Fusion of Digital and Visual Print Quality – Final Report of DigiQ Project

*Halonen, R., *Nuutinen, M., *Oittinen, P., **Leisti, T.,
**Nyman, G., ***Mettänen, M., ***Ritala, R., ****Eerola, T., ****Lensu, L.,
****Kämäräinen, J-K., ****Kälviäinen, H.

*Department of Media Technology, Aalto University School of Science and Technology, http://media.tkk.fi/en/
**Department of Psychology, University of Helsinki, http://www.helsinki.fi/psykologia/english/
***Department of Automation Science and Engineering, Tampere University of Technology, http://ae.tut.fi/

E-mail: Raisa.Halonen@tkk.fi, Mikko.Nuutinen@tkk.fi, Pirkko.Oittinen@tkk.fi, tuomas.leisti@helsinki.fi, gote.nyman@helsinki.fi, marja.mettanen@tut.fi, risto.ritala@tut.fi, Tuomas.Eerola@lut.fi, lasse.lensu@lut.fi, joni.kamarainen@lut.fi, heikki.kalviainen@lut.fi
TABLE OF CONTENTS

SUMMARY........................................................................................................................................ iv

1 INTRODUCTION........................................................................................................................... 1
   1.1 DigiQ Project ............................................................................................................................ 1
   1.2 Subjective Visual Quality in the Project Context ............................................................... 2
   1.3 Quality Related Concepts .................................................................................................... 2
   1.4 Developing the Visual Quality Index Model ....................................................................... 3
   1.5 Image Quality Modelling .................................................................................................... 4
   1.6 Modelling Human Perception with Bayesian Network ..................................................... 5

2 VISUAL QUALITY EVALUATION ............................................................................................ 7
   2.1 Subjective Visual Quality .................................................................................................... 7
       2.1.1 Project Considerations ............................................................................................... 7
       2.1.2 Concept Cases ............................................................................................................. 7
       2.1.3 Process of Subjective Visual Quality Evaluation ....................................................... 11
   2.2 Subjective Image Quality Evaluation Methodology .......................................................... 13
   2.3 Concepts from Case Studies ............................................................................................... 15
       2.3.1 High- and Low-Level Attributes ................................................................................. 15
       2.3.2 Subjective Dimensions of High and Low Quality ....................................................... 18
       2.3.3 Subjective Quality Dimensions for Different Printing Methods ............................... 18
   2.4 Consistency of a Jury .......................................................................................................... 20
   2.5 Variation in Jury Evaluations ............................................................................................. 22
   2.6 Properties of Bayesian Network Models ............................................................................. 24

3 INSTRUMENTAL CHARACTERIZATION OF DIGITAL PRINTING ........................................ 26
   3.1 Relevant Physical Characteristics of Paper ......................................................................... 26
   3.2 Test Target Measurements ................................................................................................. 27
       3.2.1 Standard Measurements .............................................................................................. 27
       3.2.2 Printing Dot Properties ............................................................................................... 27
       3.2.3 Line and Edge Features .............................................................................................. 27
       3.2.4 Colour Features .......................................................................................................... 28
       3.2.5 Mottling ....................................................................................................................... 28

4 COMPUTATIONAL CHARACTERIZATION OF DIGITALLY PRINTED IMAGES .................. 29
   4.1 Classification of Image Quality Metrics ............................................................................. 29
   4.2 Full-reference Image Quality Metrics .............................................................................. 30
       4.2.1 Introduction ................................................................................................................... 30
       4.2.2 Arbitrary Signal Fidelity Criteria ............................................................................... 31
       4.2.3 HVS Error-Based Methods ........................................................................................ 31
       4.2.4 Image Quality Metrics Using Natural Scene Statistics ............................................. 32
       4.2.5 Computing FR Metrics from Printed Images ............................................................. 32
   4.3 No-Reference Image Quality Metrics .............................................................................. 34
       4.3.1 Introduction ................................................................................................................... 34
       4.3.2 Sharpness ..................................................................................................................... 34
       4.3.3 Noise ............................................................................................................................. 36
       4.3.4 Contrast ......................................................................................................................... 38
       4.3.5 Colourfulness .............................................................................................................. 39
4.4 Reduced-Reference Quality Metrics ................................................................. 40
  4.4.1 Introduction ................................................................................................... 40
  4.4.2 Dominant Colour Contrast ........................................................................... 41
  4.4.3 Graininess ...................................................................................................... 42
  4.4.4 Sharpness ........................................................................................................ 44

5 VISUAL QUALITY MODELS – FROM OBJECTIVE MEASUREMENTS TO PREDICTING
HUMAN ASSESSMENT ................................................................................................. 47

  5.1 Full-Reference Metrics ...................................................................................... 47
    5.1.1 Introduction .................................................................................................. 47
    5.1.2 Methodology ............................................................................................... 47
    5.1.3 Results .......................................................................................................... 48
    5.1.4 Conclusions .................................................................................................. 49
    5.1.5 Further Analysis ......................................................................................... 49
  5.2 Regression Approach ....................................................................................... 50
    5.2.1 Introduction .................................................................................................. 50
    5.2.2 Methodology ............................................................................................... 51
    5.2.3 Results .......................................................................................................... 53
    5.2.4 Conclusions .................................................................................................. 56
  5.3 Content Adaptive Approach ........................................................................... 57
    5.3.1 Introduction .................................................................................................. 57
    5.3.2 Methodology ............................................................................................... 57
    5.3.3 Results .......................................................................................................... 60
    5.3.4 Conclusions .................................................................................................. 61
  5.4 Cumulative Match Score Approach ............................................................... 62
    5.4.1 Introduction .................................................................................................. 62
    5.4.2 Methodology ............................................................................................... 62
    5.4.3 Results .......................................................................................................... 63
    5.4.4 Conclusions .................................................................................................. 66
  5.5 Bayesian Network Approach .......................................................................... 67
    5.5.1 Introduction .................................................................................................. 67
    5.5.2 The Layered Structure Approach .............................................................. 67
    5.5.3 The Structure Optimisation Approach ....................................................... 73

6 DEVELOPING TEST IMAGES FOR PRINT QUALITY EVALUATION .................. 78

  6.1 Introduction ....................................................................................................... 78
  6.2 General Principles ............................................................................................ 80
  6.3 Gaze-Driven Approach for Test Image Development ..................................... 81
    6.3.1 Quality Assessment from Prints ................................................................. 81
    6.3.2 Quality Assessment from a Display ........................................................... 82
  6.4 Test Images Developed in the Project ............................................................. 83
    6.4.1 Test Image I .................................................................................................. 83
    6.4.2 Test Image II .................................................................................................. 83
    6.4.3 Test Image III .................................................................................................. 86
  6.5 Goodness of Test Images Developed .............................................................. 87
    6.5.1 Colour World of the Test Images ................................................................. 87
    6.5.2 Perceived Usefulness of Object Categories and Image Regions ................. 88
    6.5.3 Subjective Appearance of the Test Images ............................................... 91
    6.5.4 Conclusions .................................................................................................. 92

7 VISUAL QUALITY INDEX FOR DIGITALLY PRINTED IMAGES – DREAMS AND
REALITY ....................................................................................................................... 93
SUMMARY

The report is a summary of a three year (2006 – 2009) research project carried out in collaboration of four research partners and five industrial partners. The research partners had partly distinct, partly supporting research competences. The industrial partners had an interest in print quality from different standpoints such as papermaking, paper research and system development.

The research was motivated by the emergence of digital printing, both electrophotography and ink jet, as challengers of offset lithography in printing of image rich products. Examples of product types include advertising brochures, magazines, photo books and photographs. Some of these are novel as print products and offer new business opportunities for digital printing companies and papermakers.

The goals of the study in the project plan were expressed as follows:

- to develop a general framework for communication about printed image quality,
- to develop measurement and computing procedures for a visual quality index (VQI) by fusing visual and computational approaches and to create understanding of the validity and reliability of quality data for decision making and
- to find the sensitivity of the quality metrics to variation of paper structure in digital prints.

Communication about quality takes place between papermakers and professional printers on one hand and between papermakers and printing devices manufacturers on the other. The quality potential of paper is also of increasing interest to consumers who choose paper for their own printing jobs. Currently communication about paper quality to consumers uses descriptive terms which seek to give rise to positive associations without actually specifying quality. An example is “photo-quality”. The terminology used in professional settings is more rigorous, but often linked to descriptions of the paper structure and manufacturing (such as “matt-coated paper”) rather than the quality potential. Against this background the purpose was to develop a quality framework which categorizes quality concepts, their attributes and relations in a comprehensive manner.

The framework obtained as the result of the experimentation in the project is shown as Figure 2. It consists of five layers of quality concepts. The highest level is defined by the subjectively assessed overall printed image quality. The layer below defines the relevant visual attributes. The third layer defines computational measurements from printed images. Instrumental measurements on the second layer relate to basic quality factors made from test targets. The bottom layer includes specification of paper properties and printing device characteristics including the ink or toner. Overall quality is the highest level quality concept. Quality models link high level concepts to lower level concepts. High level visual concepts identified experimentally in the project include for instance realism, genuineness, naturalness, clarity and depth. Ideally, models can be found which enable overall printed image quality to be predicted from low level objective measurements. The latter include sharpness, noise, mottle, contrast, and colour gamut.

As for the second objective, the visual quality index of paper is defined as a numerical measure (on a category or interval scale with or without confidence intervals) which predicts the quality potential of paper in a given printing context in a visually verified way. It is well-known that several quality attributes influence overall quality assessments. The building blocks of VQI are models (“rules”) for combining quality attributes to a VQI and procedures for measuring the attributes or algorithms for computing them. Models are needed because determination of the visual quality by subjective tests using a sufficiently large group of people is too time consuming. Bayesian
Bayesian models are network models consisting of nodes (quality attributes) and edges (probabilistic relations) between them. Both layered models and structure optimised models were tested. The Bayesian approach proved to be highly promising, but requires a considerable amount of data for teaching the network. Once the network has been established it can be used for simulation and prediction of overall subjective visual quality and its distribution. Data collection could be made in the frame of large scale research collaboration. The models would form a solid basis for computing the visual quality index. Regression models tend not to be robust due to significant correlations between print properties, but have applicability for visual quality predictions in some well defined company context.

The issue of the reliability of subjective quality assessments was addressed in the project. According to the data, the confidence intervals of subjective visual quality assessments of 30 observers on the scale of 1 – 5 was about 0.25 for the ink jet prints and 0.33 for electrophotographic prints. For the same number of observers in two observer groups, the linear correlation coefficient between the assessments was .99. Statistical tools for analyzing the consistency of subjective quality assessments were also developed. Besides overall visual quality, the interpretation-based quality approach was employed to find the quality attributes which the naive observers used in their evaluation. Most commonly mentioned attributes include sharpness, graininess, brightness, naturalness, colourfulness, contrast and gloss.

Computation of quality attributes from printed images and test targets which were digitized after printing was one area of focus. The approaches included the no-reference approach which means that quality attributes such as sharpness, noise, contrast and colourfulness are computed without comparing to some reference image. Another approach, called reduced-reference computation, divides the image in spatial scales according to image information and computes quality attributes selectively. This is to take into consideration the visibility of artefacts and quality excellence depending on image information such as the conspicuousness of noise in flat but not in detail rich areas. The full-reference approach which requires spatial registration of digitized printed images with the corresponding digital originals was also used and a method developed for registering. The challenges of registration arise from changes in local dimensions of paper in printing due to process and toner/ink interactions.

Research in the project was structured around sets of printed samples and their characterization data by objective and subjective methods. The primary source of variation in the prints originated from the use of different paper grades. This is to say that the objective attributes and quality models strive to give expression to the variations arising from the variations in paper properties. As is well-known, printed ink-jet quality is far more sensitive to differences in paper properties than electrophotographic quality is. Computation and subjective visual quality data proved to be highly dependent on image content. To manage this, on one hand several images were used in the tests and on the other hand a reference test image was developed by a systematic process. The ease of assessing the key quality attributes was the primary criterion of development. Others included realism of the image content and pleasantness of viewing.

