are evaluated. For the same set of images the metrics are calculated, resulting in one value describing the quality of the image. The results from the IQ metrics are then compared against the results of the observers, forming the basis for the evaluation.

6.2 Existing evaluation methods

A number of existing evaluation methods are found in the literature. We will go through the methods in which IQ metrics are evaluated against perceptual data.
6.2.1 Correlation based methods

Correlation is a common method for the evaluation of IQ metrics. Correlation describes the statistical relationship between two or more variables.

6.2.1.1 Pearson product-moment correlation coefficient

The most common measure of correlation is the Pearson product-moment correlation coefficient [232], which is a linear correlation between two variables. Pearson’s correlation $r$ is defined as:

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}},$$  \hspace{1cm} (6.1)

where $X$ and $Y$ are two variables and $N$ is the number of samples. The correlation value $r$ is between $-1$ and $+1$. For evaluation of metrics, one of the variables is the results from a psychophysical experiment (for example z-scores) and the other variable is the results from an IQ metric. Depending on the scale of the metric, whether a low number is high quality or a high number is high quality, the best correlation will be $-1$ or $+1$, for easy comparison of results the correlation values of the metrics where a low number indicates high quality the results are usually multiplied with $-1$. Unless stated otherwise this is done for all results in this work. This method has been extensively used in the evaluation of metrics [113, 169, 279, 459, 495].

Significance test  Significance tests can be carried out for Pearson’s correlation by using the following equation:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}},$$  \hspace{1cm} (6.2)

where $N$ is the number of pairs of scores, and with $N-2$ degrees of freedom (number of values in the final calculation of a statistic). A $t$ table can further be used to find the probability value (sometimes called $p$, i.e. the probability of obtaining a statistic as different from or more different from the parameter specified in the null hypothesis).

Confidence intervals  Confidence intervals can also be calculated for Pearson’s correlation. One consideration is that the sampling distribution of Pearson’s $r$ is not normally distributed. Therefore, Pearson’s $r$ is converted to Fisher’s $z’$ and the confidence interval is computed using Fisher’s $z’$. The values of Fisher’s $z’$ in the confidence interval are then converted back to Pearson’s $r$.

The first step is to use the Fisher’s Z-transform:

$$z = \frac{1}{2} ln \left( \frac{1+r}{1-r} \right),$$  \hspace{1cm} (6.3)

where $r$ is the correlation coefficient. The confidence intervals for $r$ are calculated on the transformed $r$ values ($z’$). The general formulation of confidence intervals for $z’$ is

$$z’ \pm z\sigma_{z’},$$  \hspace{1cm} (6.4)
where the criterion $z$ is the desired confidence level ($1.96$ gives a 95% confidence interval), and $\sigma_z$ is defined as

$$\sigma_z = \frac{1}{\sqrt{N-3}}, \quad (6.5)$$

where $N$ is the number of correlation samples. The upper and lower limits confidence interval limits for $z'$ are found by using Equation 6.4. To translate from the $z$-space back into the $r$-space, it is necessary to invert Equation 6.3. This results in the following equation:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}. \quad (6.6)$$

The upper and lower confidence interval limits for $r$ can be found by using the upper and lower $z$ values into Equation 6.6.

Figure 6.1 shows an overview of the lowest and highest correlation limits for $N$ values from 16 to 1024 (95% confidence interval). Figure 6.1(a) show the lowest limit, and Figure 6.1(b) the highest. The higher the number of data points ($N$) the smaller the confidence interval, for a small number of data points the confidence interval is large and important to be aware of.

### 6.2.1.2 Spearman’s rank-correlation coefficient

Another correlation measure often used [253, 279, 311, 364, 366, 398, 458, 495] is the Spearman rank-correlation coefficient [232], which is a non-parametric measure of association based on the ranks of the data values that describes the relationship between the variables without making any assumptions about the frequency distribution. In the case of no tied ranks Spearman correlation ($r$) is defined as:

$$r = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}, \quad (6.7)$$

where the raw scores $X_i$ and $Y_i$ are converted into ranks $x_i$ and $y_i$, and the difference between the ranks is $d_i = x_i - y_i$. $N$ is the number of samples. In the case where ranks exist the Pearson correlation coefficient (Equation 6.1) should be used for the calculation of correlation [32]. Many report the Spearman correlation since we cannot expect a linear relation between the metric and perceived distortion due to the non-linearities in the HVS [439].

Confidence intervals can be calculated for Spearman with the same procedure as for Pearson [10], since the distribution of Spearman’s $r$ is similar to that of Pearson’s $r$.

### 6.2.1.3 Kendall tau rank correlation coefficient

A third correlation measure is also found in the literature [311, 364, 366, 495] for the evaluation of metrics. The Kendall tau rank correlation coefficient [232] is a measure of rank correlation as Spearman, but the Kendall rank correlation interpretation is easier since the correlation represents the difference between two probabilities. Kendall tau rank correlation coefficient is defined as :

$$\tau = \frac{(number \ of \ concordant \ pairs) - (number \ of \ discordant \ pairs)}{\frac{1}{2} N(N-1)}, \quad (6.8)$$
Figure 6.1: Low and high correlation limits for different $N$ values. The low limit shows the bottom of a 95% confidence interval of a correlation value, while the high limit shows the top of the 95% confidence interval. The figure shows values for $N$ between 16 and 1024, and only correlation values above 0.5. Figure b) has been rotated for better representation.
where two random variables $X$ and $Y$ and their pair $(x_i, y_i)$ and $(x_j, y_j)$ are said to be concordant if the rank of both elements agree, and they are discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. $N$ is the number of samples, and $\tau$ is the Kendall correlation coefficient.

### 6.2.2 Error methods

#### 6.2.2.1 Root mean squared error

Root Mean Squared Error (RMSE) calculation is used to quantify the error between fitted objective data and corresponding subjective data [95, 190]. RMSE is given by

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (O_i - S_i)^2}, \quad (6.9)$$

where $N$ is the number of samples, $O$ is the objective data, and $S$ the subjective data. This method has been used by a number of researchers [95, 253, 311, 396, 398, 458, 495].

#### 6.2.2.2 Mean absolute error

The Mean Absolute Error (MAE) is used to measure how close predictions are from the actual values. MAE is given by

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |e_i|, \quad (6.10)$$

where $N$ is the number of samples, and $e_i$ is the absolute error of $f_i - y_i$ between the prediction $f_i$ and the true value $y_i$. This method has been used by some researchers [279, 458].

#### 6.2.3 Outlier ratio

Outlier ratio [450, 461] is the ratio of “false” scores by an IQ metric to the total number of scores. The “false” scores are the scores that lie outside the interval $[\text{MOS} - 2\sigma, \text{MOS} + 2\sigma]$, where MOS is the mean opinion score and $\sigma$ the standard deviation of the MOS. This is a less used method, but some researchers [279, 458] calculate this in addition to other methods.

### 6.3 New method for the evaluation of image quality metrics

Existing methods, such as those based on correlation, are sensitive to scale differences between images [296] (Figure 6.2) and extreme values. Especially problematic is the use of correlation when it is calculated on data coming from differently scaled and shifted scenes. These are problems when one would like to have a value indicating the overall performance of a metric over a larger set of images. Because of this we propose a new method for evaluating the overall performance of IQ metrics (Figure 6.3). We propose to process the results from IQ metrics in the same way as we commonly do with the results from a perceptual rank order evaluation, i.e. that a metric simulates an observers. For each scene in the dataset we obtain
a ranking based on the metric results, this is then used as input to the rank order method (as described in Section 3.2.2.3). The result is z-scores for the overall performance of the metric, which can be directly compared to the z-scores from observers. One disadvantage of using rank order is the number of scenes needed, in order to provide useful results the number of scenes must be high (following the same recommendations regarding the number of observers as described in Section 3.2.4.1). The more scenes used the more accurate results. The more data provided the more accurate the results will become, and the confidence intervals in the metric’s z-score will become smaller. The confidence intervals can be used as a measure of when a reproduction method is significantly better than another, and this provides more information than Pearson’s or Spearman’s correlation coefficients.

Figure 6.2: Metric values plotted against observer values for two different scenes, and the regression line (black). The correlation within each scene (circles) is rather high, but correlation of all scenes is lower due to scale differences between the scenes. One should note that correlation is invariant to scaling and shift, while the slope and intercept of the regression line is not.

Figure 6.3: The proposed method for evaluation of IQ metrics. The rank for the IQ metrics is calculated for each scene, these rankings are used as input to the rank order method, which results in an overall z-score for the metric. Z-scores are also calculated for the observers. The two z-scores can now be compared, if the metric z-scores are similar to the observer z-scores the metric has the same overall performance as the observers, indicating that it is able to predict overall perceived IQ.
6.3.1 Example of the proposed method based on rank order

An example is given to show how the method can be used in the evaluation of IQ metrics.

6.3.1.1 Psychophysical experiment

Figure 6.4: Images from the experiment carried out by Dugay et al. [108]. These images have been gamut mapped with five different algorithms.

20 different images (Figure 6.4) were chosen for a psychophysical experiment [107, 108]. Considering the color gamut of an Ocê TCS 500 printer, the amount of out-of-gamut colors ranged from 25% to 100% with a mean of 57%. The images were reproduced using five different Gamut Mapping Algorithms (GMAs).

- HPminDE (Hue preserving minimum $\Delta E_{ab}^*$ clipping) which is a baseline GMA proposed by the CIE [80]. The algorithm does not change in-gamut colors at all, while out-of-gamut colors are mapped to the closest color on the destination gamut while preserving the hue.

- SGCK (chroma-dependent SiGmoidal lightness mapping and Cusp Knee scaling) [80] is an advanced spatially invariant sequential gamut compression algorithm. The lightness is first compressed by a chroma dependent sigmoidal scaling, resulting in high chroma colors being compressed less than neutral ones. The resulting colors are then compressed along lines toward the cusp [303] of the destination gamut using a 90% knee scaling function. For the final compression the image gamut is used as the source gamut.

- Zolliker and Simon [502] proposed a spatial GMA; its goal being to recover local contrast while preserving lightness, saturation and global contrast. A simple clipping is performed as a first step; then by using an edge-preserving high pass filter the difference between the original and gamut clipped image is filtered. The filtered image is then added to the gamut clipped image. As a last step the image is clipped in order to be in-gamut.

- Kolås and Farup [242] recently proposed a hue- and edge-preserving spatial color GMA. The image is first gamut clipped along straight lines toward the center of the gamut. A
relative compression map is then created from the original and clipped image. Using this compression map, a new image can be constructed as a linear convex combination of the original image and neutral gray image. This image is in turn filtered by an edge-preserving smoothing minimum filter. As the final step the gamut mapped image is constructed as a linear convex combination of the original image and neutral gray using the filtered map.

- Farup et al. [131] proposed a multiscale algorithm (Gatta) preserving hue and local relationship between closely related pixel colors. First a scale-space representation of the image and then gamut clipping the lowest scale is constructed. The resulting gamut compression is then applied to the image at the next smallest scale. Operators are used to reduce the effect of haloing. The process is repeated until all scales are treated. The Fourier domain is used to speed up the process.

The 20 different images have been evaluated by 20 observers in a pair comparison experiment [107, 108]. All observers had normal or corrected to normal color vision. The observers were presented with the original image in the middle of the screen, with two different reproductions on each side. The observers were asked to pick the image with the most accurate reproduction with respect to the original image. When the observer had picked one image, a new pair of reproductions was shown until all combinations were evaluated. All pairs were also shown twice in opposite order for consistency. The monitor was a Dell 2407WFP LCD display calibrated with a D65 white point and a 2.2 gamma. The viewing conditions were chosen as close to the ones described in the CIE guidelines [80] as possible. The level of ambient illumination was measured to approximately 20 lux. The observer was seated approximately 50 cm from the screen.

6.3.1.2 Z-scores

The z-scores are based on Thurstone’s law of comparative judgment [116, 437]. Data collected are transformed into interval scale data where scores represent the distance of a given image from the mean score of a set of images in units of the standard deviation of the scenes [302], and therefore being relative. The 95% confidence intervals are calculated in the same way as proposed by Morovic [302]. The error bars are then computed as

\[ \bar{X} \pm \frac{\sigma}{\sqrt{N}}, \]

where \( \bar{X} \) is the Z-score, \( \sigma \) is the standard deviation, and \( N \) is the size of the sample set. For these experiments this is the number of observers multiplied with two, because each image pair was shown twice for consistency. With this confidence interval there is a 95% estimate that the value will be within the interval, and if the confidence interval of another GMA is outside this interval the difference is significant.

6.3.1.3 Selected image quality metrics

Five IQ metrics have been chosen, \( \Delta E_{ab}^* \), S-CIELAB [499], iCAM [125], SSIM [458], and the hue angle algorithm [177]. All metrics except SSIM have a scale where closer to 0 indicate a reproduction closer to the original, while SSIM has a scale between 0 and 1, where 1 indicate an identical reproduction.
6.3.1.4 Results

From the pair comparison experiment z-scores were calculated, indicating the performance of the different GMAs. From Figure 6.5 we can see that the Gatta algorithm gets the highest score from the observers, indicating the lowest visual difference from the original. This algorithm gives statistically the same visual difference from the original as the SGCK and Zolliker algorithm. Kolás has the fourth best score, but has the same visual difference as Zolliker. HPminDE clearly gives the highest visual difference from the original, the low score here indicating a large consensus among the observers about the low performance of this algorithm.

6.3.2 Rank order based on metric order

Aiming to develop a universal IQ metric, this metric should work across multiple scenes and in different conditions. One way of evaluating the performance of IQ metrics is to check the correlation between the perceived IQ and the calculated IQ [22, 305] as seen in Figure 6.6. We can see from Figure 6.6 that the data points are very spread, and there is no correlation. Thus it is not possible to use the IQ results as a way of evaluating the performance of the best GMA. This is the case for all the metrics, where the correlation is generally low for all scenes as seen in Table 6.1 for both Pearson’s correlation and Spearman’s rank order correlation. The Spearman rank order correlation does not take into account the frequency distribution of the variables, and should therefore be less sensitive to extreme values, but as seen in Table 6.1 this is not a good measure for overall prediction of performance. The z-score for each scene does not say anything about the difference between scenes; it is based on how preferred each image is within each scene. We follow the above proposal to process the IQ metric results in the same way as we commonly do with the results from a perceptual rank order evaluation.
Figure 6.6: Z-score from observers against $\Delta E_{ab}$ values. The data points are very spread, and we get a very low correlation between the z-scores and $\Delta E_{ab}$ values. The HPminDE algorithm is rated as the best by $\Delta E_{ab}^*$, opposite of the observers (Figure 6.5).

Table 6.1: Correlation between all z-scores and all algorithms. The correlation here is low for all metrics both for Pearson and Spearman. SSIM has the highest Pearson’s correlation, however, not being higher than 0.16 it indicates a low performance. For the Spearman correlation the hue angle algorithm has the highest correlation, but still very low. The plot for $\Delta E_{ab}^*$ with a linear fitted line and calculated Pearson’s correlation is found in Figure 6.6.
The rank for each metric in the 20 scenes have been used as a basis for the overall performance of the GMAs, this correspond to 20 observers in a ranking experiment. If the results from the metrics match the overall results from the psychophysical experiment (Figure 6.5), the metrics predict perceived IQ. This method will only provide information about the order of the image samples, not information of the distance between the samples. The number of images used should follow the recommendations for the number of observers (Section 3.2.4.1) to produce reliable data. CIE [80] recommends at least 15 observers, in this case 15 images.

In principle the rank order and pair comparison approaches provide the same information [91]. The rank order data has been used to generate corresponding pair comparison data [91, 160], and the z-scores were computed as for a pair comparison experiment [116, 302]. Babcock [17] got similar score for pair comparison, rank order, and graphical rating. This implicates that scale values from one type of experiment can be directly compared to scale values from another type of experiment. The rank order z-scores in this experiment have been calculated by using the Colour Engineering Toolbox [162].

![Graph](image-url)  
(a) $\Delta E_{ab}^*$ rank order score.  
(b) Observer z-score against $\Delta E_{ab}^*$ rank order score.

*Figure 6.7: Rank order score for $\Delta E_{ab}^*$ values with a 95% CI, and these values plotted against z-scores from observers with a 95% CI with a linear regression line.*

From Figure 6.7(a) we can see results from the rank order for $\Delta E_{ab}^*$. The HPminDE have the highest z-score, but this has the lowest z-score from the observers. This GMA will clip the color to the minimum $\Delta E_{ab}^*$ and will always be rated as the best by the $\Delta E_{ab}^*$ formula. The SGCK gets a very low score in the ranking, while the observers rated this as one of the best GMAs. Figure 6.7(b) shows z-score from the observers plotted against the rank order z-score from $\Delta E_{ab}^*$. The Pearson correlation here is -0.62, indicating a low performance for the $\Delta E_{ab}^*$. This is important to notice that the HPminDE strongly influence the results, since it is judged clearly worse than the other GMAs (Figure 6.5).

Figure 6.8(a) shows the results for SSIM, as we can see the HPminDE gets the lowest score by SSIM. This is the same as the observers. The four other GMAs cannot be differentiated with a 95% confidence interval. The observers also have very small differences between these algorithms; the Koláš algorithm has a score just lower than the SGCK, Gatta, and Zolliker. The score from SSIM is very similar to the score from the observers, this is also verified with a correlation between the scores of 0.98 (Figure 6.8(b)). The correct ranking of the HPminDE GMA is the basis for the excellent correlation here, and this ranking (Figure 6.8(a)) also reflect the observers ranking (Figure 6.5).
The overall Pearson correlation is only 0.16 and Spearman correlation is 0.05 (Table 6.1) between the z-scores for each scene and SSIM scores for each scene. These are therefore not good measured of overall performance, even though high correlation can be found within each scene in both measures. The correlation within a scene can be average, but the ranking can be correct as seen on Figure 6.9. Here the Spearman correlation will perform well, while Pearson’s only perform average. In other cases the Spearman correlation will perform average, while the Pearson correlation will perform excellent. Spearman’s rank order correlation can provide low correlation where clusters of data are found, but the ranking within the cluster is not necessarily correct. When the ranking is used in the rank order method, the normal distribution is taken into account and will therefore better handle extreme values. Because of this the ranking of the results within each scene and using these as a basis for the rank order z-scores, SSIM will reflect perceived IQ. The overall Pearson correlation is only 0.16 and Spearman correlation is 0.05 (Table 6.1) between the z-scores for each scene and SSIM scores for each scene. These are therefore not good measured of overall performance, even though high correlation can be found within each scene in both measures. The correlation within a scene can be average, but the ranking can be correct as seen on Figure 6.9. Here the Spearman correlation will perform well, while Pearson’s only perform average. In other cases the Spearman correlation will perform average, while the Pearson correlation will perform excellent. Spearman’s rank order correlation can provide low correlation where clusters of data are found, but the ranking within the cluster is not necessarily correct. When the ranking is used in the rank order method, the normal distribution is taken into account and will therefore better handle extreme values. Because of this the ranking of the results within each scene and using these as a basis for the rank order z-scores, SSIM will reflect perceived IQ.

![Graphs showing SSIM and observer z-scores](image)

**Figure 6.8:** Rank order score for SSIM values with a 95% CI, and these values plotted against z-scores from observers with a 95% CI. The results here indicate a very high performance by the SSIM, the correct ranking of the HPminDE is the main reason for this, but also the similar ranking of the remaining GMAs. Since we only have five data points ($N = 5$), the CI will be [0.72, 1.00]. However, this depends strongly on the HPminDE observation.

There are large differences between the score from iCAM and the observers, mostly due to the calculated values for HPminDE in iCAM. iCAM has a spreading of the four best algorithms from the observers, and the HPminDE is rated as the best by iCAM, opposite of the observer rating. The Pearson correlation between the observer z-score and iCAM z-score is
Correlation SSIM for scene 14: 0.75

Figure 6.9: Correlation between z-score and calculated IQ for SSIM on scene 14. The visual quality differences between the images is not correctly predicted by SSIM, but the ranking is. Based on this the rank order method on the SSIM data will be a better measure.

-0.68, resulting in a low performance by iCAM.

The HPminDE receives the highest score by S-CIELAB, both the SGCK and Kolås get low scores. The results for S-CIELAB is very similar to the ones found with $\Delta E_{ab}^*$. This results in a low Pearson’s correlation of -0.62.

The hue angle algorithm has similar results to $\Delta E_{ab}^*$ and S-CIELAB, this is not surprising due to the familiarity between these metrics. The ranking of the GMAs are the same with only minor differences in the z-score values, this results in almost the same Pearson correlation between observer z-score and metric z-score as S-CIELAB, with -0.72.

Table 6.2: Rank indicates the Pearson correlation between rank order z-score and observer z-score. Mean Pearson indicate correlation between ranked metric score and observer z-score calculated as Pearson’s correlation, where the correlation for each scene has been averaged. Mean Spearman is similar to mean Pearson but for Spearman’s rank order correlation.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Rank Pearson</th>
<th>Mean Pearson</th>
<th>Mean Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta E_{ab}^*$</td>
<td>-0.62</td>
<td>0.28</td>
<td>0.17</td>
</tr>
<tr>
<td>SSIM</td>
<td><strong>0.98</strong></td>
<td><strong>0.84</strong></td>
<td><strong>0.78</strong></td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>-0.62</td>
<td>0.28</td>
<td>0.20</td>
</tr>
<tr>
<td>iCAM</td>
<td>-0.68</td>
<td>-0.03</td>
<td>-0.07</td>
</tr>
<tr>
<td>Hue angle</td>
<td>-0.72</td>
<td>0.27</td>
<td>0.15</td>
</tr>
</tbody>
</table>

6.3.2.1 Overall observations

SSIM is the IQ metric with the best fit between the observers and the algorithm z-score (Table 6.2), indicating a good prediction of perceived IQ. All other metrics have a correlation
below 0, indicating that these metrics do not predict perceived IQ well. In all metrics except SSIM the HPminDE GMA has been miscalculated, i.e. given a too high rank by the metrics. These metrics are based on $\Delta E_{ab}$ and therefore the HPminDE will be given a high rank.

### 6.4 Which methods will we use for the evaluation of image quality metrics

We will use a set of the methods presented above to evaluate IQ metrics. In order to compare our results to other researchers, and to follow the most common strategies, we adapted Pearson (Section 6.2.1.1) and Spearman (Section 6.2.1.2) correlation. Additionally, to evaluate the overall performance of the metrics we will use the proposed method based on rank order (Section 6.3). The method used will depend on many factors, such as the metrics being evaluated and the experimental data. These methods require evaluation against perceptual data gathered from a large user group, and in some cases other evaluation methods might be used to obtain a thorough evaluation.

### 6.5 Summary

In this chapter we have presented methods for evaluating IQ metrics against human observers. Further, a new method for evaluating the overall performance of metrics was proposed based on the rank order method. An example of using the proposed method is given, before we introduced the methods used in this thesis.
7 Evaluation of Image Quality

Metrics

The IQ metrics presented in Chapter 4 take different approaches in their common goal to predict perceived IQ. These metrics are created to be correlated with perceived quality. In order to assure that the metrics are correlated with human observers they need to be evaluated. This evaluation is usually carried out by comparing the results of the metrics against the results of human observers. The better correlated the metrics are the higher their performance. To describe their performance, in terms of correlation with the percept, we carried out extensive evaluations of a set of metrics (Table 7.1). The metrics selected are from the different classified groups of metrics from Chapter 4, ranging from standard metrics as the PSNR and $\Delta E_{ab}^*$ to more advanced metrics such as S-CIELAB and SSIM, but also recently published metrics. They also differ in their simulation of the HVS and whether they incorporate color information. To evaluate the performance of the these metrics we follow the methodology of Chapter 6. For this evaluation we need a set of test images with a range of distortions, which has been judged by human observers. There are a number of databases available online for evaluation of metrics (Table 7.2), which contain images with distortions and perceptual data. There is a trend that the number of images and number of observers in the databases increases, and recently experiments have also been carried out on the web [123, 367]. Web-based experiment and other large experiments have the advantage with many observers, but also the the disadvantage that controlling the viewing conditions is difficult, and therefore they might not be the best to evaluate metrics taking into account the viewing conditions. Our goal is to use IQ metrics metrics to evaluate printed images, however, to our knowledge there is no database available with printed images, therefore the evaluation in this chapter is carried out with the existing databases.

Among the few public databases providing images for evaluation of IQ metrics, we used the Tampere Image Database 2008 (TID2008) [365] and the IVC image database [58]. These two databases cover the distortions found in the other databases in Table 7.2. Additionally, we selected four datasets containing respectively luminance changed images [339, 356], JPEG and JPEG2000 compressed images [63, 410], images with global variations of contrast, lightness, and saturation [6], and gamut mapped images [107, 108]. The datasets from Pedersen et al. [339, 356] and Ajagamelle et al. [6] are related to color differences, which is an important aspect for metrics to be able to evaluate. The dataset from Simone et al. [63, 410] contains small compression artifacts, unlike the TID2008 database that has larger differences, this dataset will be able to tell if the metrics are able to evaluate just noticeable differences in terms of compression. The dataset from Dugay [107, 108] is based on gamut mapping where a number of different attributes are changed simultaneously, making it a difficult task for the metrics. These databases include a wide range of distortion, from color changes to struc-
Table 7.1: Metrics evaluated in terms of correlation with the percept. Year gives the year of the publication where the metric was proposed, Name is the (abbreviated) name of the metric, Author is the authors name, Type indicates the type of metric as given by the authors, HVS indicates if the metric has any simulation of the HVS, and Image indicates if the metric is a color or grayscale metric. For the grayscale metrics the images were converted to grayscale images using the rgb2gray function in Matlab 7.5.0.

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>Author</th>
<th>Type</th>
<th>HVS</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>MSE</td>
<td>-</td>
<td>Image difference</td>
<td>No</td>
<td>Gray</td>
</tr>
<tr>
<td>-</td>
<td>PSNR</td>
<td>-</td>
<td>Image difference</td>
<td>No</td>
<td>Gray</td>
</tr>
<tr>
<td>1976</td>
<td>$\Delta E^*_{ab}$</td>
<td>CIE</td>
<td>Color difference</td>
<td>No</td>
<td>Color</td>
</tr>
<tr>
<td>1995</td>
<td>$\Delta E_94$</td>
<td>CIE</td>
<td>Color difference</td>
<td>No</td>
<td>Color</td>
</tr>
<tr>
<td>1996</td>
<td>S-CIELAB</td>
<td>Zhang and Wandell [499]</td>
<td>Image difference</td>
<td>Yes</td>
<td>Color</td>
</tr>
<tr>
<td>2001</td>
<td>$\Delta E_{00}$</td>
<td>CIE</td>
<td>Color difference</td>
<td>No</td>
<td>Color</td>
</tr>
<tr>
<td>2002</td>
<td>UIQ</td>
<td>Johnson and Fairchild [223]</td>
<td>Image quality</td>
<td>Yes</td>
<td>Color</td>
</tr>
<tr>
<td>2002</td>
<td>Hue angle</td>
<td>Hong and Luo [177]</td>
<td>Image difference</td>
<td>No</td>
<td>Color</td>
</tr>
<tr>
<td>2003</td>
<td>$Q_{COLOR}$</td>
<td>Toet and Lucassen [439]</td>
<td>Image fidelity</td>
<td>No</td>
<td>Color</td>
</tr>
<tr>
<td>2004</td>
<td>SSIM</td>
<td>Wang et al. [458]</td>
<td>Image quality</td>
<td>No</td>
<td>Gray</td>
</tr>
<tr>
<td>2006</td>
<td>VIF</td>
<td>Sheikh and Bovik [396]</td>
<td>Image fidelity</td>
<td>Yes</td>
<td>Gray</td>
</tr>
<tr>
<td>2006</td>
<td>PSNR-HVS-M</td>
<td>Egiazarian et al. [113]</td>
<td>Image quality</td>
<td>Yes</td>
<td>Gray</td>
</tr>
<tr>
<td>2006</td>
<td>SSIM$_{IPT}$</td>
<td>Bonnier et al. [41]</td>
<td>Image difference</td>
<td>No</td>
<td>Color</td>
</tr>
<tr>
<td>2007</td>
<td>VSNR</td>
<td>Chandler and Hemami [70]</td>
<td>Image fidelity</td>
<td>Yes</td>
<td>Gray</td>
</tr>
<tr>
<td>2009</td>
<td>$\Delta E_F$</td>
<td>Oleari et al. [329]</td>
<td>Color difference</td>
<td>No</td>
<td>Color</td>
</tr>
<tr>
<td>2009</td>
<td>S-DEE</td>
<td>Simone et al. [412]</td>
<td>Image difference</td>
<td>Yes</td>
<td>Color</td>
</tr>
<tr>
<td>2009</td>
<td>SHAME</td>
<td>Pedersen and Hardeberg [351]</td>
<td>Image quality</td>
<td>Yes</td>
<td>Color</td>
</tr>
<tr>
<td>2009</td>
<td>SHAME-II</td>
<td>Pedersen and Hardeberg [351]</td>
<td>Image quality</td>
<td>Yes</td>
<td>Color</td>
</tr>
<tr>
<td>2009</td>
<td>ABF</td>
<td>Wang and Hardeberg [459]</td>
<td>Image difference</td>
<td>Yes</td>
<td>Color</td>
</tr>
<tr>
<td>2010</td>
<td>$S_{DOG}$-CIELAB</td>
<td>Ajagamelle et al. [5]</td>
<td>Image quality</td>
<td>Yes</td>
<td>Color</td>
</tr>
<tr>
<td>2010</td>
<td>$S_{DOG}$-DEE</td>
<td>Ajagamelle et al. [5]</td>
<td>Image quality</td>
<td>Yes</td>
<td>Color</td>
</tr>
</tbody>
</table>
Table 7.2: Image quality databases available online. Observers show the total number of observers on which the scores are computed.

<table>
<thead>
<tr>
<th>Database</th>
<th>Type of distortion</th>
<th>Original scenes</th>
<th>Number of distortions</th>
<th>Total number of images</th>
<th>Observers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TID2008(^1) [365]</td>
<td>JPEG, JPEG2000, blur, noise, etc.</td>
<td>25</td>
<td>17</td>
<td>1700</td>
<td>838</td>
</tr>
<tr>
<td>LIVE(^2) [399]</td>
<td>JPEG, JPEG2000, blur, noise, and bit error</td>
<td>29</td>
<td>5</td>
<td>982</td>
<td>161</td>
</tr>
<tr>
<td>A57(^3) [69]</td>
<td>Quantization, noise, JPEG, JPEG2000, blur</td>
<td>3</td>
<td>6</td>
<td>57</td>
<td>7</td>
</tr>
<tr>
<td>IVC(^4) [58]</td>
<td>JPEG, JPEG2000, LAR coding, and blurring</td>
<td>10</td>
<td>4</td>
<td>235</td>
<td>15</td>
</tr>
<tr>
<td>Toyama(^5) [179]</td>
<td>JPEG and JPEG2000</td>
<td>14</td>
<td>2</td>
<td>168</td>
<td>16</td>
</tr>
<tr>
<td>CSIQ(^6) [254, 255]</td>
<td>JPEG, JPEG2000, noise, contrast, and blur.</td>
<td>30</td>
<td>6</td>
<td>866</td>
<td>35</td>
</tr>
<tr>
<td>WIQ(^7) [119]</td>
<td>Wireless artifacts</td>
<td>7</td>
<td>Varying</td>
<td>80</td>
<td>30</td>
</tr>
</tbody>
</table>

\(^1\) 18/02/2011: http://www.ponomarenko.info/tid2008.htm
\(^2\) 18/02/2011: http://live.ece.utexas.edu/research/quality
\(^3\) 18/02/2011: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html
\(^4\) 18/02/2011: http://www2.irccyn.ec-nantes.fr/ivcdb/
\(^5\) 18/02/2011: http://mict.eng.u-toyama.ac.jp/mictdb.html
\(^6\) 18/02/2011: http://vision.okstate.edu/index.php?loc=csiq
\(^7\) 18/02/2011: http://www.bth.se/tek/rcg.nsf/pages/wiq-db

tural changes, over a large set of images, covering many different characteristics, to assure an extensive evaluation of the IQ metrics.

The performance of each metric is estimated by calculating the correlation between the perceptual quality scores from psychophysical experiments and the quality values calculated by the metric. We opted for two standard types of correlation:

- The product-moment correlation coefficient or Pearson correlation coefficient, which assumes a normal distribution in the uncertainty of the data values and that the variables are ordinal.

- The Spearman rank-correlation coefficient, which is a non-parametric measure of association, based on the ranks of the data values, which describes the relationship between the variables without making any assumptions about the frequency distribution.

### 7.1 TID2008 Database, Ponomarenko et al.

The TID2008 database contains a total of 1700 images, with 25 reference images and 17 types of distortions over 4 distortion levels (Figure 7.1 and Table 7.3) [365]. This database can be considered as an extension of the well-known LIVE database [398]. MOS are the results of 654 observers attending the experiments. No viewing distance is stated in the TID database, therefore we have used a standard viewing distance (50 cm) for the metrics requiring this setting. We have used version 1.0 of the TID2008 database, and we have taken the natural logarithm of the results of the metrics when calculating the correlation.
Table 7.3: Overview of the distortions in the TID database and how they are related to the tested subsets. The database contains 17 types of distortions over four distortion levels. The sign "+" indicates that the distortion type was used to alter the images of the subset and the sign "-" that it was not considered for this subset.

<table>
<thead>
<tr>
<th>Type of distortion</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noise</td>
</tr>
<tr>
<td>1 Additive Gaussian noise</td>
<td>+</td>
</tr>
<tr>
<td>2 Noise in color components</td>
<td>-</td>
</tr>
<tr>
<td>3 Spatially correlated noise</td>
<td>+</td>
</tr>
<tr>
<td>4 Masked noise</td>
<td>-</td>
</tr>
<tr>
<td>5 High frequency noise</td>
<td>+</td>
</tr>
<tr>
<td>6 Impulse noise</td>
<td>+</td>
</tr>
<tr>
<td>7 Quantization noise</td>
<td>+</td>
</tr>
<tr>
<td>8 Gaussian blur</td>
<td>+</td>
</tr>
<tr>
<td>9 Image denoising</td>
<td>+</td>
</tr>
<tr>
<td>10 JPEG compression</td>
<td>-</td>
</tr>
<tr>
<td>11 JPEG2000 compression</td>
<td>-</td>
</tr>
<tr>
<td>12 JPEG transmission errors</td>
<td>-</td>
</tr>
<tr>
<td>13 JPEG2000 transmission errors</td>
<td>-</td>
</tr>
<tr>
<td>14 Non eccentricity pattern noise</td>
<td>-</td>
</tr>
<tr>
<td>15 Local block-wise distortion</td>
<td>-</td>
</tr>
<tr>
<td>16 Mean shift</td>
<td>-</td>
</tr>
<tr>
<td>17 Contrast change</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 7.1: The TID2008 database contains 25 reference images with 17 types of distortions over four levels.

The correlations for the overall performance of the IQ metrics are listed in Table 7.4. Here the correlation between all values from the IQ metrics and the subjective scores are found. The best metric for the full database regarding the Pearson correlation is the VIF, followed by VSNR, and UIQ. These are metrics designed and optimized for the distortions found in the TID2008 database. The SSIM is not as good as its predecessor UIQ in terms of Pearson correlation, but it has a higher Spearman correlation indicating a more correct ranking. It is worth noting that metrics based on color difference formulae do not perform well, with S-CIELAB being the best with a correlation of 0.43. The simple color difference formulae are found on the bottom part of the list with a very low correlation for the entire database, both in Pearson and Spearman. The low correlation in many of the IQ metrics can be caused by the high number of different distortions in the TID2008 database, since there might be scale differences between the different distortions and scenes. This problem occurs when the observers rate images with different distortions to have the same quality, but the IQ metrics rate them to be different. However, looking at the same distortion the metric might have a high correlation, but taking the correlation of two or more distortions might be low due to the scale differences (For explanation of scale differences see Figure 6.2).

When looking at specific distortions (Table 7.5), PSNR-HVS-M is the best for five of the seven datasets. However, in the two last datasets PSNR-HVS-M does not perform well, with a correlation of 0.26 in both sets. The reason for this is that PSNR-HVS-M has a problem with the mean shift distortion. The PSNR-HVS-M incorporates a mean shift removal and rates images where a mean shift has occurred as similar even though the observers rate them differently. VSNR and VIF do not have this problem and therefore they are able to obtain a higher correlation for both the exotic and exotic2 dataset. SSIM has a low correlation in the exotic dataset, and a much higher correlation in exotic2. The reason for this is a couple of outliers in the exotic dataset, and with the introduction of more images in exotic2 several points are
Table 7.4: Comparison of the IQ metrics over all the images of the TID2008 database. The results are sorted from high to low for both Pearson and Spearman correlation. VIF, UIQ, and VSNR correlate rather well with subjective evaluation with a Pearson and Spearman correlation above 0.6. $S_{DOG}$-CIELAB is calculated with $R_c=1$, $R_s=2$, pyramid $=1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, ...$, and equal weighting of the level. $S_{DOG}$-DEE is calculated with $R_c=3$, $R_s=4$, pyramid $=1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, ...$, and the variance used as weighting of the level to give more importance to the high-resolution levels. For more information on $S_{DOG}$-CIELAB and $S_{DOG}$-DEE we refer to Appendix E. S-CIELAB$_J$ includes only the modified CSF. For other metrics standard settings have been used, unless otherwise is stated. $N = 1700$. For confidence limits we refer to Figure 7.2.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
<th>Metric</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>0.74</td>
<td>VIF</td>
<td>0.75</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.71</td>
<td>VSNR</td>
<td>0.72</td>
</tr>
<tr>
<td>UIQ</td>
<td>0.62</td>
<td>SSIM</td>
<td>0.63</td>
</tr>
<tr>
<td>PSNR-HVS-M</td>
<td>0.59</td>
<td>PSNR-HVS-M</td>
<td>0.61</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.55</td>
<td>MSE</td>
<td>0.57</td>
</tr>
<tr>
<td>MSE</td>
<td>0.54</td>
<td>SSIM-IPT</td>
<td>0.57</td>
</tr>
<tr>
<td>QC$_{COLOR}$</td>
<td>0.52</td>
<td>QC$_{COLOR}$</td>
<td>0.48</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.51</td>
<td>PSNR</td>
<td>0.56</td>
</tr>
<tr>
<td>SSIM-IPT</td>
<td>0.48</td>
<td>S-CIELAB</td>
<td>0.45</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.43</td>
<td>SHAME</td>
<td>0.41</td>
</tr>
<tr>
<td>SHAME</td>
<td>0.41</td>
<td>S-CIELAB$_J$</td>
<td>0.38</td>
</tr>
<tr>
<td>$S_{DOG}$-CIELAB</td>
<td>0.38</td>
<td>$\Delta E_E$</td>
<td>0.38</td>
</tr>
<tr>
<td>S-CIELAB$_J$</td>
<td>0.32</td>
<td>$S_{DOG}$-CIELAB</td>
<td>0.37</td>
</tr>
<tr>
<td>SHAME</td>
<td>0.30</td>
<td>SHAME</td>
<td>0.35</td>
</tr>
<tr>
<td>S-DEE</td>
<td>0.29</td>
<td>S-CIELAB$_J$</td>
<td>0.31</td>
</tr>
<tr>
<td>ABF</td>
<td>0.28</td>
<td>Hue angle</td>
<td>0.29</td>
</tr>
<tr>
<td>$\Delta E_E$</td>
<td>0.27</td>
<td>S-DEE</td>
<td>0.29</td>
</tr>
<tr>
<td>Hue angle</td>
<td>0.26</td>
<td>$\Delta E^*_{ab}$</td>
<td>0.28</td>
</tr>
<tr>
<td>$S_{DOG}$-DEE</td>
<td>0.26</td>
<td>ABF</td>
<td>0.26</td>
</tr>
<tr>
<td>$\Delta E^*_{ab}$</td>
<td>0.23</td>
<td>$S_{DOG}$-DEE</td>
<td>0.26</td>
</tr>
<tr>
<td>$\Delta E^*_{94}$</td>
<td>-0.06</td>
<td>$\Delta E^*_{00}$</td>
<td>0.24</td>
</tr>
<tr>
<td>$\Delta E^*_{00}$</td>
<td>-0.06</td>
<td>$\Delta E^*_{94}$</td>
<td>0.24</td>
</tr>
</tbody>
</table>
placed between the outliers and the rest increasing the correlation. Some of the problems in
the exotic dataset for SSIM stem from the contrast change distortion, but mainly from scale
differences between the distortions (i.e. one distortion has an overall smaller difference than
the other but they are judged similarly by the observers). Within a specific distortion a fairly
high correlation is found. Amongst the color difference based metrics S-CIELAB is the most
stable metric, with a high correlation in the noise, noise2, hard, and simple dataset. Since
S-CIELAB incorporates spatial filtering it will be able to modulate the frequencies that are
less perceptible, and therefore S-CIELAB shows good results for the datasets with noise. S-
CIELAB has problems in the exotic and exotic2 because of scale differences, as many of the
other metrics. The 95% CI of the correlation coefficients for all r’s are illustrated in Figure 7.2,
with a 95% CI it is a 95% probability that the true correlation coefficient will be within the
upper and lower limit.

![Figure 7.2: Pearson correlation coefficients r plotted against the Pearson correlation coefficient r ± 95% CI. N = 1700. Values calculated using the method described in Section 6.2.1.1. The blue line indicates the Pearson correlation coefficient, red the lower limit (with a 95% CI), and green the upper limit (95% CI). The CI is rather small compared to the CI in Figure 6.1, this is due to a high N.](image-url)
Table 7.5: Correlation for the IQ metrics for the seven different datasets in the TID2008 database. Each dataset is sorted descending from the highest Pearson correlation to the lowest Pearson correlation. N = 700, 800, 700, 800, 400, 400, 600 for the noise, noise2, safe, hard, simple, exotic and exotic2 dataset, respectively. For confidence limits we refer to Figure 6.1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Noise</th>
<th>Noise2</th>
<th>Safe</th>
<th>Hard</th>
<th>Simple</th>
<th>Exotic</th>
<th>Exotic2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR-HVS-M</td>
<td>0.92</td>
<td>0.92</td>
<td>0.81</td>
<td>0.94</td>
<td>0.57</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>PSNR-HVS-M</td>
<td>0.85</td>
<td>0.84</td>
<td>0.79</td>
<td>0.89</td>
<td>0.43</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>VSNR</td>
<td>0.78</td>
<td>0.82</td>
<td>0.78</td>
<td>0.86</td>
<td>0.26</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.77</td>
<td>0.75</td>
<td>0.75</td>
<td>0.83</td>
<td>0.23</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>0.74</td>
<td>0.73</td>
<td>0.74</td>
<td>0.74</td>
<td>0.81</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>PSNR</td>
<td>0.72</td>
<td>0.72</td>
<td>0.71</td>
<td>0.73</td>
<td>0.16</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>SHAMEII</td>
<td>0.63</td>
<td>0.59</td>
<td>0.71</td>
<td>0.77</td>
<td>0.16</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.57</td>
<td>0.58</td>
<td>0.67</td>
<td>0.67</td>
<td>0.12</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>UIQ</td>
<td>0.51</td>
<td>0.51</td>
<td>0.57</td>
<td>0.65</td>
<td>0.04</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>SHAME</td>
<td>0.51</td>
<td>0.51</td>
<td>0.56</td>
<td>0.67</td>
<td>0.03</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>QCOLOR</td>
<td>0.50</td>
<td>0.50</td>
<td>0.52</td>
<td>0.66</td>
<td>0.03</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>SDGOSEE</td>
<td>0.48</td>
<td>0.47</td>
<td>0.48</td>
<td>0.65</td>
<td>0.61</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>SHAME</td>
<td>0.46</td>
<td>0.46</td>
<td>0.48</td>
<td>0.65</td>
<td>0.61</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.46</td>
<td>0.46</td>
<td>0.48</td>
<td>0.65</td>
<td>0.61</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>THAISEL</td>
<td>0.45</td>
<td>0.45</td>
<td>0.44</td>
<td>0.49</td>
<td>0.49</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Hue Angle</td>
<td>0.45</td>
<td>0.43</td>
<td>0.43</td>
<td>0.49</td>
<td>0.49</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>ΔE94</td>
<td>0.44</td>
<td>0.40</td>
<td>0.42</td>
<td>0.45</td>
<td>0.48</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>SSIM</td>
<td>0.43</td>
<td>0.38</td>
<td>0.42</td>
<td>0.43</td>
<td>0.48</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>ΔE94</td>
<td>0.43</td>
<td>0.37</td>
<td>0.41</td>
<td>0.40</td>
<td>0.39</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>S-DEE</td>
<td>0.42</td>
<td>0.36</td>
<td>0.40</td>
<td>0.25</td>
<td>0.36</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>ΔE94</td>
<td>0.41</td>
<td>0.33</td>
<td>0.39</td>
<td>0.25</td>
<td>0.34</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>
7.2 Luminance changed images, Pedersen et al.

The database includes four original images (Figure 7.3) reproduced with different changes in lightness. Each scene has been altered in four ways globally and four ways locally [339, 356]. The globally changed images had an increase in luminance of 3 and 5 $\Delta E_{ab}^*$, and a decrease in luminance of 3 and 5 $\Delta E_{ab}^*$. The local regions were the front and background in two images, a horizontal line in one, and two smaller regions in the last. These regions had changes in luminance of 3 $\Delta E_{ab}^*$. 25 observers participated in the pair comparison experiment, which was carried out on a calibrated CRT monitor, LaCIE electron 22 blue II, in a gray room. The viewing distance was set to 80 cm, and the ambient light measured to 17 lux. The images were judged in a pair comparison setup, with the original in the middle and the two reproductions on each side.

For this database, a group of IQ metrics correlate reasonably well with subjective assessment (Table 7.6). The original hue angle algorithm, SHAME, SHAME-II, S-CIELAB, and ABF exhibit the best Pearson correlation coefficients, all with a Pearson correlation above 0.8. It is also interesting to notice that all of these are based on color differences, using the $\Delta E_{ab}^*$ as a color difference formula. Because the changes in this database have been carried out according to $\Delta E_{ab}^*$, it is not surprising that these metrics perform well. The metrics with a low correlation usually have a problem with the images where a small local change from the original is highly perceivable. It is also interesting to notice that the IQ metrics based on structural similarity, such as SSIM and UIQ, are outperformed by many other metrics. Simple metrics, such as MSE and PSNR, also perform quite well.

7.3 JPEG and JPEG2000 compressed images, Simone et al.

The 10 original images (Figure 7.4) of this database were corrupted by JPEG and JPEG2000 distortions [63, 410]. The images in the database was been selected according to the distortions that are being evaluated following the recommendations of Field [143] and CIE [80]. Four different levels of compression were used for JPEG and JPEG2000, resulting in eight compressed images for each scene giving a total number of 80 test images. The compression rate in Bit Per Pixel ($BPP$) was calculated as:

$$BPP = (\text{file size in bytes} \times 8 / \text{image size in pixels}).$$

(7.1)
Table 7.6: Correlation between subjective scores and metric scores for the luminance changed images from Pedersen et al. [356]. We can see that the metrics based on color difference perform well, compared to the metrics based on structural similarity. The results are sorted from high to low for both Pearson and Spearman correlation. $N = 32$, which gives a CI for instance for SHAME-II for Pearson 0.82 equal to [0.66, 0.91]. For confidence limits we refer to Figure 6.1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
<th>Metric</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHAME-II</td>
<td>0.82</td>
<td>ABF</td>
<td>0.78</td>
</tr>
<tr>
<td>SHAME</td>
<td>0.81</td>
<td>S-CIELAB</td>
<td>0.77</td>
</tr>
<tr>
<td>Hue angle</td>
<td>0.81</td>
<td>S-CIELAB$_J$</td>
<td>0.76</td>
</tr>
<tr>
<td>ABF</td>
<td>0.81</td>
<td>$\Delta E_{ab}^*$</td>
<td>0.74</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.80</td>
<td>Hue angle</td>
<td>0.73</td>
</tr>
<tr>
<td>S-CIELAB$_J$</td>
<td>0.78</td>
<td>SHAME-II</td>
<td>0.73</td>
</tr>
<tr>
<td>$\Delta E_{ab}^*$</td>
<td>0.76</td>
<td>MSE</td>
<td>0.72</td>
</tr>
<tr>
<td>MSE</td>
<td>0.67</td>
<td>PSNR</td>
<td>0.72</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.66</td>
<td>SHAME</td>
<td>0.71</td>
</tr>
<tr>
<td>$\Delta E_{94}^*$</td>
<td>0.65</td>
<td>PSNR-HVS-M</td>
<td>0.68</td>
</tr>
<tr>
<td>$\Delta E_{E}^*$</td>
<td>0.63</td>
<td>$\Delta E_{94}^*$</td>
<td>0.68</td>
</tr>
<tr>
<td>PSNR-HVS-M</td>
<td>0.63</td>
<td>$\Delta E_{00}^*$</td>
<td>0.68</td>
</tr>
<tr>
<td>$\Delta E_{00}^*$</td>
<td>0.63</td>
<td>$\Delta E_{E}^*$</td>
<td>0.64</td>
</tr>
<tr>
<td>UIQ</td>
<td>0.45</td>
<td>SSIM-IPT</td>
<td>0.51</td>
</tr>
<tr>
<td>S-DEE</td>
<td>0.39</td>
<td>S-DEE</td>
<td>0.51</td>
</tr>
<tr>
<td>VIF</td>
<td>0.39</td>
<td>UIQ</td>
<td>0.49</td>
</tr>
<tr>
<td>SSIM-IPT</td>
<td>0.30</td>
<td>SSIM</td>
<td>0.49</td>
</tr>
<tr>
<td>$Q_{COLOR}$</td>
<td>0.28</td>
<td>$Q_{COLOR}$</td>
<td>0.48</td>
</tr>
<tr>
<td>$S_{DOG}$-DEE</td>
<td>0.24</td>
<td>VIF</td>
<td>0.35</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.22</td>
<td>$S_{DOG}$-DEE</td>
<td>0.32</td>
</tr>
<tr>
<td>$S_{DOG}$-CIELAB</td>
<td>0.20</td>
<td>$S_{DOG}$-CIELAB</td>
<td>0.32</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.02</td>
<td>VSNR</td>
<td>0.02</td>
</tr>
</tbody>
</table>
For the experiment the compression rates were chosen such that JPEG and JPEG2000 have similar values of the BPP. Table 7.7 shows the BPP used to compress each scene.

The images were used in a category judgment experiment, where the original was presented on one side and the compressed on the other side. Each pair of images (original and compressed) was displayed on an Eizo ColorEdge CG241W digital LCD display. The monitor was calibrated and profiled using GretagMacbeth Eye-One Match 3. The settings on the monitor were sRGB with 40% of brightness and a resolution of 1600 × 1200 pixels. The white point was set to the D65 white point and the gamma was set to a value of 2.2. The display was placed at a viewing distance of 70 cm. A total of 18 observers participated in the experiment.

![Image links](b) Barbara (c) Cafe (d) Flower (e) House (f) Mandrian (g) Parrots (h) Picnic (i) Poster (j) Sails

Figure 7.4: The 10 images in the dataset from Simone et al. [63, 410]. These images have been compressed with JPEG and JPEG2000.

Table 7.7: Selected bit per pixels for each image for the dataset from Simone et al. [63, 410].