Several sample sets, called case studies, were produced in the project, characterized by objective and subjective methods and shared between the partners. The electrophotographic and ink jet
devices used in printing represented state-of-the-art technologies. Printing was colour managed by applying paper-specific colour profiles. The layouts in the printing tests included both simple test targets and natural images. For digitization of the samples a camera based procedure utilizing several exposures and a scanner-based procedure were used.

In conclusion, the project contributed to fundamental knowledge about the assessment of printed image quality by naive observers and about the most important attributes when quality variations arise from the paper. Modelling of overall subjective visual quality with objectively determined quality attributes as variables proved to be feasible using different approaches. The Bayesian network approach was the most comprehensive showing a lot of promise, but requiring a lot of data for establishing the model. Computation of a visual quality index for specifying the quality potential of paper in a given context is one of the applications of the models.

The primary contextual factor which influences subjective visual quality assessment is image content. Approaches to manage the influence of image content included the reduced-reference and full-reference approaches and, representing a fully different way of reasoning, development of one single test image to be a representative of the quality content space.

The final report is divided in 8 chapters. The structure reflects the reasoning that the subjective visual assessments constitute the core data and the foundation for quality measurement, computation and modelling. Chapter 1 defines the key concepts of the project. A more comprehensive list is provided as Appendix 3. Chapter 2 continues this by presenting the methods used in the subjective visual tests and Chapters 3 and 4 the instrumental and computational characterization methods. Chapter 5 is devoted to modelling of quality and the results obtained with the different methods. The procedure for developing test images for use in the project and more generally is documented in Chapter 6. Discussion of the results is presented from the standpoints of the different partners in Chapter 7 and overall conclusions in Chapter 8. A list of the publications produced by the research consortium is given in Appendix 2.
1 INTRODUCTION

1.1 DigiQ Project

In development of visual technologies such as printing, quality is a key criterion. In all printing technologies, the properties of paper exert an influence on the outcome which makes paper a key component of the technologies. The printed result is obtained as a result of interactive mechanisms of the device and the paper in addition to being dependent on device and paper parameters separately. Interactions may be physical, electrical or thermal. They have proven to be so complex that a long tradition of print quality research has emerged.

Printed quality has several dimensions. When divided in two groups, these are optical quality (including colour quality) and mechanical quality. This study is focused on the former. Print quality parameters can be measured from test targets with specific devices such as densitometers, glossmeters and spectrophotometers. These, called instrumental measurements, differ in terms of the geometrical, optical and spectral parameters of measurement. Another approach of objective measurements involves computation of quality parameters from prints (digitally or mechanically produced) which have been digitized after printing. Here we are especially focusing on computation of quality attributes from digitized printed natural images (photographs) by electrophotographic and ink jet methods.

During the last few years, research emphasis on printed image quality has shifted towards subjective visual quality. It is maintained that the end user is the final link in the quality chains and end users’ assessments of quality are of high importance. In prints, the quality of images has a central role because with text, an adequate level of quality has been achieved in all printing technologies. To find functioning quality measures for printed images and models which enable prediction of the multidimensional nature of visual assessment of quality are the primary challenges. Measures and models have application in communication about quality and its numerical specification.

Assessment of image quality is highly dependent on image content. Computational approaches to handle the role of image content are largely lacking. Even more challenging are the wider issues of the influence of image context. These include the relationship between the observer and the image, and also the use of the printed images whether for short term documentation, a family photo album or professional printing to name some examples.

The overall goal of the project was to come up with models which enable prediction of subjective visual quality attributes and image quality score from objectively determined model parameters from printed images. The goal was divided in three objectives. The first concerns the concepts of printed image quality for communication about quality. The second is related to quality models and the concept of visual quality index (VQI) for specification of the quality potential of paper in given contexts of use such as retail sales to consumers. The third objective seeks answers to the question of the sensitivity of the quality models and attributes to variation in paper properties. These issues were addressed through structured experimental activities which were a joint effort of the partners in the project.
This report is a summary of the key ideas, experimental set-ups, test image development, results and discussion. The text is partly based on material already published /listed in Appendix 2/, in part it presents unpublished data. It is divided first of all in three chapters which deal with methods – Chapter 2 subjective visual quality evaluation and Chapters 3 and 4 instrumental and computational quality measurements. Chapter 5 discusses the modelling approaches in the project and Chapter 6 the development of test images. Chapter 7 is devoted to discussion of the results taking a critical stance and presenting different viewpoints. Chapter 8 lists the conclusions shortly. To aid the reader, the abbreviations used are listed in Appendix 1 and a glossary has been assembled and presented in Appendix 3.

1.2 Subjective Visual Quality in the Project Context

In this particular project, the starting point for quality was the subjective experience of visual quality - how the users of print products really experience the quality differences between prints. Objective measures for quality were thus sought and evaluated based on data given by studying aspects of subjective quality. Often the approach has been opposite: the designs for objective measurements have been based on purely theoretical models that may have included knowledge of low-level vision, for instance, and subjective quality has only been used as a reference, against which the accuracy of the objective measurements has been evaluated.

Even though much emphasis was given to the development of different kinds of objective methods for print quality measurement in this project, the experience of quality was also thoroughly studied, as the theoretical conceptions of quality often have little resemblance of the subjective experience of quality in the print context. Most importantly, the users’ experience of quality is the factor that influences their choices as a consumer, and therefore it should be the most relevant target for prediction. It must be noted, however, that this project concentrated on measuring the image quality of the prints, not the end user experience of a final print product, which is a much more difficult target for prediction and dependent on different external factors, such as style, genre and the user segment of the print product. These aspects cannot be measured objectively and are therefore out of the scope of this project.

1.3 Quality Related Concepts

Image quality (IQ) is typically thought to comprise of components, the combination of which then defines the overall quality. What the components are is dependent on the definition of quality. The effect of one component on the overall quality is dependent on the other components, which makes image quality a complex, multidimensional phenomenon. If models for overall quality prediction are developed, it requires knowledge of all relevant aspects of quality in order to function properly. Keelan /39/ calls this multidimensional formalism.

In case of subjective quality, the components of image quality are called subjective attributes, or informally ‘nesses’, because attribute dimensions often end with ‘ness’, e.g. graininess, sharpness, naturalness. The approach of this project was largely based on examination of these attributes, meaning collecting all the relevant concepts of subjective experience of quality. Therefore each stage of the project was called a concept case, since it consisted of creation of the test samples, objective measurements and the collection of the quality ratings and subjective concepts for describing the quality.

Several contextual factors influence the evaluation of visual quality. Image content is the most well-known factor and it can affect the visibility of imaging artefact already at low level vision. Masking,
for example, has a very significant effect on the perception on noise /39/. Noise is easily detected if there is e.g. clear sky in the image, but detection of noise in foliage is far more difficult.

The psychological IQ evaluation process described in Section 2.1.3 may also be affected by image content. Some targets often attract more attention than others and the IQ attributes visible in those targets may have more weight on the overall image rating. Content can also influence the interpretation of artefacts that are seen in the image /57/. Unsharpness may give an artistic impression in certain image contents, and it is therefore not interpreted as weakness. In print context, the fluorescence of the paper may cause it look sunny or faded, depending on the interpretation.

Some properties of print quality, such as the gloss, function also as a part of product design. In that case, they become part of the style of the print product, which differentiates the print quality context from the mere informational image quality. In this sense, subjective visual print quality is an essential part of the reader experience of the print product. Stylistic aspect, however, acted a minimal role in this project, since only natural images were evaluated. This applied to the use aspect also: The observers were asked to evaluate general quality only, not to think about any use scenarios.

1.4 Developing the Visual Quality Index Model

Measurements, instrumental or subjective assessments, provide information about the state of the reality, i.e. the target of the measurement which in this case is the set of objective quality elements in an image. In order to interpret the data provided by the measurements, the measurement itself must be modelled. As the measurement information is always incomplete about the state of reality, measurements are described as conditional distributions: what is the probability distribution of obtaining a measurement result given the target of the measurement (state of reality). The measurement description may be supplemented with some pre-existing information about the target of the measurement. When such a measurement description is available, the instrumental measurement data obtained can be interpreted – based on the well-known rules of probability – as information about the state of the reality and as prediction about what the jury assessment would be. Correspondingly, based on jury assessment results, the result of instrumental measurements can be predicted. As the interpretation is in terms of probability distributions, not only the values obtained in jury assessment, respectively instrumental measurements are obtained, but also, the uncertainty of such predictions. Furthermore, the measurement model specifies the sensitivity of such predictions to changes in measurement results: we may compute how much does the probability distribution of predicted jury assessment change when a small change is made in the instrumental data. Global sensitivity is measured with the mutual information between instrumental and jury assessment data.
Thus defining the VQI on the basis of instrumental measurements amount to identifying which combination of instrumental measurements is most informative about the overall quality assessment by jury. There are many ways of selecting the framework in which the measurement model for the system /see Figure 1/ is identified. In this project, the Bayesian network has been chosen as the key modelling tool, but also other approaches have been tested.

Figure 1 What are the elements of reality?

One important question in applying this approach is: which are the elements of reality? Concept analysis provides answers to this question. In particular, the case studies performed in the project (i.e. ‘concept cases’) have supported the development of the modelling framework by pointing out the quality attributes that are often used by human evaluators to assess visual quality.

The jury provides a discrete probability distribution over the predefined discrete scale of qualities. The result from the jury is thus a probability density on classification rather than an estimate of quality. If this information is used in decision making, e.g. selection of printing paper, the decision maker’s attitude towards the uncertainty in quality estimation can be taken into account with a utility function /24/, the expectation of which is to be maximized in decisions. If a utility is shared by many decision makers in many decisions/context/content, such a utility function is the visual quality index, a scalar measure of the quality of an image.

1.5 Image Quality Modelling

An image quality model can be any mathematical model which predicts the quality rating or subjective attributes (‘nesses’) using the values of technology variables or physical image parameters. Engeldrum /16/ proposed the concept of Image Quality Circle (IQC) for helping to understand the structure, links and elements between image quality rating and technology variables.

Many image quality models for imaging devices have been developed using linear or polynomial regression models /49,56/. In addition Minkowski and related power metrics have been used successfully. The root of a Minkowski metric is from multidimensional scaling where it has been used as a distance metric. One reason for the success of Minkowski type metrics is that observers focus on the worse perceptual attributes when they make image quality judgement. The exponent of Minkowski metric captures this.

I3A /32/ proposed the quality and measurement hierarchy for camera phones. This hierarchy and the IQC proposed by Engeldrum /16/ resemble each other closely. The measurement hierarchy of I3A is divided into subjective and objective levels. The subjective level is divided into two levels and the objective level into three levels. The objective level includes low level technical properties (as technology variables), low-level technical measurements (as physical image parameters) and perceptual measurements (as customer perceptions). The subjective level includes low-level subjective attributes (as customer perception) and high-level subjective attributes (as customer quality preference or quality rating).
The structure of this report is based on the measurement hierarchy for digitally printed image shown in Figure 2. The measurement hierarchy includes the subjective and objective levels. Objective measurements will be done from printed images (Chapter 3), printed test targets (Section 2.3) and unprinted paper (Section 2.1) by instrumental measurements and computational methods. Subjective measurements will be done from the printed images using human observers (Chapter 4). The arrows added on the Figure 2 define the input values for the image quality models used in the project and described in detail in Chapters 5 and 6.

![Figure 2](image)

**Figure 2** Hierarchy of printed image measurements.

### 1.6 Modelling Human Perception with Bayesian Network

The Bayesian network (BN) has been found to be very promising for the development of an evaluation framework for overall image quality. Bayesian network is a directed acyclic graph that presents random variables and their dependencies as a probabilistic graphical model. In the graph, the nodes represent the random variables and the edges between the nodes denote conditional dependencies. The probabilistic modelling framework is in accordance with describing measurement uncertainty in more traditional measurement systems; each node actually contains a conditional probability for the node element value, given the values of the elements from which an arrow is pointing to the node in question. Thus the Bayesian model can account for the subjectivity of human assessments - as well as the uncertainties in instrumental measurements - and makes it possible to view the overall image quality as a probability distribution of states of quality.

The network can simulate the hierarchical formation of image quality by a top-down layered structure. There may be several hierarchical levels in the perception of quality. Figure 3(a) exemplifies the structure of a BN model with four levels specified by an expert. The highest level is the subjectively assessed overall image quality on the right hand side of Figure 3(a). The second
The highest level consists of the high-level quality attributes such as naturalness and usefulness (clarity of details). The lowest perception level in the model of Figure 3(a) consists of the low-level quality attributes. They are assumed to be the attributes of the image that the humans can directly assess and that their assessments—despite the non-physical nature—are objective. Hence a jury would largely have consistent opinions about them. The low-level attributes reflect the state of reality concerning the perceived image quality and thus determine statistically both the instrumental measurement values and the higher level subjective quality. Figure 3(b) shows the other approach for defining the structure of network. The network has input values (instrumental measurements), one output value (overall quality) and a local network of nodes and edges between them. In this case the structure of the network has been learnt from the measurement data using some optimization based search.

The identification of the structure and parameters of the BN requires reference measurements of the image quality at the various abstraction levels. The subjective assessments made by a jury are used as reference. The jury assessment may involve various types of inconsistencies because the target property is highly subjective. Compared to direct model, a layered model structure is more robust against these problems. In the layered BN identification, the first step is to determine the linking structure from low-level to high-level image quality attributes. This is based on general knowledge about human assessment of images rather than on jury data. As the second step the parameters of the links in the structure can be determined from jury data that contains parallel observations of the states of all nodes.

Figure 3  Graphical scheme of a Bayesian network model: a layered structure (a) and a machine learned structure (b).

The key idea in modelling the quality assessment with a Bayesian network is that once the model is identified, it can be executed again with any new set of measurements. In particular, any evidence about a (low abstraction level) node state can be propagated through the network so that the probability of all image quality attributes and overall image quality values are updated. Obtaining evidence of quality attribute values through instrumental measurements and then inferring about the overall quality is the main intended use. However, as image content plays a major role in the subjective image quality evaluation process, it may be necessary to construct separate BN models for different image contents.
2 VISUAL QUALITY EVALUATION

2.1 Subjective Visual Quality

2.1.1 Project Considerations

Historically objective quality metrics have been grounded on mathematical models of image quality that sometimes incorporate present theoretical knowledge of low-level perception (for full-reference metrics, see e.g. /15/). These objective IQ measurement values have then been compared with mean subjective ratings of human observers in order to evaluate the performance of the methods. How people really evaluate image quality has, however, largely been ignored.

In addition to creating a single quality index we aimed at studying the quality experience more thoroughly as it may help develop more precise objective measurement of image quality. Our approach is considered a top-down strategy as it starts by defining the experience of visual quality (“top”) and its features, after which it proceeds to physical properties (“down”) via subjective attributes and computational properties of an image. This is a novel approach: a psychological process that ends in a judgment of overall quality for an image is a rather unknown area.

Rating image quality requires rather complicated, high-level cognitive and perceptual processing. Also the traditional view of subjective image quality has mostly been based on the theories of detection, that is, at what probability is a signal (in case of image quality a distortion, e.g. noise) detected. This, essentially a psychophysical approach, could be described as a bottom-up strategy, as it strives to compute and predict the general quality straight from the visibility of the technical properties of an image.

2.1.2 Concept Cases

Outline

The selected top-down strategy was carried out by examining the concepts that people use for describing their quality experience. Free description has been proved to provide useful information about the quality experience /57/. The qualitative analysis of the descriptions yields subjective attributes that clarify the reasoning behind the quality rating. Usage of these attributes clarifies the aspects of quality that attract people’s attention and the interpretations that people do from the quality differences. Each round of subjective testing was called a concept case (CC). From CC2 on, printed samples were used. The following sections present the materials used, the studied image contents and the process of printing each set of samples.

Material

In CC2, five multipurpose (MP) papers were studied. The properties of the paper grades are presented in Table 1.
Table 1  The variation in paper characteristics in the test series for CC2.

<table>
<thead>
<tr>
<th>Paper / Print</th>
<th>Grammage (g/m²)</th>
<th>RMS Roughness</th>
<th>Print Gloss (Black, GU)</th>
<th>Print Density (Black, D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>80-90</td>
<td>2.3-4.9</td>
<td>1.6-2.3</td>
<td>1.36-1.41</td>
</tr>
</tbody>
</table>

For CC3 (and CC3.5), 10 multipurpose and 6 ink-jet (IJ) papers were chosen. The variation in paper and print characteristics in the series is shown in Table 2.

Table 2  The variation in paper and print characteristics in the test series for CC3.

<table>
<thead>
<tr>
<th>Paper / Print</th>
<th>Grammage (g/m²)</th>
<th>Whiteness (%)</th>
<th>Brightness (%)</th>
<th>Opacity (%)</th>
<th>Permeability (µm/Pas)</th>
<th>Permeability Bendtsen (ml/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>80-90</td>
<td>79-162</td>
<td>86-100</td>
<td>86-95</td>
<td>4.1-17</td>
<td>180-700</td>
</tr>
<tr>
<td>IJ</td>
<td>100-255</td>
<td>104-136</td>
<td>91-96</td>
<td>94-97</td>
<td>0.5-3.9</td>
<td>0-170</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paper / Print</th>
<th>Roughness PPS (µm)</th>
<th>Roughness Bendtsen (ml/min)</th>
<th>Paper Gloss (GU)</th>
<th>Print Gloss (Black, GU)</th>
<th>Print Density (Black, D)</th>
<th>Colour Gamut (a<em>b</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>3.3-6.7</td>
<td>31-230</td>
<td>4.7-8.1</td>
<td>0.9-1.4</td>
<td>1.01-1.12</td>
<td>4800-5500</td>
</tr>
<tr>
<td>IJ</td>
<td>0.6-7.0</td>
<td>0-320</td>
<td>2.2-65</td>
<td>0.1-110</td>
<td>1.33-2.22</td>
<td>7400-11900</td>
</tr>
</tbody>
</table>

The samples for CC4, which were subsequently used for CC4.5 and CC5.5 as well, consisted of 6 multipurpose and 15 electrophotographic (EPG) papers. Table 3 presents the variation in paper and print characteristics in the test series.

Table 3  The variation in paper and print characteristics in the test series for CC4, CC4.5 and CC5.5.

<table>
<thead>
<tr>
<th>Paper / Print</th>
<th>Grammage (g/m²)</th>
<th>Whiteness (%)</th>
<th>Brightness (%)</th>
<th>Opacity (%)</th>
<th>Fluorescence (%-units)</th>
<th>Permeability Bendtsen (ml/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>77-84</td>
<td>79-157</td>
<td>89-100</td>
<td>86-94</td>
<td>0-68</td>
<td>392-706</td>
</tr>
<tr>
<td>EPG</td>
<td>91-274</td>
<td>107-161</td>
<td>90-101</td>
<td>92-100</td>
<td>24-73</td>
<td>0-213</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paper / Print</th>
<th>Roughness PPS (µm)</th>
<th>Roughness Bendtsen (ml/min)</th>
<th>Paper Gloss (GU)</th>
<th>Print Gloss (Black, GU)</th>
<th>Print Density (Black, D)</th>
<th>Colour Gamut (a<em>b</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>4.5-7.2</td>
<td>91-247</td>
<td>4.0-7.4</td>
<td>9.0-20</td>
<td>1.59-1.65</td>
<td>7500-8500</td>
</tr>
<tr>
<td>EPG</td>
<td>0.7-4.1</td>
<td>0-122</td>
<td>8.2-87</td>
<td>13-53</td>
<td>1.55-1.80</td>
<td>7700-9000</td>
</tr>
</tbody>
</table>

For CC5, a total of 6 multipurpose and 15 ink-jet papers were chosen. The variation in paper and print characteristics in the test series is shown in Table 4.
Table 4  The variation in paper and print characteristics in the test series for CC5.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Grammage (g/m²)</th>
<th>Whiteness (%)</th>
<th>Brightness (%)</th>
<th>Opacity (%)</th>
<th>Fluorescence (%-units)</th>
<th>Permeability Bendtsen (ml/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>77-84</td>
<td>79-157</td>
<td>89-100</td>
<td>86-94</td>
<td>0-68</td>
<td>392-706</td>
</tr>
<tr>
<td>IJ</td>
<td>98-280</td>
<td>91-150</td>
<td>86-98</td>
<td>89-99</td>
<td>7-59</td>
<td>0-193</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paper / Print</th>
<th>Roughness PPS (µm)</th>
<th>Roughness Bendtsen (ml/min)</th>
<th>Paper Gloss (GU)</th>
<th>Print Gloss (Black, GU)</th>
<th>Print Density (Black, D)</th>
<th>Colour Gamut (a<em>b</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>4.5-7.2</td>
<td>91-247</td>
<td>4.0-7.4</td>
<td>0.6-0.9</td>
<td>1.09-1.18</td>
<td>5600-6300</td>
</tr>
<tr>
<td>IJ</td>
<td>0.6-5.7</td>
<td>0-270</td>
<td>2.2-96</td>
<td>0.1-88</td>
<td>1.32-2.11</td>
<td>8700-13800</td>
</tr>
</tbody>
</table>

**Image Contents**

Whereas the three image contents used in CC1 were chosen based on the presence of a few aspects whose reproduction is critical for image quality, such as skin colour, different shades of natural green, areas with uniform colour and small details, a more comprehensive approach was used in choosing the image contents for the printed samples. The images represented typical content types such as objects with details (*cityscape*), a human portrait (*woman*) and a landscape (*countryside*). Moreover a studio setting, which contained all the above mentioned content types, was constructed and photographed (*studio*). The cityscape presented a busy street café on a sunny day with several buildings and some sky in the background. The portrait presented a close-up of a young brown-haired woman wearing a white t-shirt and a light blue cap. The woman was in front of a grey background and was facing the camera. The countryside presented a green field under a blue sky with white cloud clusters. Some water and trees were also in the image.

The *studio* image /Figure 4/ included objects for easy recognition of print defects. The grey wall in the background and the black strip of cardboard were included in order to facilitate the evaluation of noise. Objects representing naturalness were the girl’s skin, the oranges, the sky, the grass and the roses. It was expected that these objects would be assessed using memory colours. The white cap with the black tassel was included in the image for the evaluation of dynamic range and gloss variation. The tassel of the cap and the folded newspaper facilitated the evaluation of details.

**Figure 4  Image content studio.**

For CC3, a new set of images was photographed /Figure 5/ to replace the three image contents used in addition to *studio*. The same image content types were however used: objects with details (*cactus*), a human portrait (*man*) and a landscape (*landscape*). The photographs were captured by a semi-professional photographer using a professional digital camera.

In photographing the images, some content-related requirements were considered. In the photographs, versatile memory colours, e.g. skin, sky and foliage, as well as memory shapes, e.g. a
human face and a cactus, were included. Different types of surface materials from shiny and smooth, e.g. porcelain and fabric, to detailed and textured, e.g. a zipper and rock, were also considered. All photographs had a considerable amount of details, e.g. cactus spikes, human hair and tree leaves. Moreover, perceptual and aesthetical aspects were considered.

![Figure 5](image)

*Figure 5* The different image content types used in addition to studio in CC3-CC5.5, from left to right cactus, man and lake.

From CC3 on, a set of objective test targets were included in the print layouts. The test fields consisted of full-tone CMYK patches, a large 50% gray patch, RGB and CMYK wedges with 10% increments from 10% to full tone, and vertical and horizontal lines with a width of 1-10 pixels (one pixel increments).

**Printing the Samples**

The samples of CC2 were printed using a laser printer HP Color LaserJet 5550. Due to the similarity of the papers, no ICC profiles were used. The papers differed mainly in roughness /Table 1/, but paper D had also a detectable yellowish tint.

The samples of CC3 (and consequently of CC3.5) were printed with an Epson Stylus Pro 3800 desktop ink-jet printer (CMY, light C, light M, light K, light light K and photo K or matte K). Optimal printing settings were chosen for each paper by identifying the print media type. Paper-specific ICC profiles were determined for these settings in Printopen 4.0.5. PO Standard RGB test sheet was used as the profiling target. A Spectrolino spectrophotometer attached to a SpectroScan XY table was used for measuring the printed profiling target. The test images were printed from Adobe Photoshop 7.0.1. Three replicates were printed and the best was chosen for further study.