<table>
<thead>
<tr>
<th>Image scene</th>
<th>BPP JPEG</th>
<th>BPP JPEG2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image 1: Balls</td>
<td>0.8783</td>
<td>0.9331</td>
</tr>
<tr>
<td>Image 2: Barbara</td>
<td>0.9614</td>
<td>1.0482</td>
</tr>
<tr>
<td>Image 3: Cafe</td>
<td>0.8293</td>
<td>0.9080</td>
</tr>
<tr>
<td>Image 4: Flower</td>
<td>0.6611</td>
<td>0.7226</td>
</tr>
<tr>
<td>Image 5: House</td>
<td>0.9817</td>
<td>1.0500</td>
</tr>
<tr>
<td>Image 6: Mandrian</td>
<td>1.4482</td>
<td>1.5910</td>
</tr>
<tr>
<td>Image 7: Parrots</td>
<td>0.6230</td>
<td>0.6891</td>
</tr>
<tr>
<td>Image 8: Picnic</td>
<td>1.0353</td>
<td>1.1338</td>
</tr>
<tr>
<td>Image 9: Poster</td>
<td>0.4148</td>
<td>0.4474</td>
</tr>
<tr>
<td>Image 10: Sails</td>
<td>0.7518</td>
<td>0.8198</td>
</tr>
</tbody>
</table>

We found that all the metrics have low performance for the images from this database (Table 7.8), spanning a range in correlation below 0.4. This is probably because these particular images were initially selected in order to determine the Just Noticeable Distortion (JND). Only small distortions were applied to the original images making it arduous for the observers to assess IQ, and therefore also very difficult for the IQ metrics. PSNR-HVS-M and PSNR have the best Pearson correlation, but they do not perform well with a Pearson correlation of

---

Look at the numbers and consider the implications for image compression and quality assessment. The BPP values give insight into the compression levels for each image, which is crucial for understanding how different metrics perform under these conditions. The experiment setup, including monitor calibration and viewing conditions, is essential for obtaining reliable results in image quality assessment.
Table 7.8: Correlation between subjective scores and metric scores for the JPEG and JPEG2000 compressed images from Simone et al. [63, 410]. The results are sorted from high to low for both Pearson and Spearman correlation. All IQ metrics have a low correlation, both for Pearson and Spearman correlation. However, it is interesting to notice the more correct ranking by the color difference based IQ metrics, such as S-CIELAB and SHAME. N = 80 for this database. For confidence limits we refer to Figure 6.1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
<th>Metric</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.31</td>
<td>SSIM-IPT</td>
<td>0.37</td>
</tr>
<tr>
<td>PSNR-HVS-M</td>
<td>0.31</td>
<td>SHAME</td>
<td>0.36</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.29</td>
<td>PSNR-HVS-M</td>
<td>0.36</td>
</tr>
<tr>
<td>MSE</td>
<td>0.26</td>
<td>S-CIELAB</td>
<td>0.36</td>
</tr>
<tr>
<td>SSIM-IPT</td>
<td>0.25</td>
<td>SSIM</td>
<td>0.35</td>
</tr>
<tr>
<td>S&lt;sub&gt;DOG&lt;/sub&gt;-CIELAB</td>
<td>0.23</td>
<td>S-CIELAB</td>
<td>0.33</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.22</td>
<td>ABF</td>
<td>0.33</td>
</tr>
<tr>
<td>VIF</td>
<td>0.21</td>
<td>PSNR</td>
<td>0.32</td>
</tr>
<tr>
<td>S-CIELAB&lt;sub&gt;J&lt;/sub&gt;</td>
<td>0.17</td>
<td>MSE</td>
<td>0.30</td>
</tr>
<tr>
<td>ABF</td>
<td>0.16</td>
<td>ΔE&lt;sub&gt;ab&lt;/sub&gt;</td>
<td>0.28</td>
</tr>
<tr>
<td>ΔE&lt;sub&gt;ab&lt;/sub&gt;</td>
<td>0.16</td>
<td>VIF</td>
<td>0.24</td>
</tr>
<tr>
<td>Q&lt;sub&gt;COLOR&lt;/sub&gt;</td>
<td>0.15</td>
<td>ΔE&lt;sub&gt;E&lt;/sub&gt;</td>
<td>0.24</td>
</tr>
<tr>
<td>SHAME</td>
<td>0.15</td>
<td>SHAME-II</td>
<td>0.24</td>
</tr>
<tr>
<td>ΔE&lt;sub&gt;00&lt;/sub&gt;</td>
<td>0.13</td>
<td>Hue angle</td>
<td>0.23</td>
</tr>
<tr>
<td>S-DEE</td>
<td>0.13</td>
<td>Q&lt;sub&gt;COLOR&lt;/sub&gt;</td>
<td>0.21</td>
</tr>
<tr>
<td>ΔE&lt;sub&gt;E&lt;/sub&gt;</td>
<td>0.13</td>
<td>ΔE&lt;sub&gt;00&lt;/sub&gt;</td>
<td>0.20</td>
</tr>
<tr>
<td>ΔE&lt;sub&gt;94&lt;/sub&gt;</td>
<td>0.13</td>
<td>S-DEE</td>
<td>0.19</td>
</tr>
<tr>
<td>S&lt;sub&gt;DOG&lt;/sub&gt;-DEE</td>
<td>0.11</td>
<td>ΔE&lt;sub&gt;94&lt;/sub&gt;</td>
<td>0.19</td>
</tr>
<tr>
<td>SHAME-II</td>
<td>0.09</td>
<td>S&lt;sub&gt;DOG&lt;/sub&gt;-CIELAB</td>
<td>0.18</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.08</td>
<td>S&lt;sub&gt;DOG&lt;/sub&gt;-DEE</td>
<td>0.14</td>
</tr>
<tr>
<td>Hue angle</td>
<td>0.07</td>
<td>UIQ</td>
<td>0.11</td>
</tr>
<tr>
<td>UIQ</td>
<td>0.05</td>
<td>VSNR</td>
<td>0.11</td>
</tr>
</tbody>
</table>

only 0.31. It is interesting to notice the more correct ranking by the color difference based IQ metrics, such as S-CIELAB and SHAME, but still they are not performing very well.

This dataset is similar to the simple dataset in the TID2008 dataset, that contains JPEG and JPEG2000 compressed images in addition to gaussian blur and additive gaussian noise (Table 7.3). However, there is a large difference in the performance of the metrics, for example PSNR-HVS-M had a correlation of 0.94 for the simple dataset and only a correlation 0.31 for the images from Simone et al. [63, 410]. The reason for this difference is the visual differences between the images in the dataset, in the TID2008 database the differences between the different levels of a distortion are high while in the images from Simone et al. [63, 410] they are just noticeable. This results in a difficult task for the observers and of course very difficult for an IQ metric.
7.4 IVC database, Le Callet et al.

The IVC database contains blurred images and images distorted by three types of lossy compression techniques - JPEG, JPEG2000, and Locally Adaptive Resolution (LAR) [58]. A total of 10 different images are found in this database (Figure 7.5). The viewing distance was set to 87 cm. Only the color images in the database have been computed.

The most accurate IQ metrics are UIQ, VSNR, and VIF (Table 7.9). It is interesting to note that these metrics are based on different approaches; the VIF is based on statistics, UIQ on structural similarity, and VSNR on contrast thresholds. Many of the gray scale metrics have a higher performance than the color based metrics, this could be explained by the fact that the changes in this database are not directly related to color. The IVC database is very similar to the simple dataset from the TID2008 database, since both contain blurring and compression. We can also notice that the IQ metrics with a high correlation for the simple dataset in TID2008 also have a high correlation for the IVC database, reflecting the similarities between the two datasets.

7.5 Contrast, lightness, and saturation alterations, Ajagamelle et al.

This database contains a total of 10 original images (Figure 7.6) covering a wide range of characteristics and scenes [6]. The images were modified on a global scale with separate and simultaneous variations of contrast, lightness, and saturation. The experiment was carried out as a category judgment experiment with 14 observers. The viewing distance was set to 70 cm, and the ambient illumination was 40 lux.

PSNR-HVS-M has the best Pearson correlation, while SHAME-II has the best Spearman correlation (Table 7.10). We can notice that the differences between the metrics are low, for the Spearman correlation 13 metrics have a correlation between 0.71 and 0.79. The color dif-
Table 7.9: Correlation between subjective scores and metric scores for the images from the IVC database. The results are sorted from high to low for both Pearson and Spearman correlation. $N = 185$ for the IVC database. For confidence limits we refer to Figure 6.1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
<th>Metric</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>0.88</td>
<td>VIF</td>
<td>0.90</td>
</tr>
<tr>
<td>UIQ</td>
<td>0.82</td>
<td>UIQ</td>
<td>0.83</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.77</td>
<td>VSNR</td>
<td>0.78</td>
</tr>
<tr>
<td>PSNR-HVS-M</td>
<td>0.73</td>
<td>SSIM</td>
<td>0.78</td>
</tr>
<tr>
<td>SSIM-IPT</td>
<td>0.71</td>
<td>SSIM-IPT</td>
<td>0.77</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.70</td>
<td>PSNR-HVS-M</td>
<td>0.77</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.67</td>
<td>MSE</td>
<td>0.69</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.61</td>
<td>PSNR</td>
<td>0.69</td>
</tr>
<tr>
<td>QCOLOR</td>
<td>0.61</td>
<td>SHAME</td>
<td>0.68</td>
</tr>
<tr>
<td>(\Delta E)</td>
<td>0.56</td>
<td>(\Delta E_{ab})</td>
<td>0.66</td>
</tr>
<tr>
<td>SHAME</td>
<td>0.55</td>
<td>QCOLOR</td>
<td>0.66</td>
</tr>
<tr>
<td>S_DOG-DEE</td>
<td>0.55</td>
<td>S-CIELAB</td>
<td>0.66</td>
</tr>
<tr>
<td>(\Delta E_{ab})</td>
<td>0.54</td>
<td>(\Delta E)</td>
<td>0.62</td>
</tr>
<tr>
<td>ABF</td>
<td>0.51</td>
<td>S_DOG-DEE</td>
<td>0.58</td>
</tr>
<tr>
<td>MSE</td>
<td>0.51</td>
<td>ABF</td>
<td>0.56</td>
</tr>
<tr>
<td>S-DEE</td>
<td>0.46</td>
<td>Hue angle</td>
<td>0.52</td>
</tr>
<tr>
<td>S-CIELAB_I</td>
<td>0.42</td>
<td>(\Delta E_{00})</td>
<td>0.47</td>
</tr>
<tr>
<td>(\Delta E_{94})</td>
<td>0.38</td>
<td>(\Delta E_{94})</td>
<td>0.46</td>
</tr>
<tr>
<td>(\Delta E_{94})</td>
<td>0.37</td>
<td>S-DEE</td>
<td>0.46</td>
</tr>
<tr>
<td>SHAME-II</td>
<td>0.37</td>
<td>S-CIELAB_I</td>
<td>0.43</td>
</tr>
<tr>
<td>Hue angle</td>
<td>0.35</td>
<td>SHAME-II</td>
<td>0.42</td>
</tr>
<tr>
<td>S_DOG-CIELAB</td>
<td>0.29</td>
<td>S_DOG-CIELAB</td>
<td>0.32</td>
</tr>
</tbody>
</table>
ference formulae from CIE ($\Delta E_{ab}^*$, $\Delta E_{94}$, and $\Delta E_{00}$) all perform very well, the same also with PSNR. This is different from many of the other databases, where these simple pixelwise metrics do not correlate with perceived IQ. The high correlation over some of the other databases is probably due to the global variations made by Ajagamelle et al., since the pixelwise IQ metrics weight all pixels equally. The results from this database are similar to the results from the lightness changed images from Pedersen et al. [356].

7.6 Gamut mapped images, Dugay et al.

In this dataset, 20 original images (Figure 6.4) were gamut mapped with five different algorithms [107, 108]: HPminDE, SGCK, Zolliker, Kolås, and Gatta. For more information on the gamut mapping algorithms we refer the reader to Section 6.3.1.1.

The 20 different images were evaluated by 20 observers in a pair comparison experiment. The monitor used was a Dell 2407WFP LCD display calibrated with a D65 white point and a gamma of 2.2. The level of ambient illumination on the monitor was around 20 lux. The viewing distance for the observers was approximately 50 cm.

We see from the results in Table 7.11 that most of the metrics does not correlate well with perceived quality. In gamut mapping many different attributes are changed simultaneously, and because of this the objective assessment is very complex. Previous research has also shown that IQ metrics have problems when multiple distortions occur simultaneously, as in gamut mapping [41, 169]. This is not the case for TID2008 and some of the other databases evaluated here, since usually only one attribute changes at the time.

7.7 Overall performance of the metrics

Figure 7.7 shows the performance for all IQ metrics for the six different databases. We can see that none of the 22 evaluated metrics have a high correlation for all databases. If we discard the JPEG and JPEG2000 compressed images from Carraciolo et al. and the gamut mapped images from Dugay et al. some metrics perform rather well, such as the PSNR-HVS-M and PSNR, which have a correlation of above 0.5 for the remaining databases. However, the performance seems to be database or distortion dependent, and the metrics perform better
Table 7.10: Correlation between subjective scores and metric scores for the images from Ajagamelle et al. [6]. The results are sorted from high to low for both Pearson and Spearman correlation. PSNR-HVS-M has the highest Pearson correlation, while SHAME-II has the best Spearman correlation. VSNR has the lowest correlation, which is caused by scale differences between the difference images. $N = 80$. For confidence limits we refer to Figure 6.1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
<th>Metric</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR-HVS-M</td>
<td>0.79</td>
<td>SHAME-II</td>
<td>0.79</td>
</tr>
<tr>
<td>$\Delta E_{ab}^*$</td>
<td>0.75</td>
<td>SSIM-IPT</td>
<td>0.76</td>
</tr>
<tr>
<td>$\Delta E_{94}^*$</td>
<td>0.74</td>
<td>ABF</td>
<td>0.76</td>
</tr>
<tr>
<td>$\Delta E_{00}^*$</td>
<td>0.74</td>
<td>PSNR-HVS-M</td>
<td>0.75</td>
</tr>
<tr>
<td>ABF</td>
<td>0.73</td>
<td>QCOLOR</td>
<td>0.75</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.72</td>
<td>UIQ</td>
<td>0.73</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>0.67</td>
<td>SHAME</td>
<td>0.73</td>
</tr>
<tr>
<td>SSIM-IPT</td>
<td>0.66</td>
<td>S-CIELAB</td>
<td>0.72</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.64</td>
<td>Hue Angle</td>
<td>0.72</td>
</tr>
<tr>
<td>UIQ</td>
<td>0.63</td>
<td>$\Delta E_{ab}^*$</td>
<td>0.72</td>
</tr>
<tr>
<td>Hue Angle</td>
<td>0.62</td>
<td>SSIM</td>
<td>0.72</td>
</tr>
<tr>
<td>SHAME-II</td>
<td>0.62</td>
<td>MSE</td>
<td>0.71</td>
</tr>
<tr>
<td>QCOLOR</td>
<td>0.62</td>
<td>PSNR</td>
<td>0.71</td>
</tr>
<tr>
<td>MSE</td>
<td>0.61</td>
<td>$\Delta E_{00}^*$</td>
<td>0.69</td>
</tr>
<tr>
<td>DEE</td>
<td>0.60</td>
<td>$\Delta E_{94}^*$</td>
<td>0.69</td>
</tr>
<tr>
<td>$S_{DOG-DEE}$</td>
<td>0.59</td>
<td>DEE</td>
<td>0.68</td>
</tr>
<tr>
<td>S-CIELAB$J$</td>
<td>0.58</td>
<td>S-CIELAB$J$</td>
<td>0.65</td>
</tr>
<tr>
<td>VIF</td>
<td>0.53</td>
<td>$S_{DOG-DEE}$</td>
<td>0.61</td>
</tr>
<tr>
<td>SHAME</td>
<td>0.50</td>
<td>VIF</td>
<td>0.59</td>
</tr>
<tr>
<td>$S_{DOG-CIELAB}$</td>
<td>0.41</td>
<td>$S_{DOG-CIELAB}$</td>
<td>0.48</td>
</tr>
<tr>
<td>S-DEE</td>
<td>0.40</td>
<td>S-DEE</td>
<td>0.41</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.16</td>
<td>VSNR</td>
<td>0.14</td>
</tr>
</tbody>
</table>
Table 7.11: Correlation between subjective scores and metric scores for the gamut mapped images from Dugay et al. [107, 108]. The results are sorted from high to low for both Pearson and Spearman correlation. $N = 100$. For confidence limits we refer to Figure 6.1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
<th>Metric</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>0.31</td>
<td>VIF</td>
<td>0.27</td>
</tr>
<tr>
<td>UIQ</td>
<td>0.31</td>
<td>UIQ</td>
<td>0.19</td>
</tr>
<tr>
<td>$Q_{COLOR}$</td>
<td>0.19</td>
<td>$Q_{COLOR}$</td>
<td>0.12</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.16</td>
<td>VSNR</td>
<td>0.09</td>
</tr>
<tr>
<td>MSE</td>
<td>0.09</td>
<td>SSIM</td>
<td>0.05</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.09</td>
<td>S-CIELAB$_J$</td>
<td>0.03</td>
</tr>
<tr>
<td>PSNR-HVS-M</td>
<td>0.07</td>
<td>PSNR-HVS-M</td>
<td>0.03</td>
</tr>
<tr>
<td>$S_{DOG}$-CIELAB</td>
<td>0.05</td>
<td>$AE^*_{00}$</td>
<td>0.00</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.04</td>
<td>$AE^*_{04}$</td>
<td>0.00</td>
</tr>
<tr>
<td>$AE^*_0$</td>
<td>0.03</td>
<td>SSIM-IPT</td>
<td>-0.02</td>
</tr>
<tr>
<td>$S_{DOG}$-DEE</td>
<td>0.02</td>
<td>MSE</td>
<td>-0.02</td>
</tr>
<tr>
<td>S-CIELAB$_J$</td>
<td>0.02</td>
<td>$S_{DOG}$-DEE</td>
<td>-0.02</td>
</tr>
<tr>
<td>$AE^*_0$</td>
<td>0.02</td>
<td>PSNR</td>
<td>-0.03</td>
</tr>
<tr>
<td>SSIM-IPT</td>
<td>0.01</td>
<td>$AE_E$</td>
<td>-0.06</td>
</tr>
<tr>
<td>Hue angle</td>
<td>0.00</td>
<td>$AE^*_{ab}$</td>
<td>-0.08</td>
</tr>
<tr>
<td>$AE_E$</td>
<td>-0.01</td>
<td>$S_{DOG}$-CIELAB</td>
<td>-0.08</td>
</tr>
<tr>
<td>$AE^*_{ab}$</td>
<td>-0.03</td>
<td>Hue angle</td>
<td>-0.08</td>
</tr>
<tr>
<td>S-DEE</td>
<td>-0.03</td>
<td>S-DEE</td>
<td>-0.09</td>
</tr>
<tr>
<td>SHAME</td>
<td>-0.03</td>
<td>SHAME</td>
<td>-0.09</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>-0.06</td>
<td>S-CIELAB</td>
<td>-0.11</td>
</tr>
<tr>
<td>ABF</td>
<td>-0.08</td>
<td>ABF</td>
<td>-0.11</td>
</tr>
<tr>
<td>SHAME-II</td>
<td>-0.10</td>
<td>SHAME-II</td>
<td>-0.13</td>
</tr>
</tbody>
</table>
for specific distortions, such as VIF for the IVC database with compression, and ABF and SHAME-II for the luminance changed images from Pedersen et al.

7.8 Summary

In this chapter we have evaluated IQ metrics from the survey in Section 4.2. For the evaluation different state of the art databases have been used. However, no database with printed images is available, resulting in that evaluation of metrics for printed images require new databases. The results of the evaluation of the different databases indicate that the observed performance of the different metrics can depend on both the type of distortion and on the image database that is used.
Figure 7.7: Performance of the IQ metrics across the databases. The figure shows the Pearson correlation coefficients for six different sets of test images and 22 IQ metrics.
PART III

IMAGE QUALITY METRICS FOR THE EVALUATION OF PRINTING WORKFLOWS
Subjective assessment of print quality can be carried out by asking a group of observers about the quality of the printed image. However, assessment of a printed image by IQ metrics is more complicated. The original image is of a digital format and the printed image is of an analog format, and because of this the printed image must be digitized before IQ assessment can be carried out. In this chapter we discuss the transformation from a physical reproduction to a digital reproduction with the goal of proposing a framework for using IQ metrics to evaluate the quality of color prints.

8.1 State of the art of frameworks to digitize printed images

A couple of frameworks have been proposed for using IQ metrics on printed images. All these frameworks follow the same procedure. Firstly, the printed image is scanned. Then image registration is performed to match the scanned image with the original. This is usually followed by a descreening procedure to remove halftoning patterns. Finally, an IQ metric can be applied.

Grice and Allebach [163] proposed a framework for using the metrics specified in ISO/IEC 13660 [213]. The framework is based on a test target that is printed and scanned before a set of metrics are applied. Different scanning resolutions are investigated, and the authors concluded that a scanning resolution of 400 dpi might be too low for some metrics. The metrics applied were computational metrics (i.e. not design to account for the HVS) defined by ISO [213], such as metrics for contrast, mottle, darkness, and raggedness.

A framework was proposed by Zhang et al. [498] in 1997. First, the image is scanned, and then three additional scans are performed, each with a different color filter. This results in enough information to transform the image correctly to CIEXYZ. The printed image was scanned with an Agfa Horizon flatbed scanner, and the scanning resolution was set to 1200 dpi. In this article no information about the image registration was given, or on the descreening procedure. The applied IQ metric was S-CIELAB [499], and the images were color patches.

Another framework was proposed by Yanfang et al. [483] in 2008. Before printing two control points are applied to the image, one to the upper left corner and one to the upper center, to help in the registration. The images were scanned with a Kodak i700 at 300 DPI before registration, where the two control points are used for matching the printed image with the original (Figure 8.1). Descreening was performed by the scanner at 230 LPI. No information is given regarding the scaling of the image. The applied IQ metric was S-CIELAB [499].
Recently, Eerola et al. [110, 111] proposed a new framework (Figure 8.2), which follows the same steps as the previous frameworks. The printed reproduction is scanned, and then both the original and the reproduction go through a descreening procedure, which is performed using a Gaussian low-pass filter. Further, image registration is carried out, where local points are used with a Scale-Invariant Feature Transform (SIFT). A RANd SAmple Consensus principle (RANSAC) was used to find the best homography. This is different from the previous frameworks, since it uses local points instead of control points. Scaling was performed using bicubic interpolation. Scanner resolution was set to 1250 DPI, and in Eerola et al. [110] LABMSE was the applied IQ metric. Several other metrics, as SSIM [458], Noise Quality Measure (NQM) [99], PSNR, and UIQ [455], were used in Eerola et al. [111].

Figure 8.1: Framework by Yanfang et al. [483]

8.2 A framework based on control points

We modify and propose a framework similar to the framework by Yanfang et al. [483], which performs image registration based on control points. These control points act as the basis for the image registration, where the control points are used to perform different transformation procedures. The advantage of using control points is that uniform images can be registered, opposed to frameworks based on local points.

As a first step in our framework, the image is padded with a white border and equipped with four control points prior to printing. These points are black squares placed just outside the four corners of the image. Then the image is printed with a given device, such as an ink jet printer. Further, the image is scanned with a scanner. The ICC profile of the scanner is
then assigned to the scanned image in order to achieve a correct description of the colors. The next step in the framework is to find the coordinates of the center of the control points, in both the original image and the scanned image. This is done by a simple automatic routine based on detection of squares. The scanned image can be affected by several geometric distortions, such as translation, scaling and rotation. Because of this, image registration must be carried out. The coordinates for the control points, in both images, are used to create a transformation for the registration. There are several possible transformation types for doing the registration, which are discussed below. In addition the interpolation method for scaling also has several possible methods, which are compared below. After the scanned image has been registered to match the original, a simple procedure is applied to remove the white padding and the control points. Finally, an IQ metric can be used to calculate the quality of the printed image. An overview of the framework is shown in Figure 8.3. This framework differs from the one from Eerola et al. [110], not only in the registration method, but also in the descreening. In our modified framework we do not perform any specific descreening, but we leave this to the IQ metrics in order to avoid a double filtering of the image. It is therefore recommended that the IQ metrics perform some kind of simulation of the HVS, for example spatial filtering based on contrast sensitivity.

![Figure 8.3: Overview of the proposed framework for using image quality metrics with printed images.](image)

### 8.2.1 Transformation type

The transformation type to be used with the framework must use the same or fewer number of control points added to the image. In order to find the most appropriate transformation, we have tested five different transformation types:

- **Nonreflective similarity** uses two pairs of control points, and is suitable when the image is distorted by some combination of translation, rotation, and scaling.

  \[
  \begin{bmatrix}
  u \\
  v 
  \end{bmatrix} = \begin{bmatrix} x & y & 1 \end{bmatrix} T,
  \]

  where \( x \) and \( y \) are the dimensions to be transformed, \( T \) is a 3-by-2 matrix that depends on 4 parameters, and \( u \) and \( v \) are the transformed dimensions.

- **Similarity** resembles the previous, but uses reflection in addition.

- **Affine** uses three pairs of control points, and is useful when the image exhibit shearing.

  \[
  \begin{bmatrix}
  u \\
  v 
  \end{bmatrix} = \begin{bmatrix} x & y & 1 \end{bmatrix} T,
  \]

  where \( x \) and \( y \) are the dimensions to be transformed, \( T \) is a 3-by-2 matrix that depends on 4 parameters, and \( u \) and \( v \) are the transformed dimensions.
where $T$ is a 3-by-2 matrix where all six elements can be different.

- Projective uses four pair of control points, and is commonly used when the images appears tilted.
  \[
  \begin{bmatrix}
  w_p \\
  w_p \\
  w
  \end{bmatrix} = \begin{bmatrix}
  x & y & w
  \end{bmatrix} T,
  \]
  (8.3)
  where $T$ is a 3-by-3, where all nine elements can be different.

- Piecewise linear uses four pair of control points, and is used when parts of the image appear distorted differently. This transformation method applies affine transformations separately to triangular regions of the image.

Theoretically, translation, rotation, and scaling are needed to register a scanned image. However, we wanted to test other transformation methods as well, such as to account for shearing and local distortions. In order to find the most appropriate transformation type several tests have been performed, which are discussed below.

### 8.2.2 Interpolation method

The interpolation method used for scaling the images are important, there are three main methods for doing this:

- bicubic interpolation,
- bilinear interpolation,
- nearest-neighbor interpolation.

The first method considers a $4 \times 4$ neighborhood, and accounts for the distance to the unknown pixel in the calculation. The second method considers a $2 \times 2$ neighborhood, and it takes the weighted average of these four pixels to arrive at its final interpolated value. The two first methods also incorporate an anti-aliasing filter to suppress moire patterns [451], which is a lowpass filter. The latter method simply selects the value of the nearest point, and does not consider the values of other neighboring points. To find the most appropriate interpolation method, i.e. the method introducing the least error, several tests have been carried out.

### 8.2.3 Evaluation of different transformation types

To ensure the best possible registration we evaluate the different transformation types (five different types as mentioned above) together with the different interpolation types\(^1\) (three different types), which results in totally 15 combinations. These combinations are applied to three different images shown in Figure 8.4. Before applying these combinations, the image, which is going to be registered, has been rotated and resized so that all the most relevant problems during the scanning procedure are faced. Finally, the results have been compared with the original image.

The test images are selected to have a good combination of what would be expected for printed images, containing both uniform areas and high frequency areas. The first image (Figure 8.4(a)) is a typical scenery, the second image (Figure 8.4(b)) is an image with a fairly

\(^1\)Standard parameters in Matlab have been used for the interpolation methods.
**A FRAMEWORK FOR USING METRICS WITH COLOR PRINTS**

Figure 8.4: Images used for evaluating different transformation types and interpolation methods. The images differ in detail content, from having large uniform areas to high frequency content. This will reveal differences between the methods using local points and control points.

uniform background, and the third image (Figure 8.4(c)) is an image with fine details. Insisting on the uniformity and nonuniformity of the background is because edges have a great influence on the errors when comparing the original and the registered image.

Figures 8.5 and 8.6 show the MSE between the original image and the registered image after applying the combinations of different transformation and interpolation for Figure 8.4(a) and Figure 8.4(c). Results for Figure 8.4(b) are similar to the results for Figure 8.4(a), and are not shown. It should be kept in mind that the MSE values in the figures are normalized with the highest value for the combinations.

**Figure 8.5: MSE between the original and registered image calculated for Figure 8.4(a)**

As it can be seen, using a "similarity" transformation and a "bilinear" interpolation has the lowest MSE value. Inspection of the difference between the original image and the registered image show that the errors using this combination of transformation and interpolation occur on the edges. For the image in Figure 8.4(c) containing a lot of edges and details, bilinear interpolation gives the lowest error of the interpolation methods for all five transformation types as seen in Figure 8.6. In addition to our results, research carried out by Acharya and
Tsai [2] indicates that bilinear interpolation has the least error when reducing the size of an image, supporting our findings.

8.2.4 Scanner

There are two different scanner types available; Charge Coupled Device (CCD) and Contact Image Sensor (CIS). The disadvantage with a CIS scanner is the lack of depth-of-view [338]. However, this is not a problem with prints, since the print is usually flat on the scan surface.

8.2.4.1 Scanning resolution

A central issue while scanning is to have a resolution high enough to capture the perceived details of the printed image. In order to find the appropriate resolution an image was scanned with a Microtek ScanMaker 9800XL scanner at different resolutions. The scanner was calibrated using an IT8.7 Target and the profile was created with ProfileMaker.

An image (Figure 8.7) was prepared with control points just outside the corners, and it was printed with an Océ ColorWave 600 wide format printer on Océ Red Label paper at a resolution of 150 pixels per inch. The printed image was then scanned with the following resolutions: 72, 100, 150, 300, and 600 DPI without any automatic corrections. The scanned image was used as input to the framework, where the errors were investigated using MSE and $\Delta E^*_{ab}$. From the results in Table 8.1 we see that the values fluctuate, and do not give any indication of the resolution needed. Since the printed image will be different from the original, just finding the lowest objective value is not appropriate, because of this further investigation is needed.

The next step after the objective evaluation of the scanning resolution is a subjective evaluation. Visual inspections of the images reveal a higher detail level in the 600 DPI scans, than at lower resolutions. However, at a normal viewing distance (60 cm) these details are not apparent in the printed image. The scanner resolution is also dependent on the visual angle the

Figure 8.6: MSE between the original and registered image calculated for Figure 8.4(c)
Figure 8.7: Scanned test images with control points outside the four corners.

Table 8.1: Scanning resolution test. MSE is in the 10^4.

<table>
<thead>
<tr>
<th>Resolution (dpi)</th>
<th>MSE</th>
<th>ΔE_{ab}^*</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>7.5</td>
<td>14.5</td>
</tr>
<tr>
<td>100</td>
<td>6.5</td>
<td>12.5</td>
</tr>
<tr>
<td>150</td>
<td>5.7</td>
<td>12.1</td>
</tr>
<tr>
<td>300</td>
<td>5.9</td>
<td>12.9</td>
</tr>
<tr>
<td>600</td>
<td>6.1</td>
<td>14.6</td>
</tr>
</tbody>
</table>
prints are evaluated at, and can also be dependent on the IQ metrics Grice and Allebach [163]. Metrics simulating the HVS usually perform some sort of blurring (filtering) of the image and they might require lower resolution than computational metrics that do not simulate the HVS. The edge raggedness and edge blur metrics as defined by ISO [197] requires a scanning resolution of at least 600 DPI. Lee et al.’s [262] implementation of the edge raggedness and edge blur metrics require a resolution of 1200 DPI. Lim and Mani [267] state that a resolution of 600 DPI should be sufficient to capture the details of a print. Zeise et al. [492] commented that a resolution of 400 DPI would be sufficient, but in many cases a higher resolution is need to limit aliasing artifacts from high-frequency image fluctuations. ISO/IEC 13360 standard [197] suggests a resolution of minimum 600 DPI for evaluation of binary monochrome text and graphic images. Stanek and Burns [424] scanning parameters for print analysis, and they concluded that a resolution higher than 800 DPI is not necessary when looking at the noise power spectrum of different resolutions. It should also be noted that the scanning resolution can be dependent on the evaluation, for situations where the viewing distance is large a lower resolution might be adequate since the image will be heavily filtered, for situations where the viewing distance is short higher resolutions might be required.

8.2.4.2 Scanner ICC profile

The ICC profile for the scanner should be created with a test target printed on the same printer as the image set, since the profile is dependent on the equipment [259, 260, 334]. A profile created with another setup, or a standard profile, might not match your individual device. In the case where several printers (with different ink) and a variety of papers, a profile should be generated for each combination of paper and ink/printer.

8.2.4.3 Cropping of the scan

When scanning the printed images the operator will crop the image, which might result in differences in the registered image. In order to investigate if cropping influences the registration in the framework, a printed image was scanned several times with different crop regions. 14 differences scans at different angles were carried out with three different crop regions. As a measure of variation in the registration process we use the Coefficient of Variation (CV), which is defined as:

$$CV = \frac{\sigma}{|\mu|},$$

(8.4)

where $\sigma$ is the standard deviation, and $\mu$ the mean.

For each scan the rotation angle is registered, and for the fourteen different angles we obtain an average CV as low as 0.01, indicating very small variations in the registration process.

8.2.4.4 Scanner rotation

The angle at which the printed image is scanned at could also possibly influence the registration. We use the same scans as for the previous section (8.2.4.3) to investigate if the rotation influences the registration. The printed image has been scanned 14 times at different angles, and each angle is scanned three times with a different crop, resulting in 42 different registered images. This constitutes the basis for the investigation. For every registered image we
have calculated the IQ values for SSIM [458] (Section 4.2.3.1) and S-CIELAB [499] (Section 4.2.2.1). If the rotation does not influence the final result of the metrics we would expect that the results from the metrics are stable of the different registered images. We use CV (Equation 8.4) as a measure of the variation in the metric results. CV for SSIM is 0.00003 and for S-CIELAB 0.003, both being very small, indicating that the variation between the different rotations are small. We have also plotted the average results of the metrics for the three crops against the average rotation of the three crops to reveal any relationship between them. SSIM shows a correlation of -0.31 (Figure 8.8(a)) and S-CIELAB -0.01 (Figure 8.8(b)). The results do not indicate a relationship between the rotation and the metrics.

![Graph of SSIM](image1)

![Graph of S-CIELAB](image2)

**Figure 8.8**: Metric values plotted against the rotation angle. There is no indication of a relationship between the rotation and the results of S-CIELAB and SSIM. The blue solid line is the linear regression line (r=-0.01 for S-CIELAB and r=-0.31 for SSIM), the red dashed lines indicate the 95% CI of the regression line, and the blue dotted curves indicate the total 95% CI of the linear regression estimate [418].
To ensure that there is no difference between different scanning angles, we have also computed the CV for 25 scans with a rotation less than 0.3 degrees (average rotation 0.17 degrees). The SSIM CV is 0.00005 and for S-CIELAB 0.001, both being very small and similar to the results for the images with a larger rotation. Based on these findings we cannot say that the rotation will result in variations in the final metric values.

8.2.5 A comparison against another framework

The most important aspect to consider when selecting a framework for color prints is that the errors introduced should be as small as possible. From the state of the art two different types of frameworks can be found; those who use local interest points (features), such as the one proposed by Eerola et al. [110], and those who use control points, such as the framework by Yanfang et al. [483] and the modified framework presented above. Both have disadvantages and advantages. For the local features no extra control points need to be added in the process, but local features will not work well for uniform or near-uniform surfaces, since no points can be found in these areas. Using control points can be a disadvantage, since these need to be added prior to printing, however, uniform images can be correctly registered, including color patches. We compare these two different types of IQ frameworks, the one proposed by Eerola et al. [110] and our proposed framework.

We use the images in Figure 8.4 to compare the frameworks. For the proposed modified framework we use bilinear interpolation and similarity as transformation, for the framework by Eerola et al. [110] we use the same settings as in their paper. The best framework should have the least difference between the original image and the registered image. The results for the images are shown in Figure 8.9, and we can clearly see that the proposed framework based on Control Points (CP) introduces less error than the framework based on Local Features (LF), both in terms of absolute difference and mean square error. The biggest difference is found in the image with uniform areas (Figure 8.4(b)), which is a problem for frameworks based on local features since no registrations points can be found. In images with a lot of details the difference in error is almost insignificant, but the proposed framework performs slightly better. Figure 8.9(b) shows the relative error (normalized by the maximum error), and we can see that for three different images we have the same tendency, with the CP based framework introducing much less error than the LF framework.

![Figure 8.9](image)

**Figure 8.9:** Error of the proposed framework compared to the error from the framework from Eerola et al. [110].

In addition, we also checked the computational time used for the registration. The com-
putational time for the proposed framework is stable over the three images. The other framework, which is based on local features, has varying computational time since the content of the image affects the number of registration points. Comparing the time used by the two different frameworks, the proposed framework is more than 20 times faster than the framework by Eerola et al. [110]. The time differences are mainly from the number of registration points. Furthermore, the complexity of a framework based on local features is higher than that of a framework based on control points.

Based on the results shown here, we will use the proposed modified framework based on control points to evaluate IQ.

### 8.3 Summary

In this chapter we have proposed a new framework for digitizing printed images in order to apply IQ metrics. The proposed framework is based on control points for the registration. Investigation shows that our framework outperforms a state of the art framework.
9 Evaluation of Metrics Using the Framework

We have used the framework explained in the previous chapter to evaluate a set of IQ metrics on images from a color workflow.

9.1 Experimental setup

15 images were obtained from Cardin [64] (Figure 9.1). These were processed with two different source profiles, the sRGB v2 perceptual transform and the sRGB v4 perceptual transform [185, 186]. These were further processed with four different softwares for obtaining the destination profile:

- Basic rendering (black point compensation and hue preserving minimum $\Delta E_{ab}^*$).
- LOGO colorful from ProfileMaker Pro.
- LOGO Chroma plus from ProfileMaker Pro.
- Onyx from Onyx Mediacreator.

For each image this results in eight different reproductions, as seen in Figure 9.2. These images were printed, with the Océ ColorWave 600 wide format CMYK printer on Océ Red Label paper, at a resolution of 150 pixels per inch and at a size resulting in a printed reproduction at approximately 8 by 10 inches.

The printed images were evaluated by 30 observers in a category judgment experiment with three categories. The three categories were described as:

1. the most pleasant,
2. neither more nor less pleasant,
3. less pleasant.

The observers were presented with a reference image on an EIZO ColorEdge CG221 display at a color temperature of 6500 Kelvin and luminance level of 80 cd/m$^2$, following the specifications of the sRGB. The image set was rendered for a sRGB display, and therefore a monitor capable of displaying the sRGB gamut was the most adapted reproduction device for this set of images. The printed images were presented randomly in a controlled viewing room at a
Figure 9.1: Images in the experiment. They cover a wide range of characteristics to ensure quality differences in many different quality attributes.

Figure 9.2: Overview of the workflows used in the experiment. The image is either processed using the sRGB v2 or sRGB v4. Furthermore, each of these is used as input to one of the four rendering methods. This finally results in eight different printed reproduction of the same image. Figure inspired by Cardin [64].
color temperature of 5200 Kelvins, an illuminance level of 450 ± 75 lux and a color rendering index of 96. The observers viewed the reference image and the printed image simultaneously from a distance of approximately 60 cm. The experiment was set up to follow the CIE guidelines [80] as closely as possible. More details on the experiment can be found in Cardin [64] or Bonnier et al. [39].

The 15 images printed with the eight different workflows were scanned with a Microtek ScanMaker 9800XL at 600 DPI without any automatic corrections. The scanner was characterized using a Kodak IT8.7 target, since a printed test target was not available, and the profile was built using ProfileMaker 5.0.8. Analysis of the scans showed that the scanner exhibited a slight shearing, which occurs due to mechanical issues. Because of this a transformation method which takes into account shearing must be applied. Even though these methods (affine) introduce more error than the methods that do not correct for shearing (similarity), as seen in Figures 8.5 and 8.6.

9.2 Psychophysical results

30 observers participated in the experiment, where they judged the images on a three step scale. The answers from the observers where processed with the Colour Engineering Toolbox [162] to obtain z-scores. The results are shown in Figure 9.3. Onyx V4 has the highest z-score, but cannot be differentiated from Onyx V2. It is worth noticing the small difference between the highest and lowest z-score. This indicates a low visual difference between the different workflows, and that the task was difficult for the observers. We will focus on the objective evaluation of IQ, but an in depth analysis of the subjective evaluation can be found in Cardin [64].

![Figure 9.3: Z-scores based on 30 observers for 15 images. Solid horizontal lines indicate categories. All images are within the middle category, being "neither more nor less pleasant".](image-url)
9.3 Selection of image quality metrics

A set of IQ metrics were applied to evaluate the IQ of the printed images. Since no direct descreening is applied in the registration process IQ metrics not incorporating aspects of the HVS are most likely inappropriate. A set of metrics has been selected, and a brief introduction of them is given. For further information we refer the reader to their respective references or to Chapter 4.

- S-CIELAB [499] is often used as a reference metric, and has wide acceptance.
- S-CIELAB$_J$ [223] is a modified version of the original S-CIELAB, in terms of the spatial filtering.
- S-DEE [412] extends the S-CIELAB$_J$ with a more refined color difference calculation.
- SHAME [351] combines two state of the art metrics (the hue angle measure [177] and S-CIELAB$_J$). We have included both SHAME-I and SHAME-II, which differ in terms of the spatial filtering. We refer the reader to Chapter 5 for more information.
- Adaptive Bilateral Filter (ABF) [459] uses bilateral filtering to simulate the HVS, before the color difference is calculated using $\Delta E_{ab}^*$.

Further, IQ metrics taking into account structural information might be suitable, since these are potentially good at detecting artifacts.

- SSIM [458], since this is commonly used and has received great publicity since it was proposed in 2004. This metric works on a local neighborhood, and is therefore considered as appropriate for our use.
- Cao et al. [61] proposed a metric designed to detect artifacts, which is based on the difference of saliency. Saliency usually refers to a representation of the idea that certain parts of a scene are pre-attentively distinctive and create some form of significant visual arousal [227]. This metric should be suitable since the dataset from Cardin [64] is reported to have different artifacts, such as loss of details and contouring [39]. The percentage of pixels with artifacts has been used as a measure of quality for this metric. For more information on this metric we refer the reader to Appendix C.

Evaluation of performance is done by calculating the correlation coefficient between the subjective score and the objective score, as introduced in Section 6.2.1. Three different kind of correlation are computed; the Pearson product-moment correlation coefficient, the Spearman’s rank correlation coefficient, and the Kendall tau rank correlation coefficient [232]. The first assumes that the variables are ordinal, and finds the linear relationship between variables. The second, Spearman, is a non-parametric measure of correlation that uses the ranks as basis instead of the actual values. It describes the relationship between variables without making any assumptions about the frequency distribution of the variables. The third, Kendall, is a non-parametric test used to measure the degree of correspondence between two rankings, and assessing the significance of this.

Two types of evaluation are carried out, first the overall performance of the metrics over the entire image set, and then evaluation of the metrics for each of the 15 different images.
9.3.1 Overall evaluation

We compare the overall score from the observers to the overall score by the metrics in order to evaluate the overall performance of the metrics. The most common method for this is by computing the Pearson correlation for all scores. The results from this show that all metrics have a very low Pearson correlation, approximately around zero. This indicates that the IQ metrics cannot predict perceived IQ. In addition, both the Spearman and Kendall correlation coefficients are similar to the Pearson correlation.

The low correlation is because of scale differences between the images, and because of this problem a new method [350] for evaluating overall performance of metrics was proposed in Section 6.3. In this method the scores for the metric are used to simulate an observer using the rank-order method, and results in a z-score plot as from the experiment (Figure 9.3). The results from the metric should be similar to the results from the observers if the metric exhibit a good performance.

Figure 9.4 shows the results for the S-CIELAB IQ metric with this method. A visual inspection of the relation between the results from the metric and the observers show differences, which is evidence that the metric does not predict perceived overall IQ. The z-scores from the observers are plotted against the S-CIELAB z-scores in Figure 9.5. The Pearson correlation between these are only 0.34, indicating a low correlation. The others also have a low correlation, as seen in Table 9.1.

![Figure 9.4: S-CIELAB z-score calculated with the method proposed in Section 6.3 for the same 15 images as in Figure 9.3.](image)

The small visual quality differences between the images contribute to making this a difficult task for the IQ metrics. For this experiment the IQ metrics cannot predict perceived overall IQ.
Figure 9.5: Z-scores from observers (Figure 9.3) plotted against the z-scores from S-CIELAB (Figure 9.4). This gives a correlation of 0.34. Corresponding correlation for all metrics are given in Table 9.1.

Table 9.1: Overall performance of the metrics using the method proposed in Section 6.3.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-CIELAB</td>
<td>0.34</td>
</tr>
<tr>
<td>S-CIELAB$_{f}$</td>
<td>0.14</td>
</tr>
<tr>
<td>S-DEE</td>
<td>-0.27</td>
</tr>
<tr>
<td>SHAME-I</td>
<td>-0.29</td>
</tr>
<tr>
<td>SHAME-II</td>
<td>0.19</td>
</tr>
<tr>
<td>ABF</td>
<td>-0.09</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.18</td>
</tr>
<tr>
<td>Cao et al.</td>
<td>-0.60</td>
</tr>
</tbody>
</table>
9.3.2 Image-wise evaluation

We have also analyzed the performance of the IQ metrics for each image, since previous research by Hardeberg et al. [169] has shown that some IQ metrics can be linked with certain characteristics of the image. Image-wise evaluation will reveal what causes the low overall performance of the metrics. Figure 9.6 shows the Pearson correlation for the different IQ metrics for the 15 different images. We can notice a great variation in the results for the different images, and also some variance between the IQ metrics. The Spearman and Kendall correlation coefficients follow the same tendency as the Pearson correlation.

Figure 9.6 shows that metrics such as S-CIELAB, S-CIELAB$_J$, SHAME-I, SHAME-II, and ABF perform similarly in many of the images. In images 5, 7, 8, 11, and 12 they have a good correlation. S-DEE differs a bit from the previous metrics, most likely since it is based on another color difference formula. SSIM is different from the other metrics in terms of performance, and it has generally low correlation, except in images 8, 11, and 12. It should be noted that SSIM is a gray scale metric, and does not take into consideration color information, which can explain why it is worse than the other metrics. The metric by Cao et al. performs contrary to the other metrics. This is not surprising, since it is based on artifact detection. We can see a very high correlation in image 3, where the observers mostly judged the images based on detail visibility.

In the images where there is a large difference from the original, but the difference is increasing the quality, the IQ metrics do not perform well. In the second image some of the reproductions have lost shadow details, but they have a small difference from the original. In this image observers preferred reproductions where details were preserved at the cost of larger color differences. This is also the reason why the metric from Cao et al. is very similar to the observers. The problem of when a difference contributes to increasing IQ is a difficult issue to handle for the IQ metrics, and the IQ metrics are still not good enough to predict this. However, in the images where a small difference from the original is preferred, such as image 5, 7, 8, 11, and 12, the IQ metrics perform reasonably well. In these cases IQ metrics working solely on the difference from the original are giving good results. This indicates that the metrics could be chosen based on the characteristics of the image. The results also show that one metric is not able to predict perceived IQ, and that several metrics evaluating different aspects might be required.

9.4 Summary

We have evaluated a set of IQ metrics against the percept using the framework presented in Chapter 8. The results show that none of the evaluated IQ metrics correlate well with perceived overall quality. For some of the images in the dataset certain metrics perform well, which could indicate that characteristics of the images may influence the performance of the image quality metrics.
Figure 9.6: Pearson correlation image-wise. For some images some IQ metrics have a high correlation, but the results vary over the 15 images in the dataset. The Spearman and Kendall correlation follow the same tendency as the Pearson correlation.
10 WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?

Both objective and subjective evaluation of IQ are dependent on a number of Quality Attributes (QAs), which are terms of perception [479] such as sharpness, contrast, and saturation. These QAs influence the overall IQ differently, and knowledge about their importance can be used to achieve an optimal reproduction of an image [135]. The importance of different QAs has been investigated and acknowledged by many researchers [28, 45, 93, 97, 135, 269, 304, 323, 383, 384], but there is, so far, no overall agreement on which QAs that are most important. In this chapter we focus on QAs for the evaluation of color prints.

It has been a goal in IQ evaluation for many years to develop objective measures correlated with subjective quality. The advancements in this field have been driven by the desire to reduce the dependence on human observers and minimize both time and resources needed to quantify IQ. To achieve this, the QAs used in subjective evaluation of quality and their importance need to be identified. These objective measures, such as IQ metrics (Chapter 4), can be used to help observers detect quality issues, identify where loss of quality occurs in a printing workflow or compare the quality of different printing systems.

IQ models have been created to establish a link between subjective and objective quality. These models are theories of perception that enable the prediction of IQ [115]. They are intended to describe the overall IQ of a system and to help researchers in the evaluation of IQ. IQ models are composed of QAs, and they show how QAs relate to each other and their influence on overall IQ. The goal for IQ models is to evaluate all QAs and their relationships, however this is a very difficult and complex task. Most IQ models reduce the complexity by using a subset of QAs composed of the most important QAs. By doing this, strengths and weaknesses of a given system can be meaningfully represented with a relatively small number of QAs [314]. Several IQ models have been proposed [28, 97, 230, 499], however the search for better and improved IQ models is still ongoing. The goal of this chapter is to identify and discuss the most important QAs for the evaluation of color prints. These QAs can further be used to create a link between subjective and objective IQ in an IQ model.

10.1 State of the art

A survey of QAs and IQ models is given in this section.
10.1.1 Image quality attributes

Norberg et al. [323] evaluated overall quality, as well as color rendition, sharpness, contrast, detail rendition in highlight and shadow areas, color shift, gloss, mottle, and print homogeneity in a comparison of digital and traditional print technologies. In a study by Lindberg [269] 12 different QAs (overall quality, gamut, sharpness, contrast, tone quality, detail highlights, detail shadow, gloss level, gloss variation, color shift, patchiness, mottle, and ordered noise) were used for the evaluation of color prints. Based on the evaluation performed by observers these 12 QAs were reduced to two orthogonal dimensions using factor analysis; one related to print mottle and one related to color gamut. These two dimensions accounted for almost all variation in the data set. Hardeberg and Skarsbø [170] compared the quality of digital printing presses based on colorimetric and spatial measurements, visual xpert evaluations, and panel tests. For the visual tests th observers evaluated color pleasantness, total image quality, smoothness, and detail rendition. Gast and Tse [152] evaluated six different QAs, blur, noise, banding, color rendition, tone reproduction and printer type. These QAs were evaluated in terms of preference. Nussbaum et al. [325] investigated print quality evaluation of governmental purchase decisions. In their work they evaluated text quality, color gamut, repeatability, and register for general quality, color match, visual resolution, surface texture, logo alignment, and artifacts for logo assessment, and color match, accuracy, and artifacts for copy quality. Additionally, several researchers have investigated the importance of QAs like sharpness [45], contrast [41], artifacts (for example noise [383] and banding [93]), naturalness [135], and color [41, 97, 169, 178, 304].

Research on the combined influence of QAs has been carried out as well. In 1980 Sawyer [384] investigated the influence of sharpness and graininess on perceived IQ, and their combined influence. Two years later Bartleson [28] investigated the combined influence of sharpness and graininess on color prints. Both Sawyer and Bartleson showed results where the worst QA tended to determine the quality, and a change in other QAs would not increase quality. Natale-Hoffman et al. [314] investigated the relationship between color rendition and micro uniformity on preference in 1999. This was considered by the authors as a step towards predicting preference, without dependence on human observers.

Identification of QAs has also been recognized as important for IQ metrics. Morovic and Sun [304] based an IQ metric on perceptual QAs, where the QAs were determined based on answers from observers. Lightness, hue, chromaticity, details, and contrast were found to be important. Only the three first, being the most important according to Morovic and Sun [304], were incorporated in an IQ metric (Δl_m). Later Wang and Shang [463] showed that defined QAs were beneficial for training IQ metrics.

10.1.2 Image quality models

A framework for IQ models was proposed by Bartleson [28] in 1982. His approach was divided into three parts:

1. identification of important QAs,
2. determination of relationships between scale values and objective measures,
3. combination of QA scale values to predict overall IQ.

Bartleson used this framework to investigate the combined influence of sharpness and graininess on the quality of color prints. This framework has the advantage of representing strengths
and weaknesses of a given system by a relatively small number of QAs. Because of this and its perceptual considerations, the framework has been adopted by several researchers [97, 115, 230]. We also adopt this framework in this chapter, where we mainly discuss the first part, identification of important QAs.

Dalal et al. [97] followed the framework in the creation of the document appearance characterization system, which is a two-sided appearance-based system composed of QAs: one part for the printer and one for materials and stability. For most QAs in the system, evaluation is performed by experts. The basic IQ is given by 10 QAs, for both the printer and materials and stability. These describe different aspects of the system, such as color rendition, uniformity, tone levels, and stability. The document appearance characterization system has several advantages. It uses high-level descriptors, which cover a wide range of IQ issues, like defects and sharpness. The printer is also separated from materials and stability, allowing for separate analysis. It is also a clear advantage that it is technology independent. In addition, the QAs in the system are somewhat orthogonal (i.e., they do not influence each other).

This system has some drawbacks as well, since the evaluation is mostly carried out by experts the results are influenced by the subjectivity of the expert. It might be unsuitable for non-experts due to its complexity. This is because the QAs are associated with known printing problems and technological issues. This approach is different from the approach taken by Morovic and Sun [304], where QAs were chosen based on answers from observers, resulting in more general QAs. The QAs in the model are also not adapted to IQ metrics, making it difficult to obtain a completely objective evaluation of IQ. It should be taken into account that the document appearance characterization system was not intended to use only IQ metrics, since it is made for subjective evaluation by experts. In addition, the system does not directly account for the contrast QA, which has been regarded as an important QA by other researchers [41, 269, 304, 323].

Keelan [230] also adopted the framework proposed by Bartleson [28]. He first identified important QAs, then the relationship between a subjective scale (based on just noticeable differences) and an objective metric is found. In the case where multiple QAs influence the quality of an image, the influence of each QA to overall IQ is found. Keelan adopted multivariate formalism as a tool to combine the influence of each QA, in order to obtain a value for overall IQ. QAs used in the model by Keelan are assumed to be independent, which is different from the QAs used by others [269, 304, 323]. The advantage is that QAs do not influence each other, and can be easily combined in order to achieve a value of overall IQ, which is not straightforward for dependent QAs. However, the disadvantage is that it might be very difficult to identify independent QAs. The model by Keelan also assumes that objective metrics can be readily designed. Nonetheless, a method to deal with this was proposed; rather than dealing with several QAs, the problem was approached by considering the non-independent QAs as a single QA with several facets.

Engeldrum [115] focused on the building of an IQ model, and partially adopted Bartleson’s framework. He proposed the IQ circle, which is based on four elements; customer quality preference, technology variables, physical image parameters, and customer perceptions. The last element, customer perceptions, contains the perceptual QAs (or "nesses"), being the topic for this chapter. The quality circle shows the relationship between objective and subjective quality, but does not include which "nesses" are important, or how they should be quantified. Engeldrum also states that observers most likely will not be able to perceive more than five QAs simultaneously. This statement is contradictory with the other IQ models and use of QAs, such as the document appearance characterization system, where in total 20 QAs (10 for each side of the system) are evaluated. Norberg et al. [323] evaluated overall
IQ and nine QAs. Lindberg [269] evaluated overall IQ in addition to 12 QAs, but analysis showed that these QAs could be reduced to only two QAs.

Eerola et al. [112] proposed a model based on Bayesian network which connects the objective measurements to the subjective scores. The model discovers links between the objective measures and the QAs used by human observers, and the nature of the QAs and their influence on each other.

Many other IQ models have been proposed as well [304, 458, 499]. Some of these are IQ metrics, which calculates one or several values representing IQ. One of these metrics is the S-CIELAB [499] (Section 4.2.2.1), where a spatial pre-processing of the image is carried out before the CIELAB color difference formula [82] is applied to calculate IQ. Metrics, like the S-CIELAB and others, are most often constructed to quantify either overall quality or the quality of specific QAs. These metrics usually incorporate several stages of processing. Each stage is linked to a specific aspect of IQ, where characteristics of different QAs are taken into account. There are several different approaches to measure IQ. S-CIELAB [499] and the $\Delta E_{cm}$ [304] are built on the idea that color differences are responsible for a large proportion of differences between an original and its reproduction. Another IQ metric, SSIM [458] (Section 4.2.3.1), is based on the degradation of structural information. Many of these metrics have been proposed for different purposes, such as image difference and image fidelity. For a complete overview of full-reference IQ metrics we refer the reader to Chapter 4, Pedersen and Hardeberg [354], or Appendix A.