The samples of CC4 (as well as of CC4.5 and CC5.5) were printed with a Xerox DocuColor 6060 Digital Color Press (CMYK). Optimal print settings were chosen for each paper by identifying the grammage category and whether the paper is coated. Paper-specific calibrations with an X-Rite QuickCal densitometer were also carried out. Paper-specific ICC profiles were determined in Profilemaker Pro 5.0.8. The target used for profiling was IT8.7-3 CMYK. The targets were printed with i1iO layout as they were measured with an X-Rite i1 Pro spectrophotometer attached to an X-Rite i1iO automated chart reader. A printer-specific algorithm, Dot200, was used for halftoning. The samples were printed from Mac OS 10.4.9 using Adobe Photoshop 10.0. Ten copies were made to control the print-to-print variation. The ninth was chosen for further study.

The samples of CC5 were printed with the Epson Stylus Pro 3800. Optimal print settings were chosen for each paper. Paper-specific ICC profiles were determined in Profilemaker Pro 5.0.8. The target used for profiling was TC9.18 RGB with an i1iO layout. A halftoning algorithm defined by
the printer was used. The samples were printed from Windows XP Pro using Photoshop 7. Three copies were made and the one with the best overall quality was chosen.

**Summary**

Table 5 summarises the samples used in the concept cases.

**Table 5** Paper types (MP stands for multipurpose, IJ for ink-jet and EPG for electrophotographic papers), printing methods and image contents studied in the concepts cases.

<table>
<thead>
<tr>
<th>Concept Case</th>
<th>Paper Types</th>
<th>Printing</th>
<th>Image Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC1</td>
<td>none</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>CC2</td>
<td>5 MP</td>
<td>EPG</td>
<td></td>
</tr>
<tr>
<td>CC3</td>
<td>10 MP, 6 IJ</td>
<td>IJ</td>
<td></td>
</tr>
<tr>
<td>CC3.5</td>
<td>10 MP, 6 IJ</td>
<td>IJ</td>
<td></td>
</tr>
<tr>
<td>CC4</td>
<td>6 MP, 15 EPG</td>
<td>EPG</td>
<td></td>
</tr>
<tr>
<td>CC4.5</td>
<td>6 MP, 15 EPG</td>
<td>EPG</td>
<td></td>
</tr>
<tr>
<td>CC5</td>
<td>6 MP, 15 IJ</td>
<td>IJ</td>
<td></td>
</tr>
<tr>
<td>CC5.5</td>
<td>6 MP, 15 EPG</td>
<td>EPG</td>
<td></td>
</tr>
</tbody>
</table>

**2.1.3 Process of Subjective Visual Quality Evaluation**

The detection-based, bottom-up approach /see Section 2.1.1/ is problematic in several senses. For example, all the visual information on an image is not visible at once. High quality information is only acquired from a small region around the centre of the gaze, so information about the image has to be gathered by fixating the eye on different points at the image /59/. Visual information does not therefore flow automatically to further processing, but it goes through significant filtering before any additional processing.

Attention is the mechanism that selects the visual information essential for the observer. Only some visual information can be detected without selective attention /80/. When quality attributes require selective attention, some of them may be detected and some neglected in a probabilistic manner. Importance of attention is demonstrated by the following example from CC2: Table 6 shows what targets attracted attention in the test image content (studio). In this experiment, 10 observers...
evaluated images printed on 5 different papers in pair-wise comparison setting. During the analysis of the observers’ preferences, it was noted that they showed severe inconsistencies in their preferences in three comparison conditions, which made their reasoning of image quality seem irrational. It was found out that the inconsistencies were produced because observers concentrated on different targets on the images in different conditions /Table 6/. In other words, different aspects of quality were evaluated. On the basis of this examination of observers’ preferences, it can be stated that although people seem to be irrational in the levels of pair-wise choices, a further analysis of their experience of quality may reveal that they really have well-grounded reasons for their choices.

Table 6 Frequency of target objects used to evaluate image content studio in three conditions A vs. B, A vs. E, and B vs. E (CC2), where the most salient inconsistencies of preference were noticed.

<table>
<thead>
<tr>
<th>Object</th>
<th>A vs. B</th>
<th>A vs. E</th>
<th>B vs. E</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oranges</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Red rose</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Green leaves of the rose</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Sky</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Grey background</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Foliage</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Rocks</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Face</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Tablecloth</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Newspaper</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>White cap</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Black frame</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Water</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>19</td>
<td>15</td>
<td>58</td>
</tr>
</tbody>
</table>

The physical attributes that are detected, are always interpreted somehow. Interpretation gives a meaning for physical attributes that are detected, and this meaning may differ between individuals and image contents /57/. In other words, the same physical attribute may be experienced as different subjective attribute by different observers. This is more thoroughly examined in the next section about subjective quality concepts.

Finally, after being detected and interpreted, the observer has to decide what the importance of different subjective attributes is in the final rating. There are several strategies of how people combine information about the alternatives in a choice situation, but it is not known how the subjective attributes are pooled in image quality rating. In research concerning people’s decision making behaviour in general, it is usually concluded that they use quite simple rules in choice situation, especially if the decision is not very important for the person /3/. Human information processing capacity is quite limited, and it is usually concluded that people cannot exploit all information that is available.
Understanding the processes, which result in single quality evaluation, might make it easier to understand the effect of content on image quality, for instance. One hypothesis that has been carried out during the project has been that low-level attributes may be perceptually integrated into high-level attributes that may finally be the attributes to be used in the final quality decision making.

2.2 Subjective Image Quality Evaluation Methodology

All the subjective tests were conducted in a laboratory, specifically equipped for subjective print quality evaluation experiments. Walls were covered with neutral grey curtains and the table used for the presentation of the samples was covered with a tablecloth of similar grey shade. The illumination (2000 lux) was provided with Just Normlicht Color Control Daylight, with a correlated colour temperature of 5000 Kelvin and a colour rendering index of 98. The samples were attached to grey frames in order to avoid the wear of the samples and to make their handling easier. In all subjective tests, the order of the presentation of the printed samples and contents was randomized. The observers were given a permission to take the samples in their hands holding the frames, since free handling of printed samples is essential, e.g. in the evaluation of gloss. The observers were recruited from the university e-mail lists.

Subjective quality ratings are very sensitive to several confounding factors and they demand much time and effort. The choice of the proper method is always a trade-off between generalisability, sensitivity and other aspects. In this project, several methods were tested in order to evaluate their suitability for the subjective evaluation of the print samples, for instance pair-wise comparison, category scaling, qualitative methodology, semi-structured interview and grouping. Methods used for subjective IQ evaluations are compiled in Table 7.
Table 7  Comparison of the methods between different concept cases.

<table>
<thead>
<tr>
<th>Concept case</th>
<th>Quality evaluation</th>
<th>Other methods</th>
<th>What was looked for?</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC1</td>
<td>Pair-wise</td>
<td>Interview</td>
<td>Quality, concepts</td>
<td>Digital images</td>
</tr>
<tr>
<td>CC2</td>
<td>Pair-wise</td>
<td>Interview</td>
<td>Quality, concepts</td>
<td>5 EPG papers</td>
</tr>
<tr>
<td>CC3</td>
<td>Quality scale 1-5</td>
<td>Grouping + interview</td>
<td>Perceptual similarity, concepts</td>
<td>16 IJ papers</td>
</tr>
<tr>
<td>CC3.5</td>
<td>Quality scale 1-5</td>
<td>Interview</td>
<td>Quality, hierarchy of concepts, quality decision making</td>
<td>16 IJ papers</td>
</tr>
<tr>
<td>CC4</td>
<td>Quality scale 1-5</td>
<td>Attribute scales</td>
<td>Quality, attribute scale data</td>
<td>21 EPG papers</td>
</tr>
<tr>
<td>CC4.5</td>
<td>Quality scale 1-5</td>
<td>Grouping + interview</td>
<td>Perceptual similarity, concepts</td>
<td>21 EPG papers</td>
</tr>
<tr>
<td>CC5</td>
<td>Quality scale 1-5</td>
<td>Interview</td>
<td>Quality, quality decision making</td>
<td>16 IJ papers</td>
</tr>
<tr>
<td>CC5.5</td>
<td>Quality scale 1-5</td>
<td>Interview</td>
<td>Quality, quality decision making</td>
<td>21 EPG papers</td>
</tr>
</tbody>
</table>

Each of the methods had their strengths and weaknesses. For one-dimensional quality data, the best trade-off between different aspects was achieved by the relative mean opinion score (MOS) elicited by category scaling. Category scaling usually uses verbal descriptions for the categories, but the above-mentioned method was done in relation to the best and the weakest quality sample (therefore the relative MOS). This approach was selected, because it was the easiest way to ensure that the observers were using the scale similarly and therefore giving most coherent ratings. The problem with category scales usually is that people do not share the meanings of the verbal descriptions. In this project, subjective ratings for quality were given in scale from 1 to 5, where 1 denotes the poorest quality sample in the certain printing method and 5 denoted the highest quality. The sample set was also thought to be representative, that is, in a certain printing context; this range from 1 to 5 consists of the best and the weakest sample available.

In addition to single quality rating, also other aspects that relate to quality evaluation were studied, such as the perceptual similarity of the samples (quantitative and qualitative), by using a grouping task and verbal descriptions. Similarity evaluation is closely related to the quality evaluation. Third quantitative method that was used in the project was pair-wise comparison. It however had some weaknesses (needed too much time, observers gave inconsistent responses).

In addition to quality rating, grouping methodology gained larger usage. In this method, observers were simply asked to group similar samples into the same groups. The purpose of this methodology was to get a similarity measure between samples. The similarity measure was based on how many times samples were grouped into the same groups. Similarity data can be described as multidimensional quality data whereas quality ratings serve as one-dimensional quality data. However, it turned out that in case of paper, quality is mostly one-dimensional and related to the surface structure of the paper. This is apparently due to the fact that the properties of the studied paper grades have high correlations with each other.

The quantitative methods describe above were supplemented with qualitative methodology. The purpose was to gain information about the relevant subjective attributes that influence the visual quality evaluation. In practice, the qualitative approach meant that with each evaluation, the
observers were asked to give the reasons behind the evaluation. They were free to answer anything that comes to their mind, but in all other respects, the experiment had the same controlled nature as the methods described above. Analysis of the qualitative data was done by collecting the attributes from the citations that observers use to describe the samples. Figure 6 further clarifies the coding process.

The category scaling method was also applied to the subjective attributes of the samples in CC4. Properties like sharpness, graininess and contrast were evaluated on 5-point scales. The choice of the attributes was based on a qualitative analysis of the grouping data, which consisted of people’s description of visual differences between the samples.

The quality ratings done in the project were quite reliable, the correlation between two different evaluation settings and observer groups being higher than .99. Reliability of subjective tests can be improved using sufficient amount of observers (preferably 20 – 30, at least), proper test images and controlled experimental conditions.

2.3 Concepts from Case Studies

2.3.1 High- and Low-Level Attributes

Already from studies preceding the current project, it was noticed that the abstractness of the quality attributes differed significantly: There are concrete (low-level) subjective attributes that relate to simpler physical properties of an image, and there are abstract (high-level) subjective attributes that describe image quality by the means of clarity and naturalness of an image, for instance, which may be related to several physical properties of an image. These two attributes were further used in the project.

On the whole, we found out five major higher level attributes. The most abstract and demanding attribute was realism /see Table 8; CC3/, which refers to the observers’ feeling of being in the scene depicted by the image or the feeling of missing the mediating device between the scene being seen. Naturalness is related to the perception of the objects or the scene ‘as they are in real life’, or their perception without any distortions. Clarity refers to how easily the objects of the images can be
discriminated: a clear image is easy to watch. Finally, there are some attributes that are just associated with good quality, like gloss. It is also interesting that good image quality made the images look more three-dimensional.

Table 8  High level attributes used to reason meaning of the low level attributes for the quality rating, their frequencies, description and image quality concepts that resemble those attributes (CC3 samples).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Freq.</th>
<th>Description</th>
<th>Related concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realism, genuineness etc</td>
<td>73</td>
<td>The feeling of being absorbed into image, the feeling of being in the image</td>
<td>Presence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The feeling there is no medium between the world of the image and the real world</td>
<td></td>
</tr>
<tr>
<td>Naturalness</td>
<td>69</td>
<td>The objects being reproduced as they are</td>
<td>Naturalness</td>
</tr>
<tr>
<td>Clarity</td>
<td>66</td>
<td>The image being informative</td>
<td>Usefulness, clearness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The objects are easy to discriminate</td>
<td></td>
</tr>
<tr>
<td>Depth</td>
<td>35</td>
<td>Feeling of 3-dimensionality, objects coming forward from the image</td>
<td></td>
</tr>
<tr>
<td>Quality associations</td>
<td>34</td>
<td>Image can resemble e.g. image of the high class magazine or photograph</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>People associate certain low-level attributes with good quality</td>
<td></td>
</tr>
</tbody>
</table>

All these concepts are overlapping to some extent with each other. Realism, naturalness and depth, for example, are all very much interrelated. Therefore only two attributes, naturalness and clarity were further examined.

Table 9 describes how the concrete and abstract attributes relate to each other. It shows the results from the analysis of the observers’ verbal descriptions of their quality experience of ink-jet prints. The cross-tabulation of the Table 9 was further analysed by correspondence analysis, which showed that brightness, sharpness and larger contrast made the image seem clearer. Having no graininess, good and bright colours added realism and naturalness to the image. When examining the low-level quality attributes, they had different associations to the high level attributes. For example, bright colours were associated to realism and naturalness in high quality, but faded colours were associated with unclarity. Also graininess seemed to be the main reason for the appearance of an image being perceived as unreal, unnatural and flat.

These results offer some knowledge on how the quality as an experience is constructed, but generalizing results should, however, be done with caution, since the experiment was done only with ink-jet prints.
Table 9  Relation between subjective high-level and low-level attributes (CC3).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Unreal</th>
<th>Natural</th>
<th>Unnatural</th>
<th>Clear</th>
<th>Unclear</th>
<th>Good quality</th>
<th>Bad quality</th>
<th>More depth</th>
<th>Less depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brightness of colours</td>
<td>Bright</td>
<td>18</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Faded</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Sharpness</td>
<td>Sharp</td>
<td>13</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Unsharp</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Graininess</td>
<td>Grainy</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Not grainy</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>White dots</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Bright</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Faded</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Colour quality</td>
<td>Good colours</td>
<td>6</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Bad colours</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Gloss</td>
<td>Glossy</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Contrast</td>
<td>Large contrast</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lightness</td>
<td>Light</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

There is also a third class of attributes that does not fit the above-mentioned categories. These attributes may have minor importance for image quality evaluation, but they still illustrate the experiential manner of the evaluation. They may also have a great influence on the impression a print product gives to an observer. Table 10 shows the frequencies of these attributes that do not share the concrete nature of low-level attributes and are not general in the way high-level attributes are. The experiential nature of image quality is represented in these attributes highly dependent on the interpretation of quality differences in certain contexts.

Table 10  Other abstract attributes used in quality evaluation (CC3).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fresh vs. smudgy</td>
<td>13</td>
</tr>
<tr>
<td>Vividness</td>
<td>11</td>
</tr>
<tr>
<td>Soft vs. hard</td>
<td>10</td>
</tr>
<tr>
<td>Summer vs. autumn</td>
<td>9</td>
</tr>
<tr>
<td>Fogginess</td>
<td>8</td>
</tr>
<tr>
<td>Happiness</td>
<td>7</td>
</tr>
<tr>
<td>Warm vs. cold</td>
<td>7</td>
</tr>
<tr>
<td>Inviting</td>
<td>3</td>
</tr>
</tbody>
</table>
2.3.2 Subjective Dimensions of High and Low Quality

The attribute dimensions that make an image poor may differ from attributes that make image excellent. Table 11 shows the most important attributes that were used with ink-jet prints: MOS column shows the mean quality rating that was used for each attribute dimension. Attribute dimensions consist of polarities, such as sharp and unsharp. From the table, it can be seen that attribute dimensions like depth, glossiness and vividness were only used when evaluating high quality images, fadedness and white dots were used in low quality images.

Table 11  Usage of different attribute dimensions.

<table>
<thead>
<tr>
<th>Attribute dimension</th>
<th>MOS</th>
<th>Difference between dimension polarities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>3.71</td>
<td>1.34</td>
</tr>
<tr>
<td>Glossiness</td>
<td>3.61</td>
<td>0.83</td>
</tr>
<tr>
<td>Vividness</td>
<td>3.48</td>
<td>1.58</td>
</tr>
<tr>
<td>Even printed surface</td>
<td>3.32</td>
<td>0.97</td>
</tr>
<tr>
<td>Naturalness</td>
<td>3.31</td>
<td>1.81</td>
</tr>
<tr>
<td>Brightness</td>
<td>3.28</td>
<td>0.42</td>
</tr>
<tr>
<td>Contrast</td>
<td>3.23</td>
<td>1.28</td>
</tr>
<tr>
<td>Evenness</td>
<td>3.20</td>
<td>0.21</td>
</tr>
<tr>
<td>Sharpness</td>
<td>3.19</td>
<td>1.48</td>
</tr>
<tr>
<td>Colourfulness</td>
<td>3.09</td>
<td>1.34</td>
</tr>
<tr>
<td>Colour reproduction</td>
<td>3.09</td>
<td>1.15</td>
</tr>
<tr>
<td>Clearness of colours</td>
<td>3.04</td>
<td>1.28</td>
</tr>
<tr>
<td>Clarity</td>
<td>3.00</td>
<td>1.78</td>
</tr>
<tr>
<td>Identical to the best or to the worst</td>
<td>2.98</td>
<td>3.96</td>
</tr>
<tr>
<td>Photography-likeness</td>
<td>2.96</td>
<td>2.59</td>
</tr>
<tr>
<td>Lightness</td>
<td>2.86</td>
<td>0.75</td>
</tr>
<tr>
<td>Graininess</td>
<td>2.80</td>
<td>1.26</td>
</tr>
<tr>
<td>Well-lighted vs. dim</td>
<td>2.79</td>
<td>0.01</td>
</tr>
<tr>
<td>Naturalness of colours</td>
<td>2.76</td>
<td>2.12</td>
</tr>
<tr>
<td>Smudginess</td>
<td>2.68</td>
<td>1.23</td>
</tr>
<tr>
<td>Greyness</td>
<td>2.64</td>
<td>0.33</td>
</tr>
<tr>
<td>Fadedness</td>
<td>2.45</td>
<td>0.7</td>
</tr>
<tr>
<td>White dots</td>
<td>2.44</td>
<td>1.66</td>
</tr>
</tbody>
</table>

2.3.3 Subjective Quality Dimensions for Different Printing Methods

Although there are several attributes that describe the quality of an image, the subjective visual dimensions that differentiate between samples may be few in the case of printed samples. The paper grades studied were commercial products; many of the paper properties were highly correlated. The subjective dimensions that resulted from analyses were therefore dependent on both the correlation
between the physical properties of the paper grades and the subjective perception of those properties.

Three methods were used to reveal the subjective quality dimensions from the experimental test data: Multidimensional scaling, factor analysis and correspondence analysis.

**Electrophotography**

In electrophotographic printing, there seem to be two or three dimensions that visually differentiate images printed on different papers from each other. The most important dimension is associated with graininess, sharpness and contrast according to the factor analysis of the attribute scales /Figure 7; CC4/. Correspondence analysis revealed attributes such as “white dots” and naturalness to be also related to that dimension. The physical property that is most closely associated with this dimension is paper roughness and it is the only dimension that is directly related to quality.

The second most important dimension is related to subjective lightness and colourfulness in the image. It seems to be associated with paper whiteness or fluorescence. Correspondence analysis revealed that in this dimension more lightness also meant worse contrast, colder colours, a less vivid and pale impression. Multidimensional scaling gave indications that observers could also separate whiteness and colourfulness dimensions from each other, but it may not be relevant in case of quality.

**Ink-Jet**

In the ink-jet context, the paper surface seems to have two dimensions that affect subjective quality of prints. The first is related to the paper roughness that makes the print seem to have small white dots and therefore being not uniform, the other is related to gloss. In the accompanying perceptual map of the correspondence analysis of the CC4 /Figure 8/, these dimensions (glossiness and uniformity) are rotated 45 degrees from dimensions 1 and 2 and are marked by black arrows. The third dimension is related to the tint of the paper and separated the slightly yellowish paper number 16 from other papers.
2.4 Consistency of a Jury

To detect and analyze the interrelations between the perceptual image quality and the instrumentally measurable quality properties, a reference observation of the quality is needed. The obvious - and only - choice is to use as the reference data the perceptual quality judged by a group of human observers, i.e. a jury. A survey of using jury as a reference for instrumental measurements of image quality was made in this project /110/. This section presents an overview of the topic.

The main issue in using jury as a reference is rationality, i.e. the consistency of the decisions made by the jury members and the jury as a whole. The ambiguity of quality attributes and assessment scales and the subjectivity of perceived quality often make the jury or its members seem irrational. The lack of rationality increases the uncertainty in jury assessment and thus decreases the reliability and informativeness of the results.

According to Arrow’s impossibility theory /24/, there is no way to combine the preferences of the jury members so that the decision is guaranteed to be rational. For instance, the comparison of three or more samples provides the possibility of transitive irrationality in the form of a circular preference. A circular preference appears if a jury (or a juror) decides, given the options A, B and C, that A > B (i.e., option A is better than option B), B > C and C > A. Now the preferences are in a circle: A > B > C > A. It is possible to circumvent the circular preferences by considering the options equal, denoted as A ~ B ~ C. Obviously, the informativeness of such jury decision as a reference may be questioned. In this project we have seen cases where individual members of the jury are irrational, but when the results of the jurors are averaged to form the preference of the jury, then the jury as a group is rational. In forming a quality reference from the jury assessments, it may be justified to discard the samples or jury members that cause circular preferences.

Two measures were introduced /see 110/ to characterize the consistency of a jury: alpha rationality and the normalized sum of rational partial juries (NSRPJ). Alpha rationality can also be applied to a single juror. It is a measure of the objectivity of the assessments. When the alpha rationality of a single person gets its maximum value, 1, the person is considered fully rational and thus he/she makes only coherent decisions that do not form circular groups. Smaller values of alpha correspond
to less rational decision making. On the other hand, the minimum alpha value, 0, does not guarantee the irrationality of the decision maker. It does, however, make it possible that the utilities of the rank ordered samples are not coherent with the rank ordering. To clarify this idea, let us assume that a set of $M$ samples has been evaluated by a jury member. The preference order of the samples for this person can be presented by the index, $i$ ($i = 1...M$), so that $i = 1$ corresponds to the lowest rank and $i = M$ represents the best sample according to this person. The utility of a sample for each such jury member can then be calculated as

$$E_i = \frac{\alpha \cdot i}{M} + \left(1 - \alpha \cdot \frac{1}{M}\right) \cdot S,$$

where $S$ is an uniformly distributed random number between zero and one. The utility values, $E_i$, then vary between zero and $1+1/M$.

When alpha rationality is applied to a jury, it is typically assumed that each jury member has the same alpha value. Thus the alpha rationality of a jury describes the mutual consistency of the jury decisions.

Another way to describe the uniformity of a jury’s decision is to inspect the rationality of all the possible partial juries. The fraction of rational partial juries is denoted as the *normalized sum of rational partial juries* (NSRPJ). The partial juries consist of an odd number of evaluators, three or more. For a jury of $N$ members and the partial jury size $n_p$ members ($n_p < N$) the number of partial jury combinations is

$$S_p = \binom{N}{n_p} = \frac{N!}{n_p!(N-n_p)!}.$$  

In the project, the number of members in partial juries was always three, $n_p = 3$. For image quality evaluation there are often about 15 to 30 evaluators, so the number of three-member partial juries are 455 and 4060, respectively. This is still feasible from the computational point of view, but for instance a 100-member jury can form 161700 three-member partial juries, and each has to be examined for circular preferences. It can thus be stated that NSRPJ is a feasible rationality measure for juries that have less than a hundred evaluators.