10.2 Investigation and selection of important quality attributes

As a first step towards an IQ model relevant and important QAs must be identified. In order to do this we have taken the approach of doing a survey of the existing literature. Numerous QAs have been considered as important and evaluated by researchers to quantify IQ. In order not to exclude QAs in this part of the investigation, we have included QAs based on both technology and perception, and QAs used with different intentions. These QAs include, for example, lightness [230, 304], sharpness [45, 239, 269, 320, 323], blur [152], contrast [41, 269, 304, 323], contouring [161], noise/graininess [28, 152, 292, 320, 327, 383, 384], banding [23, 24, 93, 152], details [41, 169, 304, 320, 323], naturalness [135], color [41, 97, 169], hue [178, 304], chroma [304], saturation [41], color rendition [97, 152], process color gamut [97], artifacts [41], mottle [239, 269, 379], gloss [245, 269, 323], color reproduction [292], tone reproduction [152, 292], color shift [239, 323], ordered noise [323], patchiness [323], line quality [97, 445], text quality [97], gamut size [12], adjacency [97], printer type [152], effective resolution [97], effective tone levels [97], gloss uniformity [97], skin color [169], paper roughness [239, 480], paper flatness [97], paper whiteness [238, 239], perceived gray value [320], structure changes [320], micro uniformity [97], macro uniformity [97], structure properties [320], color gamut [239], correctness of hue [182], colorfulness proportional to the original [182], correctness of lightness [182], edge sharpness [445], and edge raggedness [445].

When reducing these QAs found in the literature, there are several important issues to consider, as mentioned previously, such as the intention of how QAs should be used, and their origin. A long term goal of this research is to create a link between subjective and objective IQ of color prints. With this intention the QAs should be based on perception and account for technological printing issues. The QAs should be general enough to be evaluated by observers,
and in order not to exclude novice observers the QAs should be somewhat straightforward to evaluate. In addition, the QAs should be suitable for IQ metrics, being the intended objective method. The existing sets of QAs and models do not fulfill all of these requirements, and therefore a new set of QAs is needed.

Many of the QAs listed above are similar and have common denominators, which enables them to be grouped within more general QAs in order to reduce the dimensionality and create a more manageable evaluation of IQ. There is usually a compromise between generality and accuracy when it comes to dimensionality. A small set of general QAs results in lower accuracy, but low complexity, while a higher dimensionality offers accuracy, but higher complexity. We have linked most of above QAs to six different dimensions, considered as important for the evaluation of IQ. This results in a reasonable compromise between accuracy and complexity. We are also close to the statement by Engeldrum [115] that observers will not perceive more than five QAs simultaneously. We have reduced the QAs found in the literature to the following Color Printing Quality Attributes (CPQAs) six:

- **Color** contains aspects related to color, such as hue, saturation, and color rendition, except lightness.
- **Lightness** is considered so perceptually important that it is beneficial to separate it from the color CPQA [230]. Lightness will range from "light" to "dark" [479].
- **Contrast** can be described as the perceived magnitude of visually meaningful differences, global and local, in lightness and chromaticity within the image.
- **Sharpness** is related to the clarity of details [45] and definition of edges [66, 134].
- In color printing some artifacts can be perceived in the resulting image. These artifacts, like noise, contouring, and banding, contribute to degrading the quality of an image if detectable [55, 441].
- The physical CPQA contains all physical parameters that affect quality, such as paper properties and gloss.

The six dimensions are concise, yet comprehensive, general high-level descriptors, being either artifactual, i.e., those which degrade the quality if detectable [55, 441], or preferential, i.e., those which are always visible in an image and have preferred positions [441].

We have turned to Venn diagrams to create a simple and intuitive illustration of the CPQAs and their influence on overall IQ. Venn diagrams may be used to show possible logical relations between a set of attributes. However, it is not possible to create a simple Venn diagram with six fold symmetry [164]. Therefore we illustrate the CPQAs using only five folds, leaving the physical CPQA out. This does not mean that the physical CPQA is less important than the other CPQAs.

The Venn diagram of Figure 10.1 illustrates how the overall IQ is influenced by one, two, three, four, or five of the CPQAs. Many of the CPQAs are interdependent [336], making quality a multidimensional issue [319], in this case five dimensions. These CPQAs can influence the overall quality differently, and therefore the ellipses may not have equal sizes or the same positions in all situations. It is difficult to reduce all of the CPQAs used in the literature to six dimensions while preserving independence together with having perceptual CPQAs. It is very difficult, if not impossible, to do this while accounting for most of the aspects of IQ for color prints. This is a disadvantage since it leads to a more complex analysis of the results. However, methods have been proposed to deal with this problem [230].
WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?

The six CPQAs are a good starting point for the quality evaluation of color prints, and they can be adapted to different situations. Each of these CPQAs can be divided into sub-QAs for adaptation to specific issues, and thereby increase the accuracy of the QAs. The artifact QA can, for example, be divided into three sub-QAs: noise, contouring, and banding. Separate analysis of these can be advantageous since it allows for specific analysis either by experts or IQ metrics. Splitting into sub-QAs can also prove useful when there is skewness in distribution of CPQAs, resulting in better balance between the CPQAs. Furthermore, some sub-QAs can be placed under several main CPQAs, such as uniformity. This sub-QA can be placed under color, but also under artifacts since lack of uniformity can be thought of as an artifact. The placement of these sub-QAs must be done where it is most appropriate. Additionally, all CPQAs might not be used in a given evaluation of IQ, in this case CPQAs can be excluded. By being able to exclude CPQAs and divide CPQAs into sub-QAs, QAs used by other researchers can be thought of as special cases of the proposed CPQAs.

In the following parts we take a closer look at the six different CPQAs, where links to QAs from the literature are identified.

10.2.1 Color

Color is a sensation. It is the result of the perception of light by the HVS [149]. The color CPQAs includes color related issues, like hue, saturation, and color rendition. Lightness is not a part of our color CPQAs. Since our HVS process lightness and chromaticity information differently, it is convenient to treat these as separate CPQAs [230].

Many of the QAs used in the literature can be linked with one of the six proposed CPQAs. Within our color CPQAs we can find several QAs used by other researchers, such as color [41, 97, 169], hue [178, 304], chroma [304], saturation [41], color rendition [97, 152], process color gamut [97], color reproduction [292], color shift [239, 323], gamut size [12], skin color [169], color gamut [239], correctness of hue [182], and colorfulness proportional to the original [182]. Many of these can also be connected to other QAs, which are discussed below.
10.2.2 Lightness

A common definition of lightness is the visual sensation according to which the area in which the visual stimulus is presented appears to emit more or less light in proportion to that emitted by similarly illuminated areas perceived as a “white” stimulus [479]. Variations in lightness range from “light” to “dark” [479].

Many QAs used by other researchers can be included within our lightness QA, like tone reproduction [152, 292], perceived gray value [320], correctness of lightness [182], and lightness [230, 304]. Other QAs are more difficult to link with just one QA, such as color shift [239, 323], gamut size [12], color gamut [239], process color gamut [97], and color rendition [97, 152]. All of these have ties to lightness, but also to the color CPQA. Other QAs will influence lightness, but cannot be accounted for within the lightness QA, for example paper flatness and gloss level.

10.2.3 Contrast

Contrast is a difficult CPQA since there are many different definitions of contrast [357, 362, 375]. Michelson [287] defined contrast as:

$$\frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}$$

where $I_{\text{max}}$ and $I_{\text{min}}$ represent the highest and lowest luminance. The Weber contrast [473] is defined as:

$$\frac{I - I_b}{I_b}$$

where $I$ represents luminance of features and $I_b$ the background luminance. Root-Mean-Square (RMS) contrast is another common way to define contrast:

$$\text{RMS} = \left[ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}$$

where $x_i$ is a normalized gray level value and $\bar{x}$ is the mean normalized gray level. Contrast can also be defined as the visual property that makes an object distinguishable. This definition is useful to express readability of prints. Another used definition of contrast is the lightness ratio between two areas in an image. Fedorovskaya [134] defines contrast as an integrated impression of differences in lightness, or lightness variation observed within the whole picture. Keelan [230] defines contrast, in the context of color and tone reproduction, as the relationship between original scene lightness perceived by a photographer and final image (reproduced) lightness perceived by an observer.

Contrast is clearly difficult to define, and it will change according to the application. Even so, the literature distinctly shows some common characteristics of contrast. It is related to chromaticity [415, 443], and because of this a definition of contrast based solely on lightness cannot describe perceived contrast in color prints. As well as being correlated with color, contrast is also related to local and global impressions [415, 443]. Therefore definitions working on a global aspect are unsuitable for defining contrast in color prints.

The proposed CPQA is a perceptual QA, so a definition should relate to the HVS. To
account for these characteristics of contrast, we define contrast for color prints as the perceived magnitude of visually meaningful differences, global and local, in lightness and chromaticity within the image. This definition takes into account the characteristics of contrast, and is applicable to color prints.

Our contrast CPQA can be linked with the use of contrast by several researchers [41, 269, 304, 323]. Due to our definition of contrast, it will have ties with many different other QAs, such as chroma, saturation, and lightness.

**10.2.4 Sharpness**

Caviedes and Oberti [66] relate the perception of sharpness to the clarity of detail and edge definition of an image. Bouzit and MacDonald [45] follow a similar thinking, and relate sharpness to details and edges. Fedorovskaya [134] defines overall sharpness as the overall impression of clarity of edges observed within the whole picture. Sharpness is related to both details and edges, and because of this our sharpness CPQAs is defined as the clarity of details and edge definition of an image.

QAs that are suitable to group within the sharpness CPQAs are diverse and many, including sharpness [45, 239, 269, 320, 323], details [41, 169, 304, 320, 323], line quality [97, 445], adjacency [97], blur [152], effective resolution [97], edge sharpness [445], and edge raggedness [445].

**10.2.5 Artifacts**

Different printing artifacts are found in this CPQA, such as noise, contouring, and banding. All of these have in common that if they are detectable they contribute to degrade the quality of an image [441]. In this QA, definitions of sub-QAs can be useful. For printing, noise, contouring, and banding are often considered as important artifacts. Image noise can be defined as random variations in brightness or color in an image. Contouring can be characterized as the perceived color change within a small region exceeding a threshold [246]. This results in perceived discontinuity. Banding is the presence of extraneous lines or bands in a printed page [24, 93], these appear as nonuniform light or dark lines across the page. This CPQA differ from the other since it has a zero point (i.e. no artifacts are presents), while the others have an optimum.

We can link the proposed artifacts CPQA with several of the QAs used in the literature, for example contouring [161], noise/graininess [28, 152, 292, 320, 327, 383, 384], banding [23, 24, 93, 152], artifacts [41], effective tone levels [97], mottle [239, 269], ordered noise [323], patchiness [323], structure changes [320], and structure properties [320]. All of these will degrade quality if detectable.

**10.2.6 Physical**

This CPQA is important since the first five CPQAs cannot account for physical QAs, such as paper roughness and gloss level. Investigation of the literature has shown these physical QAs to be very important for overall IQ, and should be accounted for in the evaluation of IQ. There are several QAs used by researchers that fit within this attribute, like paper roughness [239, 480], paper flatness [97], gloss [269, 323], and printer type [152].
10.2.7 Relations between quality attributes

The six proposed CPQAs are not necessarily independent, and in order to calculate overall IQ it is important to know which QAs influence other QAs and the magnitude of their influence. Investigation of the literature shows many of the relationships between CPQAs. Color can be linked to a number of other attributes, one of these is contrast [357, 441]. Color differences can also be linked to different artifacts, as contouring [246] and banding [79]. One way to preserve details in gamut mapping is to introduce a slight color difference [41], relating color to the sharpness QA, where details are included. Other relations to sharpness can be found as well; in the case of sharpening a color difference can be introduced [76, 368]. In extreme cases, sharpening can result in halo artifacts caused by color differences [230].

The last two relations for color can also occur due to changes in lightness, creating relations to sharpness and artifacts. Lightness can be linked to contrast [56, 134, 357], but also to artifacts, such as banding [79] and contouring [246].

Contrast has already been linked to color [357, 441] and lightness [56, 134, 357]. It can be linked to sharpness [26, 56, 57, 217, 221, 222, 392] as well, since an increase in contrast generally increases sharpness [222]. In the literature we also find relations to contouring [448] and banding [52] artifacts.

Above, sharpness has been linked with color [41, 76, 368], artifacts [230], and contrast [26, 56, 57, 217, 221, 222, 392]. Another artifact that is commonly mentioned regarding sharpness is noise, and the relationship between these CPQAs has been extensively examined in the literature [55, 104, 221, 222, 395].

Artifacts can be related to a number of different CPQAs. For noise it has already been mentioned the relations to sharpness [55, 104, 221, 222, 395], and for halos to lightness and color [230]. While contouring is linked to contrast [448], banding can be related to both color and lightness, since the bands can be caused by lightness and (or) color variations [79], but also to contrast [52]. The relations for artifacts will change according to the different artifacts evaluated.

For the physical CPQA many relations can be found. Paper characteristics can for example influence color [238] and artifacts (as lack of smoothness [239]), while paper coating can affect artifacts (for example lack of uniformity [323]).

There might also be situations where an increase in the quality of a CPQA will reduce the quality of another CPQA, one example can be the trade-off between noise and sharpness, which has been investigated in the literature [28].

The relations mentioned here do not address the magnitude of influence between CPQAs.

10.3 Validation of the quality attributes - experiment I

In this section we will take a closer look at the three first CPQAs; color, lightness, and contrast as seen on Figure 10.2. In order to investigate quality issues in a color workflow, and to confirm the proposed QAs in the previous section, an experiment was carried out. A set of images was reproduced and investigated by 15 observers to find the most important QAs, and to learn which QAs observers use in the quality evaluation of a color workflow. The observers were both male and female, and ranged from experts to non-experts.

The images were reproduced using the ICC perceptual rendering intent, which adjusts color appearance to achieve the most attractive result on a medium different from the original [187]. In the evaluation of this color workflow, observers evaluate different QAs, and the
Figure 10.2: A psychophysical experiment was carried out to investigate CPQAs in a color workflow. In this experiment a subset of the main CPQAs is considered to affect overall IQ: color, lightness, and contrast.

influence that these QAs have on overall IQ affects the observer’s judgment of quality. For some QAs the quality decreases, for other QAs the quality might increase, while some QAs neither increase nor decrease quality. Investigating only the QAs that influence quality might not give a correct representation of which QAs that are important. All QAs being evaluated by observers should be investigated, because of this the instructions given to the observers are crucial for obtaining correct results.

10.3.1 Experimental setup

10.3.1.1 Test images

Several guidelines have been given in the literature for the selection of images (Section 3.2.4.6), in the context of investigating IQ issues. Holm et al. [176] recommend the use of a broad range of natural images as well as test charts to reveal the quality issues. The Commission Internationale de l’Éclairage [80] suggests to include images with the following characteristics: high-key, low-key, low lightness contrast, leaves and sky, no neutrals, no white point, no black point, heavy cast, few hues, business graphic, and flesh tones. Büring et al. [54] propose to use natural images, as well as saturated colors. To achieve this we followed the recommendations of Field [143] and CIE [80], where the test images were chosen based on the following criteria:

- Low, medium, and high levels of \textit{lightness},
- Low, medium, and high levels of \textit{saturation},
- \textit{Hue} primaries,
- Low, medium, and high \textit{contrast},
- \textit{Larger areas of the same color},
- \textit{Fine details},
- \textit{Memory colors} as skin tones, grass, and sky blue,
WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?

- Color transitions,
- Neutral gray.

Most of the images were pictorial with a wide range of scenes like landscapes, portraits, and personal items (such as jewelry, books, and clothes). This helps to characterize the impacts for QAs [230], and ensures that the observer examines a wide variety of QAs. In addition to the pictorial images a set of test charts were included, since these are content-free and have a selection of “interest area” colors suitable for evaluation of different aspects of IQ [143].

A total of 56 images, as seen in the lower part of Figure 10.4, were used in this experiment. 7 images from ISO [199], 2 images from CIE [80], 3 test charts, and 44 other images captured by the authors, were included. The 44 images captured by the authors were in RAW format, which were converted to sRGB using Camera Raw 5.0 in Adobe PhotoShop CS4 with a resolution of 150 DPI and a 16-bits encoding. The RAW images were manipulated to look optimal by a professional, and then they were verified by two other professionals.

The images were printed at a resolution of 150 pixels per inch and at a size resulting in a printed reproduction at approximately 8 by 10 inches.

10.3.1.2 Color workflow

The first step in the color workflow was to re-render the set of images from sRGB to the perceptual reference medium gamut [209] using the sRGB v4 perceptual transform. The output profile (Figure 10.3) was generated using the TC3.5 CMYK test target, measured using a GretagMacbeth Eye-One Pro spectrophotometer and ProfileMaker Pro 5.0.8. Then, as a second step, linear lightness scaling compensation in the CIE XYZ color space plus the hue preserving minimum \( \Delta E \) clipping gamut mapping algorithm [80] was applied to the image to re-render from the perceptual reference medium gamut to the gamut of the printing system. The linear lightness scaling were made between the black point CIELAB coordinates of each images to the black point CIELAB coordinates of the printing system contained in the output profile. The third and last step was to convert the color data from the profile connection space values to CMYK values of the printing system by a relative colorimetric transform. The images were then printed with the Océ ColorWave 600 wide format CMYK printer on Océ Red Label paper.

10.3.1.3 Viewing conditions

The observers were presented with a reference image on an EIZO ColorEdge CG224 (some observers on an EIZO ColorEdge CG221 since the experiment was carried out in two locations) display at a color temperature of 6500 K and luminance level of 80 cd/m², following the specifications of the sRGB. This set was rendered for sRGB display, and therefore a monitor capable of displaying the sRGB gamut was the most adapted reproduction device for this set of images. In addition, the display was fitted with a monitor hood to prevent glare. The printed images were presented randomly in a controlled viewing room at a color temperature of 5200 K, an illuminance level of 450 ±75 lux and a color rendering index of 96. The observers viewed the reference image and the printed image simultaneously from a distance of approximately 60 cm. The experiment was set up to follow the CIE guidelines [80] as closely as possible.
WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?

Figure 10.3: Gamuts for the experiment shown with three different projections in CIELAB space using ICC3D [132] (19/07/11: http://www.colorlab.no/icc3d). sRGB gamut on the top, perceptual reference medium gamut in the middle, and the gamut of the printer on the bottom. The small outer wireframe indicate the sRGB gamut boundary, and the circle indicates 100 on the a and b axis.
10.3.1.4 Instructions given to the observers

The instructions given to the observers focused on the overall quality rating of the reproduction and which QAs the observer used in the evaluation. Instructions specified that all QAs used in the evaluation should be stated, even if they did not influence IQ. The two following questions were given to the observers:

- Is the printed image a pleasing reproduction?
- According to you, which quality attributes influence the quality of the reproduction?

For the first question a scale from 1 to 7, where 1 represented the most pleasing reproduction, was given to the observers. A description for each level was proposed to help the observers in their evaluation:

1. Most pleasing you can imagine,
2. Highly pleasing,
3. Very pleasing,
4. Fairly pleasing,
5. Moderately pleasing,
6. Poorly pleasing,
7. Least pleasing reproduction possible.

QAs used during the experiment were noted on a form, where QAs decreasing, increasing or not influencing quality were denoted with different symbols. Since we want to see how QAs used by observers fit within in the QAs proposed previously, no descriptions or proposals for QAs were given to the observer in order to prevent any influence on the QAs and vocabulary used by the observer.

10.3.2 Perceptual results

15 observers participated in the experiment. Five observers rated the whole data set, and 10 observers rated parts of the data set. A total of 452 evaluations were carried out by the observers, where a scale value was given to each image and the observers described the QAs they used in their evaluation. The evaluation was carried out in several sessions to prevent observer fatigue.

The overall average pleasantness of all images, based on the seven level scale, has been found to be between fairly and very pleasing (Figure 10.4). The whole scale was used in the experiment, 10 observers used at least one of the extremes on the seven step scale. Analysis of the ratings given by the observers indicates that images with a majority of large areas with the same color (especially shadow areas) and images with color transitions (both test charts and natural images) are rated as least pleasing. Analysis of the answers from the observers indicates that contouring is found in the images with color transitions, and since the contouring is highly perceivable in these images the pleasantness is low. In some images with shadow areas details were lost or they were less visible, mainly due to the difference between the input
What quality attributes are the most important?

and output gamut as seen in Figure 10.3. In images with color transitions color breaks or loss of smoothness occur reducing the pleasantness of the image, which is a result of the gamut clipping algorithm. The images rated to be most pleasant have average saturation, lightness, and contrast. Analysis of the results from the observers also reveals that the most pleasant images are equal or more colorful than the original and have equal or better contrast than the original. We will focus on the QAs used by the observers, however, an evaluation and comparison of quality between the sRGB v4 perceptual transform and sRGB v2 transform has been carried out elsewhere [39].

10.3.3 Fitting the quality attributes to the color printing quality attributes

For all observers a total of more than 50 QAs have been used in the evaluation, with an average of 10 different QAs for each observer. For each image an average of 2.95 QAs were used, with a minimum of one and a maximum of eight. This indicates that observers do not consider a high number of QAs in their evaluation of IQ. Many of the QAs used on the experiment overlap, such as lightness, brightness, luminance, and darkness. All of these are connected to the lightness of the image, and have been grouped within the QA lightness. Similar grouping is done for other QAs to fit within the previously proposed set of QAs (Figure 10.1).

Color is the most frequently used QA by the observers. As seen in Figure 10.5 it has been used to describe the IQ of more than 70 percent of the images. This is not surprising since a color workflow was investigated, and the color QA is fairly wide, containing sub-QAs like hue, saturation, and colorfulness. These three sub-QAs are commonly used by the observers, and often used together. The second most used QA, sharpness, mainly contains two sub-QAs; edges and details. Details, both in highlights and shadows, have been frequently used by the observers. Some observers also commented that loss of contrast led to a loss of perceived sharpness, since edges and details were less prominent. This phenomenon has also been acknowledged in the literature [56, 57, 230]. Contrast, the third most frequently used QA, is a narrower QA than color and sharpness. Because of this it is interesting to notice the frequent use of contrast in the evaluation of color prints, and it confirms the need for a contrast QA in the evaluation of IQ.

In the images sub-QAs as noise, contouring, banding and so on could be perceived. The term artifacts, or sub-QAs of this, were used in approximately 40 percent of the images. Lightness is considered in more than 30 percent of the images. Even though this is the least frequently used QA, it should be noted that some observers used the more general term color rather than separating lightness and chromaticity.

10.3.4 Relations between the attributes

Analysis of the relations between QAs has also been carried out, using cross-tabulation and chi-square tests. The null hypothesis, $H_0$, is that there is no relationship between two QAs. The alternative hypothesis, $H_1$, is that there is a relationship between two QAs. $p$-values from this analysis are shown in Table 10.1. For some combinations of two QAs given a 5% significance level, $H_0$ is rejected in favor of $H_1$. This indicates dependence between color and lightness, but also between lightness and sharpness, contrast and artifacts, and artifacts and
**Figure 10.4:** Average pleasantness rating for the 56 images in the experiment sorted from the most pleasing image to the least pleasing image. Each rating is plotted with a 95% confidence interval. Thumbnails of the images below the graph are sorted in the same order, from left to right and top to bottom. The most pleasing image in the top left corner, and least pleasing in the bottom right corner.
What quality attributes are the most important?

Figure 10.5: Frequency of QAs used by the observers in the experiment. All QAs used by the observers have been fitted to one of the important QAs proposed in the previous section.

lightness. These results show that use of these QAs occur simultaneously, but not how they affect each other or the overall IQ.

Table 10.1: Results from cross-tabulation and chi-square analysis. With a significance level at 5% there is dependence between color and lightness, lightness and sharpness, contrast and artifacts and artifacts and lightness. Significance at 5% shown in bold.

<table>
<thead>
<tr>
<th></th>
<th>Sharpness</th>
<th>Color</th>
<th>Artifacts</th>
<th>Contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>0.998</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artifacts</td>
<td>0.168</td>
<td>0.219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td>0.239</td>
<td>0.723</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>Lightness</td>
<td>0.005</td>
<td>&lt;0.001</td>
<td>0.008</td>
<td>0.141</td>
</tr>
</tbody>
</table>

In the experiment observers distinguished between QAs that decreased IQ, did not influence IQ, or increased IQ. Observers have marked more QAs to decrease quality than the two other groups, and more QAs to increase quality than QAs that do not influence IQ. Figure 10.6 shows the distribution of the three groups for each of the five CPQAs. For the artifacts attribute, some observers stated that lack of artifacts increased quality. Observers have not considered artifacts where it does not influence IQ, indicating that artifacts only are considered when they are perceivable, or not present in areas where observers expect to find artifacts. In the sharpness CPQAs lack of details in several of the reproductions contribute to decreased IQ. Therefore we do not have an equal distribution between increase and decrease in IQ as for color, contrast, and lightness.

We have investigated the dependence between image characteristics and the usage of the three levels (decreasing, not influencing, and increasing quality) for each CPQA. Cross-tabulation shows dependence between the use of artifacts that decrease quality and images classified as test charts. This is not surprising since contouring was most perceptible in these images. Dependence between fine details and increase in quality due to contrast can be found as well, this indicates that contrast is important for the perception of details. Low lightness and increase in IQ due to lightness also shows dependence between each other.
WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?

![Normalized frequency distribution of QAs that increase, decrease, and do not influence IQ.](image)

Figure 10.6: Normalized frequency distribution of QAs that increase, decrease, and do not influence IQ.

In this experiment the physical CPQA was not considered, and was therefore not a part of the analysis. There are also some QAs used by the observers that are difficult to link with one of the five important CPQAs, for example naturalness and warmness. These CPQAs can be linked with changes in other CPQAs, such as color and lightness.

10.4 Validation of the quality attributes - experiment II

In the previous experiment almost all the QAs used by a group of observers were grouped within the proposed CPQAs [345, 346]. This validation was done by subjectively grouping the numerous QAs used by the observers to one of the six CPQAs. However, since it was carried out subjectively by the authors, it does not fully validate the aspects on which the CPQAs were selected. Additional validation is therefore required to ensure that they satisfy the intended needs, and to make sure that the proposed CPQAs are suitable to assess IQ. The goal of this section is to validate the CPQAs experimentally.

10.4.1 How to validate quality attributes?

The validation should be adapted to the criteria on which the CPQAs were selected. The validation can be achieved by comparing data to the CPQAs, and analyzing the correspondence between the data and the CPQAs. Requirements need to be set to validate the CPQAs. Using the aspects on which they were selected (Section 10.2) we can derive the important requirements that the CPQAs should fulfill. For the CPQAs to be useful for the evaluation of IQ, to be perceptual, and account for technological issues, they should be able to cover the entire field of IQ. All issues encountered in the evaluation of color prints should be described using the CPQAs, making this one of the requirements to validate. As many as possible of the QAs used by the observers should be accounted for within one of the CPQAs, and not overlap several CPQAs. Minimum overlapping is considered as one of the requirements the CPQAs should fulfill. The CPQAs were selected to keep the number of QAs to a minimum,
What quality attributes are the most important?

This is important for usability of the QAs, and for the CPQAs to be straightforward to use. Therefore dimensionality should be one of the requirements. For the CPQAs to be suitable for IQ metrics and straightforward to use, it is important to keep independence.

Summarized, we have four different requirements the CPQAs should fulfill in order to be validated:

- the CPQAs should cover the entire field of IQ,
- few QAs should overlap the CPQAs (i.e. most of the QAs can be assigned to only one of the proposed CPQAs),
- dimensionality should be kept to a minimum,
- low or no dependence should occur between CPQAs.

There are several ways to carry out the validation for these requirements. The validation can be carried out subjectively or objectively. The drawback of the previous validation [345, 346] in Section 10.3 of the CPQAs was subjectivity. In order to minimize the subjective influence, and to have an independent validation of the QAs; objective validation methods have been investigated. It is preferable to have a fully objective method, where data, for example from an experiment, can be compared to the CPQAs. This requires a database containing all QAs, categorization of them, and their relations. To our knowledge such a database does not exist, and this method is therefore inapplicable. Another possible method is to use existing definitions of QAs to create relations between the QAs, resulting in a data structure, which can be visualized as a tree or data structure. This method is not completely objective, but it keeps the subjectivity to a minimum. Therefore, we will adopt this method for the validation of the CPQAs.

Since the CPQAs are based on human visual perception, subjective data is required for the validation. In order to validate if the CPQAs cover the entire field of IQ it is required that the observers use a wide variety QAs. Expert observers have been shown to be more precise than non-experts [108] and they have a wider vocabulary. Therefore expert observers should be recruited for such experiments. In addition, the color workflow on which the data is collected should guarantee many different quality issues. The image set should also include a wide variety of characteristics to ensure that many different IQ issues are encountered.

There are various ways to carry out such an experiment. One way is to provide the CPQAs and their definitions to the observers, and ask the observers to use them in their judgment of IQ. If the observers only use the CPQAs, one could argue that they cover all aspects of IQ. However, this experimental setup can restrict the observers to the CPQAs, and prevent them from using other QAs they normally would use. Another option is to record the QAs used by the observers during the experiment, where the observers do not have prior knowledge of the CPQAs. This last option does not restrict the observers to the CPQAs. Therefore as it has the best potential to validate the CPQAs, we adopt this method.

10.4.2 Experimental setup

10.4.2.1 Images

We have followed the same guidelines as given in Section 10.3.1.1 to select test images for the validation. 25 images have been selected (Figure 10.7), where six are the same as those in the first validation (Figure 10.4).
To address the customer segment of Océ, we have also included 3D models, maps, posters, presentations, and pdf-like documents. The images have been collected from different sources. One image from ISO [204], two from CIE [80], ten images provided by the authors, one image from MapTube [283], three images from ESA [121], four images from Google 3D Warehouse [157], one image reproduced with permission from Ole Jakob Skattum, and one image from Halonen et al. [167]. The images were 150 dpi 16-bit sRGB, saved as tiff files without compression.

![Image](image.png)

*Figure 10.7: The 25 images used in the experiment to validate the quality attributes.*

### 10.4.2.2 Color workflow

The images were printed on an Océ Colorwave 600 CMYK wide format printer on Océ Red Label (LFM054) plain uncoated paper. The profile of the printer was generated using a GretagMacbeth TC3.5 CMYK + Calibration test chart in ProfileMaker Pro 5.0.8. A round trip test was carried out to ensure a correct profile as suggested by Sharma [393], and we performed a visual inspection of color gradients to verify that no artifacts occurred. The images were printed with three different rendering intents: perceptual, relative colorimetric, and relative colorimetric with black point compensation.

### 10.4.2.3 Viewing conditions

The viewing conditions were identical to those presented in Section 10.3.1.3. Except that the reference image was presented on an EIZO ColorEdge CG221.
10.4.2.4 Instructions

The instructions given to the observers focused both on the overall quality rating and on the QAs used in the evaluation:

- Rank the reproductions according to quality.
- Elaborate on the attributes you use and quality issues you observe, i.e. all attributes you consider.
- If possible try to give an indication of the importance of the issues and attributes, and important areas.

The entire experiment was filmed, and the observers were encouraged to describe and talk about their observations. The video enabled the authors to better capture the attributes used by the observers than if they were to write down the attributes, since observers are usually more articulate orally.

10.4.3 Fitting the quality attribute data to the color printing quality attributes

Four observers, all considered to be experts, were recruited for the experiment. This resulted in a total of 100 observations by the four observers for the 25 images, and more than six hours of video were recorded. The video was transcribed by the authors with focus on the QAs used by the observers. Numerous QAs, more than 750 in total and more than 350 different QAs, were used by the expert observers. This data constitutes the basis for the validation of the CPQAs. Figure 10.8 shows a tag cloud of the top 25 words from the raw transcribed data [432].

![Tag cloud with the 25 top words from the raw transcribed data. The font size of a tag in the tag cloud is determined by the number of times the word has been used by the observers. Similar words have been grouped, such that details and detail are counted together as detail.](image)

Since many of the words and phrases from the observers are similar and some synonyms, the QAs from the experiment need to be categorized. Similar words and phrases should be grouped together, and relations between terms found. We have chosen to use existing definitions to accomplish this, and two different approaches can be taken with this method; top-down or bottom-up. In the top-down approach the relations are built from the most general QAs and downwards to the most specific QAs. This requires building a full tree structure with all relations, and then comparing it to the QAs used by the observers. In the bottom-up approach, the starting points are the QAs used by the observers. These QAs are grouped into more general attributes till the most general QA is reached. The advantage is that it does not require building a full tree structure prior to the analysis. Therefore, the bottom-up approach was chosen to validate the CPQAs.
An example of how the analysis is carried out; the observer has used the QA hue shift, this QA belongs to the more general QA hue. Using the relations of Pedersen et al. [345, 346] and the definition by Wyszecki and Styles [479], hue is considered a part of the more general color QA, which is one of the CPQAs (Figure 10.9).

![Figure 10.9: Bottom-up procedure for the attribute hue shift, which belongs to the hue attribute, which in turn belongs to the color attribute.](image)

In the experiment observers described quality issues in the reproductions, differences between the original and the reproductions, and differences between the reproductions. Since the physical CPQA was not changed in the experiment, we limit the discussion to five of the six CPQAs, excluding the physical CPQA.

The bottom-up approach described above has been used to generate a tree for all the images and observers in the experiment (Figure 10.10). In the following, we will show how the QAs from the expert observers have been grouped and fitted to the CPQAs, but will only present a part.

The observers have used many specific terms regarding color, such as red hue shift, yellow hue shift, and blue hue shift. All of these terms indicate a hue shift, which is a child of the hue attribute. A drift in hue also indicated a hue shift. Color dominance and color cast were used to indicate a general hue shift.

The observers specifically indicated which colors had an increase in saturation. They also tended to indicate which image was more saturated than another, rather than the other way around. Also, saturation loss was used in a more general way to indicate a global loss, while saturation increase was often used for a color or a region. Increase and loss of saturation are considered by the authors to be a shift in saturation. Shift in saturation is a sub-QA of saturation. Observers also used the following terms to describe saturation: intensity, chroma, purity, colorfulness, and vividness. Chroma is used for saturation in the Munsell color system [33], in the ISCC-NBS lexicon vividness is a level of saturation [33], purity is also a synonym for saturation [453], colorfulness is considered the same as chroma [154, 453], and intensity is used about saturation density [425]. Based on these definitions these terms are considered as equal to saturation.

Saturation and hue are considered as children of the color attribute. For the general color attribute observers used terms as color shift, color reproduction, and color rendering. In addition, the observers used the term chromaticity. Since the definition of chromaticity by Oleari [328] contains both hue and saturation we can set this attribute as equal to the color attribute.

### 10.4.3.1 Discussion on the fitting of quality attributes

Several issues were encountered while fitting the QAs.

**Overlapping QAs** Some of the QAs used by the observers are difficult to group within only one of the CPQAs. Naturalness is one of these attributes. We have argued earlier
WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?

Figure 10.10: The QA tree generated from the attributes used by four expert observers. Each level of a sub-tree has been sorted from left to right based on frequency (high to low).
(Section 10.3) that naturalness could be accounted for by using several of the main or sub-
attributes [345, 346]. In this experiment the observers used several QAs together with natu-
ralness, very often a change in one or several of the other QAs lead to the impression of an
unnatural image. In the five observations, where naturalness was used, the observers used the
term color in all of them, contrast in three of the five, and memory colors in four of the five
observations. In addition, it has been shown that naturalness depends on chroma and color-
fulness [217], contrast [217], and memory colors [488]. Because of this, naturalness is most
likely accounted for if these QAs are of reasonable quality.

The word gamut was also used by the observers, which is defined as the range of a set
of colors [395]. Gamut cannot be listed as a sub-QA under one of the CPQAs, since it is
dependent on both the color CPQA and the lightness CPQA. In the three observations where
gamut was used, both the lightness and the color QAs were used. Therefore, gamut can be
accounted for using the color and lightness CPQAs.

Readability and legibility are two terms from the experiment, which have been found to
be related in the literature [504], and they are often used about textual information. Research
has shown that contrast is important for text readability [240] and text legibility [265]. These
terms will also be influenced by sharpness. In five of the eleven observations where legibil-
ity and readability were used, the observers also used contrast and sharpness, in the remaining
six observations either sharpness or contrast was used. This indicates that legibility and read-
ability most likely can be accounted for with the CPQAs.

Memory colors are placed under the color CPQA, as the observers only specified color
changes (saturation and hue) for these, and not changes in lightness. However, there might
be situations where lightness should be considered as well, and memory colors in terms of
lightness will become a sub-QA of lightness.

**Independence** Dynamic range is considered as a sub-QA of lightness, but it has also been
shown to influence contrast [62]. In the two observations with dynamic range, the observers
indicated a relation to visibility of details. This issue is linked to the use of the phrase "too
dark", which was often used together with "detail loss". In these cases, the observers per-
ceived the regions where shadow details were lost, as larger dark uniform areas compared to
the original, and used the term "too dark" or "darker".

The experimental data indicates that contrast is influenced by saturation and lightness, but
also that contrast is linked to detail. Since the definition of contrast contains both color and
lightness it is perhaps the least independent QA. Furthermore, the experimental data shows
that the observers often use the contrast attribute separated from the color and lightness attri-
butes, making contrast a very complex QA. Contrast is also important to account for both
naturalness and readability. Without the contrast CPQA we would not cover the whole field
of quality, and it is therefore required in order to fulfill the criteria on which the CPQAs were
selected, even at the expense of independence.

We carried out a cross-tabulation and chi-square analysis to investigate the dependence
between the CPQAs. The null hypothesis $H_0$ was that there was no relationship between
two CPQAs. The alternative hypothesis $H_1$ was that there was a relationship between two
CPQAs. The p-values from this analysis are shown in Table 10.2. For some combinations
of two CPQAs given a 5% significance level, $H_0$ was rejected in favor of $H_1$. The input data to
the analysis was whether or not one of the five CPQAs was used by the observers for each
of the 25 images. The disadvantage of this analysis is that it does not give any information
on the nature of the relationship between the CPQAs, it only gives information about when
two CPQAs are used together. However, from the results we see a dependence between ar-
WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?

Table 10.2: P-values from cross-tabulation and chi-square analysis. With a significance level at 5%, there is a dependence between artifacts and lightness, contrast and lightness, and barely between artifacts and sharpness.

<table>
<thead>
<tr>
<th></th>
<th>Sharpness</th>
<th>Lightness</th>
<th>Artifacts</th>
<th>Contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharpness</td>
<td>0.405</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lightness</td>
<td>0.566</td>
<td>0.198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artifacts</td>
<td>0.781</td>
<td>0.048</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td>0.423</td>
<td>0.230</td>
<td>0.047</td>
<td>0.764</td>
</tr>
</tbody>
</table>

tifacts and lightness, which was also found earlier (Section 10.3.4) [345, 346]. There is also dependence between artifacts and sharpness, and contrast and lightness. The observers indicated a relation between contrast and dynamic range, one of the sub-QAs of lightness. The dependence analysis between dynamic range and contrast did not reveal a relation, neither for detail visibility and dynamic range, nor for detail loss and dark. The reason for this might be the amount of data, four observers is too low for these specific QAs (Figure 10.12). Anyhow, since this analysis does not cover the nature of the relations, further experiments are required to investigate the dependence between the CPQAs, where information on the rating of each CPQA is needed. This information could not be gathered in our experiment since the observers did not have prior knowledge about the CPQAs.

Global and local issues During the experiment the observers looked at both global and local quality issues. The QAs above can be divided into global and local attributes, and this differentiation can be important for the assessment of IQ, but also in the method used to combine results from different QAs.

One child with several own children In the tree structure (Figure 10.10) some QAs have only one child, and this child has several own children. It could then be argued that these QAs could be discarded, and the QA below could be linked to the parent of the removed QA. For example, saturation could be removed and replaced with saturation shift. However, observers have used these terms, as saturation alone, without further specification, which indicates that these levels are important and should be kept. Furthermore, in other situations there might be several children, such as for the edge QA, where one could suggest having two children as for the detail QA, one for edge loss and another for edge enhancement.

Skewness The experimental data identifies skewness in terms of the number of sub-QAs between the different CPQAs, which was also found in Section 10.3.3 or see Pedersen et al. [345, 346]. Our experimental data shows that the color CPQA has significantly more sub-QAs than the other CPQAs. This can be used as an additional argument for separating lightness from the color CPQA, in order to reduce skewness. Additionally, separating these enable the CPQAs to be easily adapted for the evaluation of grayscale images. The disadvantage of skewness between the CPQAs is that it is not straightforward to combine IQ values from the different CPQAs to one overall IQ value, since the CPQAs might have unequal weights.

Dimensionality Since all of the CPQAs have been used by the observers in the experiment, none of the CPQA can be removed directly to reduce the number of CPQAs. However, the color and lightness CPQA could be merged, but at the cost of increased skewness. There were
39 observations where the observers used both color and lightness, indicating that the observers differentiate between these CPQA. There were also nine observations where lightness was addressed without color.

It is not unlikely that the dimensionality can be reduced for specific workflows with specific documents. Therefore we have also looked at the usage of the CPQAs for the non-photographic documents. For these images all the CPQAs have been used, and for this workflow none of the CPQAs can be removed. However, there might be situations where only a part of the CPQAs are used to evaluate IQ.

### 10.4.4 Observations on the color printing quality attributes

The experimental data also leads to different observations on the CPQAs, which can be valuable in the assessment of IQ. Figure 10.11 shows the frequency of use of the CPQAs in the experiment. Color is the CPQA used most frequently by the observers, closely followed by sharpness. Artifacts is the least used CPQA by the experts. The results here indicate that the color and sharpness CPQAs should be evaluated in all images for IQ assessment. The low number of observations regarding artifacts could indicate that this CPQA only needs to be accounted for in specific images, since the artifacts might be dependent on the characteristics of the image. One example is banding, which has been perceived by some observers in the images with large uniform areas, but not in the other images.

![Figure 10.11: Frequency of use for the CPQAs for the 100 observations. Color is the CPQA used the most by the observers, closely followed by sharpness. Lightness is the third most used CPQA, with contrast being the fourth most used. Artifacts is the least used CPQA.](image)

It is also interesting to look at the distribution of sub-QAs within the CPQAs. Detail is the most often used sub-QA (Figure 10.12), closely followed by hue. Since the observers paid much attention to loss of detail in the shadow regions, and since the rendering intents reproduced these regions differently, detail is not surprisingly the most used sub-QA. The perceptual rendering intent gave a slight hue shift in some of the images, which was often noticed by the observers, resulting in the frequent use of this attribute. The sub-QAs of artifacts are the least used, most likely since these are very specific. The artifact CPQA will contain many sub-QAs since it will cover many different artifacts.

It has been suggested that the first quality issue noticed by the observer is likely to be the most important. We have analyzed this aspect. Figure 10.13 shows that color is by far the most frequent first attribute used by the observers, dominating more than in the frequency table for the whole experiment (Figure 10.11)
Figure 10.12: Distribution of sub-QAs within the CPQAs. Detail is most used.

Figure 10.13: Number of observations based on the first CPQA used by the observers.
10.4.5 Dependence between the attributes

In order to assess dependence between the QAs a second experiment was carried out. The observers rated overall quality and five different quality attributes on a seven step scale, where the ratings were used as basis for the analysis. First the correlation values between the ratings for the attributes were calculated (Table 10.3). As seen in the table non of the attributes have a very high correlation between them, the highest correlation values are found for sharpness and contrast (0.52) and lightness and color (0.51). The highest correlation values are found between the attributes and overall quality, which is expected since the observers base their decision of quality on the attributes. Artifacts is the attribute with the lowest correlation against overall quality, but the experimental data showed little differences in this attribute compared to the other attributes.

Table 10.3: Correlation values between the attributes based on the ratings from the observers.

<table>
<thead>
<tr>
<th></th>
<th>Quality</th>
<th>Color</th>
<th>Lightness</th>
<th>Contrast</th>
<th>Sharpness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lightness</td>
<td>0.63</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td>0.65</td>
<td>0.41</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sharpness</td>
<td>0.59</td>
<td>0.40</td>
<td>0.38</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>Artifacts</td>
<td>0.45</td>
<td>0.37</td>
<td>0.36</td>
<td>0.37</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Additionally, we have calculated the z-scores based on the category ratings from the observers for each image for overall quality and for the five quality attributes. These z-scores have been used to create a contingency table based on the correlation between between the z-scores of the attributes. For each of the 25 images in the experiment z-scores have been obtained, and the z-score of one attribute is compared against the z-scores of the other attributes, where correlation between them is used as a measure of similarity. The mean over the 25 images for each combination has been computed and reported in Table 10.4. A high mean correlation value would indicate a relationship between the attributes. Between the attributes the highest correlation is 0.54, which is found between sharpness and color, and sharpness and contrast. The highest correlation values are found between overall quality and the attributes, which reflects the results from Table 10.3.

Table 10.4: Correlation values between the attributes based on the z-scores from the observers.

<table>
<thead>
<tr>
<th></th>
<th>Quality</th>
<th>Color</th>
<th>Lightness</th>
<th>Contrast</th>
<th>Sharpness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lightness</td>
<td>0.68</td>
<td>0.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td>0.51</td>
<td>0.06</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sharpness</td>
<td>0.40</td>
<td>0.54</td>
<td>0.38</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Artifacts</td>
<td>0.26</td>
<td>0.14</td>
<td>0.42</td>
<td>0.14</td>
<td>-0.16</td>
</tr>
</tbody>
</table>

A PCA analysis of the ratings from the observers for each CPQAs show that all attributes contributes, and that none of the attributes can be removed (Table 10.5). By using only one CPQA 54% of the variance is accounted for, while using four CPQAs 92% is accounted for, while the remaining 8% is accounted for by using all CPQAs.
What quality attributes are the most important?

Table 10.5: Proportion each principal components (CPQAs) contributes the representation of total variance.

<table>
<thead>
<tr>
<th>Total variance proportion</th>
<th>One CPQA</th>
<th>Two CPQAs</th>
<th>Three CPQAs</th>
<th>Four CPQAs</th>
<th>Five CPQAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>One CPQA</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two CPQAs</td>
<td></td>
<td>0.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three CPQAs</td>
<td></td>
<td></td>
<td>0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Four CPQAs</td>
<td></td>
<td></td>
<td></td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Five CPQAs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

We can also analyze the linear correlation between the CPQAs, as independent variables, to the overall quality ratings, dependent variable, by using multiple linear regression. However, this is under the assumption that the relationship between the variables is linear. The correlation coefficients from this analysis shows a low correlation between the single attributes and the overall image quality ratings. The $R^2$ gives a value of 0.66, which indicates how well the regression line fits the data points. All CPQAs contribute to a positive fit of the a linear combination of them, indicating that none can be removed. The analysis also indicates that color and lightness contributes most to the fitting.

10.4.6 Validation summary

Prior to the experiment, we specified four requirements for the CPQAs to be validated. First the CPQAs should cover the entire field of IQ. This is fulfilled if all the QAs recorded in the experiment can be fitted within one or several of the CPQAs. This requirement is satisfied, and all the recorded QAs are accounted for within the CPQAs, either directly as a CPQA or as a sub-QA, or by using two or more of the CPQAs.

The second requirement was to have as few overlapping QAs as possible. Some of the recorded QAs overlap, such as naturalness, gamut, readability, and legibility. These overlapping QAs have been used totally 15 times, only a small percentage of the total number of QAs used. The overlapping QAs can be accounted for using two or more of the CPQAs. We consider the number of overlapping QAs to be acceptable, and the overlapping QAs are not frequently used by the observers. Thus the CPQAs satisfy the second requirement.

The third requirement was to keep the dimensionality to a minimum. None of the CPQAs can be directly removed, and all CPQAs have been used by the observers. However, as discussed above the division between color and lightness has advantages and disadvantages. They could possibly be combined into a single QA. Nonetheless, without merging color and lightness, and considering the use of only lightness by the observers, the third requirement is also satisfied.

The last requirement regarded dependence. The experimental results show some dependence between CPQAs, but as stated previously [345, 346] (Section 10.2.7) the CPQAs are not fully independent, because it is very difficult, if not impossible, to account for all quality issues while maintaining a low dimensionality. The experimental results indicate that contrast is the least independent CPQA. However, contrast cannot be removed since it is often used by the observer. For that reason we consider the dependence found to be acceptable, but care must be taken if the quality values for each CPQA are combined into an overall IQ value.
10.5 Summary

This chapter has investigated QAs for the evaluation of IQ. Existing QAs used in the literature have been identified, and based on these a subset of attributes have been proposed, named the Color Printing Quality Attributes (CPQAs). The CPQAs consist of six attributes covering the field of IQ. These attributes have been validated in two different experiments, and have been deemed as a good starting point for evaluating IQ for color prints.
WHAT QUALITY ATTRIBUTES ARE THE MOST IMPORTANT?
11 IMAGE QUALITY METRICS TO MEASURE QUALITY ATTRIBUTES

With a framework for using IQ metrics with prints and a set of attributes, the next step is to select suitable IQ metrics for the CPQAs. Then these metrics should be evaluated to investigate their correspondence with the percept for each CPQA. We will investigate five of the six CPQAs, leaving out the physical CPQA since this was not evaluated by the observers.

11.1 Selection of image quality metrics for the color printing quality attributes

Numerous IQ metrics have been proposed in the literature, as seen in Section 4. These take into account different aspects of IQ, and therefore a single metric might not be suitable to measure all CPQAs. Because of this we will first discuss the properties that the metrics should have for each CPQA, before a subset can be selected for further evaluation.

11.1.1 Sharpness

Sharpness is related to edges and details, and IQ metrics for sharpness should account for these two factors. Most of the metrics for sharpness found in the literature work by the principle to find edges in the image, calculating quality in a local region at the edges, and then combining the values from the local regions into an overall value representing sharpness quality [229]. Some of these metrics, such as those proposed by Zhang et al. [493], correlate well with perceived sharpness. However, these methods do not directly take into account details. Another approach, which better takes into account details, is based on structural similarity, where the metrics usually work on local neighborhoods, such as SSIM [458]. There are also other metrics that are based on the visibility of salient objects to assess detail preservation, see for example Cao et al. [61] or Appendix C. Previous analysis of sharpness indicates a relation to contrast [346], therefore metrics accounting for local contrast could be suitable to assess sharpness. Research has also shown that perceived sharpness (blur) is dominated by luminance, and more important than chromatic information [226]. This is also supported by Webster et al. [471], who found that observers are very sensitive to luminance blur, and sensitivity to chromatic blur is weak. Therefore, it is not required that the metrics take into account chromatic information.
11.1.2 Color

There are many potential IQ metrics for the color CPQA. First of all the metrics for this CPQA should account for color information, making all metrics based on color differences possible candidates. However, applying color difference formulas directly, such as the $\Delta^{*}_{ab}$, will most likely not predict perceived quality since the impression of quality is influenced by the viewing conditions. Therefore, the metrics should incorporate a simulation of the HVS, such as the S-CIELAB [499] that performs a spatial filtering before applying the CIELAB color difference formula. These color difference formulas, which many IQ metrics use, usually consist of three parts, one for lightness and two for chromaticity. Since lightness and color are separate CPQAs, they should be evaluated separately. As a result the chromaticity part of these metrics will most likely be more suitable than using all three parts. Additionally, it has also been argued in the literature that some regions are more important than others. Therefore, metrics such as SHAME [351], which use different weighting functions are potentially suitable metrics. For more information on SHAME we refer to Chapter 5.

11.1.3 Lightness

Metrics for the lightness CPQA should mainly follow the same principles as for the color CPQA. IQ metrics based on color differences commonly calculate lightness quality separated from color, such as S-CIELAB [499]. Metrics working only on grayscale can also be suitable if they analyze lightness, such as SSIM [458] that performs a comparison of lightness between the original and the reproduction. However, metrics analyzing specific aspects in the lightness channel are most likely not suitable.

11.1.4 Contrast

The definition of the contrast CPQA states that contrast is both global and local, as well as dependent on lightness and chromaticity. Therefore metrics for the contrast CPQA should account for these aspects. Metrics computing contrast over a local neighborhood for further combining the local values into a global value for contrast are potentially suitable metrics, such as $\Delta$Local Contrast ($\Delta LC$) [25] or SSIM [458]. Nevertheless, most contrast metrics do not account for chromaticity, and therefore they might not be optimal to measure contrast. One of the metrics that uses chromaticity to calculate contrast is the Weighted multi-Level Framework (WLF) [417], which also takes into account locality and globality. This metric is explained more in details in Appendix B.

11.1.5 Artifacts

The artifacts CPQA contains many different and specific QAs. Because of this it might be difficult to find an overall metric to evaluate all aspects of this CPQA. However, artifacts have some common denominators, if they are not detectable, they do not influence image quality. Therefore, the metrics used to measure artifacts should account for the sensitivity of the human visual system, such as VSNR [70] that has a threshold for when artifacts are perceptible. The characteristics of artifacts is also an important issue, for noise metrics based on local contrast might be suitable, as $\Delta LC$ [25], since noise affects local contrast. Artifacts like banding can be detected by metrics that are using edge-preserving filters, for example the ABF metric [459],
opposite of metrics performing non-edge preserving filtering, as the S-CIELAB [499]. For some sub-artifacts specific metrics might be required, such as for error diffusion worms that have specific characteristics specially designed metrics as the Error Diffusion Worm Measure (Appendix D) are the most suitable.

11.2 Experiment I

11.2.1 Selected image quality metrics

We cannot evaluate all of the IQ metrics in the literature for all the CPQAs, and because of this a sub-set of metrics were selected, as shown in Table 11.1. The selection is based on the results from previous evaluation [169, 343] (Chapters 7 and 9), the criteria on which the metrics were created, the discussion above, and their popularity. Many of the metrics have been created to assess some aspects of IQ, and therefore only the suitable metrics for each CPQA will be evaluated. Furthermore, for specific CPQAs we also evaluate parts of the metrics. For example, S-CIELAB combines the lightness and color differences to obtain an overall value. When suitable, we will evaluate these separately in addition to the full metric.

Table 11.1: Selected IQ metrics for the evaluation of CPQAs.