There is a positive non-linear relation between the alpha rationality and NSRPJ. The relation can be utilized to obtain the alpha rationality as a function of NSRPJ or vice versa. For the simulations of jury evaluation, it is easier to apply alpha rationality than to construct all partial juries with certain circular preferences. Figure 9 illustrates the dependence between NSRPJ and alpha rationality and their relation to the number of samples in the evaluation, according to simulations reported by Pulla /110/. It can be seen that NSRPJ increased when alpha increased. When the level of alpha was fixed, NSRPJ decreased as the number of samples increased because the chance of ending up with circular preferences is high if the sample set to evaluate is large. Also, a large number of samples implies samples with attribute values close to one another. A further simulation indicated that the size of the jury did not affect the rationality of partial juries, which was a natural and expected result.
After studying several jury evaluation cases with the NSRPJ criterion, both through simulations and with real data, the conclusion was that measurement descriptions could be identified with jury as reference only when the conditioning quality aspects were quite unambiguous and hence circular preferences in jury assessment were few. If circular preferences occur, the measurement description must be modelled by combining knowledge about human factors to jury and measurement data.

### 2.5 Variation in Jury Evaluations

While the two rationality measures presented above, alpha rationality and NSRPJ, gave a general view of the consistency of the jury’s decisions, there are also methods for more detailed analysis of subjective evaluation data. The computational measures of Tukey’s honestly significant difference (Tukey’s HSD) and just noticeable differences (JNDs) are briefly introduced in the following.

Tukey’s HSD is a comparison test for determining statistically significant differences between the mean values of sample groups by the analysis of variance /29/. In case the jury has given ranking points to the samples, Tukey’s HSD is applied to the ranking points to assess the reliability of jury decision. In case of pair-wise comparison data, the results are transformed into ranking points through the utilities of the samples. The mean values of the sample groups are computed and then two groups can be considered significantly different if their mean values differ by at least the value $HSD_p$, where the sub-index $p$ denotes the level of significance, e.g. $p = 0.95$. For mathematical details of computing the HSD levels, see Pulla /110/ and Lowry /42/.

An example of an HSD test conducted in the project consisted of five sample groups that were based on a fixed image content being printed on five different paper types. Each group contained the ranking evaluations of the print quality evaluated by ten persons. Averaging the ten rank values within each group (paper type) might result in results such as those given in Table 12.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mean(rank)</td>
<td>2.85</td>
<td>0.85</td>
<td>1.95</td>
<td>1.60</td>
<td>2.75</td>
</tr>
</tbody>
</table>

**Figure 9** Interrelation between NSRPJ and alpha rationality, and their dependence on the number of samples.

**Table 12** Example of possible rank values given to five different types of paper.
Apart from the critical value that is based on the tabulated values of Student’s distribution, the computation of the HSD levels uses the ranking points and the degrees of freedom (df) of the whole evaluation set. In this example, the number of groups is five and there are ten evaluators. Thus the df for the image evaluation data sets is originally 50, but the calculation of the mean values of the groups during the process reduces the degrees of freedom by five, hence df = 50 − 5 = 45. Finally the HSD values for three typical significance levels are as those given in Table 13.

Table 13  Honestly significant difference (HSD) levels corresponding to the example in Table 12.

<table>
<thead>
<tr>
<th></th>
<th>HSD_{0.99}</th>
<th>HSD_{0.95}</th>
<th>HSD_{0.90}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.6829</td>
<td>1.3835</td>
<td>1.2355</td>
</tr>
</tbody>
</table>

Based on the HSD results and the average rank values given above, the comparison between papers A and B, and B and E, had a significant difference at the 99 % significance level (p = 0.99). At the significance level of p = 0.9 also the difference between papers A and D was significant, but for instance paper C could not be distinguished from the other papers even at this significance level.

Just noticeable difference (JND) is another numerical tool for analyzing the variation in subjective evaluations. It is used for paired comparisons and it is particularly useful when the perceptual differences between the samples are very small or hardly distinguishable. The JNDs make it possible to conduct not only the preference order for the samples but also a scale that measures, how much the samples differ from each other. The basic assumption behind JNDs is that perception is probabilistic and can be presented as a distribution /39/. The members of the jury are expected to either indicate their preference in the paired comparison, or guess the preference order if they cannot distinguish the samples from each other. This leads into the concept of certainty of detection, denoted by \( p_d \), and attached to the JND as a certainty level. For example, a 50 % detection probability, \( p_d = 0.5 \), means that every second answer is a guess. This is a typical certainty level attached to JND computations. In a paired comparison between options A and B, the correct response is assumed to be the preference that most of the observations support. The (theoretical) probability of the correct response, \( p_c \), is related to the certainty of detection as \( p_c = (1 + p_d)/2 \), which results in \( p_c = 0.75 \) when \( p_d = 0.5 \). To evaluate the perceptual difference between two samples as JND units, we need to define the difference between them by converting the probability of the correct response, \( p \), into sample difference in deviate units

\[
z_a(p) = \sqrt{2\pi} \cdot (\sin^{-1}(\sqrt{p}) - \frac{\pi}{4}),
\]

where the subscript \( a \) refers to angular distribution. The above equation is the inverse of the cumulative distribution function (CDF) of the angular distribution. The measured fraction of correct answers, \( p_p \), from the paired comparison results, is proportioned with the chosen certainty level, \( p_d \) (through \( p_c \)), which results in JND units as

\[
JND = \frac{z_a(p_p)}{z_a(p_c)}.
\]

With \( p_c = 0.75 \) and the definition of the angular deviate, the equation becomes

\[
JND = \frac{12}{\pi} \sin^{-1}(\sqrt{p_p}) - 3.
\]
The JND values are conveniently on a common scale and they can be used to form the preference order of the samples by sorting the JND values into ascending or descending order. However, if the JND values for two comparisons are almost equal, the preference of the samples must be interpreted with caution. For example, Table 14 gives the 50% JND values (i.e., \( p_d = 0.5 \)) between paper A and the rest of the papers with the same “Girl” image content as presented in Table 12. The results are based on pair-wise comparisons made in the CC2.

**Table 14** Example of 50% JND values from a pair-wise comparison of five different types of paper.

<table>
<thead>
<tr>
<th></th>
<th>A vs. B</th>
<th>A vs. C</th>
<th>A vs. D</th>
<th>A vs. E</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% JND</td>
<td>1.77</td>
<td>0.38</td>
<td>0.79</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Corresponding to the HSD results presented above, paper A is significantly preferred over paper B, because the sample difference is greater than one 50% JND. All the entries in Table 14 are positive, which means that, on average, paper A has been judged better than any other paper type. However, based on these results, preference cannot be made between papers D and E since their distances to paper A are equal. Furthermore, the small difference of 0.38 50% JNDs between papers A and C implies that the jury has hardly any preference between these papers.

### 2.6 Properties of Bayesian Network Models

The general description of the Bayesian network model does not limit the number of edges chosen for the model, but as the number of edges increase, the model becomes more complex. Greater complexity means that a larger jury data is needed for the model identification.

The complexity of the joint distribution of a node and its parent nodes grows exponentially in proportion to the number of parent nodes. In general for the constant number, \( N \), of discrete states in each node, the number of possible states of a joint distribution is \( S = N^{M+1} \), where \( M \) is the number of parent nodes. For example, if usefulness, naturalness and overall quality are discrete nodes with \( N \) states, the number of states of their joint distribution is then \( S = N^{5+1} \). The number of jury assessments needed for model identification is in proportion to the maximum number of states of the joint distribution in a discrete model.

Simulations have been performed to assess the sufficient number of jury assessments (or samples) for estimating the parameters of a BN model with a reasonable accuracy. In the simulated case /105/, the reference Bayesian network consisted of eleven nodes (3 instrumental, 5 low-level attributes, 2 high-level attributes, and the overall quality). The instrumental measurement nodes had three possible discrete states and the rest of the nodes had five discrete states. Each node had three parents or less. Simulated evaluations were produced according to the state probabilities in the reference BN model (for each attribute), and new model parameters were estimated from the simulated data. Data sets of 10, 100, 300, 500, 800, 1000, 1500, and 2000 samples were drawn from the reference model separately. The sizes of the data sets correspond to approximately 1, 4, 11, 19, 30, 37, 56, and 74 human evaluators, respectively. The data sets of each size were drawn 100 times from the reference model to take account of the variation between the sampled data sets. The original and simulated models were compared through evaluating the root mean square error (RMSE) between the state probabilities of the nodes of the original and the simulated model. The conclusion of the test was that, to make the new model parameters converge sufficiently towards the original parameters, more than 500 samples are needed.
Sensitivity analysis refers to identifying the most important parameters so that unimportant parameters may be discarded from the model. Sensitivity analysis in Bayesian networks is broadly concerned with understanding the relationship between local network parameters and global conclusions drawn based on the network.

One way of performing sensitivity analysis of Bayesian networks is to compute the conditional probabilities of a target node in the network when some evidence values of nodes are available and to observe how sensitive these conditional probabilities are to small changes in the parameters or evidence values. Of course not all parameters are equally sensitive since they all have different effects on the network’s performance.

In a one-way sensitivity analysis, the values of a single parameter are changed one at a time to compute the conditional probabilities of a target node, keeping the values of all other parameters fixed. Single parameter changes are easy to compute and visualize the effects in the network.
3 INSTRUMENTAL CHARACTERIZATION OF DIGITAL PRINTING

In this chapter we introduce a set of standard and well-known and another set of less standard and well-known but otherwise relevant measures which were used in the experiments conducted in the project. The standard measures are only briefly described as they can be found from literature. The more exotic measures are also available in literature, but have also been further developed during the course of work. These measures are briefly reviewed here and full details can be found from the corresponding scientific publications by the authors.

3.1 Relevant Physical Characteristics of Paper

Physical properties of paper are typically divided into 7 groups: basic (e.g., thickness), strength (e.g., tensile, bursting and surface strength), stiffness (e.g., elasticity), structural (e.g., air permeance), surface (e.g., roughness), absorption (e.g., amount of water absorbed by the paper) and optical (e.g., gloss). The related measurements can be implemented by mechanical or optical devices, and they have been shown to contribute to print quality.

Grammage is the mass of a square meter area of paper. ISO brightness measures the reflectance of blue light with a filter peak transmission at 457 nm. Calculation of whiteness considers the entire visible spectrum. Whiteness consists of both lightness and hue. Measurement of whiteness is standardised in CIE 145. Opacity measures the ability of paper to obstruct the passage of light. If opacity is too low, printed text is visible on the opposite side of the paper. Opacity is commonly determined by the ratio of perceived lightness (Y value in CIE xyY colour space) of a single sheet against a black background and Y value of the paper. Gloss measures the shininess of paper surface. In the paper industry, the specular reflectance at 75° angle is a common measure for gloss. Permeability of paper measures the flow of air through a defined area of the paper. Roughness of paper is usually defined by measuring the ability of the paper surface to resist an air stream flowing between the paper surface and a measuring head pressed against it. An alternative roughness measure is to study the paper profile.

In this project we have examined the relationship between the small-scale 2D measurements of the reflectance of printed paper and the surface topography (i.e., 2D profile) of unprinted paper. This is not a true reflectance measurement, as described in Rance /58/, but rather a photographic image of the paper surface. The pixel size is 10 µm by 10 µm. Due to the non-deterministic nature of the dependences we apply the conditional probability distributions as models of measurements. This probabilistic approach naturally connects to the Bayesian network idea of analyzing the print quality. Whereas the dependence between surface topography and print reflectance is typically weak in the total imaged area, significant dependences can be found in the points that are responsible for the tails of the probability density functions (pdfs) of the measured variables. Results with newspapers and SC paper are reported in Mettänen et al. /98/ and /99/, respectively. Although these analysis methods have been developed and tested only with offset and gravure printed paper samples, the question that the methods try to answer is common to digital printing as well: How does the observation of an exceptional (topography) point on the unprinted paper change our information about whether the print quality attainable at that point will be exceptional or not?
3.2 Test Target Measurements

3.2.1 Standard Measurements

Test target measurements carried out at TKK included the density of full-tone black, RMS noise and colour gamut. These are some of the standard assessments measured using standard tools. The absolute density of full-tone black was measured with a Macbeth RD918 densitometer. The average of the six measurements was calculated. RMS noise of a 50% gray field in the test layout was measured by digitizing a 30 mm × 30 mm area of a K50 field and then computing the pixel-based standard deviation in the lightness component of the digitized area. Colour gamut was determined by first measuring La*b* values of full-tone RGBCMY patches with the i1 Pro spectrophotometer. The average of the six measurements was calculated. Finally, the colour gamut was defined as the a*b* area of the hexagon formed when the La*b* values were plotted in an a*b* chart.

In the following sections we describe a set of more exotic print quality measures which require special computing (machine vision and image processing). All the following measures are measured from some standardized print pattern and therefore their measurement can be automated (an automatic measurement system was implemented in the project).

3.2.2 Printing Dot Properties

Computational print quality measures are motivated by the fact that most printed products are generated by printing dot patterns (halftoning). If a human evaluator cannot see the unevenness of plain paper, and the effects of lightning and image content are not taken into account, then the observer's experiences induced by the print are caused by the printing dots. Therefore, it is theoretically possible to estimate the human perception from the properties of individual dots. Furthermore, if the printing process can be considered as sufficiently stable, the dots within a small printed region can represent the average behaviour of the whole print. The dot measures are based on geometrical, edge and size properties of the print dots. The measures are based on an ideal dot which is defined as a sharp and perfectly round dot without any extraneous spots (satellites) outside its area.

3.2.3 Line and Edge Features

Analogous to printing dots, edges of lines also become distorted in printing. An ideal edge of a printed line is assumed to be sharp, even and not spread. Measures to estimate the line quality can be formed based on these requirements. Two line quality measures, edge raggedness and edge sharpness, were used in the experiments.

Edge raggedness is determined by the displacement of a black-white boundary from an ideal boundary. The black-white boundary is a measured boundary of a printed line. The ideal boundary is determined by calculating the best fit line through boundary points. Edge raggedness can be defined as the standard deviation of displacements of a measured boundary.

An edge becomes smoother when ink or toner is spread. Sharpness is a line quality measure which is quantified by analysing the profile of the black-white boundary itself. Sharpness can be determined by measuring the average distance from the dark area to the background.
3.2.4  Colour Features

When a colour is printed on two papers with different physical properties (e.g., gloss or whiteness), the perceived colours are not identical. In theory, this can be avoided by using colour profiles defined for the papers. In practice, however, different papers have different gamuts (a subset of colours a paper can reproduce with the available inks) and variation within a single paper grade makes accurate profiling impossible. The colours of an image printed on two different papers will, therefore, not match despite the use of profiling. The perceived colours have significant influence on the experienced naturalness of the image, and thus, on the whole quality experience.

CIE 1976 L*a*b* colour space is designed to be perceptually linear meaning that the colour change in the colour space corresponds to the perceived colour change. The Euclidean distance in the CIE 1976 L*a*b* space denoted as $\Delta E$ has a high correlation with the perceived colour difference.

3.2.5  Mottling

Mottling is defined as the unwanted unevenness of solid print. It is a subjective measure, and therefore, hard to define accurately. There are several known types of mottling, but the reason behind the problem is still not fully understood. The mottle size range is usually from 0.5 mm to 10.0 mm. If the diameter of a mottle spot is smaller, it cannot be seen with the naked eye from a usual reading distance, and if it is larger, it does not disturb or it is considered as a different problem.

Several methods for automated mottling evaluation have been proposed. These methods use different image processing methods to evaluate mottling. Some of these methods only sort samples according to their mottling level, but others provide more detailed information. In the project, mottling was computed separately for L*, a* and b* colour components using the band pass method and the final mottling index was defined as $M = (M_{L*}^2 + M_{a*}^2 + M_{b*}^2)^{1/2}$. 
4 COMPUTATIONAL CHARACTERIZATION OF DIGITALLY PRINTED IMAGES

Previous efforts to automate print quality evaluation using computer and machine vision have focused on automating the current manually-performed assessments, or measuring single quantities closely connected to the mechanics of printing. For example, the KDY ImageXpert method examines selected parts of a printed test pattern and returns various values, such as roundness of a dot, or edge raggedness of a line /81/. The ISO 19751 standard /6/, currently under development, proposes the following characteristics be measured: micro-uniformity, macro-uniformity, colour rendition, text and line quality, gloss, sharpness, and spatial adjacency.

The above-mentioned and other similar methods measure technical quantities and artefacts, but do not usually consider how humans actually experience quality. The proposed measures can be inspired by the human visual system (HVS) and model some parts of it, such as the automatic evaluation of subjective unevenness in solid printed regions /e.g., 1,36,60,63,68/. Such methods, however, measure visual quality experience only partially. In digital image quality assessment research, several quality metrics have been presented to measure the overall quality of images in a way that it is consistent with the subjective human evaluation. The image quality assessment algorithms can be classified into full-reference (FR), no-reference (NR) and reduced-reference (RR) metrics based on the availability of an original image, a reference. The following sections present the three types of IQ metrics in more detail.

4.1 Classification of Image Quality Metrics

The classification of an image quality metric can be based on the availability of a reference, which is considered to be of perfect quality or free of distortions. Most image quality metrics assume the availability of an original /74/. These metrics are called full-reference image quality metrics. A model that requires both the original and the processed image to make the quality prediction can also be called a double-ended image quality model /46/.

According to Martens /46/, a significant part of developed image quality metrics are based on two images as inputs, the other one being the original. For the comparison, the images must be perfectly aligned. Typically a measure of the perceived distance is pursued. It is often assumed that the distance from the original image measures distortion, and that image quality decreases linearly with the distance from the original. FR image quality metrics typically estimate image similarity or fidelity, the measurement of which requires an original reference image /41,65/.

In many cases, there is no reference image available for quality assessment /74/. The metrics developed for this kind of ‘blind assessment’ are called no-reference image quality metrics, or single-ended image quality models /46/. A single-ended image quality metric based on multidimensional analysis of image quality needs to address three problems: deciding the distortions to be modelled, developing methods to estimate the distortions with the help of the distorted image only, and determining the model for image quality as a combination of the individual distortions.

NR image quality assessment attempts to imitate the quality assessment of a human observer. The problem is how can a metric assess the quality of an image and estimate the degradation in image quality?
information without understanding it. A slightly different philosophy has been proposed to simplify
the question at hand /65/. If all images are considered perfect, regardless of content, unless distorted
during acquisition, processing, or reproduction, the task of a NR metric is reduced to blindly
estimating the distortion and then calibrating the estimated value against the quality judgments of
human observers.

In the third type of image quality assessment, the reference image is partially available. Certain
features are extracted from the reference image and employed as side information to help the
quality assessment. This is referred to as reduced-reference image quality assessment /74/. A
diagram of a RR image quality assessment system is presented in Figure 10.

\[ Figure 10 \quad A \text{ diagram of a reduced-reference image quality assessment system /74/}. \]

RR metrics provide a trade-off between NR and FR methods, as they are good tools for applications
where only restricted information is available from the reference image. It does not require full
access to the reference image; it only needs a set of extracted features. The amount of extracted
feature information depends on the application type and can be optimized by finding the most
relevant features of the application /27/.

RR metrics have been used both for distortion specific measurements, like NR metrics, and for
overall quality measurements, like FR metrics. In addition RR metrics have been used for
measuring the overall quality by summing up the values of the distortion specific metrics in a single
value.

### 4.2 Full-reference Image Quality Metrics

#### 4.2.1 Introduction

Several approaches for full-reference evaluation have been proposed: mathematical distance
formulations (e.g., PSNR and fuzzy similarity measures, HVS-inspired models (e.g., Sarnoff
JNDmetrix), HVS models combined with application specific modelling (e.g., DCTune), structural
approaches (structural similarity metric), and information theoretic models (visual information
fidelity). FR image quality metrics are commonly used in investigating the effects of image
compression, data transmission, and display technology. These results are, however, very specific to
the application, their subjective test material is limited, and most importantly, the results are not
fully transferable to be applied as a FR overall image quality metric in print quality evaluation.

In Section 3 we described a set of technical and computational (image processing based) quality
measures which can be applied to technical test fields. In this section we describe their alternatives,
i.e. quality measures based on free form natural images. Full-reference metrics have been developed
in the context of digital images, but during the project a complete framework /see Section 4.2.5/
was developed where these measures can also be applied to printed images. In the experimental part of this work we point out how these measures are well in accordance with the other more adopted print quality measures and are even good candidates to replace some of them.

Several approaches to develop full-reference (FR) image quality metrics have been proposed. Generally, FR metrics can be divided into two groups: arbitrary signal fidelity criteria and HVS error-based methods /66/. The first group mainly contains mathematical distance formulations that are applied to image quality assessment hoping that they correlate well with perceptual quality. The second group consists of computationally heavy methods that model the HVS.

### 4.2.2 Arbitrary Signal Fidelity Criteria

Several mathematical distance formulations that compute similarity or dissimilarity between two matrices (images) have been proposed. The most widely used metrics are the mean squared error (MSE) and peak signal-to-noise ratio (PSNR). These methods have several advantages: they are computationally efficient and have a clear physical meaning.

The universal quality index (UQI) was introduced by Wang and Bovik /73/, and a further improvement was presented in Wang et al. /78/ in the form of the structural similarity metric (SSIM). The basic idea of these metrics is to measure the loss of image structure, i.e., pixels near to each other have strong dependencies which carry information about the structure of the objects in the visual scene. The HVS is assumed to be highly adapted to structural information /77/, and structural distortions should be treated in a different manner than distortions arising from variations in lightning, such as brightness and contrast changes.

### 4.2.3 HVS Error-Based Methods

A distorted image can be divided into an undistorted reference signal and an error signal. A typical HVS image quality metric is based on the assumption that the loss of perceptual quality is directly related to the visibility of the error signal. These metrics operate by weighting different aspects of the error signal based on their visibility. The approach was first introduced by Mannos and Sakrison /45/. Other popular HVS error based methods are the Visual Difference Predictor (VDP) by Daly /11/ and the Sarnoff Visual Discrimination Model /43/.

A typical scheme for computing HVS error-based metrics consists of the following steps: pre-processing, contrast sensitivity function (CSF) filtering, channel decomposition, error-normalisation and error pooling /78/. The pre-processing step includes, for example, colour space transforms and low-pass filtering to simulate the point-spread function of the eye optics. In the CSF filtering step, the image is weighted according to the sensitivity of the HVS to different spatial and temporal frequencies. In the channel decomposition step, the image is separated into sub bands (channels) using, for example, the discrete cosine transform or a wavelet transform. In the next step, the error (the difference between the reference and input image) is computed for each channel and weighted to convert the errors into units of just noticeable difference (JND). Finally, the errors in different channels are combined into a single scalar using, for example, the Minkowski distance /78/.

The HVS model of VDP /11/ is a typical example containing three main steps: amplitude non-linearity, CSF weighting and a series of detection mechanisms. First, each image is passed through a non-linear response function to simulate the adaptation and response of retinal neurons. Second, the images are weighted with the CSF in the frequency domain and converted to local contrast.
information. Next, the images are split into 31 channels (5 spatial frequency bands combined with 6 orientation bands and 1 orientation-independent band) and transformed back to the spatial domain. A masking function is applied to each channel separately, and finally, error pooling is performed to sum the probabilities of visible differences in all channels to a single map of detection probabilities, which characterises the regions in the input image that are visually different from the original image.

There are several limitations on the HVS error-based metrics /78/. In brief, it is not clear that the fundamental assumptions of the HVS error-based metrics, i.e., the error visibility is equal to the loss of quality, and that the vision models derived from psychophysical experiments using simple test patterns are generalisable to image quality assessment of complex natural images, are correct.