<table>
<thead>
<tr>
<th>Metric</th>
<th>CPQA</th>
<th>Sharpness</th>
<th>Color</th>
<th>Lightness</th>
<th>Contrast</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABF [459]</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Busyness [331]</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blurMetric [89]</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cao [61]</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>CW-SSIM[466]</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ΔLC [25]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LinLab [243]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MS-SSIM [467]</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>M-SVD [402]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF-SDM [495]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>S-CIELAB [497]</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>S-DEE [412]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAME [351]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM [458]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>VSNR [70]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLF [417]</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>YCXCzLab [244]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11.2.2 Evaluation of the selected image quality metrics

The IQ metrics proposed for the CPQAs should be evaluated in order to find the best metric for each CPQA. For doing this we compare the results from the IQ metrics against the subjective scores from an experiment. We will here only show the results for the best performing metrics.
11.2.2.1 Experimental setup

Two experimental phases were carried out for the evaluation of the metrics. 15 naïve observers participated in the first experimental phase, where they judged overall quality and the different CPQAs on a seven step scale for a set of images. In the second phase, four expert observers ranked the quality of a set of images and elaborated on different quality issues. We will give a brief introduction of the experimental setup, for more information see Section 10.4.2 or Pedersen et al. [347, 348].

We selected the same 25 images as in Figure 10.7, identically processed and printed. Two sets of images were printed at the same time, one set for the first phase and one set for the second phase. For more details on the processing of the images we refer the reader to Section 10.4.2.

The observers were presented with a reference image on an EIZO ColorEdge CG224 display for the first phase, and an EIZO ColorEdge CG221 for the second phase. The color temperature was set to 6500 K and the white luminance level to 80 \( cd/m^2 \), following the specifications of the sRGB. The printed images were presented in random order to the observers in a controlled viewing room at a color temperature of 5200 K, at an illuminance level of 450 ± 75 lux and a color rendering index of 96. The observers viewed the reference image and the printed image simultaneously from a distance of approximately 60 cm.

11.2.2.2 Evaluation of image quality metrics

We will compare the results of the metrics to the results of the observers, and the metric that correlates the best with the observers is the most suitable metric. However, the first step is to turn the printed images into a digital format. This is done with the framework presented above in Chapter 8. A HP ScanJet G4050 was used for scanning the images from the first experimental phase, while an Epson 10000XL for the second phase. The resolution was set to 300 DPI. The scanners were characterized with the same test target as used to generate the printer profile.

Since both experimental phases were carried out under mixed illumination, the CIECAM02 chromatic adaptation transform [81] was used to ensure consistency in the calculations for the metrics. The measured reference white point of the monitor and the media were used as input to the adaptation transform, following the CIE guidelines [81].

The evaluation of the metrics has been divided into two phases, one for the naïve observers and one for the expert observers. Each phase contained different methods for the evaluation adapted to the task given to the observers.

Phase 1: naïve observers In the first experimental phase each observer judged overall quality and the five CPQAs for each image, which enabled us to compute z-scores [116] for each of these. For the first phase 24 of the 25 images (Figure 10.7) were used in the experiment (the bride and groom image left out).

To assess the performance of the evaluated metrics we have adopted several different methods. In order to achieve extensive evaluation we will investigate the performance of the IQ metrics both image by image, and the overall performance over the entire set of images. The Pearson correlation [232] is used for the image-wise evaluation, comparing the calculated quality and observed quality. The mean of the correlation for each image in the dataset and the percentage of images with a correlation above 0.6 is used as a measure of performance. While for the overall performance, we will use the rank order method [350] presented in Sec-
tion 6.3, where the correlation between the z-scores from the observers and the z-scores of the metric is the indication of performance. However, for the rank order correlation one should consider that we only have three data points, and therefore it is also important to perform a visual comparison of the z-scores.

**Sharpness:** The results of the selected metrics for the sharpness CPQA are shown in Table 11.2. SSIM has a mean correlation of 0.29, but it has a correlation above 0.6 in 50% of the images. The rank order method used to evaluate the overall performance calculates z-scores for the metric, which can be compared against the z-scores from the observers. A metric capable of correctly measuring the CPQA will have z-scores similar to the z-scores from the observers. The correlation between the z-scores is used as a performance measure, and SSIM shows an excellent correlation (1.00) with a low p-value (0.03). Visual investigation of the z-scores from SSIM and the observers shows a striking resemblance, therefore SSIM seems to be a suitable metric for the sharpness CPQA. Other versions of the SSIM, as the Multi Scale-SSIM (MS-SSIM) and Complex Wavelet-SSIM (CW-SSIM) increase the performance. Since these account better for the viewing conditions they might be more robust than the single scale SSIM. Other metrics as the ΔLC and the Riesz-transform based Feature SIMilarity metric (RFSIM) also perform very well for this CPQA. We can also see from Table 11.2 that the metrics perform similar in terms of rank order correlation, and that for an overall evaluation of sharpness these metrics produce similar results.

Table 11.2: Performance of the metrics for the sharpness CPQA. Mean correlation implies that the correlation has been calculated for each image in the dataset, and then averaged over the 24 images. Percentage above 0.6 is the percentage of images where the correlation is higher than 0.6. The rank order correlation indicates the correlation between the metric’s z-scores computed with the rank order method (Section 6.3) and the observer’s z-scores for the CPQA, in addition the p-value for the correlation is found in the parenthesis.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW-SSIM</td>
<td>0.36</td>
<td>63</td>
<td>1.00 (0.05)</td>
</tr>
<tr>
<td>ΔLC</td>
<td>0.26</td>
<td>50</td>
<td>0.97 (0.14)</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.29</td>
<td>58</td>
<td>0.97 (0.15)</td>
</tr>
<tr>
<td>RFSIM</td>
<td>0.34</td>
<td>63</td>
<td>0.99 (0.09)</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.29</td>
<td>50</td>
<td>1.00 (0.03)</td>
</tr>
</tbody>
</table>

**Color:** None of the evaluated IQ metrics perform significantly better than the others, in terms of the mean correlation, percentage above 0.6 and rank order correlation (Table 11.3). The results here indicate that none of the evaluated metrics can accurately measure the color CPQA. It is interesting to notice that all of these metrics are based on color differences, which might indicate that to find a suitable metric for color one should look of metrics based on another principle than color differences.

**Lightness:** MS-SSIM shows the highest mean correlation for the evaluated IQ metrics (Table 11.4), it also has the highest percentage of images above 0.6 in correlation, and it has an excellent rank order correlation with a low p-value. However, the single scale SSIM also performs well. The other metrics in Table 11.4 also have a high rank order correlation, indicating that many metrics have an overall similarity to the observers rating.
Table 11.3: Performance of the metrics for the color CPQA. For further explanation see Table 11.2.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINLAB</td>
<td>-0.25</td>
<td>17</td>
<td>-0.93 (0.24)</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>-0.29</td>
<td>13</td>
<td>-0.95 (0.19)</td>
</tr>
<tr>
<td>S-DEE</td>
<td>-0.34</td>
<td>13</td>
<td>-0.92 (0.25)</td>
</tr>
<tr>
<td>SHAME</td>
<td>-0.04</td>
<td>21</td>
<td>-0.25 (0.84)</td>
</tr>
</tbody>
</table>

Table 11.4: Performance of the metrics for the lightness CPQA. For further explanation see Table 11.2. The subscript Lightness indicates only the lightness part of the metric.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔLC</td>
<td>0.31</td>
<td>50</td>
<td>0.94 (0.22)</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.50</td>
<td>63</td>
<td>0.99 (0.08)</td>
</tr>
<tr>
<td>S-CIELABLightness</td>
<td>0.14</td>
<td>46</td>
<td>1.00 (0.01)</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.32</td>
<td>58</td>
<td>1.00 (0.05)</td>
</tr>
<tr>
<td>VIF</td>
<td>0.34</td>
<td>54</td>
<td>0.99 (0.09)</td>
</tr>
</tbody>
</table>

Contrast: SSIM also performs well for this CPQA as seen in Table 11.5. It has a mean correlation of 0.32, 50% of the images have a correlation above 0.6, and the rank order correlation is fairly high. It is worth noticing that using just the contrast calculation in SSIM the number of images with a correlation above 0.6 is increased. Also by using MS-SSIM the number of images with a correlation above 0.6 is slightly increased compared to the single scale SSIM. ALC has the highest mean correlation, and in 58% of the images a correlation above 0.6, and the rank order correlation is high. It should be noticed that all metrics have fairly high p-values for the rank order correlation. The differences between the metrics are small for all performance measures, and therefore the results do not give a clear indication of which metric is the best.

Table 11.5: Performance of the metrics for the contrast CPQA. The subscript Contrast indicates only the contrast part of the metric. For further explanation see Table 11.2.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔLC</td>
<td>0.34</td>
<td>58</td>
<td>0.97 (0.17)</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.30</td>
<td>58</td>
<td>0.74 (0.47)</td>
</tr>
<tr>
<td>RFSIM</td>
<td>0.32</td>
<td>54</td>
<td>0.94 (0.22)</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.32</td>
<td>50</td>
<td>0.86 (0.34)</td>
</tr>
<tr>
<td>SSIMContrast</td>
<td>0.32</td>
<td>54</td>
<td>0.84 (0.37)</td>
</tr>
</tbody>
</table>

Artifacts: The observers gave similar results for the different rendering intents in terms of artifacts, because of this it is a very demanding task for the IQ metrics. The evaluation shows that RFSIM has the most images with a correlation above 0.6 together with MS-SSIM, but MS-SSIM has the highest mean correlation (Table 11.6). MS-SSIM and WLF have the highest rank order correlation, but they do also have high p-values. Therefore the results do not give a clear indication of a suitable metric. One should also consider that the artifacts CPQA contains many different sub-QAs, therefore it could be difficult to find just one IQ metric to measure overall artifact quality, and several metrics might be required.
Table 11.6: Performance of the metrics for the artifacts CPQA. For further explanation see Table 11.2.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-SSIM</td>
<td>0.23</td>
<td>46</td>
<td>0.61 (0.59)</td>
</tr>
<tr>
<td>RFSIM</td>
<td>0.14</td>
<td>46</td>
<td>0.26 (0.83)</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.09</td>
<td>29</td>
<td>0.44 (0.71)</td>
</tr>
<tr>
<td>WLF</td>
<td>0.02</td>
<td>33</td>
<td>0.76 (0.45)</td>
</tr>
</tbody>
</table>

**Phase 2: expert observers** In the second experimental phase a group of expert observers commented on quality issues in the reproductions. A video of the experiment was used by the authors to extract regions where the observers perceived quality issues. This enabled us to perform an in-depth evaluation of the IQ metrics, which ensures that the metrics are capable of measuring the different CPQAs. We will only include the metrics that performed well in the first evaluation phase, since these are the ones most likely to be suitable for the CPQAs.

We will use the picnic image (Figure 11.1) to evaluate the IQ metrics. The observers indicated that this image contained a wide variety of QAs and different quality issues. These quality issues are the important issues for the IQ metrics to detect. Based on the comments from the observers important regions have been found, each containing different quality issues:

- Tree: mainly details, but also lightness and contrast issues.
- Shoe: loss of details perceived in one of the reproductions.
- White shirt: a hue shift in one of the reproductions.
- Hair: a hue shift in the hair of the Asian girl in the middle.
- Pink shirt: one reproduction was too saturated.
- Grass: detail and saturation issues.
- Skin: a hue shift found in some reproductions.
- Cloth: one reproduction had a lighter red cloth than the others.
- Blanket: lightness issues.
- Sky: saturation and detail issues.

To evaluate the IQ metrics we compare the rank of the metrics, based on the mean value of each region, to the rank of the observers for each region. A mask for each region was created based on the comments from the observers (example shown in Figure 11.1(b)), and this mask was used to obtain the ranking from the metrics. The observers did not rank all reproductions for all regions or quality issues, but instead they indicated which one was the best or the worst. We consider it to be important for the IQ metrics to predict which reproduction that is clearly better or worse. In addition to the ranking of the metrics, a visual inspection of the quality maps from each IQ metric has been carried out by the authors. This visual inspection will reveal more information about the performance of the metrics than the mean value.
Table 11.7: Ranking for the different regions in the image where observers commented on quality issues. $P$ = perceptual rendering intent, $R$ = relative colorimetric rendering intent, and $B$ = relative colorimetric rendering intent with BPC. If $(R,P) > B$, then $B$ was ranked as the worst, but the observers did not rank the two other reproductions. () for the metric side indicates that the mean values are not significantly different with a 95% confidence level. A mask was created based on the comments from the observers, and the mean of the results from the IQ metric was used a basis for the ranking.

(a) SSIM

<table>
<thead>
<tr>
<th>Region</th>
<th>Observers</th>
<th>SSIM</th>
<th>Correct rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree</td>
<td>$P &gt; R$  ($B$)</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Shoe</td>
<td>$P &gt; R$  ($B$)</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>White shirt</td>
<td>$P &gt; B$  ($R$)</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Hair</td>
<td>($P,B$)  $&gt; R$</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Pink shirt</td>
<td>($P,B$)  $&gt; R$</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Grass</td>
<td>$P &gt; (R,B)$</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Skin</td>
<td>$R &gt; B &gt; P$</td>
<td>$P &gt; B &gt; R$</td>
<td>No</td>
</tr>
<tr>
<td>Cloth</td>
<td>$(B,R) &gt; P$</td>
<td>$P &gt; B &gt; R$</td>
<td>No</td>
</tr>
<tr>
<td>Blanket</td>
<td>$(R,B) &gt; P$</td>
<td>$P &gt; B &gt; R$</td>
<td>No</td>
</tr>
<tr>
<td>Sky</td>
<td>$P &gt; (R,B)$</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
</tbody>
</table>

(b) $\Delta LC$

<table>
<thead>
<tr>
<th>Region</th>
<th>Observers</th>
<th>$\Delta LC$</th>
<th>Correct rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree</td>
<td>$P &gt; R$  ($B$)</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Shoe</td>
<td>$P &gt; R$  ($B$)</td>
<td>$B &gt; R &gt; P$</td>
<td>No</td>
</tr>
<tr>
<td>White shirt</td>
<td>$P &gt; B$  ($R$)</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Hair</td>
<td>($P,B$)  $&gt; R$</td>
<td>$P &gt; B (B,R)$</td>
<td>No</td>
</tr>
<tr>
<td>Pink shirt</td>
<td>($P,B$)  $&gt; R$</td>
<td>($B,P) &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Grass</td>
<td>$P &gt; (R,B)$</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
<tr>
<td>Skin</td>
<td>$R &gt; B &gt; P$</td>
<td>$P &gt; B &gt; R$</td>
<td>No</td>
</tr>
<tr>
<td>Cloth</td>
<td>$(B,R) &gt; P$</td>
<td>$P &gt; B &gt; R$</td>
<td>No</td>
</tr>
<tr>
<td>Blanket</td>
<td>$(R,B) &gt; P$</td>
<td>$P &gt; B &gt; R$</td>
<td>No</td>
</tr>
<tr>
<td>Sky</td>
<td>$P &gt; (R,B)$</td>
<td>$P &gt; B &gt; R$</td>
<td>Yes</td>
</tr>
</tbody>
</table>
SSIM: For the first experimental phase SSIM had a high performance for the sharpness, lightness, and contrast CPQAs, and performed fairly well for the artifact CPQA. In the second phase SSIM was able to detect the correct order regarding details, and gives results similar to the observers, as seen from the tree, grass, and shoe regions in Table 11.7(a). The visual inspection supported this, and revealed that SSIM is able to detect even small loss of details. These findings correspond well with the results from the first experimental phase where SSIM was one of the best performing metrics for the sharpness CPQA. SSIM also correctly detected an area with a hue shift (hair), since this area in addition had a lightness shift. In the cloth region, where lightness differences were perceived by the observers, SSIM gave the correct ranking. SSIM also gave the same ranking as the observers in the tree region, where lightness and contrast were used by the observers. This shows that SSIM can be suitable to measure both lightness and contrast, but further analysis is required to ensure that SSIM is able to measure these CPQAs. We can also notice that SSIM gives similar ranking for all regions in the image.

ΔLC: In the first experimental phase ΔLC had one of the best performances for the sharpness CPQA. It is able to detect the detail issues in the grass and tree regions as seen in Table 11.7(b). This is also the reason why it performs quite well for the rank order correlation for the sharpness CPQA in the first experiment phase. ΔLC also gave good results for the lightness CPQA, but it does not predict the lightness issues commented on by the expert observers. For the contrast CPQA ΔLC correlated with the observers from the first experimental phase. The experts did not directly comment on contrast, but in the pink shirt one of the reproductions was too saturated, which is one of the aspects looked at when evaluating contrast. ΔLC correctly detects this issue, even though it does account for chromaticity. Additionally, it correctly detects the detail issues as discussed above, which is also one of the aspects of contrast. However, since ΔLC does not account for chromaticity, it might not be suitable to evaluate all aspects of the contrast CPQA.

MS-SSIM: MS-SSIM is one of the best performing metrics for the artifact CPQA. The analysis for the second experimental phase is difficult, since MS-SSIM is a multilevel approach it is difficult to compute a value for each region. However, investigation of the SSIM maps for the different levels reveals that MS-SSIM is able to detect banding in several of the
levels. The reason for the higher performance for the artifact CPQA and other CPQAs compared to the single scale SSIM might be that the multi-scale version takes better into account the viewing conditions, due to the subsampling of the image. The analysis for SSIM above is also partly valid for MS-SSIM, since the first level of MS-SSIM is the same as SSIM.

**CW-SSIM:** CW-SSIM does not compute a map like SSIM and some of the other metrics. But it is based on some principles that makes it suitable for some aspects. Linear and uniform changes correspond to lightness and contrast differences to which CW-SSIM is not sensitive because the structure is not changed, this is the reason why CW-SSIM most likely does not work well for contrast and lightness. However, in the sharpness CPQA where details are lost in several regions, the structure is changed and therefore it is suitable for sharpness.

**RFSIM:** This metric does not produce a quality map either, and therefore it is difficult to assess its performance in the second experimental phase. However, RFSIM uses an edge map to detect key locations in the image. A result of this is that changes that occur at edges are the main component for the quality calculation, which explains why RFSIM gives good results for sharpness, since sharpness is related to the definition of edges [66]. This would also make RFSIM suitable for certain artifacts, such as banding, and could also explain why it is one of the best metrics for the artifacts CPQA. However, the artifacts CPQA contains many different artifacts, and therefore one cannot conclude that RFSIM is the best metric to measure the artifacts CPQA.

11.2.3 Investigation of image characteristics

The input image to a printing system has different Image Characteristics (ICs), which can be defined as a collection of properties, that can represent an aspect of an image [429], for example, image gamut, image sharpness, or image contrast. These ICs will influence the quality of the image differently, and they are dependent on the processing of the image. Information about ICs can be used for deciding the optimal processing of the image, such as the method and parameters, but also to decide which IQ metrics that should be used in the evaluation of the image after it has been processed. As indicated in Chapters 9 and 10 and by Hardeberg et al. [169] the performance of metrics might be dependent on ICs. Therefore, in this section we investigate the influence of a set of characteristics on the performance of IQ metrics.

11.2.3.1 Selected image characteristics

Many different ICs are considered to influence IQ, such as out of gamut pixels [108, 155, 280, 475], contrast [20, 38, 155], edges [41, 166], colorfulness [92, 94, 171], saturation [426], chroma [20, 155], image resolution [27], gamut size [12, 306, 475], gradients [49, 75], memory colors [96, 147, 488], uniform areas [19], detail level [19, 41, 155], image lightness [20, 50, 155, 306], and dominant color [50]. We will investigate some of these, more precisely lightness, colorfulness, dominant color, and details. Lightness is often considered as having a major impact on the processing and the final quality [20, 50, 155, 306], and it is therefore one of the most important ICs. Colorfulness has also shown to greatly influence IQ [94, 171], being one of the ICs considered. Many images have a dominant color, and preservation of this color is important for the impression of quality [50]. Details is also one
Table 11.8: Dominant color distribution based on the classification by two expert observers.

<table>
<thead>
<tr>
<th>Dominant color</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>11</td>
</tr>
<tr>
<td>No</td>
<td>13</td>
</tr>
</tbody>
</table>

of the ICs that observers use to evaluate quality, and preservation of details is crucial for obtaining high quality images [41, 284, 429]. These four different ICs constitutes the basis for our analysis.

11.2.3.2 Experiment

The experimental images (Figure 10.7) are the basis for our investigation, and as done in Section 11.2.2 24 of the 25 images have been used. These images have been classified by two expert observers into a set of classes according to the selected ICs.

For dominant color the classes were "yes" and "no". For colorfulness and lightness the classes were "high", "medium", "low", and "mixed". For details the classes were "very detailed", "detailed", "smooth", "very smooth", and "both detailed and smooth".

The two experts carried out the experiment together, and were allowed to discuss each image and to come up with a decision on what class the image belonged to. The images were viewed on a calibrated EIZO ColorEdge CG224 from a distance of 60 cm, with similar viewing conditions as in Section 11.3.2.

The results for the metrics are based on the data collected in Section 11.2.2.2 for the naïve observers, where the images have been evaluated by a group of observers and metrics are calculated on the images. The observer data has then been compared against the metrics. We will calculate the performance, in terms of correlation, for images classified according to a given IC.

11.2.3.3 Dominant color

The two experts divided the images for dominant color into two more or less equally large groups (Table 11.8), 11 images were classified as having a dominant color, while the remaining 13 did not have a dominant color.

By computing the Pearson correlation between the IQ metric value and the z-scores from the observers for each image, we can analyze the performance for the images that are classified as having a dominant color to those classified as not having a dominant color. Each image has three reproductions (Section 10.4.2.2), resulting in a correlation value for each image with an uncertainty. However, the uncertainty is decreased by taking the mean correlation value of the values for one group, making the results more reliable results.

One could hypothesize that in images with a dominant color, one would prefer to have a color as close as possible to the original color (smallest color difference), since larger regions of the same color should be considered as more important than smaller regions [177].

Our starting point is the S-CIELAB metric [499] as introduced in Section 4.2.2.1, since this is one of the most commonly used metrics in the literature. The observer data is based on the color attribute. Figure 11.2(a) shows the correlation values for the images in the two classes, when the metric values are compared to the color quality values. The mean correlation for the 11 images classified as having a dominant color is 0.11 (standard deviation 0.66), while
for the 13 images classified as not having a dominant color the correlation is 0.44 (standard deviation 0.59). This could be an indication that the hypothesis is not valid.

We then keep the same hypothesis that if a dominant color is apparent in the image, it is probably important to maintain that color, and for images without a dominant color this is not as important. For the SSIM metric that performs well for sharpness, one could argue that for images without a dominant color SSIM would work better. We have computed the results for SSIM for the two classes (Figure 11.2(b)). The sharpness evaluation done by the observers are the basis of the correlation. SSIM performs quite well for the images classified as not having a dominant color, and for those having a dominant color we see larger within-class variations. The mean correlation for images not having a dominant color is 0.53 (standard deviation 0.73), while for the other class 0.01 (standard deviation 0.57).

Figure 11.2: Performance of the metrics for images classified as having a dominant color (red) and images classified as not having a dominant color (blue).
Table 11.9: Colorfulness distribution based on the classification by two expert observers.

<table>
<thead>
<tr>
<th>Colorfulness</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>10</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

### 11.2.3.4 Colorfulness

The experts also classified the images according to colorfulness, which enables us to carry out a similar analysis as for dominant color (Table 11.9).

We investigate the performance of S-CIELAB for the different classes of colorfulness, using the results from the observers for the color CPQA. From Figure 11.3(a) we can see that S-CIELAB performs well for the images classified as having medium colorfulness, while the correlation for the other groups are varying (Figure 11.3(a)). Even by combining the low and medium groups we have a higher correlation (0.34) compared to the high group (0.24). The mean correlation values are found in Table 11.10.

Table 11.10: Mean correlation for S-CIELAB for images classified according to colorfulness.

<table>
<thead>
<tr>
<th>Colorfulness</th>
<th>Correlation</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.01</td>
<td>0.65</td>
</tr>
<tr>
<td>Medium</td>
<td>0.48</td>
<td>0.55</td>
</tr>
<tr>
<td>High</td>
<td>0.24</td>
<td>0.65</td>
</tr>
<tr>
<td>Mixed</td>
<td>0.13</td>
<td>1.22</td>
</tr>
</tbody>
</table>

SSIM performed well for the sharpness CPQA, but in some images it does not have a high correlation. We investigate the performance of SSIM based on the colorfulness. Since SSIM is a grayscale metric, and it does not take into account color information, one can hypothesize that it performs better for images with low colorfulness than those with high colorfulness. This is to a certain extent seen in Figure 11.3(b), where SSIM performs excellent for the four images classified as having low colorfulness. It also performs well for those with mixed colorfulness, while for those classified as medium and high we find some high correlation values, but also some poor ones (Table 11.11). Observer data used for this analysis are the results for the sharpness CPQA.

Table 11.11: Mean correlation for the SSIM metric for images classified according to colorfulness.

<table>
<thead>
<tr>
<th>Colorfulness</th>
<th>Correlation</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.94</td>
<td>0.05</td>
</tr>
<tr>
<td>Medium</td>
<td>-0.05</td>
<td>0.77</td>
</tr>
<tr>
<td>High</td>
<td>0.27</td>
<td>0.53</td>
</tr>
<tr>
<td>Mixed</td>
<td>0.80</td>
<td>0.13</td>
</tr>
</tbody>
</table>

### 11.2.3.5 Lightness

Lightness was classified into low, medium, high, and mixed (Table 11.12). When we look at the results for the SSIM values against the sharpness CPQA results from the observers
Figure 11.3: Colorfulness classified images and the performance of the metrics.
Table 11.12: Lightness distribution based on the classification by two expert observers.

<table>
<thead>
<tr>
<th>Lightness</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lightness</td>
<td>6</td>
<td>13</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

for these classes we see that SSIM performs well for the images with low lightness (Figure 11.4 and Table 11.13). This is not surprising, mainly because images with pixels below the black point of the printer can lose details in these regions, especially the images that have been processed with the relative colorimetric rendering intent using a gamut clipping algorithm. These clipping algorithms usually preserve saturation, but clip image details [42].

Figure 11.4: SSIM performance for images classified with different lightness levels.

Table 11.13: Mean correlation for the SSIM metric for images classified according to lightness.

<table>
<thead>
<tr>
<th>Lightness</th>
<th>Correlation</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.92</td>
<td>0.11</td>
</tr>
<tr>
<td>Medium</td>
<td>-0.01</td>
<td>0.68</td>
</tr>
<tr>
<td>High</td>
<td>0.35</td>
<td>0.73</td>
</tr>
<tr>
<td>Mixed</td>
<td>0.26</td>
<td>0.73</td>
</tr>
</tbody>
</table>

11.2.3.6 Details

For the details IC we have grouped the answers for "very smooth" and "smooth" together, and also for "detailed" and "very detailed". For these three groups ("smooth", "both smooth and detailed", and "detailed" as seen in Table 11.14) we look at the performance of the SSIM metric. Since SSIM is based on structural similarity it should be more suitable for images with a higher detail level than those being smooth. Figure 11.5 and Table 11.15 show the performance of the SSIM metric compared against the observers on the sharpness CPQA.
**Table 11.14**: Detail distribution based on the classification by two expert observers.

<table>
<thead>
<tr>
<th>Details</th>
<th>Smooth</th>
<th>Both smooth and detailed</th>
<th>Detailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth</td>
<td>8</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Both smooth and detailed</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detailed</td>
<td></td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

For this dataset we do not see any clear indication of a dependence of detail level on the performance of SSIM.

![SSIM for images with different detail level for the sharpness CPQA.](image)

**Figure 11.5**: SSIM for images with different detail level for the sharpness CPQA.

**Table 11.15**: Mean correlation for the SSIM metric for images classified according to details.

<table>
<thead>
<tr>
<th>Colorfulness</th>
<th>Correlation</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth</td>
<td>0.39</td>
<td>0.80</td>
</tr>
<tr>
<td>Both smooth and detailed</td>
<td>0.33</td>
<td>0.60</td>
</tr>
<tr>
<td>Detailed</td>
<td>0.21</td>
<td>0.68</td>
</tr>
</tbody>
</table>

### 11.2.3.7 Overall observations

The analysis carried out in this section indicates that there might be a relationship between ICs and the performance of IQ metrics, such as for SSIM when it comes to images with low lightness. However, the number of images in the dataset, only 24, is too low to draw a conclusion. Further work with more images is required to determine the influence of ICs on the performance of IQ metrics, and other ICs than those considered here should be subject for further investigations.
11.3 Experiment II

11.3.1 Selected image quality metrics

Many metrics have been proposed (Chapter 4), and as in the previous section a selection needs to be done. For this experiment we select the same metrics as for the first experiment (Table 11.1), but we supplement with some additional metrics to ensure a thorough evaluation. Since many of the metrics are designed to account for specific aspects, only the ones suitable for a given CPQA are evaluated. An overview of the 23 metrics selected for the evaluation and the CPQAs they evaluate is found in Table 11.16.

Table 11.16: Selected IQ metrics for the evaluation of CPQAs.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Sharpness</th>
<th>Color</th>
<th>Lightness</th>
<th>Contrast</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABF [459]</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Busyness [331]</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blurMetric [89]</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cao [61]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>CW-SSIM [466]</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ΔLC [25]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>IW-SSIM [460]</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LinLab [243]</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS-SSIM [467]</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>M-SVD [402]</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>PSNR-HVS-M [366]</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>PSNR-HVS [366]</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RFSIM [495]</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RRIQA [465]</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>S-CIELAB [497]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>S-DEE [412]</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>SHAME [351]</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAME-II [351]</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SSIM [458]</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>VIF [396]</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>VSNR [70]</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>WLF [417]</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>YCXCzLab [244]</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

11.3.2 Evaluation of the selected image quality metrics

11.3.2.1 Experimental setup

We want to investigate the relationship between the percept of the CPQAs and IQ metrics. In order to do this we have carried out an experiment where human observers judge the quality of the CPQAs on a set of printed images.
Test images  Ten images (Figure 11.6) were selected from the ISO standards [199, 209]. The number of images follow the recommendation by Field [143], who recommend between five and ten images, and the CIE [80], who recommend at least four images. The images were selected to cover a wide range of characteristics, such as lightness from low to high levels, saturation from low to high levels, contrast from low to high levels, hue primaries, fine details, memory colors as skin tones. These different characteristics will ensure evaluation of many different aspects of IQ. Due to the low number of images (10) investigation of ICs will not be carried out, since the number of images are too low for classification.

Printing workflow  Firstly, the color space of all the images was changed to sRGB to define the reference images. Secondly, then the color space was changed to CMYK using the output profile that was generated using a TC3.5 CMYK test target (Section 3.1.4.4), measured with a GretagMacbeth Eye-One Pro spectrophotometer and generated with ProfileMaker Pro 5.0.8. Finally the CMYK images were printed by a HP DesignJet 10ps printer with the HP software RIP v2.1.1 using four different modes on Stora Enso Multicopy Original plain paper: the best print mode, with the resolution of 1200x1200, and the perceptual intent, the best mode and relative colorimetric intent, normal print mode, with the resolution of 600x600 and the perceptual intent, and the last with normal print mode and relative colorimetric intent. This resulted in the ten images having four different reproductions, giving a total of 40 images for the observers to judge.

Observers  Ten observers participated in the experiment, all had normal vision without visual deficits. There were 3 females and 7 males with an average age of 23 years.

Viewing conditions  The observers were presented with a reference image on an EIZO ColorEdge CG224 at a color temperature of 6500 K and luminance level of 80 cd/m2. The image set was rendered for sRGB display, and therefore a monitor capable of displaying the sRGB gamut was the most adapted reproduction device for the set of images. A hood was fitted to the monitor to prevent glare. The printed images were presented randomly in a controlled viewing room at a color temperature of 5200 K, an illuminance level of 450 ±75 lux and a color rendering index of 96. The observers viewed the reference image and the printed image simultaneously from a distance of approximately 60 cm. The experiment followed the CIE guidelines [80] as closely as possible.
**Experiment procedure**  The observers were asked to compare one image selected from the ten images at random to its four prints. Sharpness quality, color quality, lightness quality, contrast quality, artifacts quality, and the quality of the main characteristics were evaluated on a five step scale, where 1 indicated best quality and 5 the worst quality. The physical CPQA was not evaluated since no physical parameter was changed.

11.3.2.2 Experimental results

From the experiment z-scores were calculated using the color engineering toolbox [162], which indicated the perceived differences between the four reproductions. These z-scores were calculated for each CPQA and the main characteristics, both image-wise and for the complete dataset.

It has been suggested in the literature that some regions of the image are more important than others [253, 356, 457]. In order to investigate the relationship between the CPQAs and different regions of the image, we have calculated the Pearson correlation coefficients [232] between the main characteristics and the CPQAs. This analysis would reveal if the quality of the CPQAs are related to the quality of main characteristics (region-of-interest). From Table 11.17 we can see that in the different reproductions the main characteristics have varying correlation coefficients with the CPQAs. This indicates that the quality of the CPQAs are not directly linked with main characteristics, but that other characteristics are important for the impression of quality of most CPQAs. However, for some CPQAs and printing modes we see a high correlation between the main characteristics and the CPQAs, this might indicate that IQ metrics performing a weighting of regions, such as SHAME, could be more suitable than those assigning equal weight to the entire image.

**Table 11.17:** Pearson correlation between z-scores of the main characteristics and the z-scores of the CPQAs for each printing mode and for all modes. N is the number of images used in the calculation of the correlation values.

<table>
<thead>
<tr>
<th>Mode</th>
<th>N</th>
<th>Color</th>
<th>Lightness</th>
<th>Sharpness</th>
<th>Contrast</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best perceptual</td>
<td>10</td>
<td>0.85</td>
<td>0.47</td>
<td>0.55</td>
<td>0.92</td>
<td>0.28</td>
</tr>
<tr>
<td>Best Relative</td>
<td>10</td>
<td>0.72</td>
<td>0.45</td>
<td>0.48</td>
<td>0.78</td>
<td>0.55</td>
</tr>
<tr>
<td>Normal perceptual</td>
<td>10</td>
<td>-0.02</td>
<td>0.60</td>
<td>0.30</td>
<td>0.61</td>
<td>0.71</td>
</tr>
<tr>
<td>Normal relative</td>
<td>10</td>
<td>0.31</td>
<td>0.29</td>
<td>0.31</td>
<td>0.88</td>
<td>0.60</td>
</tr>
<tr>
<td>All</td>
<td>40</td>
<td>0.79</td>
<td>0.77</td>
<td>0.71</td>
<td>0.89</td>
<td>0.77</td>
</tr>
</tbody>
</table>

11.3.2.3 Evaluation of image quality metrics

In this part we evaluate the selected set of IQ metrics (Table 11.1) for each CPQA against the perceptual data from the experiment.

**Preparation of the printed images**  The printed images cannot be directly used with IQ metrics, since the metrics require a digital input. Because of this the images need to be digitized. To perform this we have adopted the framework [343] presented in Chapter 8. First the images were scanned at a resolution of 600 DPI using an HP ScanJet G4050. The scanner was characterized with the same test target as used to generate the printer profile. Since the
experiment was carried out under mixed illumination, the CIECAM02 chromatic adaptation transform [81] was used to ensure consistency in the calculations for the metrics. The CIE guidelines were followed [81], using the measured reference white point of the monitor and the media were used as input to the adaptation transform.

Evaluation method  Three different methods were adopted for the evaluation of the IQ metrics. In order to evaluate all aspects of the metrics we will investigate the performance of the IQ metrics both image by image, and the overall performance over the entire set of images. The Pearson correlation [232] is used for the image-wise evaluation, comparing the calculated quality and observed quality. The mean of the correlation for each image in the dataset and the percentage of images with a correlation above 0.6 is used as a measure of performance. Overall performance is also an important aspect, and for this evaluation we will use the rank order method [350], where the correlation between the z-scores from the observers and the z-scores of the metric is the indication of performance. With only four data points it is important to carry out visual inspections of the z-scores to validate the correlation values. This evaluation is similar to one carried out for experiment I in Section 11.2.2.

Evaluation results  Due to many IQ metrics and several CPQAs we will only show the results of the best performing metrics for each CPQA.

Sharpness  For sharpness the Structural SIMilarity (SSIM) based metrics perform well (Table 11.18). The MS-SSIM has the highest mean correlation with 0.73 and the highest number of images with a correlation above 0.6. It also performs among the best for the rank order correlation. The results show that metrics based on structural similarity are well-suited to measure perceived sharpness quality. However, other approaches as the ΔLC and the RFSIM have very good performance, indicating that these might be suitable as well.

\[
\begin{array}{|l|c|c|c|c|}
\hline
\text{Metric} & \text{Mean correlation} & \text{Percentage above 0.6} & \text{Rank order} & \text{Correlation} & \text{p-value} \\
\hline
\text{CW-SSIM} & 0.66 & 70 & 0.94 & 0.06 \\
\text{ΔLC} & 0.43 & 50 & 1.00 & 0.00 \\
\text{IW-SSIM} & 0.56 & 70 & 0.89 & 0.11 \\
\text{MS-SSIM} & 0.73 & 80 & 0.94 & 0.06 \\
\text{RFSIM} & 0.61 & 70 & 0.97 & 0.03 \\
\text{SSIM} & 0.66 & 80 & 0.96 & 0.04 \\
\hline
\end{array}
\]

Color  For the color CPQA none of the evaluated metrics perform well (Table 11.19). It should be noted that all of these metrics are based on color differences, and this might be an indication that using only the color difference from the original is not enough to predict perceived color quality. The color CPQA had a fairly high correlation for all modes between the main characteristic and perceived IQ (Table 11.17), which might indicate that metrics giving more importance to certain regions, such as SHAME and SHAME-II, could perform better than the metrics that equally weight the entire image. The experimental results in Table 11.19 show that these metrics do not outperform other metrics.
Table 11.19: Evaluation of metrics for the color CPQA. Color indicates the color part of the metric.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABF</td>
<td>0.07</td>
<td>0</td>
<td>0.23</td>
</tr>
<tr>
<td>LinLab</td>
<td>-0.09</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>-0.27</td>
<td>0</td>
<td>-0.24</td>
</tr>
<tr>
<td>S-DEE&lt;sub&gt;Color&lt;/sub&gt;</td>
<td>-0.38</td>
<td>0</td>
<td>-0.35</td>
</tr>
<tr>
<td>SHAME</td>
<td>0.01</td>
<td>10</td>
<td>0.10</td>
</tr>
<tr>
<td>SHAME&lt;sub&gt;Color&lt;/sub&gt;</td>
<td>0.05</td>
<td>20</td>
<td>0.12</td>
</tr>
<tr>
<td>SHAME-II</td>
<td>0.23</td>
<td>30</td>
<td>0.27</td>
</tr>
<tr>
<td>YCxCzLab</td>
<td>0.24</td>
<td>30</td>
<td>0.33</td>
</tr>
</tbody>
</table>

**Lightness** The SSIM based metrics perform very well for the lightness attribute (Table 11.20), the CW-SSIM has a mean correlation 0.86 and all images have a correlation above 0.6. However, other metrics also perform well, such as the RFSIM, ΔLC, S-DEE with only the lightness part (S-DEE<sub>Lightness</sub>) and ABF with only the lightness part (ABF<sub>Lightness</sub>). The results indicate that any of these are appropriate to measure lightness quality. These metrics take different approaches to measure lightness quality, indicating that different strategies are suitable.

Table 11.20: Evaluation of metrics for the lightness CPQA. Lightness indicates the lightness part of the metric.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABF&lt;sub&gt;Lightness&lt;/sub&gt;</td>
<td>0.69</td>
<td>80</td>
<td>0.87</td>
</tr>
<tr>
<td>CW-SSIM</td>
<td>0.86</td>
<td>100</td>
<td>0.93</td>
</tr>
<tr>
<td>ΔLC</td>
<td>0.69</td>
<td>80</td>
<td>0.99</td>
</tr>
<tr>
<td>IW-SSIM</td>
<td>0.85</td>
<td>80</td>
<td>0.95</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.82</td>
<td>90</td>
<td>0.93</td>
</tr>
<tr>
<td>RFSIM</td>
<td>0.86</td>
<td>90</td>
<td>1.00</td>
</tr>
<tr>
<td>S-DEE&lt;sub&gt;Lightness&lt;/sub&gt;</td>
<td>0.80</td>
<td>90</td>
<td>0.89</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.63</td>
<td>70</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Contrast** Many metrics perform well for the contrast CPQA (Table 11.21). The SSIM based metrics all have a correlation above 0.6 in more than 70% of the images, they also have a high mean correlation and excellent rank order correlation. The RFSIM has a similar performance to the SSIM based metrics. All of these metrics would be appropriate for measuring contrast. However, one should notice that all of the well performing metrics for contrast are based on lightness, and none of them take color information into account. This might make them inappropriate to measure contrast in images where color strongly contributes to the impression of contrast.
Table 11.21: Evaluation of metrics for the contrast CPQA.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order Correlation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW-SSIM</td>
<td>0.72</td>
<td>90</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>IW-SSIM</td>
<td>0.59</td>
<td>70</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.72</td>
<td>80</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>RFSIM</td>
<td>0.67</td>
<td>80</td>
<td>0.96</td>
<td>0.04</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.65</td>
<td>70</td>
<td>0.99</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 11.22: Evaluation of metrics for the artifacts CPQA.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean correlation</th>
<th>Percentage above 0.6</th>
<th>Rank order Correlation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW-SSIM</td>
<td>0.83</td>
<td>90</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>ΔLC</td>
<td>0.72</td>
<td>70</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>IW-SSIM</td>
<td>0.83</td>
<td>90</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.77</td>
<td>90</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>RFSIM</td>
<td>0.82</td>
<td>90</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.60</td>
<td>70</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Artifacts  The performance for the artifacts CPQA (Table 11.22) follow the results of many of the other CPQAs. The SSIM based metrics perform well together with ΔLC and RFSIM. There are only minor differences between these, and any of them seem to be suitable to measure artifacts. However, artifacts can vary significantly and to measure specific artifacts specially designed metrics might be required.

11.4 Summary

A set of IQ metrics for each CPQA has been evaluated in this chapter. The results indicate that metrics based on structural similarity are suited to evaluate the sharpness CPQA. For the color CPQA none of the evaluated metrics perform well, and more work is needed to find a suitable metric for this attribute. Structural similarity based metrics also perform well for the lightness CPQA, but other metrics based on color differences also perform well. For the contrast CPQA the results are similar to sharpness, with structural similarity based metrics being the best. Good performance can also be found for the artifact CPQA, but this attribute contains many different sub-attributes which might require metrics adapted to the properties of the sub-artifact.
12 Improving Performance of Metrics by Pooling

An IQ metric has several different stages as mentioned in Chapter 4 (Figure 12.1), where each stage plays a critical role in the assessment of quality. In this chapter we extend the evaluation carried out in Chapter 11, and we investigate how pooling methods can improve the assessment of print quality through quality attributes.

Pooling is a term commonly used in resource management and refers to the grouping of resources to maximize advantages and minimize risk. It has also been used in vision research where it describes the combination of responses of feature detectors in nearby locations into some statistic that summarizes the distribution of the features over some region of interest [43]. In medical research pooling often refers to the statistical process of generating a summary of data [333]. For IQ metrics, pooling is the technique used to reduce the number of quality values to a more manageable number, most often one value. The main motivation for doing pooling is that IQ metrics usually calculate quality locally (either pixel-wise or in a local neighborhood) producing an IQ value for each pixel, resulting in an extensive amount of data, which is difficult to analyze and process. Therefore, pooling is needed to reduce the number of values, usually to a single IQ value, since it is easier to manage one value than many. Previous research has shown that pooling is very important for achieving an IQ metric correlated with the percept [458, 463]. To demonstrate the importance and impact of pooling we have selected one of the images (Figure 12.2) from the Dugay et al. [107, 108] dataset (Section 7.6). This image has been gamut mapped using five different gamut mapping algorithms. In the reproduction shown in Figure 12.2(b) we can notice a loss of details in certain regions, while other parts of the image have contrast changes. These are local changes, that observers notice and use in the judgment of quality, and since locality is important, the pooling strategies should account for this. The ranking based on three different pooling strategies can be seen in Table 12.1, and they all produce different ranking results, one close to the

Figure 12.1: Stages of IQ metrics. In this chapter we focus on pooling.
observers ranking (strategy 1) and one far from it (strategy 3). This shows that the pooling strategy chosen may have a great impact on the performance of the metric.

Table 12.1: Difference between various pooling strategies. The ranking of the five gamut mapped images using three pooling strategies are different, and none of them completely match the observers ranking.

<table>
<thead>
<tr>
<th>Image</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooling strategy 1</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Pooling strategy 2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Pooling strategy 3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Observers rank</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Pooling can usually be applied in three different stages, which can be referred to as spatial pooling, channel pooling and quality attribute pooling [258, 463]. Many IQ metrics produce a quality value for each pixel in the image; all of these values are usually referred to as a quality map. To combine the values in the quality map in the image domain is called spatial pooling. Most metrics also calculates quality in different channels, for example in a color space such as CIELAB. Pooling values from different channels is referred to as channel pooling, which usually results in a quality map. The last type of pooling, quality attribute pooling, is used to combine several quality maps generated from different quality attributes (i.e., color, lightness, contrast, etc.). Spatial pooling is always needed, while channel pooling and quality attribute pooling are needed only when the image is decomposed into different channels or different quality maps are calculated for each quality attribute. In this chapter, we focus on spatial pooling methods.

The simplest spatial pooling method is to compute the average, which assumes that each quality value is of equally significance. However, since the image features are highly spatially non-stationary, it is intuitively obvious that each region in an image may not bear the same importance as others [298]. This is also shown in Figure 12.1 where local changes occur in the reproduction. As a result, giving appropriate weights to each quality value before computing the average is reasonable. Thus, the main issue is to determine the weights.
12.1 State of the art

A lot of work has been done concerning spatial pooling approaches. There are two rationales behind these approaches. The first is that some regions in the image are more perceptually significant (region of interest) so that distortions in these areas might drastically decrease IQ. These methods differ in their modeling of region significance. The second is that human subjects tend to pay more attention to the low quality regions. Consequently, the low quality regions are more important and should be weighted more heavily. The first method refers to local content of the image, while the latter refers to the local value of a quality map that is derived as a feature image. We call these two kinds of spatial pooling methods; content-based and quality-based.

Existing pooling methods can be divided into these two groups with different subgroups:

- Quality based pooling (Section 12.1.2)
  1. Minkowski pooling [102, 463] (Section 12.1.2.1)
  2. Local quality/distortion-weighted pooling [297, 298, 463] (Section 12.1.2.2)
     (a) Monotonic function pooling [463]
     (b) Percentile pooling [297, 298]

- Content based pooling (Section 12.1.3)
  1. Information content-weighted pooling [463] (Section 12.1.3.1)
  2. Gaze based pooling (Section 12.1.3.2)
     (a) Gaze-attentive fixation finding engine [297, 298]
     (b) Saliency map computed by fixation number and duration [321]
     (c) Gaussian weighted sum model [270]
  3. Computational saliency model based pooling (Section 12.1.3.3)
     (a) Itti’s saliency model [137, 214]
     (b) Spectral residual model [180, 277]

12.1.1 General formulation

A general form of a spatial pooling approach is given by

\[ M = \frac{\sum_{i=1}^{n} w_i m_i}{\sum_{i=1}^{n} w_i}, \]

where \( w_i \) is the weight given to the \( i \)th location and \( m_i \) is the quality measure of the \( i \)th location. \( M \) is the pooled quality value. Most spatial pooling methods could be formulated in this way. In a simple average pooling method, \( w_i \) is the same over the image space. Actually, \( w_i \) might be related to quality \( m_i \) value at each position or some other factors which are currently unknown to us.

Obviously, spatial pooling methods mainly consist of two steps. The first step is to find out what factors might be related to \( w_i \). The second step is to determine the exact function that maps the assumed factors to \( w_i \). The quality based methods assume that the weights \( w_i \) might
be related to the quality value \( m_i \), and the content based methods assume that the weights \( w_i \) might be related to image content in the local region around the \( i \)th pixel.

We will first review quality based pooling methods, before we go through the content based methods.

### 12.1.2 Quality based pooling

Wang and Moorthy [297] and Wang and Shang [463] assume that \( w_i \) is determined by the IQ value at the \( i \)th location in the quality map, i.e.,

\[
w_i = f(m_i). \tag{12.2}
\]

The methods belonging to quality based pooling follow the principle that low quality values should be weighted more heavily compared to higher quality values.

Nielsen et al. [317] classified weighting functions for gray-level cooccurrence matrices into two classes depending on how the information is extracted. The first class is based on the position in the matrix, and the other on the value of the matrix elements. We can also adapt this classification to pooling, where the weighting in Equation 12.1 can fall into these two classes:

- **Weighting based on position.**
  - This is a weighting based on the quality feature value, which covers Minkowski and Monotonic Function pooling.

- **Weighting based on value.**
  - This is a weighting based on the frequency of occurrence of a given quality feature value, which covers percentile pooling in a general way.

With this grouping, different than that of dividing Minkowski from local quality/distortion-weighted pooling, show a difference between the quality based pooling methods when they are formulated in terms of a histogram. This grouping shows the difference between Percentile pooling on one hand and Minkowski and Monotonic Function pooling, which has only a difference in the factor \( m \), on the other.

#### 12.1.2.1 Minkowski pooling

Wang and Shang [463] use Minkowski pooling [102] as a spatial pooling method, which was previously used by de Ridder [102] as a combination rule for digital-image-coding impairments. The Minkowski average is defined as

\[
M = \frac{1}{N} \sum_{i=1}^{N} m_i^p, \tag{12.3}
\]

where \( m_i \) is the quality measure at the \( i \)th position of the quality map and \( N \) is the total number of values in the map. We can easily see that Equation 12.3 can be reformulated according to
Equation 12.1, which is defined as

\[ M = \frac{1}{N} \sum_{i=1}^{N} m_i^{p-1} m_i, \]  

(12.4)

which could be considered as the quality value at the \( i \)th position is weighted by \( w_i = m_i^{p-1} \) without normalization. If \( p = 1 \), it reduces to the arithmetic mean. If \( p > 1 \), \( w_i \) is a monotonically increasing function of \( m_i \). If \( m_i \) is a distortion measure, regions of low quality will be weighted more heavily. More emphasis will be put at highly distorted regions as \( p \) increases.

Since only the quality value is involved in Minkowski pooling, and not the position, the summation may either be performed as a summation over all positions of the (2D) image as implied by Equations 12.3 and 12.4:

\[ M = \frac{1}{N} \sum_{x} \sum_{y} (m(x,y))^p, \]  

(12.5)

or as a summation of the the (1D) histogram \( h(m) \) of IQ values \( m \):

\[ M = \frac{1}{N} \sum_{i} h(m_i)m_i^p. \]  

(12.6)

### 12.1.2.2 Local quality/distortion-weighted pooling

#### Monotonic function pooling

Wang and Shang [463] define \( w_i \) in Equation 12.1 as

\[ w_i = m_i^p, \]  

(12.7)

where \( p \) ranges from 1/8 to 8 when \( f(m_i) \) is a distortion measure, which makes sure that \( w_i \) is a monotonically increasing function of \( m_i \). When \( m_i \) is a quality measure, \( p \) ranges from -8 to -1/8 so that \( f(m_i) \) is a monotonically decreasing function of \( m_i \).

#### Percentile pooling

Moorthy and Bovik [297, 298] consider the statistical principle of heavily weighting the lower-percentile scores instead of an arbitrary monotonic function of quality. They rank the quality values in a quality map first and then merely weight the lowest \( p\% \) quality values more heavily. In this method, \( f(m_i) \) is a step function with an adaptive threshold \( T \) computed in a statistical manner. Assuming that \( m_i \) is the distortion, \( f(m_i) \) could be formulated as

\[ w_i = f(m_i) = \begin{cases} r & : m_i > T \\ 1 & : m_i < T \end{cases}, \]  

(12.8)

where \( T \) is determined by the quality map and \( p \).
12.1.3 Content based pooling

Many researchers [137, 180, 214, 256, 270, 277, 297, 298, 321, 463] hold the opinion that the weight $w_i$ is determined by the image content of the local region around the $i$th location, i.e.,

$$w_i = c_i,$$

(12.9)

where $c_i$ is a measure of perceptual significance of image content in the local region around the $i$th location. They assume that an error that appears on a perceptually significant region is much more annoying than a distortion appearing in an inconspicuous area. The main issue is how to model perceptual significance based on the image content.

12.1.3.1 Information content-weighted pooling

Wang and Shang [463] compute information content from local image regions to measure perceptual significance. Regions with higher information content are considered as more perceptually significant, which implies that people will pay more attention to the regions with higher information content. Hence, regions of higher information content should have larger weights.

The authors quantify information content ($I$) as the number of bits that can be received from a Gaussian image information source model which passes through a visual channel with additive Gaussian noise. The information received is defined as

$$I = \frac{1}{2} \log_2 \left( 1 + \frac{S}{C} \right),$$

(12.10)

where $S$ is the power of the source model and $C$ is the power of noise channel. We can see that more information is obtained if the ratio between source power and noise power is larger. The final information content is given by the information in a local region in the reference image and the distorted image,

$$I = \frac{1}{2} \log_2 \left[ \left( 1 + \frac{S_r}{C} \right) \left( 1 + \frac{S_d}{C} \right) \right],$$

(12.11)

where $S_r$ is the source power of a local region in the reference image, which could be calculated as $\sigma_x^2$. $S_d$ is the source power of a local region in the distorted image, which could be calculated as $\sigma_y^2$. $\sigma_x^2$ and $\sigma_y^2$ are the signal variances of a local region in the reference and distorted image, respectively.

12.1.3.2 Gaze based pooling

These pooling methods use saliency maps generated from data gathered in eye tracking experiments. Different ways of doing the experiment and different methods used to calculate the saliency maps result in different results. However, it has been reported that visual fixation data improves IQ assessment [18, 356].
**Gaze-attentive fixation finding engine**  When shown an image, a human tends to fixate at certain points on the image. It is likely that these points are more perceptually significant when human beings assess IQ. Taking this into account, Moorthy and Bovik use a Gaze-Attentive Fixation Finding Engine (GAFFE) developed by Rajashekar [369] to predict ten fixations points in each image and then weight local regions around these points by a factor \( k \), which is set to 265 by Moorthy and Bovik [297, 298]. Other pixels in the image are weighted by 1.

**Saliency map computed by fixation number and duration**  Ninassi et al. [321] compute saliency maps in two different ways for each observer and for each picture. The first way is the fixation number for each spatial location. The second way is the fixation duration for each spatial location. To determine the most visually important regions, all the saliency maps are merged yielding an average saliency map.

**Gaussian weighted sum model**  Liu et al. [270] assume that each fixation location gives rise to a patch whose activity is Gaussian distributed. The width \( \sigma \) of the Gaussian patch approximates the size of the fovea. The \((k,l)\)th value in the saliency map is defined as

\[
S(k,l) = \sum_{j=1}^{T} \exp\left[-\frac{(x_j-k)^2 + (y_j-l)^2}{\sigma^2}\right],
\]

where \( S(k,l) \) is the saliency value in the position \((k,l)\). \( x_j, y_j \) is the position of \( j \)th fixation and \( T \) is the total number of fixations.

**12.1.3.3 Computational saliency model**

These methods do not rely on visual fixation data, but model saliency directly, which might be inspired by the behavior of the HVS. Compared to eye tracking based saliency maps, these methods are more demanding because they could be directly used in IQ assessment methods. However, none of these saliency models capture all the aspects that might contribute to saliency detection in the HVS.

**Itti’s saliency model**  Feng et al. [137] proposed a saliency based method for assessing video sequences affected by packet losses. Itti’s bottom-up Saliency based Visual Attention Model (SVAM) [214] was used to calculate the saliency maps. The model is inspired by the behavior and the neuronal architecture of the early primate visual system, which demonstrates good performance in many vision applications. Feng et al. hypothesize that an error appearing on a saliency region is more annoying and show the improvement in video quality assessment using SVAM.

**Spectral residual model**  Ma et al. [277] employ the spectral residual model to detect saliency, which is used as weights in the pooling strategy. To compute the spectral residual \( R \), an image should firstly be transformed to the Fourier domain. We denote the amplitude spectrum by \( A(f) \) and phase spectrum by \( P(f) \). The spectral residual is defined as

\[
R(f) = \log(A(f)) - h \ast \log(A(f)),
\]

where \( h \) is a weighting function.
where \( h \) is an averaging filter. The saliency map \((SM)\) is generated by Inverse Fourier Transform (IFT):

\[
SM = |IFT(exp(R(f)) + jP(f))|^2
\]  

(12.14)

### 12.2 Experimental setup

In order to evaluate the existing pooling methods we will compare the results from perceptual experiments against the results of a set of IQ metrics using different pooling strategies. Two different test sets have been selected for this.

#### 12.2.1 Test sets

##### 12.2.1.1 Test set 1

The first test is the same as in Section 11.2.2. The database contains 24 of the 25 reference images, leaving out the bride and groom image as in Section 11.2.2.2. These images (Figure 10.7) have been chosen according to several image characteristics, such as lightness, contrast, saturation, etc.

The images were printed on an Océ Colorwave 600 CMYK wide format printer on Océ Red Label (LFM054) plain uncoated paper using three different rendering intents: perceptual, relative colorimetric, and relative colorimetric with black point compensation. The CPQAs for each printed image were judged by 15 observers. The results are then quantified as z-scores for these printed images. Finally, each reference image has three reproductions whose subjective quality is represented as z-scores [116] of the five CPQAs. We will evaluate spatial pooling methods for each CPQA, since pooling might be attribute dependent.

In order to apply objective image quality metrics to the printed images, these images are scanned into digital images and stored without compression using the framework proposed in Chapter 8. For more details on this dataset we refer the reader to Section 11.2.2 or 10.4, or Pedersen et al. [347].

##### 12.2.1.2 Test set 2

The experiment is also conducted on the ten image dataset (Figure 11.6), being the same as in Section 11.3.2. The images were printed by a HP DesignJet 10ps printer using four different modes: the best print mode and the perceptual intent, the best mode and relative colorimetric intent, normal print mode and the perceptual intent, and the last with normal print mode and relative colorimetric intent. Ten observers participated in the evaluation of the different CPQAs, from which the results were calculated into z-scores. For details about the dataset we refer the reader to Section 11.3.2 or Pedersen et al. [359].

#### 12.2.2 Image quality metrics

Six IQ metrics have been chosen for the evaluation based on the selection made in Sections 11.2.1 and 11.3.1, namely, SSIM [458], S-CIELAB [499], S-DEE [412], WLF [417], ABF [459] and ΔLC [25]. All of these metrics produce a quality map, making them suitable
for evaluation of spatial pooling strategies. We will go through each CPQA to show the performance of the pooling methods. Since all metrics are not designed to evaluate all CPQAs, we only evaluate the suitable metrics for each CPQA, where the selection is based on the guidelines found in Section 11.1 and Pedersen et al. [349] (Table 12.2).

Table 12.2: Selected IQ metrics for the evaluation of the CPQAs and pooling strategies.

<table>
<thead>
<tr>
<th>Metric</th>
<th>CPQA</th>
<th>Sharpness</th>
<th>Color</th>
<th>Lightness</th>
<th>Contrast</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSIM [458]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ABF [459]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔLC [25]</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-CIELAB [499]</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-DEE [412]</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLF [417]</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

12.2.3 Pooling methods

For each metric, we test Spatial Average Pooling, Minkowski Pooling [463], Monotonic Function Weighted pooling [463], Information Content Based Pooling [463], IG Saliency Model Based Pooling [1], and Nonparametric Bottom-up (NB) Saliency Model Based Pooling [389, 390]. In the two saliency based methods the saliency map is used as a weighting map using the general formulation in Equation 12.1.

12.2.4 Evaluation methods

We use three evaluation methods to compare the performance of different spatial pooling methods for each IQ metric.

- Pearson Linear Correlation Coefficient (PLCC) as explained in Section 6.2.1.1. The mean PLCC of the whole database is calculated as the average of PLCCs of each image.

- The Percentage of Images (POI) with PLCC higher than 60% is also examined.

- Rank Correlation (RC), as presented in Section 6.3 and Pedersen and Hardeberg [350], which is the PLCC correlation between objective rank order z-score and subjective z-score to measure the order consistency between objective and subjective assessment. To compute the objective rank order z-score, we firstly rank the original objective scores of each scene’s three reproductions and then compute objective rank order z-score using the rank order function in the Colour Imaging Toolbox [162].

12.3 Experimental results

We compare the results from the IQ metrics with the different pooling methods to the perceptual results.
12.3.1 Sharpness

12.3.1.1 Test set 1

For the ΔLC metric the NB saliency model has the highest PLCC value, improving the PLCC by 0.1 compared to average pooling as seen in Table 12.3. But POI values between these remains the same because the NB pooling method is not able to improve the low correlation values to high ones, although it improves some correlation values slightly. On the other hand, Minkowski pooling with \( p = \frac{1}{2} \) and \( p = 2 \), and the Monotonic Function pooling with \( p = \frac{1}{4} \) and \( p = \frac{1}{2} \) change one bad correlation value to a very good one, but they decrease the correlation values of other images. This is why POI improves while PLCC slightly increases or even decreases. The above pooling methods show similar performance in terms of RC compared to average pooling. For ΔLC we also see that the correlation decreases when the \( p \) for both Monotonic Function and Minkowski is high.

For the SSIM metric, Monotonic pooling with \( p = 1 \) improves both PLCC and POI while it maintains a high RC. The RC for Minkowski pooling is very dependent on the \( p \) value, decreasing the performance with increasing \( p \) values.

For both metrics the selected parameter in the pooling methods is important, and a wrong selection can decrease the performance of the metric. However, SSIM seems to be more dependent on the pooling method and parameters than ΔLC.

Table 12.3: Performance of the spatial pooling methods for the sharpness CPQA on Test Set 1. The ΔLC and SSIM metrics are selected for the evaluation. The best results for each performance measure and metric are shown in bold.
12.3.1.2 Test set 2

For $\Delta \text{LC}$, the spatial average pooling gives an excellent rank order correlation of 1.0 (Table 12.4). NB Saliency Model Based Pooling improves the performance in terms of mean correlation and percentage of images above 0.6, with a rank order correlation comparable to average pooling. We can also notice that using Minkowski or Monotonic Function pooling with a correct choice of $p$ will increase the POI up to 70%. For SSIM, Minkowski Pooling with $p = 8$ gives the best results concerning mean correlation and percentage of images above 0.6. However, the rank order correlation drops a from 0.96 for average pooling to 0.7. The spatial average pooling seems the best for SSIM, since it has the highest RC while maintaining good results for PLCC and ROI. The result is to some extent consistent with that of the first dataset.

Table 12.4: Performance of the spatial pooling methods for the sharpness CPQA on Test Set 2. $\Delta \text{LC}$ and SSIM are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>p</th>
<th>$\Delta \text{LC}$</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLCC</td>
<td>POI</td>
</tr>
<tr>
<td>Average</td>
<td>0.43</td>
<td>50%</td>
<td>1.00</td>
</tr>
<tr>
<td>Minkowski</td>
<td>0.42</td>
<td>40%</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>40%</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.43</td>
<td>50%</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.44</td>
<td>60%</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>0.46</td>
<td>70%</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>0.31</td>
<td>50%</td>
<td>0.77</td>
</tr>
<tr>
<td>Monotonic Function</td>
<td>0.42</td>
<td>50%</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>60%</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>60%</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>0.43</td>
<td>70%</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>0.47</td>
<td>70%</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>20%</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>-0.26</td>
<td>10%</td>
<td>-0.73</td>
</tr>
<tr>
<td>IC</td>
<td>0.39</td>
<td>50%</td>
<td>1.00</td>
</tr>
<tr>
<td>IG</td>
<td>0.42</td>
<td>60%</td>
<td>0.89</td>
</tr>
<tr>
<td>NB</td>
<td>0.51</td>
<td>70%</td>
<td>0.99</td>
</tr>
</tbody>
</table>

12.3.2 Color

12.3.2.1 Test set 1

For the color CPQA, the three selected metrics all give very poor results (Table 12.5). This might be a consequence of the fact that quality does not monotonically decrease with color difference between the reference image and printed images. Human observers might prefer images that are more saturated, even though they might have a larger color difference than less saturated images. Table 12.5 shows that the pooling methods contribute little to the results...
when the metrics are very poor. The Monotonic Function pooling strategy improves the results slightly, possibly because local regions with large color difference that are really annoying are given larger weights.

Table 12.5: Performance of the spatial pooling methods for the color CPQA on Test Set 1. ABF, S-CIELAB, and S-DEE are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>$p$</th>
<th>ABF</th>
<th>S-CIELAB</th>
<th>S-DEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
<td>PLCC</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.39</td>
<td>8%</td>
<td>-0.99</td>
<td>-0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minkowski</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\frac{1}{3}$</td>
<td>-0.40</td>
<td>4%</td>
<td>-0.99</td>
<td>-0.25</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>-0.40</td>
<td>4%</td>
<td>-0.99</td>
<td>-0.25</td>
</tr>
<tr>
<td>2</td>
<td>-0.40</td>
<td>4%</td>
<td>-0.98</td>
<td>-0.26</td>
</tr>
<tr>
<td>4</td>
<td>-0.36</td>
<td>13%</td>
<td>-0.99</td>
<td>-0.32</td>
</tr>
<tr>
<td>8</td>
<td>-0.18</td>
<td>8%</td>
<td>-1.00</td>
<td>-0.34</td>
</tr>
<tr>
<td></td>
<td>-0.07</td>
<td>21%</td>
<td>-0.99</td>
<td>-0.20</td>
</tr>
<tr>
<td>Monotonic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\frac{1}{3}$</td>
<td>-0.39</td>
<td>13%</td>
<td>-0.99</td>
<td>-0.30</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>-0.38</td>
<td>13%</td>
<td>-0.99</td>
<td>-0.31</td>
</tr>
<tr>
<td>1</td>
<td>-0.36</td>
<td>13%</td>
<td>-0.98</td>
<td>-0.33</td>
</tr>
<tr>
<td>2</td>
<td>-0.29</td>
<td>13%</td>
<td>-1.00</td>
<td>-0.13</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>21%</td>
<td>-0.94</td>
<td>0.11</td>
</tr>
<tr>
<td>8</td>
<td>0.27</td>
<td>29%</td>
<td>-0.97</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.27</td>
<td>29%</td>
<td>-0.97</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td>-0.38</td>
<td>4%</td>
<td>-0.99</td>
<td>-0.27</td>
</tr>
<tr>
<td>IC</td>
<td>-0.22</td>
<td>13%</td>
<td>-0.94</td>
<td>-0.23</td>
</tr>
<tr>
<td>NB</td>
<td>-0.27</td>
<td>13%</td>
<td>-0.99</td>
<td>-0.21</td>
</tr>
</tbody>
</table>

12.3.2.2 Test set 2

None of evaluated metrics shows satisfactory results for the color CPQA for most of the pooling methods (Table 12.6), which is similar to the results on the first dataset. However, the Monotonic Function Based Pooling with $p = 2$ significantly improves the performance on test set 2 for the ABF metric, giving a PLCC of 0.66 (Figure 12.3), 70% of the images above 0.6, and a RC of 0.98. The improved performance could be caused by the fact that large color differences strongly contributes to the impression of quality, and with a relative large $p$ value these are given higher importance. We also see a similar increase for S-CIELAB and S-DEE but the maximum performance is reached at $p = 4$ (Figure 12.3). The results for the different metrics also indicate that the selection of the pooling method and parameters is very important for the performance of the metric.
Table 12.6: Performance of the spatial pooling methods for the color CPQA on Test Set 2. ABF, S-CIELAB, and S-DEE are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>p</th>
<th>ABF</th>
<th>S-CIELAB</th>
<th>S-DEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.07</td>
<td>0%</td>
<td>0.23</td>
</tr>
<tr>
<td>Minkowski</td>
<td></td>
<td>-0.14</td>
<td>0%</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.12</td>
<td>0%</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.06</td>
<td>0%</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.30</td>
<td>20%</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.57</td>
<td>60%</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.32</td>
<td>20%</td>
<td>0.38</td>
</tr>
<tr>
<td>Monotonic Function</td>
<td></td>
<td>0.15</td>
<td>10%</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.22</td>
<td>10%</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>40%</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.52</td>
<td>50%</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.66</td>
<td>70%</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.29</td>
<td>20%</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-0.16</td>
<td>0%</td>
<td>-0.30</td>
</tr>
<tr>
<td>IC</td>
<td></td>
<td>0.07</td>
<td>0%</td>
<td>0.23</td>
</tr>
<tr>
<td>IG</td>
<td></td>
<td>-0.01</td>
<td>0%</td>
<td>0.27</td>
</tr>
<tr>
<td>NB</td>
<td></td>
<td>0.06</td>
<td>10%</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Figure 12.3: PLCC performance of Monotonic Function pooling for the color attribute.
12.3.3 Lightness

12.3.3.1 Test set 1

SSIM and ΔLC give slightly better results than ABF, S-CIELAB, and S-DEE in terms of PLCC and POI (Table 12.7). The RC of SSIM and ΔLC is very high, while for the other metrics it is low. For ABF, S-CIELAB, and S-DEE, the IG pooling method shows the best performance regarding PLCC and ROI, which is comparable to the best performance of SSIM and ΔLC. However, none of the spatial pooling methods is able to raise the low RC coefficients of ABF, S-CIELAB, and S-DEE to satisfactory values. For ΔLC, none of the pooling methods stands out in any of the performance measures, but a slight increase can be found for specific methods. For SSIM, the Monotonic Function pooling with $p = \frac{1}{4}$ improves the performance to the best one among all the combinations of evaluated metrics and pooling methods (see Table 12.7). However, the parameter for Minkowski and Monotonic function pooling is important for the performance.

12.3.3.2 Test set 2

For the lightness CPQA for the second test set, ΔLC and SSIM performs quite well for average pooling (Table 12.8), while ABF, S-CIELAB, and S-DEE show poor results with spatial average pooling (Table 12.8). For ABF, S-CIELAB and S-DEE, Monotonic Function pooling with $p = 2$ or $p = 4$ gives better results, but the selection of the parameter is important. This increase in performance was not found in the first data set, which might indicate that pooling is image dependent. From Table 12.8 and Figure 12.5 we can see that SSIM and ΔLC has decreasing performance in terms of PLCC for increasing $p$ values for Monotonic Function pooling, while for the color difference metrics (S-CIELAB, ABF, and S-DEE) the PLCC is increasing up to a certain $p$ value. For Minkowski pooling ABF peaks at $p = 4$ (Table 12.8), while ΔLC and SSIM are high and more robust to changes in $p$ compared with Monotonic Function pooling (Figure 12.4). For ΔLC and SSIM, spatial pooling methods other than average gives only slight improvements in the performance. For SSIM, Monotonic Function pooling with $p = 2$ shows the best performance.

![Figure 12.4: PLCC performance of Minkowski pooling for the lightness attribute.](image-url)
Table 12.7: Performance of the spatial pooling methods for the lightness CPQA on Test Set 1. ABF, S-CIELAB, S-DEE, ΔLC, and SSIM are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>p</th>
<th>ABF</th>
<th>S-CIELAB</th>
<th>S-DEE</th>
<th>ΔLC</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
<td>PLCC</td>
<td>POI</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.12</td>
<td>38%</td>
<td>0.11</td>
<td>0.29</td>
<td>50%</td>
</tr>
<tr>
<td>Minkowski</td>
<td></td>
<td>0.14</td>
<td>38%</td>
<td>0.05</td>
<td>0.31</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.14</td>
<td>38%</td>
<td>0.12</td>
<td>0.31</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.09</td>
<td>38%</td>
<td>0.11</td>
<td>0.25</td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.06</td>
<td>33%</td>
<td>0.01</td>
<td>0.25</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.20</td>
<td>46%</td>
<td>-0.18</td>
<td>0.17</td>
<td>42%</td>
</tr>
<tr>
<td>Monotonic</td>
<td></td>
<td>0.12</td>
<td>38%</td>
<td>0.11</td>
<td>0.28</td>
<td>50%</td>
</tr>
<tr>
<td>Function</td>
<td></td>
<td>0.11</td>
<td>38%</td>
<td>0.11</td>
<td>0.26</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.08</td>
<td>38%</td>
<td>0.12</td>
<td>0.23</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.05</td>
<td>38%</td>
<td>-0.12</td>
<td>0.14</td>
<td>42%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.12</td>
<td>33%</td>
<td>-0.39</td>
<td>-0.08</td>
<td>29%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.05</td>
<td>38%</td>
<td>0.18</td>
<td>0.09</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.05</td>
<td>38%</td>
<td>0.18</td>
<td>0.17</td>
<td>46%</td>
</tr>
<tr>
<td>IC</td>
<td></td>
<td>0.14</td>
<td>38%</td>
<td>0.11</td>
<td>0.30</td>
<td>50%</td>
</tr>
<tr>
<td>IG</td>
<td></td>
<td>0.35</td>
<td>50%</td>
<td>0.27</td>
<td>0.39</td>
<td>54%</td>
</tr>
<tr>
<td>NB</td>
<td></td>
<td>0.16</td>
<td>42%</td>
<td>0.04</td>
<td>0.35</td>
<td>50%</td>
</tr>
</tbody>
</table>
Table 12.8: Performance of the spatial pooling methods for the lightness CPQA on Test Set 2. ABF, S-CIELAB, S-DEE, ΔLC, and SSIM are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>ABF</th>
<th>S-CIELAB</th>
<th>S-DEE</th>
<th>ΔLC</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average</strong></td>
<td>-0.08</td>
<td>-0.03</td>
<td>-0.04</td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Minkowski</strong></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Monotonic</strong></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Affinity</strong></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>-0.08</td>
<td>-0.03</td>
<td>-0.04</td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Minkowski</strong></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Monotonic</strong></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Affinity</strong></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

For the evaluation, the best results for each performance measure and metric are shown in bold.
12.3.4 Contrast

12.3.4.1 Test set 1

For SSIM, Minkowski pooling with \( p = \frac{1}{8} \) and Monotonic Function pooling with \( p = 1 \) perform better than the average in terms of all three performance measures (Table 12.9). This is possibly because regions of larger contrast difference are more perceivable by human observers. For the \( \Delta LC \) metric, Minkowski pooling with \( p = \frac{1}{8} \) and \( p = \frac{1}{4} \) slightly improve PLCC and RC at the cost of a lower POI. This means that two or three images get lower correlation, but the overall correlation increases. The result for WLF is similar to that for \( \Delta LC \), but WLF seems to be less dependent on the parameters and pooling method for PLCC and POI than \( \Delta LC \) and SSIM. However, the differences for the three metrics between the best pooling method and other methods are small, and therefore the results for the first data set indicate that there is no pooling method that is overall better than others.

12.3.4.2 Test set 2

For the contrast CPQA for the second test set, SSIM and \( \Delta LC \) have a good performance for average pooling, while WLF shows poor results (Table 12.10). Some improvement for SSIM and \( \Delta LC \) can be found by changing the pooling method and the parameters, where Minkowski pooling provides more stable results for different \( p \) values than Monotonic Function. For the WLF a large increase in performance can be found compared to the average when using Minkowski pooling, and they are more or less stable over different \( p \) values as seen in first data set (Table 12.9). However, Monotonic Function pooling does not give acceptable results as in the first data set (Table 12.9), which might indicate that pooling is image/database dependent. In general Minkowski pooling gives the highest performance of the pooling methods regardless of the IQ metric for the second dataset.
Table 12.9: Performance of the spatial pooling methods for the contrast CPQA on Test Set 1. SSIM, $\Delta$L and WLF are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>$p$</th>
<th>SSIM</th>
<th>$\Delta$L</th>
<th>WLF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>0.32</td>
<td>50%</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Minkowski</strong></td>
<td>1</td>
<td>0.40</td>
<td><strong>58%</strong></td>
<td><strong>1.00</strong></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.36</td>
<td>50%</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.33</td>
<td>46%</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.16</td>
<td>42%</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.07</td>
<td>33%</td>
<td>-0.26</td>
</tr>
<tr>
<td><strong>Monotonic Function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.31</td>
<td>54%</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.32</td>
<td>54%</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.36</td>
<td><strong>58%</strong></td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td><strong>0.40</strong></td>
<td><strong>58%</strong></td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.17</td>
<td>50%</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.05</td>
<td>38%</td>
<td><strong>1.00</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.13</td>
<td>38%</td>
<td><strong>1.00</strong></td>
</tr>
<tr>
<td><strong>IC</strong></td>
<td></td>
<td>0.34</td>
<td>54%</td>
<td>0.97</td>
</tr>
<tr>
<td><strong>IG</strong></td>
<td></td>
<td>0.21</td>
<td>46%</td>
<td>0.59</td>
</tr>
<tr>
<td><strong>NB</strong></td>
<td></td>
<td>0.33</td>
<td>54%</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 12.10: Performance of the spatial pooling methods for the contrast CPQA on Test Set 2. SSIM, $\Delta$L and WLF are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>$p$</th>
<th>SSIM</th>
<th>$\Delta$L</th>
<th>WLF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>0.65</td>
<td>70%</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Minkowski</strong></td>
<td>1</td>
<td>0.60</td>
<td>80%</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.61</td>
<td>80%</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.63</td>
<td>70%</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td><strong>0.77</strong></td>
<td>70%</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>Monotonic Function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.64</td>
<td>70%</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.62</td>
<td>80%</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.57</td>
<td>80%</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.03</td>
<td>40%</td>
<td>-0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.12</td>
<td>20%</td>
<td>-0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.27</td>
<td>10%</td>
<td>-0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.37</td>
<td>0%</td>
<td>-1.00</td>
</tr>
<tr>
<td><strong>IC</strong></td>
<td></td>
<td>0.64</td>
<td>80%</td>
<td>0.97</td>
</tr>
<tr>
<td><strong>IG</strong></td>
<td></td>
<td>0.68</td>
<td><strong>90%</strong></td>
<td>0.96</td>
</tr>
<tr>
<td><strong>NB</strong></td>
<td></td>
<td>0.70</td>
<td>80%</td>
<td>0.92</td>
</tr>
</tbody>
</table>
12.3.5 Artifacts

12.3.5.1 Test set 1

SSIM combined with IG pooling gives the best POI for artifacts with 50% of the images above 0.6, but with a low PLCC of 0.28 (Table 12.11). ABF, S-DEE, and S-CIELAB give very poor results regardless of what pooling methods we use, with a PLCC around 0 and always a negative RC. For the WLF metric, Monotonic Function pooling with $p = \frac{1}{8}$ performs the best in terms of PLCC and RC, but the results are only marginally better than SSIM. For $\Delta L_C$, Monotonic Function pooling gives better results for PLCC and RC than average pooling, however the POI remains low, indicating that the pooling methods work well for only a few images in the dataset.

Table 12.11: Performance of the spatial pooling methods for the artifact CPQA on Test Set 1. All the six metrics are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>$p$</th>
<th>ABF</th>
<th>S-CIELAB</th>
<th>S-DEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
<td>PLCC</td>
</tr>
<tr>
<td>Average</td>
<td>-0.04</td>
<td>21%</td>
<td>-0.81</td>
<td>-0.01</td>
</tr>
<tr>
<td>Minkowski</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\frac{1}{8}$</td>
<td>-0.02</td>
<td>25%</td>
<td>-0.84</td>
<td>-0.02</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>-0.02</td>
<td>25%</td>
<td>-0.84</td>
<td>-0.02</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>-0.02</td>
<td>25%</td>
<td>-0.80</td>
<td>-0.02</td>
</tr>
<tr>
<td>2</td>
<td>-0.08</td>
<td>21%</td>
<td>-0.81</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>-0.09</td>
<td>17%</td>
<td>-0.86</td>
<td>0.01</td>
</tr>
<tr>
<td>8</td>
<td>0.05</td>
<td>38%</td>
<td>-0.94</td>
<td>-0.07</td>
</tr>
<tr>
<td>Monotonic Function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\frac{1}{8}$</td>
<td>-0.04</td>
<td>21%</td>
<td>-0.81</td>
<td>-0.01</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>-0.06</td>
<td>21%</td>
<td>-0.81</td>
<td>-0.00</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>-0.08</td>
<td>21%</td>
<td>-0.80</td>
<td>0.02</td>
</tr>
<tr>
<td>1</td>
<td>-0.09</td>
<td>21%</td>
<td>-0.92</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>-0.07</td>
<td>17%</td>
<td>-0.99</td>
<td>-0.09</td>
</tr>
<tr>
<td>4</td>
<td>-0.00</td>
<td>29%</td>
<td>-0.76</td>
<td>0.01</td>
</tr>
<tr>
<td>8</td>
<td>-0.00</td>
<td>29%</td>
<td>-0.76</td>
<td>0.04</td>
</tr>
<tr>
<td>IC</td>
<td>-0.04</td>
<td>25%</td>
<td>-0.81</td>
<td>-0.02</td>
</tr>
<tr>
<td>IG</td>
<td>-0.11</td>
<td>29%</td>
<td>-0.70</td>
<td>-0.05</td>
</tr>
<tr>
<td>NB</td>
<td>-0.12</td>
<td>25%</td>
<td>-0.84</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

12.3.5.2 Test set 2

For the artifact CPQA for the second data set, SSIM and $\Delta L_C$ perform quite well with average pooling. However, the other four metrics show poor performance with average pooling. For ABF, S-CIELAB, and S-DEE Monotonic Function pooling with a large $p$ gives much better results, which could be because severe artifacts are weighted higher. ABF combined with Monotonic Function pooling ($p = 2$) greatly improves the performance of this metric, and with this setting it shows comparable results to the best settings for SSIM. Minkowski pooling ($p = \frac{1}{8}, \frac{1}{4}$) makes the very poor results of WLF to very good ones in terms all three performance measures. However, increasing the $p$ value for WLF decreases the performance, showing the
Table 12.12: Performance of the spatial pooling methods for the artifact CPQA on Test Set 1. All the six metrics are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>p</th>
<th>SSIM PLCC</th>
<th>POI</th>
<th>RC</th>
<th>WLF PLCC</th>
<th>POI</th>
<th>RC</th>
<th>ΔLC PLCC</th>
<th>POI</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.09</td>
<td>29%</td>
<td>0.44</td>
<td>0.28</td>
<td>29%</td>
<td>0.77</td>
<td>-0.02</td>
<td>25%</td>
<td>0.18</td>
</tr>
<tr>
<td>Minkowski</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>0.06</td>
<td>29%</td>
<td>0.00</td>
<td>0.23</td>
<td>33%</td>
<td>0.66</td>
<td>0.04</td>
<td>29%</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.10</td>
<td>33%</td>
<td>0.07</td>
<td>0.23</td>
<td>38%</td>
<td>0.66</td>
<td>0.03</td>
<td>25%</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.06</td>
<td>29%</td>
<td>0.06</td>
<td>0.23</td>
<td>38%</td>
<td>0.66</td>
<td>0.01</td>
<td>25%</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.06</td>
<td>38%</td>
<td>-0.21</td>
<td>0.17</td>
<td>29%</td>
<td>0.80</td>
<td>-0.04</td>
<td>25%</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.06</td>
<td>42%</td>
<td>-0.94</td>
<td>0.08</td>
<td>33%</td>
<td>0.53</td>
<td>0.02</td>
<td>29%</td>
<td>0.55</td>
</tr>
<tr>
<td>Monotonic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>1/2</td>
<td>0.12</td>
<td>29%</td>
<td>0.52</td>
<td>0.06</td>
<td>33%</td>
<td>0.88</td>
<td>-0.06</td>
<td>21%</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.13</td>
<td>25%</td>
<td>0.44</td>
<td>0.25</td>
<td>29%</td>
<td>0.88</td>
<td>-0.08</td>
<td>21%</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.20</td>
<td>42%</td>
<td>0.33</td>
<td>0.12</td>
<td>29%</td>
<td>0.77</td>
<td>-0.07</td>
<td>25%</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.04</td>
<td>33%</td>
<td>0.29</td>
<td>0.02</td>
<td>29%</td>
<td>0.38</td>
<td>0.04</td>
<td>38%</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-0.01</td>
<td>33%</td>
<td>-0.01</td>
<td>-0.03</td>
<td>38%</td>
<td>-0.00</td>
<td>0.14</td>
<td>42%</td>
<td>0.52</td>
</tr>
<tr>
<td>IC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.06</td>
<td>29%</td>
<td>0.18</td>
<td>0.26</td>
<td>33%</td>
<td>0.81</td>
<td>-0.05</td>
<td>21%</td>
<td>0.11</td>
</tr>
<tr>
<td>IG</td>
<td></td>
<td>0.28</td>
<td>50%</td>
<td>0.76</td>
<td>0.21</td>
<td>33%</td>
<td>0.62</td>
<td>0.02</td>
<td>21%</td>
<td>0.04</td>
</tr>
<tr>
<td>NB</td>
<td></td>
<td>0.08</td>
<td>33%</td>
<td>0.36</td>
<td>0.10</td>
<td>25%</td>
<td>0.67</td>
<td>0.02</td>
<td>25%</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Importance of a correct parameter. ΔLC gave good results for average pooling, and by using a low p value with Minkowski pooling the POI can be increased from 70% with average pooling to 90%, but the PLCC and RC are still similar to high values for average pooling.

12.3.6 Overall observations

In general for the methods using parameters the selection of the parameter is important for the performance, and should be set with care. The results also indicate that pooling is metric dependent, and also that the parameters for a pooling strategy is metric dependent.

For sharpness, SSIM and ΔLC perform quite well with Average pooling, but a slight increase in performance can be achieved by using more advanced pooling strategies. For the color attribute the performance of the metrics can also be improved by using either Minkowski or Monotonic Function pooling compared to average pooling. However, the selection of the p values are important for obtaining the best performance as seen for Monotonic Function pooling in Figure 12.3. For the lightness attribute the results differ between the databases but also between the metrics, indicating that pooling can be both image and metric dependent. For the contrast attribute the second dataset show more variation between the pooling methods and between parameters, and the method together with the parameter is important for high performance. For artifact, the saliency based pooling improves the performance of SSIM on both data sets. For other metrics, improved performance can be found, such as for the ABF metric using Monotonic Function on the second test set.
Table 12.13: Performance of the spatial pooling methods for the artifact CPQA on Test Set 2. All the six metrics are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>p</th>
<th>ABF</th>
<th>S-CIELAB</th>
<th>S-DEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.06</td>
<td>20%</td>
<td>0.12</td>
</tr>
<tr>
<td>Minkowski</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/3</td>
<td></td>
<td>-0.11</td>
<td>10%</td>
<td>0.01</td>
</tr>
<tr>
<td>1/2</td>
<td></td>
<td>-0.09</td>
<td>10%</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>-0.05</td>
<td>10%</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.32</td>
<td>20%</td>
<td>0.59</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.65</td>
<td>50%</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.22</td>
<td>20%</td>
<td>0.34</td>
</tr>
<tr>
<td>Monotonic Function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/3</td>
<td></td>
<td>0.14</td>
<td>20%</td>
<td>0.35</td>
</tr>
<tr>
<td>1/2</td>
<td></td>
<td>0.23</td>
<td>20%</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.38</td>
<td>20%</td>
<td>0.61</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.62</td>
<td>60%</td>
<td>0.81</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.74</td>
<td>60%</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.22</td>
<td>20%</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td></td>
<td>-0.18</td>
<td>10%</td>
<td>-0.34</td>
</tr>
<tr>
<td>IG</td>
<td></td>
<td>0.07</td>
<td>20%</td>
<td>0.12</td>
</tr>
<tr>
<td>NB</td>
<td></td>
<td>0.11</td>
<td>20%</td>
<td>0.34</td>
</tr>
</tbody>
</table>

12.4 Summary

In this chapter pooling methods have been investigated with the goal of improving the performance of IQ metrics. Many different pooling methods have been evaluated for different CPQAs, and their correlation with the percept has been used as a performance measure. The overall results indicate that the parameters for the pooling methods are important, but also that pooling is metric dependent. Improved performance can be found for specific metrics for specific attributes.
Table 12.14: Performance of the spatial pooling methods for the artifact CPQA on Test Set 2. All the six metrics are selected for the evaluation. The best results for each performance measure and metric are shown in bold.

<table>
<thead>
<tr>
<th>Pooling Methods</th>
<th>p</th>
<th>SSIM</th>
<th>WLF</th>
<th>ΔLC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLCC</td>
<td>POI</td>
<td>RC</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.60</td>
<td>70%</td>
<td>1.00</td>
</tr>
<tr>
<td>Minkowski</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.55</td>
<td>50%</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.56</td>
<td>60%</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.63</td>
<td>70%</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.60</td>
<td>70%</td>
<td>0.92</td>
</tr>
<tr>
<td>Monotonic Function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.58</td>
<td>60%</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.56</td>
<td>50%</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.51</td>
<td>50%</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.03</td>
<td>20%</td>
<td>-0.14</td>
</tr>
<tr>
<td>IC</td>
<td></td>
<td>-0.17</td>
<td>0%</td>
<td>-0.64</td>
</tr>
<tr>
<td>IG</td>
<td></td>
<td>-0.13</td>
<td>10%</td>
<td>-0.92</td>
</tr>
<tr>
<td>NB</td>
<td></td>
<td>-0.21</td>
<td>10%</td>
<td>-0.99</td>
</tr>
</tbody>
</table>

214
13 A NEW METRIC FOR THE COLOR QUALITY ATTRIBUTE

The evaluation carried out in the previous chapters (11 and 12) has shown that none of the current IQ metrics are suitable to measure the color CPQA. In this chapter we proposed a new metric designed for the color CPQA, where the focus is on the human visual system models (spatial filtering) and the quality calculation (third and fourth stage of Figure 13.1).

Since the introduction of Total Variation (TV) in image processing in 1992 by Rudin et al. [376], TV has become increasingly popular. The use of variational image processing has been extended to several areas of computer vision, such as inpainting, segmentation, and deblurring. Originally developed for grayscale images, TV has been extended to color images by Blomgreen and Chan [36]. Furthermore, in the last decades several efforts have been done for developing fast and robust TV solvers such as the one by Chen and Tai [73]. We refer the reader to Chan and Shen [68] for a detailed overview of variational image processing methods. In this chapter we introduce TV in the field of IQ metrics.

Human observers are sensitive to various frequencies of a visual stimuli; the Contrast Sensitivity Function (CSF) tells us how sensitive. If the frequency of visual stimuli is too high, we will not be able to differentiate between stimuli patterns (Figure 13.2). The use of CSFs has been popular in image quality metrics, such as the S-CIELAB [499] (Section 4.2.2.1) and S-DEE [412] (Section 4.2.2.2). In these metrics, the CSFs are commonly used to modulate frequencies that are less perceptible [224]. The common way to do this is to use convolution kernels to ”blur” the spatial frequencies that observers cannot perceive [499]. This method is fast, but does not result in the most precise filtering of the image [223]. Recent studies have shown that contrast is one of the most relevant perceptual and IQ factors [346]. The history of contrast is one century long, and measuring perceived contrast is not a trivial task [410]. An important milestone was given by Peli [362] in 1990, who defined a local band-limited

![Flowchart](image)

*Figure 13.1: Workflow of image quality metrics. In this chapter we focus on the human visual system models and the quality calculation.*
A NEW METRIC FOR THE COLOR QUALITY ATTRIBUTE

Figure 13.2: Contrast sensitivity threshold indicated by the curved line, if the contrast at a given spatial frequency is above the line it is invisible, while if below the line it is visible. Figure reproduced from http://www.cns.nyu.edu/~david/courses/perception/lecturenotes/channels/channels.html, visited 12/07/11.

contrast for complex images. This work will be explained in details later in the chapter as it will be relevant for our proposal of a new IQ metric for the color attribute.

13.1 Background

Peli [362] introduced a method to simulate the human visual system, where contrast at each point in an image is calculated separately to account for variations across the image, and since contrast sensitivity depends on frequency, contrast is also calculated for different frequency bands.

Peli [362] proposes the idea of a pyramidal image-contrast structure where for each frequency band, the contrast is defined as the ratio of the bandpass-filtered image at that frequency to the low-pass image filtered to an octave below the same frequency (local luminance mean).

To define local band-limited contrast for a complex image, he obtains a band-limited version of the image in the frequency domain $A(u,v)$:

$$A(u,v) \equiv A(r,\theta) \equiv F(r,\theta)G(r),$$  \hspace{1cm} (13.1)

where $u$ and $v$ are the respective horizontal and vertical spatial frequency coordinates, $G(r)$ is a band-pass filter, $r$ and $\theta$ represent the respective polar spatial frequency coordinates: $r = \sqrt{u^2 + v^2}$, $\theta = \tan^{-1}(u/v)$, and $F(r,\theta)$ is the Fourier transform of the image $I(x,y)$.

In the spatial domain the filtered image $a(x,y)$ can be represented similarly, that is, as:

$$a(x,y) = I(x,y) \ast g(x,y),$$  \hspace{1cm} (13.2)

where $\ast$ is the convolution, and $g(x,y)$ is the inverse Fourier transform of $G(r)$. 

216
In Peli’s approach of measuring local contrast, the pyramid is obtained as follows:

\[ A_i(u,v) \equiv A_i(r,\theta) \equiv F(r,\theta)G_i(r), \]

where \( G_i(r) \) is a cosinelog filter centered at frequency of \( 2^i \) cycles/picture, expressed as:

\[ G_i(r) = \frac{1}{2} \left( 1 + \cos \left( \pi \log_2 r - \pi i \right) \right). \]

The resulting contrast at the band of spatial frequencies can be represented as a two-dimensional array \( c_i(x,y) \):

\[ c_i(x,y) = \frac{a_i(x,y)}{l_i(x,y)}, \]

where \( a_i(x,y) \) is the corresponding local luminance mean image and \( l_i(x,y) \) is a low-pass-filtered version of the image containing all energy below the band.

This filtering differs from other types of filtering because suprathreshold features retain contrast and are not washed out [362] as seen in Figure 13.3.

### 13.2 The new color image quality metric

We propose a new color IQ metric based on contrast filtering and TV. First the original \( I_O \) and reproduction \( I_R \) are converted into the CIEXYZ color space. For each channel independently, the contrast of each pixel is calculated as described in Equation 13.5 in Section 13.1. The contrast \( c \) of each pixel is then compared against the contrast sensitivity threshold \( (T) \) for the corresponding channel for each band. If the contrast is suprathreshold the information is perceptible and should be kept, if the contrast is subthreshold the information is discarded (Contrast threshold axis of Figure 13.2). The contrast of each pixel is calculated for each band \( L_i(x,y) \):

\[ L_i(x,y) = \begin{cases} c(x,y) & \text{if } c(x,y) > T \\ 0 & \text{else} \end{cases}. \]
The final filtered image \( L_f \) is then the sum over the \( n \) different bands:

\[
L_f(x, y) = \sum_{i=1}^{n} L_i(x, y). \tag{13.7}
\]

For the luminance contrast sensitivity thresholds we use the same as Peli [363] while for the chrominance thresholds we use the ones from Johnson and Fairchild [223] (also the same as used for SHAME in Chapter 5), which are defined as:

\[
CSF_{\text{chroma}}(p) = a_1 \cdot e^{-b_1 \cdot p^1} + a_2 \cdot e^{-b_2 \cdot p^2}, \tag{13.8}
\]

where different parameters for \( a_1, a_2, b_1, b_2, c_1, \) and \( c_2 \) have been used as seen in Table 13.1.

Since the CIE \( \text{XYZ} \) color space is not orthogonal, i.e. the \( X \) and \( Z \) channels contain luminance information, we separate these channels into a color part and a luminance part, filtered with their respective contrast sensitivity thresholds. To obtain the luminance bandpass information in the color channel \( (X_{BL}) \), the lowpass information in the color channel \( (X_L) \) is divided by the lowpass information in the luminance channel \( (Y_L) \), and further multiplied with the bandpass information in the luminance channel \( (Y_B) \).

\[
X_{BL} = \frac{X_L}{Y_L} Y_B. \tag{13.9}
\]

The color information in the color channel \( (X_{BC}) \) is found by subtracting the luminance bandpass information in the color channel \( (C_{BL}) \) from the bandpass information in the same color channel \( (X_B) \).

\[
X_{BC} = X_B - X_{BL}. \tag{13.10}
\]

After the filtering, the original and reproduction are converted to the log-compressed OSA-UCS color space as proposed by Oleari et al. [330]. Euclidean color differences in the OSA-UCS color space are shown to correlate well with perceived differences [335].

The new Total Variation of Difference (TVD) metric, given the original contrast filtered image \( L_O \) and its filtered reproduction \( L_R \), is defined as follows:

\[
TVD = \sqrt{\sum_j \left( \int_{\Omega} \left| \nabla L_{O_j} - \nabla L_{R_j} \right| dA \right)^2} + \lambda \int_{\Omega} \sqrt{\sum_j (L_{O_j} - L_{R_j})^2} dA, \tag{13.11}
\]

where \( \sum_j \int_{\Omega} \left| \nabla L_{O_j} - \nabla L_{R_j} \right| dA \) is the TV term, while \( \lambda \int_{\Omega} \sqrt{\sum_j (L_{O_j} - L_{R_j})^2} dA \) is the Color Difference (CD) term. \( \nabla \) is the gradient of the color channel \( j \), \( \Omega \) represent the image domain, and \( \lambda \) is the weighting parameter for the CD term following the notation used by Blomgren and Chan [36]. The TV term is similar to the Color TV defined by Blomgren and Chan [36], except that we take the gradient of the difference between the original and reproduction \( \nabla (L_{O_j} - L_{R_j}) \) instead of just the gradient of the image \( \nabla L_{O_j} \). The CD term is the Euclidean color difference, and is identical to the \( \Delta E \) from Oleari et al. [329] (Section 4.2.1.3) since we calculate the difference in the OSA-UCS color space.

We will also investigate other methods to reduce the number of IQ values into a sin-
Table 13.1: The parameters for the contrast sensitivity thresholds for the chrominance channels.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>X channel</th>
<th>Z channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>109.14130</td>
<td>7.032845</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-0.00038</td>
<td>-0.000004</td>
</tr>
<tr>
<td>$c_1$</td>
<td>3.42436</td>
<td>4.258205</td>
</tr>
<tr>
<td>$a_2$</td>
<td>93.59711</td>
<td>40.690950</td>
</tr>
<tr>
<td>$b_2$</td>
<td>-0.00367</td>
<td>-0.103909</td>
</tr>
<tr>
<td>$c_2$</td>
<td>2.16771</td>
<td>1.648658</td>
</tr>
</tbody>
</table>

gle number representing quality, so called pooling strategies (as introduced in Chapter 12). For the TV term (Equation 13.11) we will replace the standard outer norm (L2) over the color channels with the L1 norm, minimum, median and maximum. For the CD term (Equation 13.11) we will replace the standard outer norm (average) over the image space with Minkowski (M) [463] (Equation 12.3), Monotonic Function (MF) [463] (Equation 12.7), information content [463], and two different information content based pooling methods based on saliency. The saliency weighting maps are used with Equation 12.1, where the saliency models from Achanta et al. [1] and Seo and Milanfar [389] are tested. We will also test 100 $\lambda$ values from 0 to 5, with equal steps.

13.3 Evaluation

Evaluation of the proposed metric is important to quantify its performance and agreement with the human observers. We will evaluate the performance of the TVD metric by a comparison with the results of observers. Two different data sets have been selected for the evaluation. Comparison against the existing metrics evaluated in Chapter 11 will show the potential of the new metric.

13.3.1 Test data sets

The first test set [347] is the same as presented in Section 10.4.2, and contains 24 of the 25 reference images in Figure 10.7. The images were printed on an Oce Colorwave 600 CMYK wide format printer on Oce Red Label (LFM054) plain uncoated paper using three different rendering intents: perceptual, relative colorimetric, and relative colorimetric with black point compensation. Each printed image were judged by 15 observers. For details we refer the reader to Section 10.4.2 or Pedersen et al. [347].

The second test consists of ten images (Figure 11.6) is the same as in Section 11.3.2. The images were printed by a HP DesignJet 10ps printer on Stora Enso Multicopy Original plain paper. The images were printed using four different modes: the best print mode and the perceptual intent, the best mode and relative colorimetric intent, normal print mode and the perceptual intent, and the last with normal print mode and relative colorimetric intent. Ten observers judged the images according to color quality, from which $z$-scores were calculated. For details we refer to Section 11.3.2 or Pedersen et al. [359].

In order to apply objective IQ metrics to these printed images, the images are scanned into digital images and stored without compression using the framework proposed in Chapter 8.
In both data sets the observers were asked to judge the color quality of the images, and the ratings have been quantified as z-scores [116].

13.3.2 Evaluation procedure

Three state of the art IQ metrics have been chosen for comparison in the evaluation: S-CIELAB [499], S-DEE [412], and ABF [459]. We use three same evaluation methods to compare the performance of different IQ metrics as in Section 12.2.4: Pearson Linear Correlation Coefficient (PLCC) is calculated for each image between the metric scores and subjective z-scores, and the first measure of performance is the mean PLCC of the whole database calculated as the average of PLCCs of each image. The second measure is the Percentage Of Images (POI) with PLCC higher than 60%. The last measure is the Rank Correlation (RC), which is the PLCC correlation between objective rank order z-score and subjective z-score. For more information on this evaluation method we refer the reader to Section 6.3 or Pedersen and Hardeberg [350].

13.3.3 Results and discussion

We will show the results for the following configurations of the TVD metric; TV term with L1 and L2 pooling, $\lambda$ equal to 0, 1, and the best $\lambda$, and for the color term we will show the mean pooling together with the best pooling method. The results from the evaluation can be seen in Tables 13.2 and 13.3 for the first and second dataset. For the first test set a combination of the TV and CD terms, where the L1 pooling for the TV term, $\lambda = 4$ and the MF pooling with $p = 4$ gives the highest correlation with the perceived color quality. With a $\lambda = 1$ we obtain results similar to existing metrics. Average pooling has the highest PLCC for $\lambda = 0$, an increase in performance can be found with both Minkowski pooling (Figure 13.5) and Monotonic Function pooling (Figure 13.6). We can see that most pooling methods produce better results than average pooling, and that in many cases using a $\lambda = 0$ will generate the a higher PLCC than using a $\lambda > 0$. However, there are $p$ values together with $\lambda$ values that produce higher PLCC values then $\lambda = 0$, such as $p = 8$ in Minkowski pooling (Figure 13.5) and Monotonic Function pooling with $p = 8$, $p = 4$, and $p = 1$. It is also noticeable that the TVD metric has less variable PLCC values for different $\lambda$ values than Monotonic Function pooling.

TV shows the results for the TV term (Equation 13.11), without the spatial filtering and the CD term. We see that it has similar performance for PLCC and POI, but a higher RC. Nonetheless, the performance of the new metric is not great, most likely since the visual differences of the first test is small, making the task very difficult for IQ metrics.

For the second test set an equal weighting of the TV and CD term gives similar results to the state of the art metrics. However, by reducing the importance or removing the CD term ($\lambda = 0$) the performance of the TVD greatly improves (Table 13.3). Average pooling shows the highest PLCC for lambda = 0 (Figure 13.7). The same can also be seen for Minkowski pooling (Figure 13.8) and for Monotonic Function pooling (Figure 13.9). The most stable pooling method is Minkowski pooling with $p = 1/8$ over different $\lambda$ values (Figure 13.8). TV without the spatial filtering and without the CD term (last line in Table 13.3), gives slightly lower performance, indicating that the spatial filtering adds value to the TVD metric.
Table 13.2: Results for the first test set. The highest PLCC and POI is found with L1 pooling for the TV term, $\lambda = 4$, and using the Monotonic Function (MF) pooling with a $p = 4$ for the color term.

<table>
<thead>
<tr>
<th>Metric</th>
<th>PLCC</th>
<th>POI</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-CIELAB</td>
<td>$-0.29$</td>
<td>13%</td>
<td>$-0.95$</td>
</tr>
<tr>
<td>S-DEE</td>
<td>$-0.34$</td>
<td>13%</td>
<td>$-0.92$</td>
</tr>
<tr>
<td>ABF</td>
<td>$-0.39$</td>
<td>8%</td>
<td>$-0.99$</td>
</tr>
<tr>
<td>TVD (L1/$\lambda = 1/\text{mean}$)</td>
<td>$-0.31$</td>
<td>8%</td>
<td>$-0.94$</td>
</tr>
<tr>
<td>TVD (L2/$\lambda = 1/\text{mean}$)</td>
<td>$-0.30$</td>
<td>13%</td>
<td>$-0.44$</td>
</tr>
<tr>
<td>TVD (L1/$\lambda = 0$)</td>
<td>$-0.15$</td>
<td>13%</td>
<td>$-1.00$</td>
</tr>
<tr>
<td>TVD (L2/$\lambda = 0$)</td>
<td>$-0.16$</td>
<td>13%</td>
<td>$-0.95$</td>
</tr>
<tr>
<td>TVD (L1/$\lambda = 4/\text{MF}$, $p=4$)</td>
<td>$0.18$</td>
<td>29%</td>
<td>$-0.93$</td>
</tr>
<tr>
<td>TV (L1)</td>
<td>$-0.26$</td>
<td>21%</td>
<td>$-0.12$</td>
</tr>
</tbody>
</table>

Figure 13.4: Performance of the TVD metric with average pooling for varying $\lambda$ values for the first dataset.

Table 13.3: Results for the second test set. The best results are found with a $\lambda = 0$, and with a $\lambda = 0.5$ and using the Minkowski (M) pooling with a $p = 1/8$.

<table>
<thead>
<tr>
<th>Metric</th>
<th>PLCC</th>
<th>POI</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-CIELAB</td>
<td>$-0.27$</td>
<td>0%</td>
<td>$-0.23$</td>
</tr>
<tr>
<td>S-DEE</td>
<td>$-0.42$</td>
<td>0%</td>
<td>$-0.42$</td>
</tr>
<tr>
<td>ABF</td>
<td>$0.07$</td>
<td>0%</td>
<td>0.23</td>
</tr>
<tr>
<td>TVD (L1/$\lambda = 1/\text{mean}$)</td>
<td>$-0.25$</td>
<td>0%</td>
<td>$-0.15$</td>
</tr>
<tr>
<td>TVD (L2/$\lambda = 1/\text{mean}$)</td>
<td>$-0.31$</td>
<td>0%</td>
<td>$-0.29$</td>
</tr>
<tr>
<td>TVD (L1/$\lambda = 0$)</td>
<td>$0.59$</td>
<td>70%</td>
<td>0.98</td>
</tr>
<tr>
<td>TVD (L2/$\lambda = 0$)</td>
<td>$0.56$</td>
<td>60%</td>
<td>0.92</td>
</tr>
<tr>
<td>TVD (L1/$\lambda = 0.5/\text{M}$, $p=1/8$)</td>
<td>$0.59$</td>
<td>70%</td>
<td>0.98</td>
</tr>
<tr>
<td>TV (L1)</td>
<td>$0.53$</td>
<td>60%</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Figure 13.5: Performance of the TVD metric with Minkowski pooling for varying λ values for the first dataset.

Figure 13.6: Performance of the TVD metric with Monotonic Function pooling for varying λ values for the first dataset.
Figure 13.7: Performance of the TVD metric with average pooling for varying $\lambda$ values for the second dataset.

Figure 13.8: Performance of the TVD metric with Minkowski pooling for varying $\lambda$ values for the second dataset.

Figure 13.9: Performance of the TVD metric with Monotonic Function pooling for varying $\lambda$ values for the second dataset.
13.4 Investigation of image characteristics

We will carry out a similar analysis as that of Section 11.2.3 for the TVD metric. This will indicate if there is any relationship between the characteristics of the image and the performance of the TVD metric.

13.4.1 Experiment

The experimental images (Figure 10.7) and results for TVD from the first test set are the basis for our investigation. These images have been classified by two expert observers into a set of classes, being the same as for Section 11.2.3. We examine the dominant color IC, as well as the colorfulness IC. For dominant color the classes were "yes" and "no", while for colorfulness and lightness the classes were "high", "medium", "low", and "mixed".

13.4.2 Dominant color

For the calculation we use the best parameters from Table 13.2; $L_1$ norm for the TV term, and MF pooling for the CD term, and $\lambda = 4$. The observer data is based on the color attribute. Figure 13.10 shows the correlation values for the images in the two classes, when the metric values are compared to the color quality values. The mean correlation for the 11 images classified as having a dominant color is $-0.13$ (standard deviation 0.64), while for the 13 images classified as not having a dominant color the correlation is $-0.43$ (standard deviation 0.59). This is the same tendency as for S-CIELAB in Section 11.2.3.3.

We also test the result for the same metric with only the CD term (only color differences), using a mean pooling (Figure 13.10(b)). For the images classified as having a dominant color we obtain a mean correlation of $-0.12$ (standard deviation 0.68), and for those classified as not having a dominant color a correlation of $-0.41$ (standard deviation 0.77). The mean correlation for all images is $-0.28$ (standard deviation 0.60). These results indicate that a separation of images when evaluating color quality might be useful, but only in terms of color differences. This is also in agreement with the hypothesis that when a dominant color is present preservation of this color is important. However, based on the small number of samples further work is needed to test the hypothesis.

13.4.3 Colorfulness

We also investigate the performance of the TVD metric for the different classes of colorfulness, using the results from the observers for the color CPQA. From Figure 13.11(a) we can see that the TVD metric performs well for the four images classified as having low colorfulness, and poor for those with mixed colorfulness. However, it is important to notice that these two classes are also those with the fewest number of images. The mean correlation values are found in Table 13.4.

We have also investigated the color difference part of the TVD metric, as for dominant color, and the results are shown in Figure 13.11(b). We have as for the TVD used the response from the observers for the color CPQA. The metric performs well for the two images classified as having mixed colorfulness, opposite of the TVD metric. The mean correlation values are found in Table 13.5.
Figure 13.10: Performance of the metrics for images classified as having a dominant color (red) and images classified as not having a dominant color (blue).

<table>
<thead>
<tr>
<th>Colorfulness</th>
<th>Correlation</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.85</td>
<td>0.17</td>
</tr>
<tr>
<td>Medium</td>
<td>0.09</td>
<td>0.58</td>
</tr>
<tr>
<td>High</td>
<td>0.15</td>
<td>0.72</td>
</tr>
<tr>
<td>Mixed</td>
<td>-0.64</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 13.4: Mean correlation for the TVD metric for images classified according to colorfulness.
A NEW METRIC FOR THE COLOR QUALITY ATTRIBUTE

Figure 13.11: Colorfulness classified images and the performance of the metrics.

Table 13.5: Mean correlation for the TVD metric with only the CD term for images classified according to colorfulness.

<table>
<thead>
<tr>
<th>Colorfulness</th>
<th>Correlation</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>-0.80</td>
<td>0.08</td>
</tr>
<tr>
<td>Medium</td>
<td>-0.38</td>
<td>0.53</td>
</tr>
<tr>
<td>High</td>
<td>-0.17</td>
<td>0.79</td>
</tr>
<tr>
<td>Mixed</td>
<td>0.85</td>
<td>0.22</td>
</tr>
</tbody>
</table>
13.5 Summary

A new IQ metric, Total Variation of Difference (TVD), designed for the color CPQA has been proposed in this chapter. The new metric incorporates a spatial filtering method based on the work by Peli [362] and the quality calculation is based on total variation. By a careful and improved selection of the parameters of the metric an increase in performance is obtained, in terms of correlation with human observations, as compared to state of the art metrics.
14 QUALITY ASSISTANT - TOOL FOR EVALUATING PRINT QUALITY

In this chapter we combine the work carried out in the previous chapters to develop a quality toolbox for the evaluation of print quality. With the toolbox, named Quality Assistant, one is able to quantify the quality of different quality attributes for a given printer using IQ metrics, and compare them against other printers. Such tools have several potential uses as pointed out by Hardeberg and Skarsbø [170] and Hardeberg [168]: trade-off analysis of speed and implementation cost versus IQ in algorithm development, benchmarking printers, and documentation of IQ improvements resulting from efforts spent on optimization of technology parameters. Tool also incorporates more advanced functions, allowing the user for example to see where in the color space quality issues occur or in which regions problems occur. This information can be used to optimize quality, or in the development of new algorithms (for example halftoning or gamut mapping), or printing techniques.

14.1 Pre-metric operations

This section presents the operations needed before IQ metrics can be applied.

14.1.1 Padding the images

The first step to use the Quality Assistant is to prepare the images for printing. This is done by adding registration points outside the printed image. These points are black squares placed a small distance from the corner of the image (Figure 14.1). These control points are used in the registration of the scanned image in order to have a match to the digital original, as explained in Chapter 8. After the images have been padded with control points, they are printed on a given device.

14.1.2 Scanning

After the image has been printed it needs to be scanned in order to obtain a digital representation. To ensure a correct representation of colors a profile for the scanner needs to be generated. A profile for each media and print mode should be created since the profile is dependent on the equipment [259, 260, 334]. The profile is built on a TC 3.5 CMYK test target (Figure 14.2 and Section 3.1.4.4) from GretagMachbeth. This contains a total of 432 patches.
covering a wide range of CMYK combinations. This test target is printed on the same printer as the other images, and then measured using a measurement device (for example a X-rite GretagMacbeth Eye-One IO). Then the printed test target is scanned, and compared against the measured values of the test target.

A round trip test should be carried out to measure the accuracy of the profile as suggested by Sharma [393], as done in Chapter 10.4.2.2. We have followed the same procedure as Sharma [393]: the raw scan of the test target is opened in Adobe Photoshop. The scanner profile is selected (using Image→Mode→Assign Profile). Further, the images are processed to LAB using Image→Mode→Convert to Profile where the destination space should be LAB Color. Absolute colorimetric is used as the rendering intent, and the Color Management Module (CMM) is ACE. This results in a LAB image, and for each patch in the test target the color difference can be calculated. \(\Delta E_{ab}^*\) (Section 4.2.1.2) is computed between the measured values and the values in the processed image. For further details on this method we refer the reader to Sharma [393] and Sharma and Fleming [394]. Standard statistical values as mean and maximum \(\Delta E_{ab}^*\) can be used as an indication of the accuracy of the profile.

### 14.1.3 Registration

When the images have been scanned the next step is to perform registration as described in Chapter 8. Registration is needed since the IQ metrics require the reproductions to be of the same size and orientation as the original. First the control points added prior to printing are found by a simple algorithm in both the original and the reproduction. Then the transformation between the points of the original and the points of the reproduction is found using an affine transformation. Further, the transformation is applied to the scanned image with bilinear interpolation. For more information on the registration we refer the reader to Chapter 8.

### 14.1.4 Registration validation

To ensure that the error introduced in the registration is minimal, validation of the registration is performed. Even though there was no clear indication that the angle at which the image is...
scanned influences the metric values (Section 8.2.4.4 and Figure 8.8) it is recommended that the angle is kept to a minimum. The rotation performed in the registration is calculated, and images placed more than one confidence interval (95%) from the mean rotation angle (of all registered images) should be rescanned with better precision. This is not a strict validation, but gives an indication of outliers among the scanned images.

14.2 Measuring attributes using image quality metrics

We will measure the IQ of different CPQAs using IQ metrics. Based on interaction with Océ we focus on specific attributes, namely sharpness, two artifacts (noise and banding), and color.

14.2.1 Sharpness

For the sharpness CPQA we measure both perceptual sharpness and computational sharpness.

14.2.1.1 Perceptual sharpness

Metric For the sharpness attribute the SSIM [458] metric is applied (Section 4.2.3.1). This metric has shown good correspondence with observers for the sharpness CPQA (Chapter 11). The original and registered image are both transformed to the CIEXYZ color space. The original by using standard color transformations from the sRGB original, and the reproduction by
using the scanner profile (scanner RGB to CIELAB to CIEXYZ or scanner RGB to CIEXYZ, depending on the Profile Connection Space (PCS) of the profile). The SSIM is calculated based on the Y channel in the CIEXYZ color space, the chrominance channels are not included since it has been shown that sharpness is mainly influenced by luminance [89, 226, 471], and luminance accounts for almost all the variability. Spatial average pooling is used to obtain one value for each image from the quality map produced by SSIM (Section 12.3.1). This results in a map and a mean value for each image in the dataset. The overall value for all images is the mean of the mean value of all images.

**Test images** A set of five images is proposed to be used with the SSIM metric to evaluate perceptual sharpness (Figure 14.3). These images contain the different characteristics needed for a complete evaluation of sharpness, namely edges and details. All test images are 240 dpi saved as uncompressed tif in the sRGB color space unless stated otherwise.

![Figure 14.3: Images for the evaluation of perceptual sharpness.](image)

We also investigate sharpness for maps (Figure 14.4), since these usually contains fine lines and sharp edges. These maps are 600 dpi. The evaluation is carried out for three different regions in one map and two in the other maps, where the regions represent different characteristics.

### 14.2.1.2 Computational sharpness

**Metric** Two specific metrics are used to measure computational sharpness: edge blurriness [197] and edge raggedness [197].

**Edge blurriness** Edge Blurriness (EB) is defined as the appearance of being hazy or indistinct in outline; a noticeable transition of blackness from background to character [197]. For the blurriness metric the reflectance factors at the 30% ($R_{30}$) and 90% ($R_{90}$) boundaries are computed (Figure 14.5). The $R_{30}$ and $R_{90}$ positions are calculated and the average distance between the $R_{30}$ and $R_{90}$ contours ($DIS_{30-90}$) is calculated. EB is then defined as [491]:

$$EB = \frac{DIS_{30-90}}{\sqrt{LD}},$$

where $LD$ is the average density of a solid printed area.
(a) Full maps

(b) Regions

Figure 14.4: Maps for the evaluation of perceptual sharpness. Top figure shows the full maps, while the cropped regions used for the metrics are shown in the bottom figure. The three first from the left are from the first map, while the remaining two regions on the right are from the second map.

Figure 14.5: Illustration of edge blurriness. Figure reproduced from Lee et al. [262].
**Edge raggedness**  Edge Raggedness (ER) is the appearance of geometric distortion of an edge from its ideal position [197]. ER is calculated by computing the $R_{60}$ and fitting a line to this ($\bar{R}_{60}$) as seen in Figure 14.6. Then the standard deviation of the distance from the fitted line to the $R_{60}$ boundary edge is calculated:

$$ER = \sqrt{\frac{\sum_{i=1}^{N}(R_{i60} - \bar{R}_{60})^2}{N - 1}},$$  \hspace{1cm} (14.2)

where $R_{i60}$ is the individual positions ($i$) for the $N$ aperture scans from the fitted line [491].

![R60 boundary](image)

*Figure 14.6: Illustration of edge raggedness. Cyan line indicates the $R_{60}$ boundary. Figure reproduced from Lee et al. [262].*

**Test images**  For computational sharpness a specific test target (Figure 3.14 and Section 3.1.4.4) is used, which consists of three square printed solid regions oriented at 0, 8, and 24 degrees from vertical [491]. These targets are printed, then scanned before being evaluated by the raggedness and blurriness metrics [197].

**14.2.2 Artifacts**

The artifacts CPQA is divided into sub-attributes, since different artifacts have distinctive properties. We focus on noise and banding.

**14.2.2.1 Noise**

Noise is, as sharpness, divided into perceptual and computational aspects.
Perceptual noise

Metric  The perceptual aspect is calculated using the Noise Quality Metric (NQM) [99]. This metric has been shown to correlate well with perceived noise [35, 70, 99, 255, 266, 398]. The NQM is built on Peli’s contrast pyramid, as explained in Chapter 13. However, some changes are done in order to account for contrast masking and to incorporate information about viewing medium and ambient luminance. The SNR is taken for the two filtered images (original and reproduction). The metric is defined as

\[ NQM(db) = 10\log_{10} \left( \frac{\sum_x \sum_y O^2(x,y)}{\sum_x \sum_y(O(x,y) - I(x,y))^2} \right), \]  

(14.3)

where \( O \) and \( I \) are the filtered original and reproduction, and \( x \) and \( y \) indicate the image width and height. NQM is calculated for each color channel since noise in the color channels have been shown to influence perceived noise [225]. We use the CIEXYZ color space as for sharpness, where the average of the three mean values for each color channel is used as a measure of noise.

Test images  For noise it is important to have images with low, medium, and high frequency content, since perceived noise is dependent on frequency [225]. Knowing that noise also is dependent on color [225] having a range of the most important hues is important. We use the same images (Figure 14.3) as for the evaluation of perceptual sharpness to evaluate perceptual noise. These images fulfill our requirements for evaluation of noise.

Computational noise

Metric  Standard deviation is often used to describe noise [174, 271], and we also apply this estimate here. Noise can be both in the luminance and in the chrominance channels, and both types influence IQ [225]. To achieve a good indication of noise we use a test target to cover different locations in the color space. This results in calculating the standard deviation for each patch for each channel of the color space of a test target. Since the CIELAB color space has been shown to give good results for noise measures [404] we also adopt this as the working color space. We calculate Computational Noise (CN) as:

\[ CN_i = \frac{1}{3} \sum_{j=1}^3 \sum_{i=1}^N \sqrt{(X_{ij} - \mu_{ij})^2} \]  

(14.4)

where \( N \) is the number of patches, \( X_i \) is patch number \( i \), \( \mu_i \) is the reference value of patch \( i \), and \( j \) indicate the color channel. This is then calculated for each patch, and the overall noise is given as

\[ CN = \frac{1}{N} \sum_i N_i, \]  

(14.5)

where \( i \) is the patch number, \( N \) the number of patches, and CN the overall noise.

Test images  For the computational noise calculation we use a TC3.5 CMYK test target (Figure 14.2), since the metric requires color patches. This test target contains a total of 432
patches covering a wide range of CMYK combinations, enabling noise estimation in different locations of the color space.

### 14.2.2.2 Banding

**Metric**  Banding is evaluated using the Total Variation of Difference (TVD) metric (Chapter 13). The image quality map from TVD (from the color difference term) is subject to post-processing to detect banding lines. Banding has been shown to be dependent on frequency [93], and this is taken into account due to the spatial filtering carried out in the TVD metric. To measure banding the TVD color difference map is summed to obtain a 1 dimensional representation horizontally and vertically, based on the assumption that banding occurs periodically. On each of these we look for re-occurring differences by using autocorrelation, which is the cross-correlation of the signal with itself. Autocorrelation is a common statistics to find repeating patterns, and has also been used for estimation of banding [105, 371]. We use the sum of the mean of the autocorrelation horizontally and vertically as the measure of banding, the higher correlation the more banding.

**Test images**  Two different types of test images are used for banding: pictorial images and test targets. For the pictorial images we select images with uniform regions, since banding is most visible in these regions [371], but also images with some regions with higher frequencies since banding is dependent on frequency [93]. The pictorial images for banding are shown in Figure 14.7.

![Figure 14.7: Pictorial images for the evaluation of banding.](image)

We have selected two different test targets, a grayscale gradient (Figure 14.8(a)) and the Granger rainbow (Figure 14.8(b)).

![Figure 14.8: Test targets for the evaluation of banding.](image)
14.2.3 Color

**Metric**  The color attribute is calculated using the TVD metric, as defined in Chapter 13. Given an original filtered image $L_O$ and its filtered reproduction $L_R$, TVD is calculated as

$$TVD = \sqrt{\sum_j \left( \int_{\Omega} |\nabla L_{Oj} - \nabla L_{Rj}| \, dA \right)^2 + \lambda \int_{\Omega} \sum_j (L_{Oj} - L_{Rj})^2 \, dA},$$  \hspace{1cm} (14.6)

where $\sqrt{\sum_i (\int_{\Omega} |\nabla L_{Oj} - \nabla L_{Rj}| \, dA)^2}$ is the TV term, while $\lambda \int_{\Omega} \sum_i (L_{Oj} - L_{Rj})^2 \, dA$ is the Color Difference (CD) term. $\nabla$ is the gradient of the color channel $j$, $\Omega$ represents the image domain, and $\lambda$ is the weighting parameter for the CD term, and $j$ indicates the color channel. OSA-UCS is the used color space. $\lambda$ is set to 0.5, and we use the L1 norm for the TV term, and Minkowski pooling with a $p$ of 1/8, since these parameters gave a TVD metric well correlated with the perceived color quality in Chapter 13. It is also important to notice that the lightness channel is used here in addition to the color channels. The results shown here are based on both lightness and color, but in the toolbox these results are also shown separately as suggested in Chapter 10.

**Test images**  Images for the color attribute should contain the most important hues, include saturated colors, and skin tones. The images should also have regions with varying frequencies, i.e. having both smooth regions and detail-rich regions. Three images having these properties are shown in Figure 14.9.

*Figure 14.9: Images for the evaluation of the perceptual color attribute.*

14.3 Visualization of results

The Quality Assistant visualizes results differently depending on the attribute and the results. The visualizations create a dynamic tool adapted to be used in the field, and to provide a simple and intuitive way to understand the results.

14.3.1 Spider and bar plots

To visualize the results we mainly use spider plots (also known as radar charts, web charts, polar charts, or kiviat diagrams). It is a graphical method to display multivariate data (i.e. the results from different IQ metrics) as a two-dimensional chart of three or more variables, where each variable is represented on the axes starting from the same point (center). These spider plots have been used by researchers to visualize IQ results [97, 126, 128]. An example
of a spider plot is shown in Figure 14.10. We use spider plots to visualize results with three or more values. For data with one or two values we will use a standard histogram (bar plot) as seen in Figure 14.11.

![Spider Plot Example](image.png)

*Figure 14.10: Example of a spider plot, each axes indicates an image, and the colored lines indicate different printer. The higher the value the higher the quality.*

### 14.3.2 Quality maps

Some IQ metrics provide IQ maps (Figure 14.12), where each pixel of the image is represented by a quality value. These quality maps are scaled and displayed to the full range of a color map. When quality maps for the same image and same metric is compared they are displayed on the same scale for easy comparison.

### 14.3.2.1 Box plots

For metrics generating a quality map different statistics can be computed. A common method in descriptive statistics to depict groups of numerical data is to use box plots (also known as a box-and-whisker diagram). A box plot (Figure 14.13) shows lower quartile (Q1), median (Q2), and upper quartile (Q3). The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data (1.5 times the inter quartile range). Notches around the median indicates the 95% confidence interval (i.e. if the notches of two groups do not overlap the median of the groups differ at a 5% significance level).
Figure 14.11: Standard bar plot.

Figure 14.12: Example of a quality map. The colorbar indicates quality, where a higher value represent higher quality.

Figure 14.13: Boxplot example.
14.3.2.2 3 and 4 dimensional plots

For some types of attributes plots with more than two dimensions can provide an intuitive way to visualize data. Noise is one example, where one can visualize noise in a three dimensional space (for example in CIELAB), making it easy to see in what regions of the color space noise is most prominent. We will use three dimensional plots where it is suitable, and extend it to four dimensions by changing the size of the data points in the three dimensional space. Figure 14.14 shows an example of such a plot for noise evaluation. Higher values and larger symbols indicate more noise shown in the CIELAB color space.

![Figure 14.14: Four dimensional plot for noise evaluation. The noise estimate is plotted in 3D in the CIELAB color space, where the fourth dimension is the size of the points (larger size equals more noise).](image)

14.3.2.3 Color histograms

Four dimensional plots gives an intuitive method to visualize data. However, it is up to the viewer to make a decision on which colors that have the highest amount of error (quality loss). This decision might be subjective and vary depending on the observer. Therefore, we adopt the color histogram proposed by Sun and Zheng [430], where we separate the color space into unique 27 blocks (Figure 14.15), and show the quality values for each block. We use the standard blocks from Sun and Zheng [430], where the color space (JCh) is segmented into eight segments in hue (0, 45, 135, 180, 225, 270, and 315 degrees), three sections for chroma (C = 20 and 50 as the threshold), and three lightness sections (J = 40 and 70) for low chroma, two for mid-chroma (J = 55), and none for high lightness. This results in 27 different blocks, representing different colors. For each block a quality value can be calculated, and then visualized as seen in Figure 14.16.
Figure 14.15: 3d color histogram. The color space is segmented into 27 blocks. Figure reproduced from Sun and Zheng [430].

Figure 14.16: Example color histogram segmentation for noise evaluation. Each bin (segment) corresponds to one of the regions in Figure 14.15. Each bin is also shown with a colored circle representing the average color of that segment.
14.3.3 Scale differences

The metrics used in the calculation have different scales, and in some cases opposite scales. SSIM has a scale from 0 to 1 (ascending scale), where 1 indicates high quality, while a color difference based metric (as TVD) has a scale where 0 is highest quality (descending scale). In order to plot the results from different metrics together they need to have the same direction. There are several options for achieving this. One possibility is the multiplicative inverse:

$$x' = \frac{1}{x},$$

(14.7)

where \(x\) is the metric value and \(x'\) is the multiplicative inverse of \(x\). Using the multiplicative inverse has the advantage that it is both right and left inverse. This is a non-linear mapping.

Another possibility is to use a linear mapping

$$x' = -x + \max(x),$$

(14.8)

where \(x\) is the metric value and \(x'\) is the inverse of \(x\).

Figure 14.17 shows the difference between the non-linear mapping and the linear mapping. A metric value of 0 will be mapped to infinity using the multiplicative inverse, and values from 1 to the highest value will be mapped between 1 and 0. The linear mapping avoids this problem, and we will therefore use a linear mapping.

![Figure 14.17: Comparison of multiplicative inverse and linear mapping.](image)

The linear mapping will result in the lowest quality value being mapped to 0. Since we use spider plots for visualization of the results, this is unfortunate since the point is hardly perceptible. Therefore, we use an \(\varepsilon\) argument to solve this:

$$x' = -x + \max(x) + \varepsilon,$$

(14.9)

where \(\varepsilon\) is 1 percent of the maximum value, and thereby avoiding a value of 0.
This is done for all spider plots resulting in a higher value being higher quality, but for quality maps or 3d plots where it is common to use a descending scale the linear mapping inverse is not applied.

14.4 Overview of the quality assistant

The Quality Assistant consists of several functions, as shown in Figure 14.18. These functions are found in the first window of the Quality Assistant (Figure 14.19). The user can also see which printers that are found in the center column, with the possibility to open the folder where the toolbox is located. We will below give an introduction to the main features (Figure 14.18) of the toolbox.

![Figure 14.18: Overview of the quality assistant.]

14.4.1 Show test images

The show test images function displays the test images in the toolbox. The entire image set can be displayed or images for specific attributes. In the main window these are shown as thumbnails, with the option to click on any image to display it at full size. A screenshot of the function is shown in Figure 14.20.

14.4.2 Pad images

The pad images function pads the images with the control points, which are used for the registration. A padded image is shown in Figure 14.1.
14.4.3 Scan profile evaluation

The scan profile evaluation function carries out evaluation of the scanner profile as described in Section 14.1.2. The result of the evaluation is shown as a color difference map, in terms of $\Delta E^*_{ab}$, for each color patch of the test target (Figure 14.21). Additional statistics from the evaluation is shown below the color difference map, stating the mean, maximum, minimum, median, and variance of the results. Together with the window shown in Figure 14.21 the test target is also shown, in order for the user to identify which patches that have the largest and smallest differences.

For the example shown in Figure 14.21 a mean value of 2.7 $\Delta E^*_{ab}$ is found, with a maximum of 7.2. Comparing against the results of Sharma [393], the mean and maximum are comparable. Sharma [393] states that a mean $\Delta E^*_{ab}$ of less than 3 is very good and likely to produce excellent results in printer-based workflows. However, it is also important to keep in mind that a large maximum value can potentially create problems in particular colors.

![Figure 14.21: Screenshot of "scan profile evaluation".](image-url)
14.4.4 Image registration

*Image registration* carries out registration of the scanned images in order to have them matched with the original image. The process of registration is explained in details in Chapter 8.

14.4.5 Image registration validation

The *image registration validation* function performs a validation of the registration. It analyses the rotation at which the print is scanned, and if above a threshold the user is recommended to re-scan the print. The function also allows the user to see which images are above the threshold. A screenshot of the *image registration validation* is shown in Figure 14.22. This is not a strict validation, and it is recommended that the user compares the results of one image to the overall trend. If an outlier is found this image should most likely be re-scanned.

![Figure 14.22: Screenshot of the validation of the registration. The threshold of showing the recommendation if images should be re-scanned is shown as a red line.](image)

14.4.5.1 Visual rotation

The *visual rotation* function displays a representation of the average rotation and maximum rotation against an ideal rotation (i.e. no rotation). This is a visual tool to show the differences between the scans. A screenshot of this function is shown in Figure 14.23.

![Figure 14.23: Screenshot of the "visual rotation" function.](image)
14.4.6 Calculate results

The *calculate results* function calculates the results for the IQ metrics, either for all printers or for a specific printer. The results are saved for further analysis.

14.4.7 Show results

The *show results* function shows the results from the IQ metrics. The user has several options when visualizing the results, showing the overall results for all quality attributes or the results for specific attributes.

The results are visualized using the methods as introduced in Section 14.3. In each view the user has several options depending on the attribute. The name of the metric used for the calculation of the quality is given in the top left corner, together with the quality axis. In Figure 14.24 the TVD metric is used, together with a quality axis where 0 is the worst quality of the printers evaluated and maximum is on the outskirts of the spider plot. Other metrics might have other scales, depending on the construction of the metric. Legend of the printers in the quality evaluation is given in the top right corner. In the bottom right corner the average of the quality values for each printer in the spider plot is given as an indication of the overall performance of the printers, where 1 is the best printer. On the bottom of the figure several options are given for further investigation of the quality of the different printers. These are presented in the next subsections.

In the example in Figure 14.24 the metric TVD gives printer D the highest score in all three images. printer C is the worst printer, and due to the linear mapping done (Section 14.3.3) this is placed in the center of the plot. This does not imply that this is the lowest quality possible, but that the printer has the lowest quality among the ones tested. For the visual evaluation of
the plots it important to look at the differences between the printers, and for doing this having a reference printer will most likely be helpful. The average over the images, as shown in the bottom right corner of Figure 14.24, can be used as an indication of the overall performance of the printers for the given attribute. For the example printer D has the highest average, which is also confirmed by the spider plot. It is importance to notice that even though the values are different, they might not be visually different, since no information on just noticeable differences are given.

14.4.7.1 Show test images

Shows the test images for the displayed attribute in a similar manner to what is done for the show test images, as described in Section 14.4.1. Thumbnails are displayed of the images, with the possibility to click on them for full-size view.

14.4.7.2 XLS export

This function exports the values from the spider plot to Microsoft Excel sheet. For the attribute displayed the function creates a new sheet in the Excel file where the data is written. An example of the exported values into a Excel sheet is shown in Figure 14.25. This export enables the user to carry out additional analysis in other statistical softwares.

![Figure 14.25: Example of the exported values into a Excel sheet.](image)

14.4.7.3 All images printer

This function displays the quality maps (Figure 14.12) of a given printer for all images. Each quality map can be clicked on for further information, where the quality map is displayed at a larger size and with a frame with detailed statistics on the quality map. The displayed information is the mean, maximum, minimum, and median values, in addition the standard deviation of the map is given. Three more statistics are given, the Delta J, Delta C, and Delta h, representing the lightness, chroma and hue difference between the original and the reproduction calculated in the CIECAM02 JCh color space [86, 301]. A screenshot from the Quality Assistant is shown in Figure 14.26.

We can analyze the data from the example in Figure 14.26. In this case since the metric is TVD a higher values indicates lower quality, and TVD maps are shown as color difference maps. We have a fairly high mean value, with a maximum close to the mean. This indicates that most of the values in the map have the same quality, which is also indicated by the low standard deviation. The minimum value is very low, which indicates that there are areas in the image having a very high quality. Using the Delta J, C, h we can also extract information about
the reason for the quality loss. We have a lower Delta J than Delta C, which indicates that
the quality loss is larger in the chroma channel (indicated by C) than in lightness (indicated
by L). Delta h can be used to indicate if hue shifts occur compared to the original, and Delta
h is high in the given example. Similar statistics are found for other metrics as well, but the
specific statistics depend on the metric.

14.4.7.4 All printers image

This function is similar to that of the previous function, but it displays the quality maps for all
printers for a given image. The quality maps are also clickable, giving the same information
as in Section 14.4.7.3 and Figure 14.26. In addition to showing the quality maps a boxplot (as
described in Section 14.3.2.1) based on the quality values is displayed (Figure 14.27), with
the same statistics as in Section 14.4.7.3.

14.4.7.5 4d plot

The 4d plot function displays quality values in a 3d plot with the fourth dimension being the
size of the symbols. A screenshot of the 4d plot for the color attribute is shown in Figure 14.28.
In this screenshot color differences are displayed in the CIELAB color space, where the user
has the possibility to rotate the plot. Such a plot can reveal where in the color space a printer
has the largest color differences or most severe quality issues. In the case of Figure 14.28 most
of the errors can be found on the lower L scale, indicating that most of the color difference
is located in darker colors. Note that only the worst 5% of the points are shown in the plot.

A similar plot can also be seen for the noise artifact for the test target (Figure 14.2),
but with additional statistics giving the mean, maximum, minimum, median, and standard
deviation. A screenshot of this plot is shown in Figure 14.29. We can notice that the highest
standard deviation is found in a few specific colors close to the neutral axis and fairly high on
Figure 14.27: Screenshot of the boxplot view in the Quality Assistant.

Figure 14.28: Screenshot of the 4d plot view in the Quality Assistant.
the L axis. However, the highest density of noise is found in the darker colors. Pure colors, being in the outskirts have less noise than those towards the neutral axis. Such an analysis is easier to carry out using the Quality Assistant since panning, zooming, and rotation is available.

14.5 Evaluation of the Quality Assistant

We perform a final evaluation of the perceptual metrics in the Quality Assistant. The evaluation will be carried out for the perceptual metrics, namely SSIM for sharpness, NQM for noise, and TVD for color and banding. Computational metrics are not evaluated, since these might require special evaluation techniques, additionally for EB and ER a compliance standard has been published in ISO 13660 [197].

First we introduce the experimental setup, then we show the results of the evaluation.

14.5.1 Experimental setup

14.5.1.1 Printers and print settings

The images have been printed on different printers and with different printer settings (Table 14.1), resulting four different reproductions. The printers handled color management with
the default profile, and without automatic enhancement. The rendering intent was set to rela-
tive colorimetric. For each printer and setting a complete set of images was printed.

Table 14.1: Printers and details for the evaluation of the Quality Assistant.

<table>
<thead>
<tr>
<th>Printer</th>
<th>Description</th>
<th>Paper</th>
<th>Print mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>printer A</td>
<td>Low-end CMYKcm inkjet printer</td>
<td>Copy paper (80 g/m²)</td>
<td>Normal</td>
</tr>
<tr>
<td>printer B</td>
<td>Low-end CMYK inkjet printer</td>
<td>Copy paper (80 g/m²)</td>
<td>Best</td>
</tr>
<tr>
<td>printer C</td>
<td>CMYKcm inkjet plotter</td>
<td>Coated paper (130 g/m²)</td>
<td>Productivity</td>
</tr>
<tr>
<td>printer D</td>
<td>12-channel inkjet plotter</td>
<td>Coated paper (130 g/m²)</td>
<td>Standard</td>
</tr>
</tbody>
</table>

14.5.1.2 Observers and instructions

Three expert observers, having print quality experience, were recruited for the experiment,
being both male and female. All observers passed the Ishihara color deficiency test (Sec-
tion 3.2.4.3).

The observers were told to judge the quality of the printed images given the reference
according to each attribute on the five point ITU quality scale (Excellent, Good, Fair, Poor,
Bad), as introduced in Figure 3.30. Together with the instructions the following definitions of
the quality attributes were given:

- **Sharpness** is related to the clarity of details and definition of edges.
- **Color** contains aspects related to color, such as hue, saturation, lightness, and color
  rendition.
- **Artifacts**
  - **Noise** is unwanted variations [205], i.e. fluctuations in color and luminance.
  - **Banding** appears as quasi-periodic density fluctuations in bands that extend across
    the printed image horizontally or vertically [371, 372].

14.5.1.3 Viewing conditions

The original image for each scene was shown on a calibrated Eizo CG241W. The display
had a color temperature of 6500 Kelvins and luminance level of 80 cd/m², following the
specifications of the sRGB. A monitor hood was fitted to the display to prevent glare. The
illuminance level of the viewing room was 450 ± 75 lux, with a color temperature of 5200
Kelvins and a color rendering index of 96. The observers viewed the reference image and the
printed image simultaneously from a distance of approximately 60 cm. The experiment was
set up to follow the CIE guidelines [80] as closely as possible.

14.5.1.4 Scanning and registration

In order to use IQ metrics to quantify quality, the printed images needed to be scanned. A HP
ScanJet G4050 was used for scanning the images at 1200 dpi. A scanner profile was made for
each printer and each setting based on the printed test target. After scanning the images have
been registered using the framework presented in Chapter 8.
14.5.2 Results and discussion

The observers results were calculated as Mean Opinion Score (MOS), and not $z$-scores due to the low number of observers. The observer scores have been inverted and scaled between 0 and 1. This will give spider plots similar to the IQ metrics, which enables easy comparison.

14.5.2.1 Sharpness

For sharpness the observers have judged printer $D$ to have the highest sharpness quality for all six images (Figure 14.30(b)). SSIM has also given printer $D$ the highest score for all six images (Figure 14.30(a)). We can also notice that printer $B$ is by the observers judged to be better than printer $A$ for all but one image, and the same tendency is also found for SSIM. However, the distinctive indication by SSIM that printer $B$ is better than printer $A$ and printer $C$ is not shared by the observers. We also evaluate the overall results of the observers to the overall results of the metrics. The overall ranking by the experts were printer $D$, printer $C$, printer $B$, and printer $A$. For SSIM the mean of the images are taken as the overall indicator for quality, and the following ranking is achieved: printer $D$, printer $B$, printer $A$, and printer $C$. Printer $C$ is the printer that is the difference, being placed fourth by SSIM and second by the observers. SSIM gives printer $C$ printer a very low score in image 5 (Figure 14.30(a)) being some of the reason why it is ranked last.

For the maps the observers judged the entire maps, while the metrics are calculated for different regions. Image 1-3 are three regions of the first map, and images 4 and 5 are two regions in the second map. For the visualization we have plotted the scores from the observers similarly to the metric scores (images 1-3 are the scores for the first map and images 4 and 5 for the second map). Figure 14.31(b) shows that the observers judge printer $D$ to have highest quality of the evaluated printers. SSIM agrees for some images with this judgment, such as for images 3 and 5. Only in image 3 has SSIM the large difference between printer $D$ and the other printers, as the observer have.

For the overall judgment the observers have ranked printer $D$ first, then printer $B$, printer $A$, and printer $C$, being the same ranking as for the images. Despite the differences for the
individual images SSIM ranks the printers identically to the observers for the overall evaluation.

### 14.5.2.2 Noise

Figure 14.32(b) shows that the observers prefer printer D, being the printer with the least amount of noise for all images. SSIM is in agreement with the observers (Figure 14.32(a)), except for image 5 where printer C has a slightly better score. The observers judge printer A to have the most noise (lowest quality) in five of the six images. SSIM has a similar score, giving printer A the lowest score in all six images.

The observers and the metric has the same overall ranking for noise, with printer D being the printer with the least amount of noise, followed by printer C, printer B, and printer A.

### 14.5.2.3 Banding

Figure 14.33 shows the results for the metric (TVD with post-processing) and the observers for the images in the banding evaluation. For the first image the metric correctly ranks printer C as better than printer D, followed by printer B and printer A. For the second image the metric estimates printer C to be slightly better than printer D, before a difference down to the two others, where printer A is ranked before printer B. The ranking of the two worst ones for this image is the same as the observers. For the third image the metric ranks printer C to be better than the three others, but the difference between the printers is very small. The observers also indicate that the banding in this image is similar between the printers, but with printer B as better. For the overall ranking the observers prefer printer D over printer C, followed by a tie between printer A and printer B. The metric ranks printer C before printer D, and printer A before printer B.

For the targets the differences between the printers are small for the observers, but for the first image observers judge printer A as better than printer B with the two other printers tied last. TVD correctly ranks printer A and printer B over printer C and printer D, with
Figure 14.32: Spider plots for NQM and observers for noise images.

Figure 14.33: Spider plots for TVD and observers for banding images.
printers A ranked before printer B. For the second image observer rank printer D as the best slightly before the three other printers that get the same score. TVD on the other hand ranks printer D as the worst one, with printer C first. However, the differences in metric scores are very small. Additionally, it should be noted that the visual differences in the scores from the observers between the targets are small, and therefore making it very difficult for the metric. Overall the observers prefer printer A over printer B and printer D, being tied, and printer C at last. However, the scores are very similar and with only three observers they are not statistically significantly different. The metric ranks printer A over printer B, with printer C slightly before printer D.

14.5.2.4 Color

The observers prefer printer D in terms of color for two of the three images for the color attribute (Figure 14.35(b)). TVD has printer D as the best printer for all three images. However, in the second image the difference between printer D and printer B is small. TVD ranks printer B as the second best in two of the three images. We can also notice that printer C and printer A are rated as the two printers having the lowest quality both by the observers, and by TVD for the two last images. In general TVD produces stable results over the three images, while the observers have slight differences from one image to the other.

The observers have given the following overall rank: printer D first, then printer B, printer C, and printer A at last. TVD has almost the same ranking, but switching the two last printers: printer D first, then printer B, printer A, and printer C.

14.5.3 Overall observations

Overall the metrics in the Quality Assistant agrees with the observers. However, in some attributes the scores from the observers are similar resulting in a difficult task for the metrics. The metrics in the Quality Assistant are most likely not suited to measure very small differences, but they can be a valuable tool to indicate the performance of printers with larger differences.

14.6 Summary

In this chapter we have introduced the Quality Assistant, a tool for the industry to evaluate quality of printed images. The functions of the Quality Assistant have been presented and
Figure 14.35: Spider plots for TVD and observers for color images.

explained. An evaluation of the perceptual IQ metrics in the Quality Assistant has also been carried out.
15 Conclusion

The goal of this work was to adapt, use, and evaluate image quality metrics for measuring perceived image quality, in the applied field of color printing. Estimating perceived quality of prints is not a trivial task, since quality is inherently multidimensional and subjective. Nevertheless, our aim was to have image quality metrics correlated with perceived quality, which is a complex and difficult task.

First, we introduced the concept of image quality and the existing methods to measure quality in Part I. Thereafter, in Part II we investigated existing image quality metrics in the literature. First we classified existing metrics, resulting in a better understanding of existing metrics and how they are constructed. Based on the knowledge gathered from this classification we proposed a new image quality metric in Chapter 5 taking into account region of interest and known aspects of the human visual system. Chapter 6 consisted of an overview of existing methods for the evaluation of image quality metrics, which revealed that a method for evaluating overall performance was missing. Therefore, we proposed a new method for the overall evaluation of image quality metrics based on rank order. In the final chapter of Part II, we selected image quality metrics for evaluations against human observers using the methods based on correlation for evaluating the performance of the image quality metrics.

Part III focused on the use of image quality metrics for color printing. Applying image quality metrics to printed images is not straightforward, since the metrics require corresponding digital images. In Chapter 8 we proposed a framework for converting printed images into digital images, including scanning and registration. Evaluation of state of the art metrics using the framework in Chapter 9 showed that none of the metrics correlated well with perceived overall image quality. Therefore, Chapter 10 dealt with breaking overall IQ into quality attributes, reducing the complexity and dimensionality of IQ. A set of six color printing quality attributes were proposed as the most important for the evaluation of quality of printed images, which were experimentally validated. Further, in Chapter 11, IQ metrics were selected for the different color printing quality attributes, and evaluated against experimental data from human observers. This evaluation showed that there is indeed room for improvement, and in Chapter 12 we investigated how one of the stages in image quality metrics, pooling, influenced the performance of image quality metrics. Our findings showed that the pooling method is metric dependent and dependent on parameters. Even with the effort carried out with the pooling strategies none of the metrics performed well for the color quality attribute, which led to a new metric for the color quality attribute that was introduced in Chapter 13. Chapter 14 combines the work carried out in this thesis into a tool for objective quality assessment of color prints, the Quality Assistant.
**Main contributions**

- Classification and survey of image quality metrics, which organizes them andidentifies differences between metrics. Metrics are classified into four groups: mathematically based metrics, low-level based metrics, high-level based metrics, and other metrics. The classification contributes to understanding the relationship between metrics, and can be used in the development of new metrics.

- Proposed a new method for evaluating the overall performance of image quality metrics. The new method is based on the rank order method, commonly used to statistically analyze subjective data. The new method takes the image quality metric scores and inputs them to the rank order method, which generates z-scores enabling convenient comparison against z-scores from subjective experiments. The method is less sensitive to scale differences and extreme values than most existing methods.

- Extensive evaluation of image quality metrics were carried out. 22 image quality metrics were evaluated against subjective data from six state of the art databases. The results show that the performance of the image quality metrics depend on the type of distortion, but also on the database. The results also indicate that the image quality metrics perform worse for images with multiple distortions.

- Two new image quality metrics have been developed: the Spatial Hue Angle MEtric (SHAME) and the Total Variation of Difference (TVD) metric. SHAME takes into account the importance of local regions and simulation of the HVS. TVD merges total variation and contrast filtering to create a unique IQ metric well-correlated with perceived image quality.

- A framework for applying image quality metrics to printed images has been introduced. The framework uses control points to register scanned prints to match the original images. The framework outperformed a state of the art framework, and enables the use of metrics on printed images.

- Proposal of the Color Printing Quality Attributes (CPQAs), which is a set of quality attributes for the evaluation of color prints. The attributes included are the following: sharpness, contrast, color, lightness, artifacts, and physical. These have been validated in two experiments, and are considered to be a good foundation for the evaluation of print quality.

- Image quality metrics have been suggested and evaluated for the CPQA, which constituted the basis for selecting the most appropriate metric for each CPQA. The results show that metrics based on structural similarity perform well for sharpness, lightness, and contrast. Many metrics perform well for artifacts, depending on the artifact. For color none of the evaluated metrics perform well.

- Evaluation of pooling strategies for image quality metrics identified that pooling is metric, parameter, and attribute dependent. The pooling strategy and its parameters should therefore be chosen carefully depending on the metric and attribute.

- The Quality Assistant has been developed to assist researchers and developers in the printing industry to evaluate quality. The Quality Assistant consists of all functions needed to evaluate quality, including: a test image suite, a framework for digitizing the prints, a set of image quality metrics, and visualization tools.
In conclusion, our work has investigated the use of image quality metrics to measure the quality of color prints. New frameworks, metrics, and tools for measuring quality have been proposed, and the work carried out in this project will be useful for the industry and for future research and development in the image quality area.
16 FUTURE WORK

In the course of the research carried out in this thesis several possibilities for further research have been identified. The thoughts on future work can be divided into several sections, related to different parts of the thesis.

16.1 Quality Assistant

We have proposed IQ metrics for some of the quality attributes in the Quality Assistant, but there are still more attributes that should be included to have a complete evaluation of quality. There is also room for improvement for the metrics in the Quality Assistant, they are still not perfectly correlated with perceived quality.

We have focused on inkjet printing and halftoned images, but the Quality Assistant should be applicable for other printing technologies, so future work should include evaluation for other technologies, for example laser printing. However, it might be that technology-specific attributes need to be included. For laser printing other sub-artifacts should be used, such as ghosting (i.e. a vestigial image repeated at regular intervals of a page and appearing as light or dark areas [51]).

To make the Quality Assistant even more usable, investigation should be carried out on the possibility of using the scanners in the multi function printers. This would make the Quality Assistant easier to use in the field, since a separate scanner would not be needed.

The Quality Assistant uses full-reference metrics, and the disadvantage of this is that the reference image is needed. Extension of the Quality Assistant to no-reference IQ metrics would make it independent of the reference image, making it applicable to any printed image. For some attributes incorporation of no-reference metrics would be easier than others, especially those attributes where specific distortions occurs, such as noise and banding.

Usability testing of the Quality Assistant is another factor that could be investigated, in order to ensure its ease of use, efficiency, and learnability. This could include testing the Quality Assistant in use in the field, or more controlled usability testing in a lab.

The spider plots for the metrics do not indicate any uncertainty, and it can be difficult to draw a conclusion whether two different printers are perceptually different or not. Investigation of Just Noticeable Difference (JND) for the different metrics could solve this problem, and give an indication of a visual difference between two quality values.

Parametric as well as non-parametric statistical decision tools could either be appended or linked to the Quality Assistant in order to rank or present similarity measures of the results in the Quality Assistant.
16.2 Quality attributes

Even though the CPQAs have been made for color printing, they are general and they could possibly be used in other fields such as for video or display quality. We believe that the use of the CPQAs could be beneficial in these fields as they have been for printing, but they might be subject to adjustments to fit the field.

In this work the physical CPQA has not been investigated, and work is needed to find suitable metrics to measure this attribute.

In Chapter 12 pooling was discussed to reduce the number of quality values from the metric. However, pooling is also an important issue when combining quality values from different quality attributes. In order to obtain a value representing overall quality, pooling of quality attributes is required.

The dependence between attributes is important in order to pool quality attributes, and should be one of the steps in further work.

16.3 Image quality metrics

The spatial filtering method introduced in Chapter 13 is interesting, and further work should be carried out to inspect the differences from other filtering methods and to evaluate its performance. In this work the TVD metric proposed in Chapter 13 has been evaluated using printed images, and another possible next step would be to evaluate it on other databases, such as TID2008 [365] and LIVE [399].

Pooling was a central part of this work, and our findings show that the pooling strategy can influence the performance of IQ metrics. We believe that further investigation of pooling strategies would reveal more advanced and better pooling strategies, that could improve IQ metrics for prints but also in general for IQ.

With IQ metrics correlated with perceived IQ they can also be used to optimize quality. The gamut mapping algorithms proposed by Alsam and Farup [8, 9] minimizes according to a constraint, where the constraint possibly could be set by IQ metrics. Similar approaches can also be investigated for other processes, for example halftoning.

16.4 Image characteristics

In this work we also investigated the relationship between image characteristics and the performance of IQ metrics. It was mentioned in Section 11.2.3 that a larger dataset is required to draw a conclusion, and therefore a natural next phase would be to investigate characteristics on a large number of test stimuli. Other characteristics than those used in this work would also be interesting to look at, both for printed images and for other areas, such as compression.

Image characteristics could possibly be used to select the most suitable metrics that should be used to evaluate an image. Image characteristics has the potential to be used to adapt the processing of images. Some work has been done adapting processing steps to characteristics of the image, for example to improve halftoning [166, 247, 448] or assist in the gamut mapping [75, 155, 181, 430]. We believe that image characteristics can be used as a decision tool for the different processing steps.
NOMENCLATURE

ABF  Adaptive Bilateral Filter
ANSI American National Standards Institute
bdvm Der Bundesverband Druck und Medien e.V.
BPP Bit Per Pixel
CCD Charge Coupled Device
CCFL Cold Cathode Fluorescent Lamp
CCT Correlated Color Temperature
CI Confidence Interval
CIE Commission Internationale de l’Éclairage
CIS Contact Image Sensor
CISM Color Image Similarity Measure
CMM Color Management Module
CN Computational Noise
Colorlab Norwegian Color Research Laboratory
CP Control Points
CPD Cycles Per Degree
CPQA Color Printing Quality Attribute
CRT Cathode Ray Tube
CSF Contrast Sensitivity Function
CSNR Contrast Signal to Noise Ratio
CV Coefficient of Variation
CW-SSIM Complex Wavelet Structural SIMilarity
DMOS Difference Mean Opinion Score
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOG</td>
<td>Difference of Gaussians</td>
</tr>
<tr>
<td>DPI</td>
<td>Dots pr inch</td>
</tr>
<tr>
<td>DSCQS</td>
<td>Double Stimulus Continuous Quality Scale</td>
</tr>
<tr>
<td>DSIS</td>
<td>Double Stimulus Impairment Scale</td>
</tr>
<tr>
<td>EB</td>
<td>Edge blurriness</td>
</tr>
<tr>
<td>ECA</td>
<td>European Color Initiative</td>
</tr>
<tr>
<td>EDWM</td>
<td>Error Diffusion Worm Measure</td>
</tr>
<tr>
<td>ER</td>
<td>Edge Raggedness</td>
</tr>
<tr>
<td>ERA</td>
<td>European Rotogravure Association</td>
</tr>
<tr>
<td>Fogra</td>
<td>Fogra Graphic Technology Research Association</td>
</tr>
<tr>
<td>GAFFE</td>
<td>Gaze-Attentive Fixation Finding Engine</td>
</tr>
<tr>
<td>GCA</td>
<td>Graphic Communications Association</td>
</tr>
<tr>
<td>GMA</td>
<td>Gamut Mapping Algorithm</td>
</tr>
<tr>
<td>GRACoL</td>
<td>General Requirements and Applications for Commercial Offset Lithography</td>
</tr>
<tr>
<td>GU</td>
<td>Gloss Units</td>
</tr>
<tr>
<td>HP</td>
<td>Hewlett-Packard</td>
</tr>
<tr>
<td>HPminDE</td>
<td>Hue preserving minimum $\Delta E^*_{ab}$ clipping</td>
</tr>
<tr>
<td>HVS</td>
<td>Human Visual System</td>
</tr>
<tr>
<td>I3A</td>
<td>International Imaging Industry Association</td>
</tr>
<tr>
<td>IAS</td>
<td>Image Quality Analysis</td>
</tr>
<tr>
<td>IC</td>
<td>Image Characteristic</td>
</tr>
<tr>
<td>IC</td>
<td>Information Content Based Pooling</td>
</tr>
<tr>
<td>ICC</td>
<td>International Color Consortium</td>
</tr>
<tr>
<td>IFC</td>
<td>Information Fidelity Criterion</td>
</tr>
<tr>
<td>IPS</td>
<td>In-Plane Switching</td>
</tr>
<tr>
<td>IQ</td>
<td>Image Quality</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunication Union</td>
</tr>
<tr>
<td>IU</td>
<td>Interval of Uncertainty</td>
</tr>
</tbody>
</table>

264
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>JND</td>
<td>Just Noticeable Difference</td>
</tr>
<tr>
<td>JPEG</td>
<td>Joint Photographic Experts Group</td>
</tr>
<tr>
<td>LAR</td>
<td>Locally Adaptive Resolution</td>
</tr>
<tr>
<td>LCD</td>
<td>Liquid Crystal Display</td>
</tr>
<tr>
<td>LF</td>
<td>Local Features</td>
</tr>
<tr>
<td>LFM</td>
<td>Logistic Function Matrix</td>
</tr>
<tr>
<td>LL</td>
<td>Lower Limit</td>
</tr>
<tr>
<td>LPI</td>
<td>Lines pr inch</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>Multi Scale Structural SIMilarity</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>NB</td>
<td>Nonparametric Bottom-up Saliency Model Based Pooling</td>
</tr>
<tr>
<td>NQM</td>
<td>Noise Quality Measure</td>
</tr>
<tr>
<td>NQM</td>
<td>Noise Quality Metric</td>
</tr>
<tr>
<td>OSA-UCS</td>
<td>Optical Society of America’s Committee on Uniform Color Scales</td>
</tr>
<tr>
<td>PCS</td>
<td>Profile Connection Space</td>
</tr>
<tr>
<td>PDF</td>
<td>Portable Document Format</td>
</tr>
<tr>
<td>PEST</td>
<td>Parameter Estimation by Sequential Testing</td>
</tr>
<tr>
<td>PLCC</td>
<td>Pearson Linear Correlation Coefficient</td>
</tr>
<tr>
<td>POI</td>
<td>Percentage of Images</td>
</tr>
<tr>
<td>PSE</td>
<td>Point of Subjective Equality</td>
</tr>
<tr>
<td>PSNR</td>
<td>Peak Signal to Noise Ratio</td>
</tr>
<tr>
<td>PSO</td>
<td>Process-Standard Offset</td>
</tr>
<tr>
<td>QA</td>
<td>Quality Attribute</td>
</tr>
<tr>
<td>QEA</td>
<td>Quality Engineering Associates</td>
</tr>
<tr>
<td>RANSAC</td>
<td>RANd SAmple Consensus</td>
</tr>
<tr>
<td>RC</td>
<td>Rank Correlation</td>
</tr>
<tr>
<td>RFSIM</td>
<td>Riesz-transfrom based Feature SIMilarity</td>
</tr>
</tbody>
</table>
## Future work

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS</td>
<td>Root Mean Square</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Squared Error</td>
</tr>
<tr>
<td>RSC</td>
<td>Retinal-like Subsampling Contrast</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>Spatial CIELAB</td>
</tr>
<tr>
<td>S-DEE</td>
<td>Spatial-DEE</td>
</tr>
<tr>
<td>SGCK</td>
<td>chroma-dependent SiGmoidal lightness mapping and Cusp Knee scaling</td>
</tr>
<tr>
<td>SHAME</td>
<td>Spatial Hue Angle MEtric</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale-Invariant Feature Transform</td>
</tr>
<tr>
<td>SQS</td>
<td>Standard Quality Scale</td>
</tr>
<tr>
<td>SSIM</td>
<td>Structural SIMilarity</td>
</tr>
<tr>
<td>STD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SWAM</td>
<td>Saliency based Visual Attention Model</td>
</tr>
<tr>
<td>SWOP</td>
<td>Specifications for Web Offset Publications</td>
</tr>
<tr>
<td>TID</td>
<td>Tampere Image Database</td>
</tr>
<tr>
<td>TV</td>
<td>Total Variation</td>
</tr>
<tr>
<td>TVD</td>
<td>Total Variation of Difference</td>
</tr>
<tr>
<td>UIQ</td>
<td>Universal Image Quality</td>
</tr>
<tr>
<td>UL</td>
<td>Upper Limit</td>
</tr>
<tr>
<td>VA</td>
<td>Vertical Alignment</td>
</tr>
<tr>
<td>VIF</td>
<td>Visual Image Fidelity</td>
</tr>
<tr>
<td>VSD</td>
<td>Verband der Schweizer Druckindustrie</td>
</tr>
<tr>
<td>VSNR</td>
<td>Visual Signal to Noise Ratio</td>
</tr>
<tr>
<td>WLF</td>
<td>Weighted multi-Level Framework</td>
</tr>
</tbody>
</table>
AUTHOR INDEX

Achanta, R. 201, 219, 281, 339
Acharya, T. 126, 281
Adelson, E. H. 281, 334, 336
Agaian, S. S. 310, 329
Ajagamelle, S. A. 6, 9, 70, 71, 99, 100, 111, 113, 114, 281, 330, 364
Akenine-Möller, T. 303, 329
Albregtsen, F. 5, 7, 9, 157, 158, 161, 163–165, 168, 171, 174, 196, 200, 201, 215, 219, 303, 305, 331
Alers, H. 49, 307
Alford, J. R. 301, 323
Algazi, R. 301, 323, 324
Algazi, V. R. 282, 325
Algeri, T. 307, 334–336, 363
Allen, A. R. 89, 288
Allen, E. 173, 187, 281, 304, 329
Alsam, A. 262, 281
Altman, D. G. 87, 281
Amirshahi, S. A. 5, 9, 173, 189, 305
An, K. 282, 328
Anderson, C. H. 281, 334, 336
Andersson, M. 144, 146, 147, 180, 282
Aoki, N. 141, 142, 144, 148, 308
Arce, G. R. 299, 347, 348, 351
Arditi, A. 163, 298
Arndt, D. 51, 288
Asada, T. 79, 289
Asikainen, R. 54, 65, 159, 293
Autrusseau, F. 99, 101, 111, 285
Avadhanam, N. 282, 325
Avcibas, I. 62, 63, 282, 326
Axelson, P. 282, 351
Ayed, M. A. B. 282, 308, 326, 328
Babcock, J. S. 75, 95, 282
Baddeley, A. J. 315, 324
Bai, J. 70, 80, 198, 282
Bakke, A. M. 152, 180, 282, 290
Bala, R. 128, 229, 299
Balasubramanian, R. 180, 282
Bando, E. 7, 67, 70, 80, 86, 93, 113, 139, 142, 144, 146, 148, 173, 180, 282, 293
Bang, Y. 48, 128, 144, 148, 233, 234, 282, 283, 297, 299
Barańczuk, Z. 6, 9, 56, 75, 136, 139,
Burns, P. D. 128, 311, 316
Burt, P. J. 281, 334, 336
Calabria, A. J. 149, 154, 285
Caldera, M. 101, 289
Callet, P. Le 63, 99, 101, 111, 285
Canny, J. 285, 349
Cao, D. 141, 142, 144, 148, 236, 287
Cao, G. 6, 9, 75, 136, 139, 171, 173, 187, 285, 330, 345
Capra, A. 163, 285
Caracciolo, V. 6, 9, 99, 107, 109, 110, 215, 285, 310, 330
Cardin, N. 133–136, 285
Carlson, C. F. 68, 180, 284
Carmel, S. 310, 331
Carnece, M. 286, 327
Castrorina, A. 163, 285
Caviedes, J. 145, 148, 180, 286
Chaddha, N. 286, 323
Chan, T. F. 215, 218, 283, 286
Cheikh, F. A. 6, 9, 99, 107, 109, 110, 215, 310, 312, 330
Chen, C. T. 284, 322
Chen, G. H. 286, 328
Chen, K. 215, 286
Chen, S. 286, 290, 330, 358
Chetouani, A. 286, 330
Cho, M-S. 67, 70, 80, 297
Cho, Y-H. 180, 262, 286
Choi, H-K. 48, 297
Choi, S. Y. 149, 286
Chou, C. 67, 286, 287, 323, 329
Chung, R. 149, 287
CIE 38, 50–53, 56, 79, 91, 92, 95, 107, 135, 150, 151, 159, 174, 188, 190, 247, 251, 287, 334
Ciocca, G. 235, 283, 330
Clarberg, P. 303, 329
Clarke, F. J. J. 287, 322
Colford, J. M. 193, 304
Comaniciu, D. 287, 343
Connah, D. 67, 70, 80, 93, 282
Cookingham, R. 145, 148, 149, 312
Corchs, S. 163, 285
Cordone, R. 147, 308, 335
Cormack, L. K. 46, 51, 101, 199, 262, 307, 309, 313
Couwenhoven, D. W. 311, 347
Cowan, W. B. 180, 311
Crean, P. A. 141–144, 146–148, 237, 288
Creelman, C. D. 33, 312
Crete, F. 173, 187, 232, 287
Crosby, F. B. 13, 14, 287
Cui, C. 45, 95, 141, 142, 144, 148, 180, 236, 287, 288
Cui, G. 67, 300, 325
Cui, L. 89, 288
Cusano, C. 180, 288
Dalal, E. N. 141–144, 146–148, 237, 288,
FUTURE WORK

303
Daly, S. 79, 288, 322
Damera-Venkata, N. 122, 235, 288, 301, 325, 347
Danielsen, H. E. 196, 303
de Freitas Zampolo, R. 288, 327
de Ridder, H. 141, 142, 144, 195, 196, 288, 291
de Veciana, G. 74, 309, 327
Deffner, G. 51, 288
Denecker, K. 124, 313
Deng, C. 317, 330
deQueiroz, R. 180, 282
Dijk, J. 149, 288
Dinet, E. 147, 312
Dolmiere, T. 173, 187, 232, 287
Donahue, K. D. 236, 288
Dong, W. 288, 329
Donohue, K. D. 236, 251, 307
Du, W. 282, 328
Dugay, F. 51, 91, 92, 99, 113, 115, 158, 180, 193, 289
Ebner, F. 68, 180, 284
ECI 23, 289
Eerola, T. 122, 123, 130, 131, 144, 289
Eidenvall, L. 141–144, 146–149, 304
El-Adawy, M. I. 87, 89, 303, 330
Elad, M. 297, 328
Elliot, S. L. 171, 232, 315
Enanoria, W. 193, 304
Endo, C. 79, 289
Engeldrum, P. G. 3, 4, 13, 14, 36, 50, 51, 54, 56, 92, 95, 141, 143, 145, 174, 200, 220, 289
Engelke, U. 101, 289
Eschbach, R. 56, 180, 282, 317
Eskicioglu, A. M. 173, 187, 309, 326
Estrada, F. 180, 201, 219, 281, 291, 339
European Space Agency 159, 290
Evans, B. L. 122, 235, 288, 301, 325, 347
Falkenstern, K. 6, 237, 290
Fan, Z. 290, 309, 347, 351
Farrell, J. E. 38, 62, 67, 70, 80, 121, 173, 187, 298, 310, 313, 316, 324, 351
Farrugia, J. P. 290, 325
Fatemi, E. 215, 308
Fechner, G. T. 27, 290
Fedorovskaya, E. A. 141, 142, 144, 145, 147–149, 290, 291
Feeny, S. 33, 291
Feng, X. 195, 198, 199, 291
Feng, X-F. 70, 80, 291, 296, 324, 326
Fernandez, S. R. 51, 291
FUTURE WORK

Hel-Or, H. 180, 262, 292
Hertel, D. W. 70, 80, 293
Herzog, P. G. 150, 285
Hemerhagen, J. T. 235, 293
Heynderickx, I. E. 49, 195, 198, 199, 300, 307
Hinke, W. H. 296, 347, 351
Holliman, N. 292, 329
Holm, J. 52, 150, 293
Hong, E. 261, 284
Hong, G. 79, 92, 100, 136, 142, 144, 146, 181, 293, 326
Horita, Y. 101, 294
Horiuchi, T. 76, 299, 329, 348, 351
Hou, X. 195, 198, 294, 339
Huertas, R. 67, 100, 218, 304, 330, 363
Hughes, C. E. 315, 328
Hull, M. 185, 284
Hung-Shing, C. 262, 294
Hunt, R. W. G. 144, 146, 147, 294
Hunt, R. W.G. 247, 302
Hussein, A. E. 87, 89, 303, 330
Huttu, I. 144, 148, 304
Imai, F. H. 294, 326
Imai, J. 3, 294
International Color Consortium 133, 149, 294
International Imaging Industry Association 14, 294
International Telecommunication Union 46, 47, 50, 51, 89, 294
Iordache, R. 294, 326
ISO/IEC 121, 296
Itti, L. 195, 198, 199, 296, 339
Ivkovic, G. 296, 327
Jacobson, R. E. 13, 14, 281, 296, 329
Janssen, T. J. W. M. 13, 149, 163, 296
Jarvis, J. F. 296, 347, 351
Jin, E. W. 70, 80, 296, 324
Johnson, G. M. 62, 70, 72, 81, 83, 92, 100, 136, 149, 215, 218, 235, 290, 296, 325, 326
Johnson, T. 52, 150, 293
Johnston, J. D. 308, 322
Judice, C. N. 296, 347, 351
Jung, E. 150, 285
Jung, J. 301, 329
Juricevic, I. 171, 232, 297
Kadir, T. 136, 297
Kaiser, P. K. 33, 291
Kälviäinen, H. 122, 123, 130, 131, 144, 289, 308
Kamarainen, J-K. 122, 123, 130, 131, 144, 289
Karunasekera, S. A. 297, 323
Katajamaki, J. 171, 297
Kaufman, A. 161, 283
Kawayoke, Y. 101, 294
Keegan, B. F. 126, 305
Keelan, B. W. 13, 14, 48–51, 141, 143–149, 151, 154, 297, 312
Kendall, M. G. 86, 87, 136, 174, 189, 190, 297, 352
Kennedy, G. 193, 304
Kennedy, R. 126, 305
Keshk, H. A. 87, 89, 303, 330
Kheiri, A. 50, 99, 307
Khriji, L. 308, 328
Kim, J-S. 67, 70, 80, 297
Kim, Y. J. 48, 297
Kim, Y-T. 180, 262, 286
Kimmel, R. 297, 328
Kingsbury, N. G. 297, 316, 323, 325
Kiorpes, L. 70, 81, 302
Kipphan, H. 16, 297
Kite, T. D. 122, 235, 288, 325
Klassen, V. 70, 80, 316, 324
Klein, A. H. 144, 148, 313
Klein, S. A. 310, 322
Knoblauch, K. 163, 298
Kobayashi, H. 141, 142, 144, 148, 308
Koch, C. 195, 198, 199, 296, 303, 339
Kodak 54, 298
Kokaram, A. C. 316, 325
Kolås, Ø. 91, 298
Kolpatzik, B. 173, 187, 298, 351
Konik, H. 312, 330
Koo, B-K. 67, 70, 80, 297
Kopperud, S. 152, 290
Kostiuk, I. V. 262, 298
Kotani, K. 235, 301, 310, 323, 324
Kotera, H. 76, 262, 294, 299, 329, 348, 351
Kress, W. C. 128, 316
Kruse, B. 144, 147, 148, 298, 303
Kuo, C. 148, 149, 298
Kuo, C.-C. J. 298, 324
Kusuma, M. 101, 289
Kuznetsov, Y. V. 262, 298
Ladret, P. 173, 187, 232, 287
Lagendijk, R. L. 315, 323
Lai, Y.-K. 298, 324
Lam, E. P. 298, 329
Lambrecht, C. 298, 324
Land, L. 171, 232, 297
Larabi, C. 51, 298
Larabi, M-C. 86, 87, 89, 300
Larson, E. C. 75, 87, 89, 101, 189, 198, 235, 298, 299
Lau, D. L. 299, 347, 348, 351
Le Callet, P. 46, 286, 303, 327
Le Meur, O. 195, 198, 199, 303
LeCun, Y. 193, 284
Lee, B-S. 128, 229, 299
Lee, H-K. 180, 262, 286
Lee, J. 76, 299, 329, 348, 351
Lee, K-Y. 128, 233, 234, 299
Lee, L. 128, 229, 304
Lee, M-S. 299, 328
Leek, M. R. 33, 299
Legge, G. E. 163, 299
Future work

Leisti, T. 122, 144, 149, 289, 307
Lemahieu, I. 124, 313
Lensu, L. 122, 123, 130, 131, 144, 289, 308
Levine, D. M. 51, 87, 283
Leynadier, C. 72, 185, 284
Li, C. 247, 302
Li, F., S. and Zhang 235, 299
Li, H. 288, 329
Li, J. 298, 324
Li, Q. 187, 314
Li, S. 195, 198, 199, 300, 316, 330
Li, Y. 286, 323
Lim, W. 128, 300
Lin, F-S. 299, 328
Lin, Q. 300, 322
Lin, W. 315, 328, 329
Lin, X. 315, 328
Lindberg, S. 141–144, 146–149, 300, 304
Liu, H. 49, 195, 198, 199, 300, 307
Liu, K. 67, 287, 329
Liu, L-Y. 299, 328
Liu, T. 195, 198, 199, 291
Locke, M. H. 315, 329
Lodge, M. A. 235, 300
Loo, K. C. 298, 329
Loulou, M. 282, 326
Love, S. 141, 142, 144, 148, 236, 251, 287, 307
Lu, L. 89, 189, 313, 314
Lu, Z. K. 315, 328
Lubin, J. 300, 323, 324
Lucassen, M. P. 74, 87, 100, 312, 326
Luebker, A. 163, 299
Lundström, C. 50, 300
Luo, M. R. 67, 70, 79, 80, 91, 92, 100, 136, 142, 144, 146, 149, 181, 247, 286, 293, 297, 300, 302, 308, 311, 314, 322, 325, 326
Ma, L. 195, 198, 199, 235, 299, 300, 316, 330
Ma, Y-F. 300, 339
Maalouf, A. 86, 87, 89, 300
MacDonald, L. 70, 80, 141, 142, 144, 145, 148, 284
MacDonald, L.W. 180, 300
Madden, T. E. 161, 292
Mahlab, H. 292, 327
Mani, S. 128, 300
Mantiuk, R. 301, 327
MapTube 159, 301
Marini, D. 307, 334–336, 363
Marini, F. 235, 283, 330
Martinez, O. 236, 251, 307
Masmoudi, N. 282, 308, 326, 328
Matkovic, K. 303, 324
McCabe, G. P. 89, 302
McCann, J. 291, 334, 336
McCann, J. J. 181, 301
McCormick-Goodhart, M. 301, 327
McCulloch, M. 193, 304
McDonald, R. 287, 322
McDonnell, K. T. 161, 313
Medeghini, G. 307, 334–336, 363
Meer, P. 287, 343
Melgosa, M. 67, 100, 218, 304, 330, 363
Menegaz, G. 68, 301
Meng, T. H. Y. 286, 323
Mensonen, A. 144, 148, 304
Meyer, G. 283, 325
Michelson, A. 147, 301
Milanfar, P. 201, 219, 309
Miller, R. 76, 312
Miller, R.L. 311, 347
Mindru, F. 301, 329
Ming, Z. 121, 122, 130, 315
Mitsa, T. 301, 322, 323, 348, 351
Miyahara, M. 301, 323, 324
Miyake, Y. 70, 79, 80, 144, 146–148, 198, 282, 289, 294, 301, 326
Miyata, K. 144, 146–148, 301
Mizokami, Y. 171, 232, 315
Monga, V. 301, 347
Montag, E. D. 41, 301
Montgomery, D. C. 48, 301
Moore, D. S. 89, 302
Moorthy, A. K. 195–199, 302
Morimoto, A. 70, 80, 291, 326
Moroney, N. 56, 68, 247, 302
Morovic, J. 40, 91–93, 95, 141–144, 146–148, 180, 300, 302, 326
Mou, X. 86, 87, 89, 173, 187, 302, 316, 330, 331
Movshon, J. A. 70, 81, 302
Mueller, K. 161, 313
Mullen, K. T. 76, 302
Munkberg, J. 303, 329
Murphy, M. 149, 285
Myszkowski, K. 301, 327
Naguib, N. A. 87, 89, 303, 330
Nakaguchi, T. 70, 80, 198, 282
Nakauchi, S. 303, 325
Nakaya, F. 141–144, 146–148, 237, 288
Näsänen, R. 76, 303, 322
Natale-Hoffman, K. 141, 142, 303
Neuhoff, D. L. 145, 304, 325, 351
Neumann, L. 303, 324
Newell, J. 296, 324
Newman, T. 247, 302
Ng, Y. 148, 149, 298
Ngan, K. N. 195, 198, 199, 235, 299, 300, 316, 330
Nicolas, M. 173, 187, 232, 287
Nicolas, R. 147, 312
Niebur, E. 195, 198, 199, 296, 303, 339
Nielsen, B. 196, 303
Nijenhuis, M. R. M. 303, 324
Nilsson, F. 144, 145, 147, 148, 303, 325, 348
Ninassi, A. 195, 198, 199, 303
Norberg, O. 141–144, 146–149, 180, 282, 304
Noyes, Y. X. 128, 229, 304
Nuutinen, M. 54, 65, 159, 293
Nyman, G. 122, 144, 149, 289, 307
Oberti, F. 145, 148, 180, 286
O’Dell, S. 145, 148, 149, 312
Ogden, J. M. 281, 334, 336
Ohno, Y. 18, 304
Oittinen, P. 54, 65, 122, 144, 148, 159, 289, 293, 304
Ojanen, H. 149, 307
Oleari, C. 67, 70, 71, 81, 100, 136, 163, 173, 187, 200, 201, 215, 218, 220, 304, 310, 330, 363
Olives, J-L. 149, 307
Omadani, M. 3, 294
Omamiuda, M. 262, 294
Ong, E. P. 315, 328, 329
Ord, J. K. 86, 87, 136, 174, 189, 190, 297, 352
O’Reilly, R. 126, 305
Orfanidou, M. 173, 187, 304, 329
Osher, S. 215, 308
Otsu, N. 304, 342
Owens, R. A. 315, 324
OAI, M. 193, 304
OAI, N. 193, 304
Pan, Y. 149, 285
Pan, Z. 128, 229, 304
Panetta, K. 310, 329
Pant, D. R. 218, 304
Pappas, T. N. 285, 328
Parish, D. H. 163, 299
Parker, K. J. 316, 324
Pascale, D. 68, 305
Pattanaik, S. N. 315, 328
Patterson, A. 126, 305
Pedersen, M. A. 20, 306
Pefferkorn, S. 306, 325
Pekarovicova, A. 144, 148, 315
Pellacini, F. 291, 326
Pelli, D. G. 33, 314
Pelz, J. B. 75, 95, 282
Peroche, B. 290, 325
Pesquet-Popescu, B. 283, 326
Philips, W. 124, 313
Piatko, C. 308, 323
Pizlo, Z. 144, 145, 148, 149, 171, 282, 283, 285, 312, 315, 316, 324, 325
Po, L. M. 286, 328
Pointer, M. R. 149, 286
Ponce, J. 193, 284
Purgathofer, W. 303, 324
Qiu, G. 50, 99, 307
Quiroga, M. 236, 251, 307
Radun, J. 149, 307
Rahardja, S. 315, 328
Rahmim, A. 235, 300
Rajashekar, U. 51, 199, 307, 313, 330
Rasmussen, D. R. 141–144, 146–148, 237, 288, 303
Raveendran, P. 64, 312
Rawashdeh, N. A. 236, 251, 307
Ray, L. 76, 312
Redi, J. 49, 307
Rees, M. 149, 287
Rigg, B. 67, 287, 300, 322, 325
Rindal, A. 152, 290
Ritala, R. 144, 289
Rohlf, F.J. 129, 311
Rudin, L. I. 215, 308
Rushmeier, H. 308, 323
Russell, R. 62, 308
Rust, B. 308, 323
Saarelma, H. 171, 297
Sabir, M. F. 48, 75, 87, 89, 101, 235, 309
Sadovnikov, A. 144, 308
Safranek, R. J. 308, 322
Saito, R. 76, 299, 329, 348, 351
Santoro, G. 292, 327
Sarkar, R. 296, 327
Sankur, B. 62, 63, 282, 326
Sano, C. 67, 308
Santoro, G. 292, 327
Sarlat, A. 135, 136, 154, 284
Sato, M. 141–144, 146–148, 237, 288, 303
Sato, T. 141, 142, 144, 148, 308
Sawyer, J. F. 141, 142, 144, 148, 308
Sayood, K. 62, 63, 282, 326
Scala, P. 163, 317
Schadenko, A. A. 262, 298
Scheermesser, T. 308, 324, 348, 351
Schettini, R. 163, 180, 235, 283, 285, 288, 330
Schmitt, F. 67, 70, 72, 74, 80, 100, 113, 142–144, 146, 148, 149, 180, 181, 185, 284, 329
Schulz, T. 21, 308
Seara, R. 288, 327
Seidel, H.-P. 301, 327
Seim, T. 308, 322
Seo, H. J. 201, 219, 309
Sequeira, J. 68, 301
Seshadrinathan, K. 64, 309
Shaked, D. 149, 297, 309, 328
Shang, X. 142, 193–198, 201, 219, 314
Sharma, A. 160, 230, 244, 309
Sharma, G. 18, 19, 31, 50, 56, 128, 149, 163, 229, 299, 309
Shen, J. 215, 286
Shiau, J. 309, 347, 351
Shibata, K. 101, 294
Shin, I-W. 251, 307
Shklyarov, D. 301, 327
Xue, W. 302, 331
Yamsang, N. 46, 315
Yanfang, X. 121, 122, 130, 315
Yang, C. L. 286, 328
Yang, D. 195, 198, 199, 291
Yang, X. K. 315, 328
Yao, S. 315, 328, 329
Yee, H. 316, 327
Yeh, E. M. 316, 325
Yendrikhovskij, S.N. 13, 163, 180, 316
Yu, Q. 316, 324
Yu, W. 121, 122, 130, 315
Yuasa, M. 51, 288
Zeise, E. K. 25–27, 128, 232, 234, 316
Zepernick, H-J. 101, 289
Zhang, B. 171, 316
Zhang, C. N. 288, 329
Zhang, D. 316, 331
Zhang, F. 316, 330
Zhang, H-J. 300, 339
Zhang, M. 302, 331
Zhao, B. 317, 330
Zheng, Y. 5, 9, 200, 219, 306, 331
Zheng, Z-W. 240, 241, 262, 312
Zolliker, P. 56, 91, 161, 172, 173, 187, 200, 201, 283, 311, 313, 317
Zuffi, S. 163, 317
Zunino, R. 49, 307


[121] European Space Agency. [http://www.esa.int](http://www.esa.int), 2010.


291


[193] ISO. ISO 12647: Graphic technology – process control for the manufacture of half-tone colour separations, proof and production prints.


303


[375] A. Rizzi, G. Simone, and R. Cordone. A modified algorithm for perceived contrast in
digital images. In *CGIV 2008 - Fourth European Conference on Color in Graphics,
Imaging and Vision*, pages 249–252, Terrassa, Spain, Jun 2008. IS&T.


[377] H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust. Comparing real and syn-
thetic images: Some ideas about metrics. In *Eurographics Workshop on Rendering

AI Memo 2001-014/CBCL Memo 201.

prior to printing. In S. P. Farnand and F. Gaykema, editors, *Image Quality and System
Performance V*, Proceedings of SPIE, pages 68080B–68080B–12, San Jose, CA, Jan
2008.

image dependent quantization and post-quantization data compression. In *International
Glasgow, UK, May 1989. IEEE.


Imaging Conference: Color Science and Engineering Systems, Technologies, Applica-
248-8.

122–123, Baltimore, MD, Sep 2005. IS&T.

[384] J. F. Sawyer. Effect of graininess and sharpness on perceived print quality. In *Pho-
Photographic Society.


[386] T. Scheermesser and O. Bryngdahl. Spatially dependent texture analysis and control in

report, System Brunner Telegram, July 2010.

[388] T. Seim and A. Valberg. Towards a uniform color space: A better formula to describe


PART IV

APPENDICES
A Overview of Image Quality Metrics

In Table A.1 we show an overview of image quality metrics. The metrics are sorted in chronological order, then the name of the metric if given by the authors, then the author names in column three. For the type column the metrics are grouped by type. ID = image difference, IQ = image quality, IF = image fidelity, CD = color difference, HT = halftoning, DP = difference predictor, VQ = video quality, IS = image similarity. For the column HVS, the metric must have a HVS model or CSF filter that simulates the HVS. The metrics like SSIM, who indirectly simulate the HVS is set to "no". MS indicates whether the metrics are multiscale. S/NS indicates whether the metric is spatial (FFS, FVS, VFS, VVS) or non-spatial (NS). We have divided the spatial metrics into 3 groups, FFS = fixed size of filter and fixed calculation, FVS = fixed size of filter and variable calculation, VFS = variable size of filter and fixed calculation, VVS = variable size of filter and variable calculation. Fixed size indicates that the filter, block or similar is fixed for the whole image, variable size indicates that the filter or block changes according to the image content. Fixed computation indicates the same calculation within the filter or block, variable calculation indicates that the calculation is dependent on the image content. C/G indicates whether the metric are for color or grayscale images. The test column indicates what kind for evaluation that has been performed. This is either objective or subjective, for the metrics where the authors has done the subjective test this is marked with (A). Scenes indicates the number of scenes used in the original work where the metric was proposed, the same for modification and observers. For scenes the first number of the number of scenes, while the second number in ( ) indicates the total number of images (originals × number of modifications). For observers the total number of observers is stated, and inside the ( ) number of experts are stated if this information is given by the author. - indicates that this information is not available or not stated by the authors. Figure A.1 shows relations between selected metrics.
### Table A.1: Overview of full-reference IQ metrics sorted according to year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>1976</td>
<td>ΔE^ab_∗_CIE</td>
<td>Clarke et al. [87]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1984</td>
<td>1984</td>
<td>CMC</td>
<td>Clarke et al. [87]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1986</td>
<td>1986</td>
<td>SVF</td>
<td>Clarke et al. [87]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>obj.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1987</td>
<td>1987</td>
<td>BFD</td>
<td>Luo and Rigg [276]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>1990</td>
<td>SQRI</td>
<td>Barten [26]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>5 (35)</td>
<td>Resolution</td>
<td>20</td>
<td>sub. data from Westerink and Roufs.</td>
</tr>
<tr>
<td>1991</td>
<td>1991</td>
<td>LIPMSE</td>
<td>Brailean et al. [48]</td>
<td>IF</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (1)</td>
<td>Gaussian blur</td>
<td>1 (A)</td>
<td>Metric used to restore a blurred image.</td>
</tr>
<tr>
<td>1992</td>
<td>1992</td>
<td>VDP</td>
<td>Mitsa and Varkur [290]</td>
<td>HT</td>
<td>Yes</td>
<td>No</td>
<td>VFS</td>
<td>Gray</td>
<td>sub.</td>
<td>11 (44)</td>
<td>Halftoning</td>
<td>12</td>
<td>3 metrics tested, same procedure but different CSFs.</td>
</tr>
<tr>
<td>1993</td>
<td>1993</td>
<td>VDP</td>
<td>Daly [98]</td>
<td>DP</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>1993</td>
<td>VDP</td>
<td>Watson [468]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>-</td>
<td>2 (8)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>1993</td>
<td>VDP</td>
<td>Silverstein and Klein [409]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Continued on Next Page…
<table>
<thead>
<tr>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>IQ</td>
<td>Chaddha and Meng [67]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>6 (48)</td>
<td>Compression artifacts</td>
<td>60</td>
<td>Also extended to video. Only the results from one scene presented.</td>
</tr>
<tr>
<td>1994</td>
<td>HT</td>
<td>Mitsa and Alford [289]</td>
<td>HT</td>
<td>Yes</td>
<td>No</td>
<td>VFS</td>
<td>Gray</td>
<td>sub.</td>
<td>11 (44)</td>
<td>Halftoning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>IQ</td>
<td>Karunasekera and Kingsbury [228]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (8)</td>
<td>Lapped Orthogonal Transform</td>
<td>8</td>
<td>Observer expertise not stated.</td>
</tr>
<tr>
<td>1994</td>
<td>IQ</td>
<td>Teo and Heeger [435]</td>
<td>IQ</td>
<td>Yes</td>
<td>Yes</td>
<td>VFS</td>
<td>Color</td>
<td>s/o</td>
<td>1 (2)</td>
<td>-</td>
<td>(A)</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>ΔE94</td>
<td>CIE</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>IF</td>
<td>Heeger and Teo [172]</td>
<td>IF</td>
<td>Yes</td>
<td>Yes</td>
<td>VFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (3)</td>
<td>JPEG</td>
<td>(A)</td>
<td>Extension of Teo and Heeger [435]</td>
</tr>
<tr>
<td>1995</td>
<td>IQ</td>
<td>Chou and Li [77]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>IS</td>
<td>Rushmeier et al. [377]</td>
<td>IS</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>Various</td>
<td>Synthetic images</td>
<td>-</td>
<td>Comparing real and synthetic images</td>
</tr>
<tr>
<td>1995</td>
<td>IQ</td>
<td>Westen et al. [472]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>VFS</td>
<td>Gray</td>
<td>sub.</td>
<td>6(105)</td>
<td>PCM, DPCM, DCT and SBC coding at different bit rates</td>
<td>7 (5)</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>IQ</td>
<td>Lubin [272]</td>
<td>IQ</td>
<td>Yes</td>
<td>Yes</td>
<td>VFS</td>
<td>Gray</td>
<td>sub/obj</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Various testing</td>
</tr>
<tr>
<td>1996</td>
<td>IQ</td>
<td>Zhang and Wandell [499]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>obj.</td>
<td>-</td>
<td>JPEG-DCT, halftoning and patterns</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>IQ</td>
<td>Miyahara et al. [291]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>5(25)</td>
<td>global and local distortion</td>
<td>9 (9)</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Metric</td>
<td>Author(s)</td>
<td>Type</td>
<td>HVS</td>
<td>MS</td>
<td>S/NS</td>
<td>C/G</td>
<td>Test</td>
<td>Scenes</td>
<td>Modification</td>
<td>Observers</td>
<td>Comment</td>
</tr>
<tr>
<td>------</td>
<td>-----------------</td>
<td>----------------------------</td>
<td>------</td>
<td>-----</td>
<td>----</td>
<td>------</td>
<td>------</td>
<td>--------------</td>
<td>--------</td>
<td>--------------</td>
<td>-----------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>1996</td>
<td>CMPSNR</td>
<td>Lambrecht and Farrell</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>VFS</td>
<td>Color</td>
<td>sub.</td>
<td>1 (400)</td>
<td>JPEG</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td></td>
<td>Scheermesser and Bryngdahl</td>
<td>HT</td>
<td>No</td>
<td>No</td>
<td>VVS</td>
<td>Gray</td>
<td>obj.</td>
<td>2 (8)</td>
<td>Halftoning</td>
<td>-</td>
<td>Also tested on 2 test pattern.</td>
</tr>
<tr>
<td>1996</td>
<td>PQS</td>
<td>Miyahara et al. [291]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>5 (25)</td>
<td>-</td>
<td>9 (9)</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>Sarnoff JND Vision Model</td>
<td>Lubin [273]</td>
<td>IQ</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>5 (15)</td>
<td>MPEG-2 with different bit-rates</td>
<td>20</td>
<td>Also tested on JPEG data.</td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td>Neumann et al. [315]</td>
<td>ID</td>
<td>No</td>
<td>No</td>
<td>VFS</td>
<td>Color</td>
<td>obj.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td>Wong [476]</td>
<td>HT</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td>Nijenhuis and Bloemmaer</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Gray</td>
<td>sub.</td>
<td>2 (25)</td>
<td>Interpolation</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td>Lai et al. [249]</td>
<td>IF</td>
<td>Yes</td>
<td>Yes</td>
<td>VFS</td>
<td>Color</td>
<td>sub.</td>
<td>1(1)</td>
<td>JPEG2000 (A)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td>Scheermesser and Bryngdahl</td>
<td>HT</td>
<td>No</td>
<td>No</td>
<td>VVS</td>
<td>Gray</td>
<td>obj.</td>
<td>2 (-)</td>
<td>Halftoning</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>Δg</td>
<td>Wilson et al. [474]</td>
<td>IS</td>
<td>No</td>
<td>No</td>
<td>VFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (5)</td>
<td>JPEG, different distortion types</td>
<td>(A)</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>IFA</td>
<td>Taylor et al. [433]</td>
<td>IF</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>CVDM</td>
<td>Jin et al. [219]</td>
<td>ID</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>obj.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td>Lai et al. [248]</td>
<td>IF</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>1(1)</td>
<td>JPEG2000 (A)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td>Yu et al. [490]</td>
<td>HT</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>6 (36)</td>
<td>Halftoning</td>
<td>8 (8)</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>KLK</td>
<td>Yu and Parker [489]</td>
<td>HT</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>5 (20)</td>
<td>Halftoning</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Continued on Next Page…
<table>
<thead>
<tr>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>Veryovka et al. [449]</td>
<td>HT</td>
<td>Yes</td>
<td>Yes</td>
<td>VFS</td>
<td>Gray</td>
<td>obj.</td>
<td>1/1 (3/3)</td>
<td>Halftoning</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>E</td>
<td>Yeh et al. [487]</td>
<td>VQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (9)</td>
<td>Block artifacts</td>
<td>8</td>
<td>Sequence of 64 frames.</td>
</tr>
<tr>
<td>1998</td>
<td>CCETT visual metric</td>
<td>Pefferkorn and Blin [361]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>6 (30)</td>
<td>MPEG-2</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>Bolin and Meyer [37]</td>
<td>DP</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>-</td>
<td>-</td>
<td>(A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>Franti [146]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>3 (42)</td>
<td>Compression</td>
<td>15-39</td>
<td>Possibilty for color images</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>WVDP</td>
<td>Bradley [46]</td>
<td>DP</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>1 (3)</td>
<td>Noise</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>PDM</td>
<td>Avadhanam and Algazi [13]</td>
<td>IF</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>5/5(50/75)</td>
<td>Compression</td>
<td>5(2)/5(2)</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>LSMB</td>
<td>Pappas and Neuhoff [336]</td>
<td>HT</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Used for halftoning optimization</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>QM_o</td>
<td>Nilsson [319]</td>
<td>HT</td>
<td>Yes</td>
<td>No</td>
<td>VVS</td>
<td>Gray</td>
<td>obj.</td>
<td>1(3)</td>
<td>Halftoning</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>PD</td>
<td>Nakauchi et al. [312]</td>
<td>ID</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>8 (48)</td>
<td>Gamut mapping</td>
<td>10</td>
<td>Used to optimize gamut mapping</td>
</tr>
<tr>
<td>2000</td>
<td>DM and NQM</td>
<td>Damera-Venkata et al. [99]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FVS</td>
<td>Gray</td>
<td>sub.</td>
<td>-</td>
<td>Various</td>
<td>(A)</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Farrugia and Peroche [130]</td>
<td>ID</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>obj.</td>
<td>4(8)</td>
<td>-</td>
<td>(A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>CIE [2001]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>CIFA</td>
<td>Wu et al. [477]</td>
<td>ID</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>1(1)</td>
<td>Hue</td>
<td>(A)</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>SCIELAB_I</td>
<td>Johnson andFairchild [223]</td>
<td>IQ, HT</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>1(72)</td>
<td>Sharpness</td>
<td>-</td>
<td>Extension of S-CIELAB Zhang and Wandell [499].</td>
</tr>
</tbody>
</table>

Continued on Next Page…
<table>
<thead>
<tr>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Imai et al. [183]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>sub.</td>
<td>6 (12)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2001</td>
<td>SNR_W</td>
<td>Iordache and Beghdadi [192]</td>
<td>IS</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (3)</td>
<td>Salt-and-pepper noise, blurring, JPEG</td>
<td>5 (0)</td>
<td>-</td>
</tr>
<tr>
<td>2002</td>
<td>iCAM</td>
<td>Fairchild and Johnson [124]</td>
<td>ID</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2002</td>
<td>Hong and Luo [177]</td>
<td>ID</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>obj.</td>
<td>2</td>
<td>Local color change</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2002</td>
<td>UIQ</td>
<td>Wang and Bovik [455]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (8)</td>
<td>Different distortion types</td>
<td>22</td>
<td>-</td>
</tr>
<tr>
<td>2002</td>
<td>Feng et al. [138]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>2 (14)</td>
<td>Halftoning</td>
<td>-</td>
<td>Extension of CVDM</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>ΔICM</td>
<td>Morovic and Sun [304]</td>
<td>ID</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>7 (32)</td>
<td>Gamut mapping</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2002</td>
<td>WMSE</td>
<td>Ayed et al. [16]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>2 (14)</td>
<td>Noise</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2002</td>
<td>HVS REAL</td>
<td>Avcibas et al. [14]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>30 (420)</td>
<td>JPEG, SPHIT, noise, and blur</td>
<td>17</td>
<td>Metric build on work by Frese et al. [148].</td>
</tr>
<tr>
<td>2003</td>
<td>Q_color</td>
<td>Toet and Lucassen [439]</td>
<td>IF</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>2 (21)</td>
<td>Quantization</td>
<td>4-16</td>
<td>-</td>
</tr>
<tr>
<td>2003</td>
<td>FDP</td>
<td>Ferwerda and Pellacini [142]</td>
<td>DP</td>
<td>Yes</td>
<td>No</td>
<td>NS</td>
<td>Gray</td>
<td>sub.</td>
<td>8 (24)</td>
<td>Computer generated images</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>2003</td>
<td>MSSIM</td>
<td>Wang et al. [467]</td>
<td>IQ</td>
<td>No</td>
<td>Yes</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>29 (344)</td>
<td>JPEG and JPEG2k</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2003</td>
<td>M – SVD</td>
<td>Shnayderman et al. [401]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>5 (30)</td>
<td>JPEG, JPEG2k, G.Noise, G.Blur, Sharpening, DC shift</td>
<td>10 (5/5)</td>
<td>Color extension possible</td>
</tr>
<tr>
<td>2003</td>
<td>SNR_WAV</td>
<td>Beghdadi and Pesquet-Popescu [30]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (3)</td>
<td>Gaussian noise, JPEG and grid pattern</td>
<td>&gt;25</td>
<td>-</td>
</tr>
</tbody>
</table>

Continued on Next Page…
<table>
<thead>
<tr>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>DQM and CQM</td>
<td>de Freitas Zampolo and Seara [100]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>4 (45)</td>
<td>Frequency distortion</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Quality Assessor</td>
<td>Carnec et al. [65]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>VFS</td>
<td>Color</td>
<td>sub.</td>
<td>- (90)</td>
<td>JPEG and JPEG2k</td>
<td>-</td>
<td>Also tested on LIVE</td>
</tr>
<tr>
<td>2004</td>
<td>B-CQM</td>
<td>de Freitas Zampolo and Seara [101]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (81)</td>
<td>Frequency distortion and noise injection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>SSIM</td>
<td>Wang et al. [458]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>29 (344)</td>
<td>JPEG and JPEG2k</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td>Ivkovic and Sankar [215]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>5 (20)</td>
<td>Contrast stretching, white noise, blur, JPEG2k</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>NNET</td>
<td>Bouzerdoum et al. [44]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>-</td>
<td>JPEG and JPEG2k</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2004</td>
<td>I*</td>
<td>McCormick-Goodhart et al. [285]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>-</td>
<td>2 printsystems</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>HDR-VDP</td>
<td>Mantiuk et al. [282]</td>
<td>DP</td>
<td>No</td>
<td>Yes</td>
<td>VFS</td>
<td>Grey</td>
<td>sub.</td>
<td>-</td>
<td>Quantization, noise</td>
<td>(A)</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>pdiff</td>
<td>Yee [486]</td>
<td>ID</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>Film production</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>IFC</td>
<td>Sheikh et al. [397]</td>
<td>IF</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>29 (344)</td>
<td>JPEG, JPEG200, Noise, Blur</td>
<td>20 - 25</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2005</td>
<td>CBM</td>
<td>Gao et al. [150]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>29 (344)</td>
<td>JPEG and JPEG2k</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2005</td>
<td>CWSSIM</td>
<td>Wang and Simoncelli [466]</td>
<td>IS</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>1 (12)</td>
<td>Distortion as JPEG, noise etc.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>PQSI</td>
<td>Guarneri et al. [165]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>-</td>
<td>Color</td>
<td>sub.</td>
<td>-</td>
<td>5 Interpolation algorithms</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>M-DWT</td>
<td>Gayle et al. [153]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>sub.</td>
<td>5 (30)</td>
<td>JPEG, JPEG2k, blur, noise, sharp and DC-shift</td>
<td>14</td>
<td>Stated as a color metric, but only operates on luminance</td>
</tr>
<tr>
<td>Year</td>
<td>Metric</td>
<td>Author(s)</td>
<td>Type</td>
<td>HVS</td>
<td>MS</td>
<td>S/NS</td>
<td>C/G</td>
<td>Test</td>
<td>Scenes</td>
<td>Modification</td>
<td>Observers</td>
<td>Comment</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>----------------------------</td>
<td>------</td>
<td>-----</td>
<td>----</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>------------</td>
<td>-------------------------------------</td>
<td>-----------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>2005</td>
<td>VQM_ESC</td>
<td>Yao et al. [485]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>- (344)</td>
<td>JPEG and JPEG2k</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2005</td>
<td>MHI</td>
<td>An et al. [11]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>1 (4)</td>
<td>JPEG, JPEG2k, gaussian noise and speckled noise</td>
<td>(A)</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Kimmel et al. [236]</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>3 (3)</td>
<td>Gamut mapping</td>
<td>(A)</td>
<td>Used for gamut mapping optimization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Xu et al. [481]</td>
<td>ID</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>sub.</td>
<td>-</td>
<td>Compression</td>
<td>(A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>NwMSE</td>
<td>Samet et al. [381]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>-</td>
<td>JPEG2k, JPEG, and blurring</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2006</td>
<td>VIF</td>
<td>Sheikh and Bovik [396]</td>
<td>IF</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>29 (344)</td>
<td>JPEG and JPEG2k</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2006</td>
<td>WCWSSIM</td>
<td>Brooks and Pappas [53]</td>
<td>VQ</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>3 (5)</td>
<td>Video compression and transmission distortion</td>
<td>-</td>
<td>Various testing of the metric.</td>
</tr>
<tr>
<td>2006</td>
<td>DTWT-SSIM</td>
<td>Lee et al. [263]</td>
<td>IS</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>10 (4860)</td>
<td>Blurring, scaling, rotation and shift</td>
<td>-</td>
<td>Tested on handwritten data as a similarity measure.</td>
</tr>
<tr>
<td>2006</td>
<td>ESSIM</td>
<td>Chen et al. [71]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>- (489)</td>
<td>JPEG2k, JPEG, and blurring</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>GSSIM</td>
<td>Chen et al. [72]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>- (489)</td>
<td>JPEG2k, JPEG, and blurring</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>UQI-HVS</td>
<td>Egiazarian et al. [113]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>2 (44)</td>
<td>Noise, blur, JPEG and JPEG2000</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>PSNR-HVS</td>
<td>Egiazarian et al. [113]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>2 (44)</td>
<td>Noise, blur, JPEG and JPEG2000</td>
<td>56</td>
<td></td>
</tr>
</tbody>
</table>

Continued on Next Page…
<table>
<thead>
<tr>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>PSNR-HVS-M</td>
<td>Egiazarian et al. [113]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>2 (44)</td>
<td>Noise, blur, JPEG and JPEG2000</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Mindru and Jung [288]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>1 (3)</td>
<td></td>
<td>Halftoning</td>
<td></td>
<td>(A)</td>
</tr>
<tr>
<td>2006</td>
<td>SSIM&lt;sub&gt;IPT&lt;/sub&gt;</td>
<td>Bonnier et al. [41]</td>
<td>ID</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>sub.</td>
<td>15 (90)</td>
<td>Gamut mapping</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>mPSNR</td>
<td>Munkberg et al. [310]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>sub.</td>
<td>16 (-)</td>
<td>HDR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>RCBM</td>
<td>Dong et al. [106]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>29 (204)</td>
<td>JPEG</td>
<td></td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2007</td>
<td>SEME</td>
<td>Silva et al. [406]</td>
<td>IS</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>(233)</td>
<td>JPEG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>P-CIELAB (ΔPE)</td>
<td>Chou and Liu [78]</td>
<td>IF</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>29 (779)</td>
<td>-</td>
<td></td>
<td>- Scenes from LIVE</td>
</tr>
<tr>
<td>2007</td>
<td>QMCS</td>
<td>Yao et al. [484]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>29 (344)</td>
<td>JPEG and JPEG2k</td>
<td></td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2007</td>
<td>DéCOR – WSNR</td>
<td>Wan et al. [452]</td>
<td>HT</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>obj.</td>
<td>1 (3)</td>
<td>Error diffusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>DP</td>
<td>Granger [159]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Color</td>
<td>obj.</td>
<td>-</td>
<td>Noise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>Lam and Loo [250]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Gray</td>
<td>sub.</td>
<td>2 (6)</td>
<td></td>
<td>Noise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>SBLC</td>
<td>Gorley and Holliman [158]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>NS</td>
<td>Gray</td>
<td>sub.</td>
<td>3 (54)</td>
<td>JPEG compression</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>“busyness”</td>
<td>Orfanidou et al. [331]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub.</td>
<td>10 (80)</td>
<td>JPEG and JPEG2k compression</td>
<td>10</td>
<td>Psychophysical data from Allen et al. [7]</td>
</tr>
</tbody>
</table>

Continued on Next Page…
<table>
<thead>
<tr>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>Spatial $\Delta E_0$</td>
<td>Chen et al. [74]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>obj.</td>
<td>1 (1)</td>
<td>Blurring</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2008</td>
<td>CED</td>
<td>Naguib et al. [311]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>Sub</td>
<td>(227)</td>
<td>Various</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2009</td>
<td>$\Delta E_c$</td>
<td>Oleari et al. [329]</td>
<td>CD</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Various testing</td>
<td>-</td>
</tr>
<tr>
<td>2009</td>
<td>S-DEE</td>
<td>Simone et al. [412]</td>
<td>ID</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>25 (1700)</td>
<td>-</td>
<td>Scenes from TID</td>
<td>-</td>
</tr>
<tr>
<td>2009</td>
<td>SHAME</td>
<td>Pedersen and Hardeberg [351]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Various testing</td>
</tr>
<tr>
<td>2009</td>
<td>SHAME-II</td>
<td>Pedersen and Hardeberg [351]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Various testing</td>
</tr>
<tr>
<td>2009</td>
<td>COMB</td>
<td>Bianco et al. [35]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>29 (779)</td>
<td>Various</td>
<td>-</td>
<td>LIVE database</td>
</tr>
<tr>
<td>2009</td>
<td>$D$</td>
<td>Rajashekar et al. [370]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>1(10)</td>
<td>Various (A)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2009</td>
<td>ABF</td>
<td>Wang and Hardeberg [459]</td>
<td>ID</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>10(420)</td>
<td>Various</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>2010</td>
<td>PSNR_V</td>
<td>Zhao and Deng [501]</td>
<td>IQ</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Gray</td>
<td>Sub</td>
<td>(169)</td>
<td>Various</td>
<td>-</td>
<td>Scenes from LIVE</td>
</tr>
<tr>
<td>2010</td>
<td>$S_{DOL-CIELAB}$</td>
<td>Ajagamelle et al. [5]</td>
<td>IQ</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>Various evaluation</td>
</tr>
<tr>
<td>2010</td>
<td>$S_{DOL-DEE}$</td>
<td>Ajagamelle et al. [5]</td>
<td>IQ</td>
<td>Yes</td>
<td>Yes</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>Various evaluation</td>
</tr>
<tr>
<td>2010</td>
<td>Cao et al. [61]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>19(38)</td>
<td>Gamut mapping</td>
<td>12</td>
<td>Expert observers</td>
<td>-</td>
</tr>
<tr>
<td>2010</td>
<td>PSNR-HVS-S</td>
<td>Tong et al. [440]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>TID2008 and LIVE database</td>
</tr>
<tr>
<td>2010</td>
<td>PSNR-HVS-M-S</td>
<td>Tong et al. [440]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>TID2008 and LIVE database</td>
</tr>
<tr>
<td>2010</td>
<td>RFSIM</td>
<td>Zhang et al. [495]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>sub</td>
<td>25 (1700)</td>
<td>-</td>
<td>838</td>
<td>TID2008</td>
</tr>
<tr>
<td>2010</td>
<td>$M_{DOL-DEE}$</td>
<td>Simone et al. [410]</td>
<td>ID</td>
<td>Yes</td>
<td>YES</td>
<td>FFS</td>
<td>Color</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Various</td>
<td>-</td>
</tr>
<tr>
<td>2010</td>
<td>Simple</td>
<td>Zhang et al. [494]</td>
<td>IQ</td>
<td>No</td>
<td>YES</td>
<td>FFS</td>
<td>Gray</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Five image quality databases</td>
<td>-</td>
</tr>
</tbody>
</table>

Continued on Next Page…
<table>
<thead>
<tr>
<th>Year</th>
<th>Metric</th>
<th>Author(s)</th>
<th>Type</th>
<th>HVS</th>
<th>MS</th>
<th>S/NS</th>
<th>C/G</th>
<th>Test</th>
<th>Scenes</th>
<th>Modification</th>
<th>Observers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>TVD</td>
<td>Pedersen et al. [358]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>Various printers</td>
<td>-</td>
<td>Printed images from Pedersen et al. [359] and Pedersen et al. [347]</td>
</tr>
<tr>
<td>2011</td>
<td>FSIM</td>
<td>Zhang et al. [496]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>Six image quality databases</td>
</tr>
<tr>
<td>2011</td>
<td>FSIM&lt;sub&gt;C&lt;/sub&gt;</td>
<td>Zhang et al. [496]</td>
<td>IQ</td>
<td>Yes</td>
<td>No</td>
<td>FFS</td>
<td>Color</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>Six image quality databases</td>
</tr>
<tr>
<td>2011</td>
<td>NSER</td>
<td>Mou et al. [307]</td>
<td>IQ</td>
<td>No</td>
<td>Yes</td>
<td>FFS</td>
<td>Gray</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>Six image quality databases</td>
</tr>
<tr>
<td>2011</td>
<td>BBCQ</td>
<td>Shoham et al. [403]</td>
<td>IQ</td>
<td>No</td>
<td>No</td>
<td>FFS</td>
<td>Gray</td>
<td>Sub</td>
<td>-</td>
<td>Various</td>
<td>-</td>
<td>Two image sets with compression</td>
</tr>
</tbody>
</table>
Figure A.1: Overview of the relationships between a selection of the IQ metrics. Pink ovals are color metrics, while squared gray boxes are grayscale metrics. An arrow from one metric to another indicates that the metric incorporates features from the parent metric.
In this appendix we present the Weighted multi-Level Framework (WLF) contrast metric. First a brief introduction to the background required for understanding the WLF metric, and then the metric is defined.

B.1 Background

B.1.1 Tadmor and Tolhurst

Tadmor and Tolhurst [431] based their analysis of contrast on the Difference of Gaussians (DOG) model, which is modified and adapted to natural images.

In the conventional DOG model, the spatial sensitivity in the center component of the receptive-fields is described by a bi-dimensional Gaussian with unit amplitude:

\[
Center(x,y) = \exp \left[ - \left( \frac{x}{rc} \right)^2 - \left( \frac{y}{rc} \right)^2 \right],
\]

(B.1)

where the radius \( rc \) represents the distance at which the sensitivity decreases to \( 1/e \) with respect to the peak level and \( (x,y) \) are the spatial coordinates where the receptive-field is placed. The surround component is represented by another Gaussian curve, with a larger radius, \( rs \):

\[
Surround(x,y) = 0.85 \left( \frac{rc}{rs} \right)^2 \exp \left[ - \left( \frac{x}{rs} \right)^2 - \left( \frac{y}{rs} \right)^2 \right].
\]

(B.2)

When the central point of the receptive-field is placed at the location \( (x,y) \), the output of the central component is calculated as:

\[
R_c(x,y) = \sum_{i=x-3r_c}^{i=x+3r_c} \sum_{j=y-3r_c}^{j=y+3r_c} Center(i-x, j-y)I(i,j),
\]

(B.3)
while the output of the surround component is:

\[
R_s(x, y) = \sum_{i=x-3r_s}^{i=x+3r_s} \sum_{j=y-3r_s}^{j=y+3r_s} \text{Surround}(i-x, j-y)I(i, j),
\]

(B.4)

where in both cases \(I(i,j)\) is the image pixel value at position \((i,j)\).

The simplest case, with \(r_c = 1\) and \(r_s = 2\), results in a \(7 \times 7\) center mask and a \(13 \times 13\) surround mask.

The result of the DOG model is obtained as:

\[
\text{DOG}(x, y) = \frac{R_c(x, y)}{R_s(x, y)}.
\]

(B.5)

The conventional DOG model assumes that the response of a neuron depends uniquely on the local luminance difference (ΔI) between the center and the surround. After the light adaptation process, the gain of the ganglion cells of the retina and the Lateral Geniculate Nucleus (LGN) neurons depends on the average local luminance \(I\). Thus the model response depends on the contrast stimulus, \(c\), and the DOG model must be modified by a division by the local mean luminance. They propose the following criterium for the measure of contrast:

\[
c^{TT}(x, y) = \frac{R_c(x, y) - R_s(x, y)}{R_c(x, y) + R_s(x, y)}
\]

(B.6)

In their experiments, using \(256 \times 256\) images, the overall image contrast is calculated as the average local contrast of 1000 pixel locations taken randomly while assuring that the center and surround masks do not exceed the edges of the image:

\[
C^{TT} = \frac{1}{1000} \sum_{n=1}^{1000} c^{TT}_n
\]

(B.7)

### B.1.2 Rizzi et al.

Rizzi et al. [374] have developed a very simple and efficient measure, able to estimate global and local components of contrast. It is based on two principles: to collect a simplified measures of difference among neighboring pixel and to do it on various frequency levels, according to some evidences, which reports that the use of multilevel as an important implementation feature to mimic HVS [3, 145]. We will refer to this contrast measure as RAMM.

It performs a pyramid subsampling of the image to various levels in the CIELAB color space [85]. A pyramidal structure is created by halving the image at each iteration. For each level, it calculates the local contrast in each pixel by taking the average difference between the lightness channel value of the pixel and the surrounding eight pixels, thus obtaining a contrast map of each level. The final overall measure is a recombination of the average contrast for each level:

\[
C^{RAMM} = \frac{1}{N_l} \sum_{l=1}^{N_l} \overline{c}_l,
\]

(B.8)
where \( N_l \) is the number of levels and, \( \overline{c}_l \) is the mean contrast in the level \( l \):

\[
\overline{c}_l = \frac{1}{i_{\text{max}} \cdot j_{\text{max}}} \sum_{i=1}^{i_{\text{max}}} \sum_{j=1}^{j_{\text{max}}} c_{i,j},
\]

where \( i_{\text{max}} \) and \( j_{\text{max}} \) indicate respectively the height and the width of the image, and \( c_{i,j} \) is the contrast of each pixel calculated as:

\[
c_{i,j} = \sum_{n \in N_8} \alpha_n \left| \text{Pixel}_{i,j} - \text{Pixel}_n \right|.
\]

The following values are used to define the weights of the neighboring pixels:

\[
\alpha_n = \frac{1}{4 + 2\sqrt{2}} \begin{bmatrix} \frac{\sqrt{2}}{2} & 1 & \frac{\sqrt{2}}{2} \\ 1 & 1 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 1 & \frac{\sqrt{2}}{2} \end{bmatrix},
\]

This measure has a computational complexity of \( \Theta(N\log N) \), where \( N \) is the number of pixels, which is lower than alternative local methods, keeping a comparable level of correctness in the contrast estimate [374].

The steps of the measure are illustrated in Figure B.1(a).

### B.1.3 Retinal-like subsampling contrast

Rizzi et al. [375] combined RAMM’s multilevel approach with Tadmor and Tolhurst’s evaluation of a color stimulus [375]. It works with the same pyramid subsampling as RAMM with the following differences: it computes in each pixel of each level the DOG contrast instead of the 8-neighborhood local contrast and it computes the DOG contrast not only for the lightness but also for the two chromatic channels. The three independent measures of each channel are then merged by a weighted linear combination. The final overall measure can be expressed by the formula:

\[
C^{\text{RSC}} = \alpha \cdot C^{\text{RSC}}_L + \beta \cdot C^{\text{RSC}}_a + \gamma \cdot C^{\text{RSC}}_b,
\]

where \( \alpha, \beta, \) and \( \gamma \) represent the weights of each color channel. The steps of the measure are described in Figure B.1(b).

The computational complexity of the RSC measure is the same as for RAMM:

\[
\Theta(N\log N),
\]

where \( N \) is the number of pixels, but with a slightly heavier multiplication constant due to the DOGs instead of the 8-neighbor difference computation.

As well as the previous presented measure, only one number of contrast is produced at the end, the averages of all the levels are averaged again among them with uniform weights.

This measure takes the name of Retinal-like Subsampling Contrast (RSC) and it derives from the fact that the DOG model has been used successfully in many studies to describe the receptive fields and responses of mammalian retinal ganglion cells and LGN neurones [431].
and from the way of building the pyramid structure [374].

**B.2 The weighted-level framework**

In this section we introduce the Weighted-Level Framework (WLF) and we address mainly two aspects: the subsampling method and the weights in the level recombination [417].

An antialiasing filter is introduced in the subsampling in order to minimize distortion artifacts at low resolutions.

As demonstrated by Frankle and McCann [145], Adelson et al. [3] and [362], each level has a different contribution to contrast so we redefine Equation B.8 as follows:

\[
C_i = \frac{1}{N_l} \sum_{i=1}^{N_l} \lambda_l \cdot \bar{c}_l, \tag{B.14}
\]

where \(N_l\) is the number of levels, \(\bar{c}_l\) is the mean contrast in the level \(l\) and \(i\) indicate the applied color channel as before and the new parameter \(\lambda_l\) is the weight assigned to each level \(l\).

The overall final measure is given by:

\[
C_{WL} = \alpha \cdot C_1 + \beta \cdot C_2 + \gamma \cdot C_3, \tag{B.15}
\]

where \(\alpha, \beta, \) and \(\gamma\) are the weights of each color channel. The measure can be extended to different color spaces such as XYZ and RGB and it is not limited to CIELAB.

The general structure of our proposed measure can be seen in Figure B.1(c) where the most important novelties red are shown in red: the antialiasing filter in the pyramid and weighted recombination of local contrast maps. In this framework the previously developed RAMM and RSC can be considered as special cases with uniform weighting levels in the CIELAB color space.

For more details and evaluation of the WLF we refer the reader to Simone et al. [414, 415, 417].
Figure B.1: Workflow for RAMM, RSC, and WLF. The difference between RAMM and RSC is found in the neighborhood calculation of local contrast, where RAMM uses a 8-neighborhood while RSC uses DOG-neighborhood. RAMM and RSC are just special cases with uniform weighting levels in CIELAB color space of WLF, which implements an antialiasing filter in the subsampling, a weighted recombination of the local contrast maps and it is extended also to RGB and XYZ color space.
C Saliency models as Gamut-Mapping Artifact Detectors

C.1 Saliency map

Visual Saliency is the perceptual quality that makes an object, person, or pixel region stand out relative to its neighbors and thus capture our attention. Visual attention results both from fast, pre-attentive, bottom-up visual saliency of the retinal input, as well as from slower, top-down memory and volition based processing that is task-dependent [316].

Saliency estimation methods can broadly be classified as biologically based, purely computational, or a combination of both. In general, all methods employ a low-level approach by determining contrast of image regions relative to their surroundings, using one or more features of intensity, color, and orientation [1]. Many models have been proposed based on these low-level features and almost all have a low resolution and ill-defined object boundary [180, 214, 278]. A recent model [1] outperforms the rest in that it is able to efficiently output full resolution saliency maps, establish well-defined boundaries of salient objects and disregard high frequencies arising from texture, noise and blocking artifacts.

In brief, this saliency detection algorithm applies the Difference-of-Gaussian (DoG) filter for band pass filtering. The DoG filter is widely used in edge detection and is suitable for detecting the artifacts due to the contouring effect in our context. The DoG filter is given by:

\[
    \text{DoG}(x, y) = \frac{1}{2\pi \sigma_1^2} e^{-\frac{x^2+y^2}{2\sigma_1^2}} - \frac{1}{2\pi \sigma_2^2} e^{-\frac{x^2+y^2}{2\sigma_2^2}} = G_1(x, y, \sigma_1) - G_2(x, y, \sigma_2),
\]

where \( \sigma_1 \) and \( \sigma_2 \) are the standard deviation of the gaussian distribution. In order to ensure the salient regions will be fully covered and not just highlighted on edges or in the center of the regions, we drive \( \sigma_1 \) to infinity. This results in a notch in frequency at DC component while retaining all other frequencies. To remove high frequency noise and textures, we use a small Gaussian kernel keeping in mind the need for computational simplicity. For small kernels, the binomial filter approximates the Gaussian very well in the discrete case. We use \( 1/16[1, 4, 6, 4, 1] \) giving \( \omega_{hc} = \pi/2.75 \).

The method of finding the saliency map \( S \) for an image \( I \) of width \( W \) and height \( H \) pixels can thus be formulated as:

\[
    S(x, y) = \| I_\mu - I_{\omega_{hc}}(x, y) \|,
\]

where \( I_\mu \) is the mean image feature vector, \( I_{\omega_{hc}}(x, y) \) is the corresponding image pixel vector value in the Gaussian blurred version (using a 5 x 5 separable binomial kernel) of the original image, and \( \| \| \) is the \( L_2 \) norm. Using the Lab color space, each pixel location is an \( [L, a, b]^T \)
SALIENCY MODELS AS GAMUT-MAPPING ARTIFACT DETECTORS

vector, and the $L_2$ norm is the Euclidean distance. We have the images transformed into CIELAB space to take the effect of lightness on saliency into account. An illustration of saliency detection over the original can be found in Figure C.1.

![Figure C.1: The original image and its corresponding saliency map.](image)

C.2 Applicability of saliency maps to gamut mapping artifacts

If an artifact is detectable in the image, it should be salient, and therefore saliency models should be applicable for detecting artifacts. If salient regions are lost, such as important details in the images, there will be a difference in saliency between the original and reproduction. Also in this case saliency models should be able to detect artifacts.

The first step to ensure that saliency maps are applicable to detect artifacts is to investigate the saliency maps from gamut mapped images. We have computed the saliency map for an original image and a gamut mapped image (Figure C.2). In the gamut mapped image significant loss of details occur, being one of the artifacts we focus on. Investigation of the saliency maps show that they are not similar, and that the saliency changes even though the image is the same. In this case the saliency maps have been computed over the entire image. The results indicate that a difference of saliency can be suitable for detecting artifacts.

C.3 Experimental framework: saliency models as artifact detectors

We propose a framework for using saliency models for the detection of gamut mapping artifacts. Two different approaches, one global and one local, are explained in details below.
Figure C.2: The illustration of different gamut-mapped images and their corresponding saliency maps. The first row shows the original image and the gamut mapped version with loss of details. In the next row the saliency map of the original is on the left, in the middle the saliency map of the gamut mapped version of the original. The right image shows the normalized absolute difference in saliency between the two normalized saliency maps. As we can see the original and the gamut mapped version have difference salient regions, indicating that a change of saliency can be used to detect artifacts.
C.3.1 Global strategy

The global approach is very straightforward as explained above. We detect saliency for the original and gamut mapped image directly and then derive the difference between the two saliency maps. An adaptive thresholding with the Otsu’s method, which chooses the threshold to minimize the intraclass variance of the black and white pixels, is used to locate the regions where the artifacts lie in the difference image [332]. Figure C.3 show the steps of the global approach. The method can also be illustrated by the following formula,

\[ D(x,y) = S(x,y) - S(x_g,y_g), \quad (C.2) \]

where \( S(x,y) \) is the saliency map derived from Equation C.1 and \( S(x_g,y_g) \) is the saliency map upon the gamut-mapped image. \( D(x,y) \) is the difference image which shows the artifacts by the global strategy.

We illustrate the global method in Figure C.4. For the images in the first row, the biggest difference of the gamut mapped image from the original is the loss of details in the eagle’s body. The saliency model is able to detect the attention shift due to the reproduction and give us the right result.
The contouring dominates the artifacts in the second row of images. However, since the heart of flower has a big color contrast from the background and attracts much of the attention, the saliency model will tell the difference in this region. It shows the inadequacy of the artifact detection using global approach and leads us to further investigate the image in specific regions. We build on the global strategy by using the saliency maps in local regions.

The artifact detection based on the global saliency gives a rough illustration of where they are, but it can be misleading since it is always difficult to set the threshold to extract the artifact regions from the background. Some regions equally attract much attention from the original to the reproduction and we cannot obtain the salient shift. Therefore we are required to pick out the errors from the specific regions.

![Image](image.png)

*Figure C.4: The original images are listed in the left column, the gamut mapped ones are in the middle and the results from the global method are shown on the right. For the first row, the global saliency detection generally tells where the artifacts lie. However, in the next row, the heart of flower stands out of both original and reproduced image. The global thresholding is no longer sufficient to detect the artifacts.*

### C.3.2 Local strategy

The local strategy is shown in Figure C.5, and is as follows: The original image is firstly transformed into the CIELAB color space considering about its lightness, and then the mean-shift clustering [88] is applied on the images. Using the CIELAB color space, each pixel is represented in \([L^* \ a^* \ b^*]\) vectors, which means the lightness and chromatic values are separated. In this way, we are able to investigate the influence of color in different lightness
levels. The mean-shift algorithm is a nonparametric clustering technique which is commonly used for edge preserving and image segmentation. It does not require prior knowledge of the number of clusters, the constrain of the shape of the clusters and it provides better boundary. The spatial band width is set as 3 and the range band width is 1.

Figure C.5: The experimental framework using the local method. By using the original image a mask is created with the meanshift algorithm in the CIELAB colorspace. This mask is used together with the saliency models to create segmented saliency maps of both the original and reproduction. Then the absolute difference of these two maps are taken to find artifacts.

With an adaptive threshold, the image could be separated into several parts allowing for the analysis of specific artifacts in different lightness level. In the case of detail loss, which is a problem in shadow and highlight areas, especially in gamut mapping, the image can be divided into two parts. One part generally contains the high lightness while the other is low lightness. Due to the shadow effect, we tend to have more focus on the regions in low lightness distribution. The adaptive thresholding is applied onto the mean-shift transformed image of the original. The same mask partitioning the original image will also be used to filter the reproduction, ensuring a correct comparison of regions. Then, the saliency detection is applied and the difference is extracted upon the segmented regions to show the artifact regions.

Compared with Equation C.2, the artifact detection based on local strategy is shown as,

\[ M = \text{Seg}(I_{lab}(x,y)), \]  
\[ D(x,y) = S(M(x,y)) - S(M_{g,x,y}), \]  

Where
where $I_{lab}(x,y)$ is the original image transformed into CIELAB color space. $Seg$ is the mean-shift filtering taking the low lightness into account, and thereby $M$ is the mask from the original image. Equation C.4 is similar to Equation C.2, except we apply the same mask on both original and gamut-mapped images.

Since the global model gives too high weight on the areas that capture the attention, local regions are now taking into account in depth. We can see that, the occurrence of artifacts is much higher in the regions with low lightness than the high lightness. Therefore, we focus on detecting the artifacts in the regions with low-lightness distribution. Figure C.6 show that we are able to obtain the regions where the artifacts occur from the model when compared against the areas agreed upon by the authors to contain artifacts.

With the local method, we are able to obtain all the artifacts in the eagle’s body and its eyes for the first set. In the second set of images, the edges of the white blobs are detected by our method.

![Figure C.6: The images in the left column illustrate the artifacts by our local saliency detection model, and the columns are the artifacts agreed upon by the authors to contain artifacts.](image)

For more details and evaluation of the metric we refer the reader to Cao et al. [61].
D DETECTION OF WORMS IN ERROR DIFFUSION HALFTONING

D.1 Introduction

Digital halftoning is the process of transforming a continuous tone image to a binary image, to allow printing or displaying on a bi-level display. Many different techniques have been proposed for the transformation. One of these being error diffusion [144] halftoning, its simplicity and ability to produce high quality reproductions has contributed to its popularity. This technique distributes the quantization error along the path of the image scan. This is done by a filter, determining the weight distribution of the error. The original four element filter by Floyd and Steinberg [144], as seen in Equation D.1, resulted in different artifacts.

One of the artifacts produced by error diffusion is worms, where black and white pixels string together are perceived as having a vermicular texture as seen on Figure D.1(a) [423]. This artifact is found in the highlights and shadows of the halftoned image. The original filter by Floyd and Steinberg suffers from this artifact, and because of this numerous error diffusion algorithms have been proposed.

\[
\begin{bmatrix}
\frac{1}{16} & 0 & 7 \\
3 & 5 & 1
\end{bmatrix}
\] (D.1)

Jarvis et al. [218] proposed a 12 element filter (Equation D.2), another 12 element filter was proposed by Stucki [427] (Equation D.3), both of these break up worm patterns, but they introduce artifacts in mid-tone areas [257]. Later Fan [129, 400] proposed a filter with a long tail (Equation D.4), this results in breaking up worm structures in the highlights and shadows while performing similar to the Floyd and Steinberg filter in the mid-tones [257]. An overview of other filters, each with different advantages and disadvantages, is given by Monga et al. [293].

\[
\begin{bmatrix}
\frac{1}{48} & 0 & 0 & 7 & 5 \\
3 & 5 & 7 & 5 & 3 \\
1 & 3 & 5 & 3 & 1
\end{bmatrix}
\] (D.2)

\[
\begin{bmatrix}
\frac{1}{42} & 0 & 0 & 8 & 4 \\
2 & 4 & 8 & 4 & 2 \\
1 & 2 & 4 & 2 & 1
\end{bmatrix}
\] (D.3)

\[
\begin{bmatrix}
\frac{1}{16} & 0 & 0 & 7 \\
1 & 3 & 5 & 0
\end{bmatrix}
\] (D.4)
DETECTION OF WORMS IN ERROR DIFFUSION HALFTONING

Figure D.1: The result of using (a) Floyd and Steinberg left-to-right scan path and (b) Floyd and Steinberg with serpentine scan path. The serpentine scan path breaks up worm patterns, and results in a visually more pleasing image as seen on Figure D.1(b).

Different approaches to scanning paths have been proposed to create visually more pleasing images. The original Floyd and Steinberg [144] algorithm used a left-to-right scan path, later a serpentine (boustrophedonic) processing was introduced where the scan is left-to-right and then right-to-left (Figure D.1). Other more creative paths have been proposed based on space filling curves, as the Peano and Hilbert paths [257] that break up worms, but introduce significant amount of noise.

Quality assessment is needed to show if a halftone technique increase the quality of an image compared to other state of the art techniques. There are basically two different ways to assess the quality of a halftoned images; subjectively and objectively. Subjective evaluation is carried out by observers, and is therefore influenced by the Human Visual System (HVS). Objective evaluation can be carried out in many ways. One typical way is to use measurement devices gathering numerical values. Another way is to use algorithms, commonly known as image quality metrics, in an attempt to quantify quality. Image quality metrics are usually developed to take into account properties of the HVS, and thus with the goal of being well correlated with subjective evaluations.

A number of these image quality metrics for halftone quality have been proposed. Lee et al. [261] proposed an image similarity measure to evaluate a hybrid error diffusion algorithm. Veryovka et al. [449] proposed a multiscale approach to analyze edges, and used it to evaluate different halftoning algorithms. Mitsa and Varkur [290] proposed an image quality measure for halftoning based on a spatial filtering of Mean Square Error (MSE). Scheermesser and Bryngdahl [385] proposed a texture metric for halftone images, that find the occurrence or absence of specific textures in quantized images. Nilsson [319] proposed an image quality metric for halftone images that incorporated models of both the printer and observer. S-CIELAB was proposed by Zhang and Wandell [499] as a spatial pre-processor of the CIELAB color difference formulae [82], and was designed to predict image difference for halftoning, distortion, etc. For a complete list of different metrics for halftoning and other areas we refer the reader to Pedersen and Hardeberg [355], Chapter 4, or Appendix A.

Even though many measures have been proposed for halftoning, many of them only work on a global scale without having a specific calculation for worms. A specific measure for
worms could be used to preserve worms where they are wanted, as in edges, but remove them in areas where they are not wanted. A specific measure for worms could also be used to select the most optimal halftoning algorithm for an image, resulting in the best possible output image. Such a measure for worms could also be included in a toolbox to detect halftoning artifacts.

### D.2 Proposed error diffusion worm measure

This section introduces a measure for detection and quantification of worms, called Error Diffusion Worm Measure (EDWM). The measure is a multi-step measure as shown in Figure D.2. It is a no-reference measure, where no information about the original is used in the calculation.

![Figure D.2: Description of the workflow of the EDWM. Highlights and shadows are extracted from the halftoned image before a Canny edge filter is applied. Single pixels are removed, before all pixels within in the edge boundary are extracted. Then a proximity detection is performed before small worms are removed. The final measure is calculated as the percentage of non-worm pixels divided by the number of pixels in the image.](image)

Since worms are found in highlights and shadows of the image, the halftoned image is filtered to remove parts without highlights and shadows. A sliding filter with width and height of 10 pixels is used, where the mean of each location is found. A filter size of $10 \times 10$ pixels was found to be sufficient to find the information needed to extract highlights and shadows. Thresholds for when a pixel is considered as highlight or shadow is set at 0.15 and 0.85 respectively, on a scale where 0 is where white pixels (no dot of ink placed) are found within the filter and 1 where all pixels are black (a drop of ink). These limits visually gives highlights and shadows. The parts of the halftoned image within the limits ($\leq 0.15$ and $\geq 0.85$ of a grayscale range from 0 to 1) are extracted for further processing.

A Canny edge filter is used to extract edges, where black pixels (a drop of ink) are found next to white pixel areas (no drop of ink), in the image [60]. This filter uses a low and a high threshold for hysteresis, and a sigma as the standard deviation for the Gaussian filter. Pixels placed close together will therefore be identified as forming an edge, and a worm can be thought of as an edge going through the image. Figure D.3 shows an example where pixels form a worm, and the Canny filter correctly detects this as an edge. In most cases a sigma of 1, resulting in a $4 \times 4$ pixel Gaussian filter, yields good results together with a heuristically chosen pair of thresholds depending on the data. Another advantage of using an edge filter is the fact that worms can be found in both highlights and shadow areas; in the highlights drops of ink make the worms, while in the shadow areas the ink-free areas make the worms. The edge filter does not differentiate between these, just the difference between dark and light areas resulting in a robust method, and no special care must be taken if it is in a highlight or shadow area.

In the cases of a single pixel without any connecting pixels the Canny filter will result in an edge around the pixel. A simple filter is applied to remove these single pixels, along with
other small pixels not connected to worms. This is done by checking the local neighbourhood around the pixel to identify whether it is connected to other pixels or not, if not connected the pixel is removed and not considered as a part of a worm.

In order to find the correct pixels that make up the worm the boundary of the object is traced, since this is a binary image a black and white boundary tracing is performed. The exterior of the object is traced along with the boundaries of the holes.

The basis of the worm structure is now found, and the next step of the worm detection is a proximity calculation. The pixels inside a worm are extracted and for each pixel the closest non-worm pixel is located, if this pixel is within a specified distance and angle it is considered as part of the worm. The angle is based on the angle from the closest detected worm pixel (Figure D.4). This process is iterative, so when a pixel is found this pixel is then further checked until all detected pixels are processed. The distance and angle can be set according to the desired sensitivity. Testing has shown that a distance of 50 pixels and angle of 40 degrees gives good results.

All worms have now been detected regardless of size. It is well known that our human visual system filters out information that we cannot process. This also applies to worm, and because of this worms that are too small for observers to notice should not be taken into account in the calculation [449]. Therefore the size of the worm is found by using dilation to increase each pixel. The worms will now ”grow” together, and the area of connected dilated pixels is computed. If the area is below a threshold, the pixels are removed from the detected worms.

The final value for the measure is calculated as the percentage of non-worm points divided by the number of pixels in the image. This will result in a value between 0 and 1, where 1 is a
worm free image. This normalization will result in the possibility to compare EDWM values across different images. Furthermore the map of worms, computed before the calculation of the final value, can be useful to detect areas where pixels can be rearranged to remove worm artifacts.

D.3 Evaluation of the error diffusion worm measure

Evaluation is an important aspect when developing a measure. An experiment was carried out to evaluate the performance of the proposed measure. Correlation between the perceived worms and predicted worms is adopted as the measure of performance. There are many ways to evaluate halftone quality, and our experiment is inspired by previous experiments carried out by Mitsa and Varkur [290], Tontelj et al. [444] and Axelson [15].

D.3.1 Experimental setup

Four different images were chosen for a psychophysical experiment as seen on Figure D.6. One artistic image (Figure D.6(b)), with both highlights and shadow areas. One gradient ramp (Figure D.6(a)), from white to black. Wang and Parker [454] stated that a gradient ramp is a basic and important evaluation in most halftone evaluations. A gradient will reveal worms, and is often used to demonstrate the performance of error diffusion algorithms [15, 243, 261, 336, 386, 400]. One highlight patch (Figure D.6(d), 2% black) and one shadow patch (Figure D.6(c), 98% black). The highlight and shadow patch are areas where we normally will have worms, and are therefore appropriate images for the evaluation error diffusion algorithms.

The images are 600 pixels wide and 200 pixels high, except the artistic image where the height is 450 pixels. These images were reproduced with five different error diffusion algorithms, Floyd and Steinberg (FS) [144], Stucki (S) [427], Fan (F) [129, 400], Jarvis, Judice and Ninke (JJN) [218] and Floyd and Steinberg Serpentine (FSS) [257, 446]. The
images were printed on a HP ColorJet 4600DN on HP Office Paper White (80 g/m²) at 600 DPI simulating 72 DPI.

The experiment was carried out in a dimmed room, with a GretagMacbeth Examolite overhead light source with D50 lighting as seen in Figure D.5. The observers were free to move, and no restrictions on viewing distance were given. 12 observers participated in the experiment, both males and females.

The experiment was divided into two parts. The first part was to rank the overall quality of the five versions of the image, while the second part was to rank the images according to the amount of worms found in the five versions. Before the second part the observers were taught how to identify worms, and where they occur. This was done to secure reliable data.

(a) Gradient. (b) Artistic image.
(c) Shadow image. (d) Highlight image.

Figure D.6: The four images used in the experiment, surrounded here by a black frame. Each was reproduced with five different error diffusion algorithms.

D.3.2 Results

This section contains results from the psychophysical experiment and the EDWM. Analysis of the results are also given. Two types of correlation will be computed for the results, the Pearson product-moment correlation coefficient and the Spearman’s rank correlation coefficient [232]. The first assumes that the variables are ordinal, and looks for the linear relationship between variables. The second, Spearman, is a non-parametric measure of correlation that uses the ranks as basis instead of the actual values. This describes the relationship between variables without making any assumptions about the frequency distribution of the variables.

D.3.2.1 Overall scores

The results from the 12 observers were combined and Z-scores were calculated using Thurstone’s law of comparative judgement [437], the scores were obtained using the Colour Engineering Toolbox [162]. As seen in Figure D.7 Floyd and Steinberg Serpentine (FSS) has the highest score both for quality and worms. For the worm artifact three algorithms cannot be differentiated by the observers, Stucki (S), Fan (F) and Jarvis, Judice and Ninke (JJN). The Floyd and Steinberg algorithm (FS) gets the lowest score, and clearly has more worms than the other error diffusion algorithms. In the quality judgement the FS also has the lowest mean Z-score, but cannot be differentiated from S and F. The JJN was rated better than FS, but cannot be differentiated from S and F.
There is also a high correlation between the quality score (Figure D.7(a)) and the worm score (Figure D.7(b)), the Pearson correlation gives a value of 0.94 (p value: 0.0197). This indicates a strong relationship between the amount of worms and the quality of the image. The images evaluated in the experiment were chosen based on areas where worms in error diffusion are present. Therefore the conclusion may not be valid in other images with less worms, i.e. without large highlights and shadow areas.

Figure D.7: Z-score for quality and worm judgement. From the figures we can see a similarity, indicating that worm artifacts are linked with the quality of error diffused images. The Pearson correlation gives a value of 0.94 (p value: 0.0197), supporting the similarities found in the figures. We can also notice that in the quality judgement the observers use less of the Z-score scale then in the worm judgement, this indicate that observers agree more on the worm judgement than in the quality judgement.

D.3.3 Gradient image

Figure D.8 and Table D.1 shows the results from the psychophysical experiment for the gradient ramp. For the quality judgement (Figure D.8(a)) JIN has the best quality, while FSS cannot be differentiated from FS. In the worm judgement (Figure D.8(b)), the result is not as clear. FSS can be differentiated from FS, but not from any other algorithms. This also indicates that observers more easily pick one "favorite" algorithm in the quality judgement (Figure D.8(a)), while in the worm judgement (Figure D.8(b)) the observers do not have a high consensus in ranking. Some observers also indicated that this scene was the hardest to rank, both when it came to quality and worms.

Figure D.9 shows a scatter plot of the EDWM and observer scores. FS is correctly given the lowest score by EDWM as the observers did, and FSS is correctly given the highest EDWM score. EDWM has ranked the S before F, the opposite of the observers, but the other error diffusion algorithms are ranked similarly to ranking of the observers. We also see that all points are within a 95% confidence interval. We get a high Pearson correlation, 0.93, indicating that EDWM can predict the amount of perceived worms for this scene.
DETECTION OF WORMS IN ERROR DIFFUSION HALFTONING

(a) $Z$-score for gradient ramp based on quality judgement. (b) $Z$-score for gradient ramp based on worm judgement.

**Figure D.8:** $Z$-score for quality and worm judgement for the gradient ramp. In the quality judgement observers agree more on the ranking than in worm judgement. This is a scene where most observers indicated to have problems with ranking according to worms.

**Table D.1:** $Z$-scores and EDWM scores for the gradient ramp. We can see that FSS has the highest EDWM, i.e. the least worms. FS has the lowest value, indicating the most worm artifacts.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>EDWM</th>
<th>$Z$-score worms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floyd and Steinberg ($FS$)</td>
<td>0.98965</td>
<td>-0.4975</td>
</tr>
<tr>
<td>Fan ($F$)</td>
<td>0.99133</td>
<td>0.0551</td>
</tr>
<tr>
<td>Stucki ($S$)</td>
<td>0.99166</td>
<td>-0.2154</td>
</tr>
<tr>
<td>Jarvis, Judice and Ninke ($JJN$)</td>
<td>0.99245</td>
<td>0.2763</td>
</tr>
<tr>
<td>Floyd and Steinberg Serpentine ($FSS$)</td>
<td>0.99318</td>
<td>0.3815</td>
</tr>
</tbody>
</table>
D.3.4 Artistic image

Figure D.10 shows the Z-scores for the artistic image, both for quality (Figure D.10(a)) and worm (Figure D.10(b)) judgement. For the quality judgement FSS cannot be differentiated from F and S by the observers, but FSS was clearly better than JIN and FS. For the worm judgement FS was judged to have more worms than the four other algorithms. These four algorithms gave almost the same amount of worms according to the observers.

Figure D.10: Z-score for quality and worm judgement for the artistic image.

Figure D.11 shows a scatter plot of the EDWM and observer score for the artistic image, while data can be found in Table D.2. We have a lower correlation than gradient ramp, but still the correlation of 0.86 \((p \text{ value}: 0.0596)\) indicates a strong relation between the predicted
Table D.2: Z-score and EDWM scores for the artistic image. S has been rated as the error diffusion algorithm with the lowest amount of worms (i.e. best rating by the observers). FSS is given the highest EDWM value, but this also has a Z-score close to S and cannot be differentiated with a 95% confidence interval.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>EDWM</th>
<th>Z-score worms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floyd and Steinberg (FS)</td>
<td>0.98398</td>
<td>-1.3830</td>
</tr>
<tr>
<td>Fan (F)</td>
<td>0.98768</td>
<td>0.2322</td>
</tr>
<tr>
<td>Stucki (S)</td>
<td>0.99389</td>
<td>0.6196</td>
</tr>
<tr>
<td>Jarvis, Judice and Ninke (JJN)</td>
<td>0.99187</td>
<td>0.0168</td>
</tr>
<tr>
<td>Floyd-Steinberg Serpentine (FSS)</td>
<td>0.99503</td>
<td>0.5144</td>
</tr>
</tbody>
</table>

Figure D.11: Correlation for artistic image between Z-score and EDWM, plotted with 95% Z-score confidence intervals.
Worms and perceived worms. Since the EDWM is a no-reference metric, it will not discriminate between worms made by edges in the image and worms in highlight/shadow areas. Because of this EDWM should be used with caution in natural images, because the worms in these images could be wanted in order to reproduce edges. The good correlation found here indicates that EDWM might be used on natural images.

**D.3.5 Highlight image**

Figure D.12 shows the results for the highlight image regarding quality (Figure D.12(a)) and worms (Figure D.12(b)), we can see that the ranking is almost similar for the quality and the worm judgement. *FFS* has the same amount of worms and quality as *JJN*, but it can be differentiated from the rest.

*Table D.3: Z-scores and EDWM scores for the highlight image. *FSS* is given the highest score by the observers, this image also receives the highest EDWM value. We can also see that the ranking of EDWM values and observer scores is the same (Figure D.12).*

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>EDWM</th>
<th>Z-score worms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floyd and Steinberg (<em>FS</em>)</td>
<td>0.98873</td>
<td>-1.5002</td>
</tr>
<tr>
<td>Fan (<em>F</em>)</td>
<td>0.99034</td>
<td>-0.7565</td>
</tr>
<tr>
<td>Stucki (<em>S</em>)</td>
<td>0.99611</td>
<td>-0.3113</td>
</tr>
<tr>
<td>Jarvis, Judice and Ninke (<em>JJN</em>)</td>
<td>0.99668</td>
<td>1.1630</td>
</tr>
<tr>
<td>Floyd and Steinberg Serpentine (<em>FSS</em>)</td>
<td>0.99854</td>
<td>1.4050</td>
</tr>
</tbody>
</table>

Figure D.13 shows a scatter plot of the EDWM and observer score for the highlight image. We can see that the ranking of the EDWM values and observer scores is the same (Table D.3), indicating that the EDWM is able to correctly rank the images according to perceived worms for this image. We also get a high correlation between the scores, indicating the high performance of EDWM.
D.3.6 Shadow image

Results for the shadow image are found in Figure D.14 and Table D.4. JJN has the lowest quality judgement (Figure D.14(a)) but cannot be differentiated from FS. For the worm judgement (Figure D.14(b)) FSS clearly gave less worms than the four other algorithms. Some observers indicated that this image was the easiest to judge compared to the three other images.

Table D.4: Z-scores and EDWM values for the shadow image. FSS is the best according to the observers and EDWM.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>EDWM</th>
<th>Z-score worms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floyd and Steinberg (FS)</td>
<td>0.99069</td>
<td>-1.1752</td>
</tr>
<tr>
<td>Fan (F)</td>
<td>0.99340</td>
<td>0.2116</td>
</tr>
<tr>
<td>Stucki (S)</td>
<td>0.99223</td>
<td>-0.5446</td>
</tr>
<tr>
<td>Jarvis, Judice and Ninke (JJN)</td>
<td>0.98933</td>
<td>-0.4714</td>
</tr>
<tr>
<td>Floyd and Steinberg Serpentine (FSS)</td>
<td>0.99854</td>
<td>1.9797</td>
</tr>
</tbody>
</table>

Figure D.15 shows a scatter plot of the EDWM values and observer score for the shadow image. A Pearson correlation of 0.92 ($p$ value: 0.0226) indicates a high performance by EDWM.

D.3.7 Comparison to other metrics

EDWM’s performance is compared against the performance of other state of the art metrics. S-CIELAB [499], Structural Content [120], Average Distance [120] and MSE have been cho-
(a) Z-score for shadow image based on quality (b) Z-score for shadow image based on worm judgement.

**Figure D.14:** Z-score for quality and worm judgement for the shadow image.

\[
\text{Shadow image correlation: 0.92 (p value: 0.0267)}
\]

**Figure D.15:** Correlation for shadow image between Z-score and EDWM, plotted with 95% Z-score confidence intervals.
DETECTION OF WORMS IN ERROR DIFFUSION HALFTONING

sen for comparison. S-CIELAB was made with the intention to predict perceived difference between an original and for example a halftoned version of the original. Because of this S-CIELAB is a suitable metric to compare the EDWM against. Structural Content compares changes in structure between two images, and a worms in the image will affect the structural content. This makes the Structural Content metric suitable for comparison against the EDWM. Average Distance included to have a comparison against traditional measures as well as new measures, such as the S-CIELAB, which simulate the HVS. The MSE is commonly used by researchers to test the performance of new metrics, because of its widespread use in image quality evaluation our results from the EDWM is compared against results from MSE.

All these metrics are full-reference metrics, i.e. they use information from the original image to calculate the difference or quality. The continuous tone image was used as a reference for all metrics. Since EDWM is a no-reference metric, this must be taken into account when analyzing the results, and also the fact that none of the other metrics are specifically made for worm artifacts.

EDWM outperforms S-CIELAB, Structural Content, Average Distance and MSE. This is clearly seen on Table D.5, where EDWM is the only measure to have a correlation above 0.85.

Table D.5: Correlation between metric score and Z-score from the worm judgement, with p value in parentheses. Gray cells indicate the best correlation. We can see that EDWM outperforms the other metrics in all images. MSE and Average Distance has a good correlation for the artistic image, but high p values (AD: 0.2985 and MSE: 0.1241) indicate that the correlation is by chance.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Artistic image</th>
<th>Gradient ramp</th>
<th>Highlight image</th>
<th>Shadow image</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDWM</td>
<td>0.86 (0.0596)</td>
<td>0.93 (0.0220)</td>
<td>0.91 (0.0311)</td>
<td>0.92 (0.0226)</td>
</tr>
<tr>
<td>S-CIELAB</td>
<td>-0.59 (0.2906)</td>
<td>0.07 (0.9143)</td>
<td>-0.48 (0.4106)</td>
<td>0.64 (0.2500)</td>
</tr>
<tr>
<td>Structural Content</td>
<td>0.08 (0.8948)</td>
<td>-0.09 (0.8836)</td>
<td>-0.23 (0.7149)</td>
<td>-0.15 (0.822)</td>
</tr>
<tr>
<td>Average Distance</td>
<td>-0.59 (0.2985)</td>
<td>-0.19 (0.7550)</td>
<td>-0.23 (0.7148)</td>
<td>-0.15 (0.8206)</td>
</tr>
<tr>
<td>MSE</td>
<td>-0.77 (0.1241)</td>
<td>-0.17 (0.7884)</td>
<td>-0.23 (0.7148)</td>
<td>0.15 (0.8206)</td>
</tr>
</tbody>
</table>

D.3.7.1 Overall performance

Since the results from EDWM are normalized we can compare results obtained for different images. Figure D.16 shows the correlation between Z-scores from the worm judgement for all images and EDWM values for all images (Tables D.1 - D.4). We can see a clear relationship between the Z-scores and EDWM values, this is also confirmed by a Pearson correlation of 0.81 (p value: 1e-005). Compared to the other metrics EDWM is significantly better (Table D.6). The other measures have a very low correlation, this is because they have big differences in the values between images, resulting in a correlation close to zero.[350] Our measure also has the highest Spearman correlation, indicating that the ranking from it corresponds to the ranking made by the observers.
Figure D.16: Correlation between all Z-score from worm judgement and all EDWM values, plotted with 95% confidence intervals. From the figure we can see a strong correlation (Pearson correlation of 0.81, p value: 1e-005). The calculated Spearman correlation is 0.77 (p value 0.0001), indicating a correct ranking of predicted worms according to the observer ranking.

Table D.6: Correlation between metric score and Z-scores from the worm judgement for all images, with p values in parentheses. We can see that EDWM outperforms the other metrics in all images with a correlation of 0.81 (p value: 1e-005), and the Spearman correlation of EDWM is also higher than for the other metrics. The other measures have a very low correlation due to big differences in the results between the images. Gray cells indicate the best correlation.
D.4 Summary

A measure for worms, EDWM, has been proposed. The measure is simple to compute, and does not require any information about the original image. It allows for comparison of results across different images, resulting in a robust measure. A psychophysical experiment was carried out to evaluate the performance of EDWM, and the results indicate a high performance. EDWM clearly outperforms other metrics.
E FROM CONTRAST TO IMAGE

DIFFERENCE: TWO NEW METRICS

Two new metrics are proposed referred as $S_{DOG}$-CIELAB and $S_{DOG}$-DEE. They are in line with the S-CIELAB approach (Section 4.2.2.1), but the spatial extension is based on the work initiated by Rizzi et al. [374] and the S-DEE metric (Section 4.2.2.2) by Simone et al.[412]. The improvements encompass a multi-level approach, the substitution of the original S-CIELAB spatial filtering with a DOG calculation, as introduced in Appendix B, and the use of the $\Delta E_E$ color-difference formula [329]. The general workflow of the metrics is as follows (Figure E.1):

- The original image and its reproduction are first converted into the CIELAB color space for $S_{DOG}$-CIELAB and into CIE XYZ for $S_{DOG}$-DEE.
- The images are subsampled to various levels afterwards. The undersampling is simple since the images are halved, and the antialiasing filtering avoids artifacts at low resolutions.
- A pixelwise neighborhood contrast calculation is executed in each level using the DOG on the lightness and on the chromatic channels separately, thus providing local contrast maps for each level and each channel.
- Local contrast errors are computed using $\Delta E_{ab}^*$ for $S_{DOG}$-CIELAB or $\Delta E_E$ for $S_{DOG}$-DEE.
- A weighted recombination of the local contrast maps is finally computed, resulting in global ID metrics.

These metrics grew out of research on contrast measures, therefore one could expect the highest correlations with images containing significant variations in contrast. The DOG model was chosen as a surrogate to the CSF filtering because it has revealed to be beneficial in the identification of edges, while the CSFs are mainly used to modulate the less perceptible frequencies. The $\Delta E_E$ formula was selected because it is statistically equivalent to CIEDE2000 in the prediction of many available empirical datasets, but with greater simplicity and clear relationships with visual processing.

Once that local contrast maps are generated for each level, how to reduce the concept of contrast from local values at each pixel location to a single number representing the global ID is an ongoing debate. The simplest strategy is taking the mean of each level and averaging all together.
These two new metrics perform a weighted recombination of the levels, given by the following equation:

\[
\text{GlobalID} = \frac{1}{N_l} \sum_{l=1}^{N_l} \lambda_l \cdot \bar{c}_l,
\]

where \(N_l\) is the number of levels, \(\bar{c}_l\) is the mean contrast in the level \(l\), and \(\lambda_l\) is the weight assigned to each level \(l\).

The metrics have several parameters for the central component \((R_c)\) and surround component \((R_s)\) of the DOG filter, the nature of the subsampling, and the type of weighting levels in the pyramid. In this work the subsampling is done using a pyramid where the pyramid is expressed by the series \(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots\). For a detailed overview of the parameters, we refer the reader to Simone et al. [413] and to Appendix B.

For more information about these two metrics we refer the reader to Ajagamelle et al. [5], Ajagamelle et al. [6], or Ajagamelle [4].

Figure E.1: Workflow of the proposed metrics. The metrics are similar to the S-CIELAB (Section 4.2.2.1) described by Zhang and Wandell [499]. The improvements encompass a multi-level approach, the substitution of the S-CIELAB spatial filtering with a DOG calculation, and the use of the \(\Delta E_E\) color-difference formula.
F Specifications

In this appendix we reproduce the specification sheets of the devices and media used in this study.

F.1 Monitors

Two different monitors have been used (Figures F.1 and F.2), and their specifications are found in Tables F.1 and F.2.

F.1.1 Eizo CG241W

Figure F.1: Eizo CG241W. Reproduced from www.puremac.de.
Table F.1: Eizo CG241W specifications.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel Size</td>
<td>24.1” / 61 cm (611 mm diagonal)</td>
</tr>
<tr>
<td>Active Display Size (H, V)</td>
<td>518.4 x 324 mm</td>
</tr>
<tr>
<td>Panel Type</td>
<td>VA (with overdrive circuit)</td>
</tr>
<tr>
<td>Viewing Angles (H, V)</td>
<td>178°, 178° (at contrast ratio of 10:1)</td>
</tr>
<tr>
<td>Brightness</td>
<td>300 cd/m² (maximum), 120 cd/m² or less (recommended)</td>
</tr>
<tr>
<td>Contrast</td>
<td>850:1</td>
</tr>
<tr>
<td>Native Resolution</td>
<td>1920 x 1200 (16:10 aspect ratio)</td>
</tr>
<tr>
<td>Pixel Pitch</td>
<td>0.270 x 0.270 mm</td>
</tr>
<tr>
<td>Display Colors</td>
<td>16.77 million from a palette of 68 billion</td>
</tr>
<tr>
<td>Wide Gamut Coverage</td>
<td>Adobe RGB: 95%, NTSC:91%, sRGB/Rec.709: 98%, EBU 98%, SMPTE-C:100%, DCI:87%</td>
</tr>
<tr>
<td>Look-Up Table</td>
<td>16 bits per color</td>
</tr>
</tbody>
</table>
F.1.2 Eizo CG211

Table F.2: Eizo CG211 Specifications

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel Size</td>
<td>22.2” / 56.4 cm (563 mm diagonal)</td>
</tr>
<tr>
<td>Active Display Size (H, V)</td>
<td>478 x 299 mm</td>
</tr>
<tr>
<td>Panel Type</td>
<td>IPS</td>
</tr>
<tr>
<td>Viewing Angles (H, V)</td>
<td>170°, 170° (at contrast ratio of 10:1)</td>
</tr>
<tr>
<td>Brightness</td>
<td>200 cd/m² (maximum), 100 cd/m² or less (recommended)</td>
</tr>
<tr>
<td>Contrast</td>
<td>400:1</td>
</tr>
<tr>
<td>Native Resolution</td>
<td>1920 x 1200 (16:10 aspect ratio)</td>
</tr>
<tr>
<td>Pixel Pitch</td>
<td>0.249 x 0.249 mm</td>
</tr>
<tr>
<td>Display Colors</td>
<td>16.77 million from a palette of 278 trillionn</td>
</tr>
<tr>
<td>Wide Gamut Coverage</td>
<td>98% of sRGB and Adobe RGB</td>
</tr>
<tr>
<td>Look-Up Table</td>
<td>16 bits per color</td>
</tr>
</tbody>
</table>

Figure F.2: Eizo CG211. Reproduced from www.eizo.pl.
F.2 Scanners

Different scanners (Figures F.5, F.3, and F.4) have been used to digitize the printed images, and their specifications are found in Tables F.3, F.4, and F.5.

F.2.1 Microtek ScanMaker 9800XL

Figure F.3: Microtek ScanMaker 9800XL. Reproduced from www.burak.pl.

Table F.3: Microtek ScanMaker 9800XL specifications.

<table>
<thead>
<tr>
<th>Type</th>
<th>Flatbed color image scanner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Sensor</td>
<td>CCD</td>
</tr>
<tr>
<td>Optical Resolution</td>
<td>1600 dpi x 3200 dpi</td>
</tr>
<tr>
<td>Hardware Resolution</td>
<td>2400 x 4800 dpi with Micro Step Drive technology</td>
</tr>
<tr>
<td>Color Bit Depth</td>
<td>48-bit internal</td>
</tr>
<tr>
<td>Optical Density</td>
<td>3.7 Dmax</td>
</tr>
<tr>
<td>Maximum Scan Area</td>
<td>12” X 16.9”</td>
</tr>
<tr>
<td>Light Source</td>
<td>Cold Cathode Fluorescent Lamp (CCFL)</td>
</tr>
</tbody>
</table>
F.2.2 Epson Expression 10000XL

Figure F.4: Epson Expression 10000XL. Reproduced from http://www.scanstation.co.uk.

Table F.4: Epson Expression 10000XL specifications

<table>
<thead>
<tr>
<th>Type</th>
<th>Flatbed color image scanner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Sensor</td>
<td>Color Epson MatrixCCD line sensor</td>
</tr>
<tr>
<td>Optical Resolution</td>
<td>2400 dpi</td>
</tr>
<tr>
<td>Hardware Resolution</td>
<td>2400 x 4800 dpi with Micro Step Drive technology</td>
</tr>
<tr>
<td>Maximum Resolution</td>
<td>12,800 x 12,800 dpi</td>
</tr>
<tr>
<td>Effective Pixels</td>
<td>87,840 pixels / line (2400 dpi)</td>
</tr>
<tr>
<td>Color Bit Depth</td>
<td>48-bit internal / external</td>
</tr>
<tr>
<td>Grayscale Bit Depth</td>
<td>16-bit internal / external</td>
</tr>
<tr>
<td>Optical Density</td>
<td>3.8 Dmax</td>
</tr>
<tr>
<td>Maximum Scan Area</td>
<td>12.2” x 17.2”</td>
</tr>
<tr>
<td>Light Source</td>
<td>Xenon gas cold cathode fluorescent lamp</td>
</tr>
<tr>
<td>Scanning Speed</td>
<td>2400 dpi in draft mode, Color 16.0 msec / line, Grayscale 5.3 msec / line, Line Art 5.3 msec / line</td>
</tr>
<tr>
<td>Focus Control</td>
<td>AutoFocus optics system (CCD and lens unit)</td>
</tr>
</tbody>
</table>
F.2.3 HP ScanJet G4050

Figure F.5: HP ScanJet G4050. Reproduced from http://www.officecritter.com.

<table>
<thead>
<tr>
<th>Type</th>
<th>Flatbed color image scanner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Sensor</td>
<td>CCD</td>
</tr>
<tr>
<td>Optical Resolution</td>
<td>4800 dpi</td>
</tr>
<tr>
<td>Hardware Resolution</td>
<td>4800 x 9600 dpi</td>
</tr>
<tr>
<td>Bit Depth</td>
<td>96 bit</td>
</tr>
<tr>
<td>Grayscale levels</td>
<td>256</td>
</tr>
<tr>
<td>Maximum Scan Area</td>
<td>21.6 x 31.1 cm</td>
</tr>
<tr>
<td>Light Source</td>
<td>Xenon gas cold cathode fluorescent lamp</td>
</tr>
</tbody>
</table>
F.3 Printers

The specifications of the printers used in this work are found below (Figures F.6, F.7, F.8 and Tables F.6, F.7, F.8).

F.3.1 Océ Colorwave 600

*Figure F.6: Océ ColorWave 600. Reproduced from www.oce.com.*

<table>
<thead>
<tr>
<th>Description</th>
<th>Wide Format TonerPearl Printer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>Océ CrystalPoint</td>
</tr>
<tr>
<td>Speed</td>
<td>B&amp;W: 31 sec/A0; Color: 34 sec/A0</td>
</tr>
<tr>
<td>Warm-up Time</td>
<td>6 minutes (from cold start-up)</td>
</tr>
<tr>
<td>Print Resolution</td>
<td>Addressable raster: 1.200 dpi</td>
</tr>
<tr>
<td>Paper Capacity</td>
<td>Up to 200 m per roll, max 6 rolls</td>
</tr>
<tr>
<td>Output Sizes Width</td>
<td>279 mm - 1067 mm</td>
</tr>
<tr>
<td>Output Sizes Length</td>
<td>210 mm - 3000 mm (longer possible, might influence side margins)</td>
</tr>
<tr>
<td>Paper Weight</td>
<td>60 g/m2 - 160 g/m2</td>
</tr>
<tr>
<td>Media Type</td>
<td>Uncoated, recycled, colored, blue back, tyvek, polyester films and more</td>
</tr>
</tbody>
</table>

*Table F.6: Océ Colorwave specifications.*
F.3.2 HP Designjet 10ps

![HP DesignJet 10ps](www.tiendatinta.com)

Figure F.7: HP DesignJet 10ps. Reproduced from www.tiendatinta.com.

<table>
<thead>
<tr>
<th>Description</th>
<th>Six-color printer (CMYKcm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>Inkjet</td>
</tr>
<tr>
<td>Speed</td>
<td>up to 4.0 ppm - Color draft, up to 10.0 ppm - Color best, up to 1.4 ppm - Color normal</td>
</tr>
<tr>
<td>Print Resolution</td>
<td>2400 x 1200 dpi</td>
</tr>
<tr>
<td>Paper Capacity</td>
<td>150.0 sheets</td>
</tr>
<tr>
<td>Max Media Size</td>
<td>13.0 in x 19.0 in</td>
</tr>
<tr>
<td>Paper Weight</td>
<td>60 g/m2 - 160 g/m2</td>
</tr>
<tr>
<td>Media Type</td>
<td>Heavy-weight coated paper, Glossy photo paper, Opaque vinyl, High-gloss photo paper, Semi-gloss photo paper, Imaging film, Plain paper, Glossy paper, Coated paper, Glossy coated paper</td>
</tr>
</tbody>
</table>
F.3.3 HP Color LaserJet 4600dn

![HP Color LaserJet 4600dn](www.hp.com)

**Table F.8: HP Color LaserJet 4600dn specifications.**

<table>
<thead>
<tr>
<th>Description</th>
<th>Color laser printer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>Laser</td>
</tr>
<tr>
<td>Speed</td>
<td>17 ppm black and color</td>
</tr>
<tr>
<td>Print Resolution</td>
<td>$600 \times 600$ dpi</td>
</tr>
<tr>
<td>Media type</td>
<td>Media type Paper (plain, glossy, film, recycled), envelopes, transparencies, labels, cardstock</td>
</tr>
</tbody>
</table>
F.4 Paper

The specification of the different papers used is found in Tables F.9, F.10, F.11, and F.12.

F.4.1 Océ LFM 050 Red Label

Table F.9: Océ LFM 050 Red Label specifications.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>75 g/m²</td>
</tr>
<tr>
<td>Whiteness</td>
<td>CIE 159</td>
</tr>
<tr>
<td>Thickness</td>
<td>99 μm</td>
</tr>
<tr>
<td>Acidity</td>
<td>7.5 pH</td>
</tr>
<tr>
<td>ISO Brightness</td>
<td>R457 + UV 108%</td>
</tr>
<tr>
<td>ISO Brightness</td>
<td>R457 - UV 88%</td>
</tr>
<tr>
<td>Opacity</td>
<td>92%</td>
</tr>
<tr>
<td>Coating</td>
<td>Not coated</td>
</tr>
</tbody>
</table>

F.4.2 Océ LFM 054 Red Label

Table F.10: Océ LFM 054 Red Label specifications.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>75 g/m²</td>
</tr>
<tr>
<td>Whiteness</td>
<td>CIE 159</td>
</tr>
<tr>
<td>Thickness</td>
<td>103 μm</td>
</tr>
<tr>
<td>Acidity</td>
<td>7.5 pH</td>
</tr>
<tr>
<td>ISO Brightness</td>
<td>R457 + UV 109%</td>
</tr>
<tr>
<td>ISO Brightness</td>
<td>R457 - UV 88%</td>
</tr>
<tr>
<td>Opacity</td>
<td>92%</td>
</tr>
<tr>
<td>Coating</td>
<td>Not coated</td>
</tr>
</tbody>
</table>
F.4.3 Stora Enso MultiCopy Original

*Table F.11: Stora Enso MultiCopy Original specifications.*

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>80 g/m²</td>
</tr>
<tr>
<td>Whiteness</td>
<td>CIE 160</td>
</tr>
<tr>
<td>Thickness</td>
<td>108 μm</td>
</tr>
<tr>
<td>Opacity</td>
<td>91.5%</td>
</tr>
<tr>
<td>Coating</td>
<td>Not coated</td>
</tr>
</tbody>
</table>

F.4.4 HP office paper

*Table F.12: HP office paper specifications.*

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>80 g/m²</td>
</tr>
<tr>
<td>Size</td>
<td>A4</td>
</tr>
</tbody>
</table>
F.5 Viewing cabinet

F.5.1 Verivide DTP

The Verivide DTP viewing cabinet has been used in this work (Figure F.9 and Table F.13)

![Figure F.9: Verivide DTP. Reproduced from www.verivide.com](image)

Table F.13: Verivide DTP specifications.

<table>
<thead>
<tr>
<th>Light source</th>
<th>D50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal color</td>
<td>Munsell grey N5</td>
</tr>
<tr>
<td>Viewing cavity</td>
<td>670×460 mm</td>
</tr>
<tr>
<td>Overall size</td>
<td>695×610×390 mm</td>
</tr>
</tbody>
</table>
The author hereof has been enabled by Océ-Technologies B.V. to perform research activities which underlies this document. This document has been written in a personal capacity. Océ-Technologies B.V. disclaims any liability for the correctness of the data, considerations and conclusions contained in this document.