### 4.2.4 Image Quality Metrics Using Natural Scene Statistics

A third way to approach the quality assessment problem is the statistical viewpoint. Natural scene statistics (NSS) refers to the statistical properties of natural images as a distinction to the statistics of artificial images such as text, paintings or computer generated graphics. It is plausible that the biological evolution of the HVS has been controlled by adaptation to natural environments, and thus, modelling the NSS and the HVS are essentially dual problems /66/.

The information fidelity criterion (IFC) based on the NSS was introduced in Sheikh et al. /66/. In the criterion, quality is evaluated by using the NSS and distortion models to find statistical information shared by the original and input images. The NSS model used is a Gaussian scale mixture /72/ in the wavelet domain. Visual information fidelity (VIF) is an extension of the IFC /64/. Since the NSS and HVS modelling are assumed to be dual problems, some parts of the HVS are already involved in the NSS model of IFC. However, e.g., contrast sensitivity and the point spread function are missing. In the VIF, the HVS model is added to include these aspects.

### 4.2.5 Computing FR Metrics from Printed Images

When the quality of a compressed image is analysed by comparing it to the original (reference) image, the FR metrics can be computed in a straightforward manner, cf., computing "distance metrics". This is possible because digital representations are in correspondence, i.e., there are no rigid, partly rigid or non-rigid (elastic) spatial shifts between the images, and the compression at least should retain photometric equivalence. This is not the case with printed media, however. In modern digital printing, a digital reference exists, but the image data undergoes various irreversible transformations, especially in printing and scanning, until another digital image for the comparison is established.

The first important consideration is related to the scanning process. Since we are interested in print quality instead of scanning quality, the scanner must be an order of magnitude better than the printing system. Fortunately, this is not difficult to achieve with the top-quality scanners available, in which the sub-pixel accuracy of the original can be achieved. It is important to use sub-pixel accuracy because it prevents the scanning distortions from affecting the registration. Furthermore, to prevent photometric errors, the scanner colour mapping should be adjusted to correspond to the original colour map. This can be achieved by using scanner profiling software accompanying the high-quality scanners. Secondly, a printed image contains halftone patterns, and therefore, descreening is needed to remove the high halftone frequencies and form a continuous tone image comparable to the reference image. Thirdly, the scanned image needs to be accurately registered.
with the original image before the FR image quality metrics or dissimilarity between the images can be computed. The registration can be assumed to be rigid since non-rigidity is a reproduction error and partly-rigid correspondence is avoided using the high scanning resolution.

Based on the discussions above, it is possible to sketch the main structure of our framework. The framework structure and the data flow are illustrated in Figure 11. First, the printed halftone image is scanned using a colour-profiled scanner. Second, the descreening is performed using a Gaussian low-pass filter (GLPF) which produces a continuous tone image. To perform the descreening in a more physiologically plausible way, the image is converted to the CIE L*a*b* colour space in which the colour channels are filtered separately. The CIE L*a*b* spans a perceptually uniform colour space and does not suffer from the problems related to, e.g., RGB, where the colour differences do not correspond to the human visual system. Moreover, the filter cut-off wavelength is limited by the printing resolution and should not be higher than 0.5 mm, which is the smallest detail visually disturbing to the human eye when the unevenness of a print is evaluated from the viewing distance of 30 cm (in ideal conditions the acuity limit of the human eye may be as small as 0.017° which corresponds to 0.1 mm). To make the input and reference images comparable, the reference image needs to be filtered with an identical cut-off wavelength. The colour profiling of the scanner provides a "photometric registration" and the descreening a "physiological registration" - in the end, a spatial registration is needed.

**Figure 11** The structure of the framework and data flow for computing full-reference image quality metrics for printed images.
Rigid image registration has been considered as a difficult problem until the invention of general interest point detectors and rotation and scale invariant descriptors. These methods provide an unparametrised approach to finding accurate and robust correspondence essential for the registration. The most popular method which combines both the interest point detection and description is David Lowe's scale-invariant feature transform (SIFT). The registration consists of the following stages: i) extracting local features from both images, ii) matching the features (correspondence), iii) finding a 2-D homography for the correspondences, and finally iv) transforming one image into another.

In the case of printed image quality assessment, FR image quality metrics have some special requirements. Although the above-mentioned registration works well, sub pixel errors do occur. Because of this, simple pixel wise distance formulations, such as the root mean square error (RMSE), do not work well. In other words, a good FR metric should not be sensitive to such small registration errors. A more notable problem emerges from the subjective tests which are carried out using printed (hardcopy) samples while the reference (original) image is in digital form. As a consequence, the reference image cannot be taken into the subjective evaluation and the evaluators do not usually see the actual reference. Therefore, a good FR image quality metric should not just compute simple similarity between the reference image and the input image, but also take into account the HVS.

4.3  No-Reference Image Quality Metrics

4.3.1  Introduction

As explained in Section 4.1, no-reference methods have typically been designed for a specific distortion. Examples of image quality metrics for specific attributes can be found for sharpness in Marziliano et al. /47/, Ferzli and Garam /20/, and Crete et al. /10/, for noise in Olsen /51/, Immerkær /33/, and Tai and Yang /69/, for contrast in Peli /54/, Bex and Makoux /4/, and Frazor and Geisler /23/, and for colourfulness in Fedorovskaya et al. /19/, Yendrikhovskij et al. /83/, and Hasler and Süsstrunk /28/. The following sections present some metrics for computing sharpness, noise, contrast and colourfulness.

4.3.2  Sharpness

The computation of sharpness has been widely studied; while some methods estimate sharpness in the spatial domain, some work in the frequency domain /38/. As image sharpness depends predominantly on the luminance content of an image instead of chrominance /31/, sharpness is usually computed from the luminance component. According to Katajamäki and Saarelma /38/, the common principle for the spatial domain techniques is to locate the edges in the image, compute a local sharpness value for each edge and finally combine these values to obtain a global sharpness value for the image.

Katajamäki and Saarelma /38/ note that due to technical and preferential reasons sharpness is often non-uniform within an image. They assume that the best focused region in an image is the basis for quality evaluation and suggest concentrating on the maximally sharp image regions. The local sharpness measure proposed by Katajamäki and Saarelma is the maximum edge gradient divided by the total edge height. The global sharpness of an image is defined as the global maximum of the local sharpness. Under reasonable assumptions, the measure is equivalent to the one-dimensional integral of the modulation transfer function.
Marziliano et al. /47/ proposed a no-reference blur metric based on the smoothing influence blur has on edges; the algorithm measures the spread of the edges in an image. Based on observations, they state that computing blur is sufficient along the vertical edges. First, the vertical edges in an image are detected by applying an edge detector, e.g. a Sobel filter. Then, each row of the image is scanned. An example of a row scan is presented in Figure 12. For each edge, the local extreme locations (minimum and maximum) closest to the edge are defined as the end and start positions of the edge. The edge width is then computed as the difference in pixels between the end and start positions and identified as the local blur measure. Finally, by averaging the local blur values over all vertical edges in the image, a blur measure for the whole image is obtained; this is to say the average edge width in the image in pixels. Marziliano et al. computed blur from the luminance component Y.

![Figure 12](image) A scanned row in an image. The points of detected edges are indicated by the green dashed lines and the points of local minima and maxima around the edges by blue dotted lines. E.g. the edge width at P3 is P4’-P4 i.e. 177 pixels – 171 pixels = 6 pixels /47/.

Crete et al. /10/ criticized edge detection based blur metrics such as the method of Marziliano et al. /47/ by stating that the metrics are sensitive to the threshold used to classify the edge as well as the presence of noise, which can mislead the edge detector. Crete et al. proposed a metric based on the discrimination between different levels of blur perceptible on the same image. The original image is blurred with a low-pass filter to get a comparison image with more blur. The low-pass filter is strong to guarantee that at least one of the images is perceived blurry by human observers. From both images, the variation in the neighbouring pixels is analyzed. A high variation between the original and the blurred image indicates a sharp original image, whereas a small variation indicates a blurry original.

The metric by Crete et al. /10/ follows three principles. Firstly, an image containing a small blurred section over a homogeneous area is perceived blurry, even if only a minor part of the image is blurred /Figure 13/. The metric takes this into account and includes only the pixels that have been altered during the blurring step. Secondly, blurring a sharp edge on a homogeneous area can create variations in the neighbouring pixels that did not exist in the original image. Due to this, the metric considers only the variations that have decreased during the blurring step. Thirdly, the metric
estimates blur in the horizontal and the vertical directions. The more annoying of these two blur values is considered as the final blur.

![Sharp or Blur?](image)

*Figure 13* A small blurred text over a large flat area /10/.

Crete et al. /10/ conducted a subjective experiment to validate the feasibility of the metric. In the tests, 15 non-expert participants evaluated twice a set of 132 images consisting of 12 original scenes and 120 modifications of the images with randomly applied levels of degradation by Gaussian filters, averaging, motion filters and an average disk filter. The metric achieved a correlation of 0.92 with the subjective assessments. Crete et al. analyzed also whether the luminance component is sufficient for estimating blur. The tests together with their earlier findings confirmed that the analysis of chrominance is almost insignificant and that blur can be estimated from the luminance channel alone.

### 4.3.3 Noise

Olsen /51/ reviewed and evaluated six methods for estimating the amount of noise in images. He assumed that the estimated noise was additive, stationary and had a mean of zero. Such noise can be determined by

\[
I(x,y) = f(x,y) + n(x,y),
\]

(6)

where \(I(x,y)\) is the image observed, \(f(x,y)\) the ideal image, and \(n(x,y)\) the noise.

Olsen /51/ evaluated the ability of the six methods to estimate the standard deviation of \(n\). The problem in estimating the amount of noise in images lies in the real image structure. The evaluated methods use two different approaches to solve the problem. Some of the methods attempt to filter off the structure in the observed image e.g. by applying the average filter (computes the sum of all pixels in the filter window and divides the sum by the number of pixels in the window) or the median filter (replaces the centre value of the filter window with the median value of all pixels in the window). Some methods, on the other hand, compute the noise of the observed image in regions that contain only little structure, e.g. one method computes the standard deviation from the average of the smallest variances measured in a set of image blocks.

Olsen found to his surprise that the methods based on pre-filtering an image performed better than the block-based methods. The most successful of the methods was based on subtracting the average filtered image from the observed image. The method utilizes an adaptive threshold value to discard too large intensity gradients. This way, the contribution of image edges to the noise estimate is avoided. In his experiment, Olsen also discovered that the ranking of the methods depended on the image content and the range of the standard deviation values.

Immerkær /33/ proposed a method for estimating similar additive zero-mean Gaussian noise in gray-scale images as defined in Equation 6. The method is based on a zero-mean operator, which is almost insensitive to the real structure of the image. Immerkær suggests using the difference
between two masks $L_1$ and $L_2$, both approximating the Laplacian of an image. The noise estimation operator $N$ with a zero mean is the mask operation using the mask

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

(7)

where $L_1$ is the mask

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

and $L_2$ mask

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

The variance of the operator output i.e. the noise variance in the observed image is estimated with the equation

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

(8)

where $W$ is the width of the image, $H$ the height of the image, $I(x,y)$ position $(x,y)$ in image $I$, and $N$ the applied mask. Immerkær reported that the metric performed well for a large range of noise variance values, but in textured images and regions it perceived thin lines as noise.

Katajamäki and Saarelma /38/ state that the assumption of noise being additive, zero-mean, Gaussian and spatially uncorrelated, i.e. white, is common due to its mathematical simplicity. They note that methods that estimate white noise based on filtering actually measure the variance of the filtered noise, so the methods may estimate perceptual noise improperly when the whiteness assumption of the noise does not hold. According to Katajamäki and Saarelma this is also the case for the method of Immerkær /33/, although the method performs very accurately when the noise is in fact white.

Tai and Yang /69/ state that to date the proposed noise metrics are still filter-based, block-based, or a combination of these two approaches. They propose a metric that has been further developed from the metric of Immerkær /33/. Tai and Yang criticize the metric of Immerkær for failing with certain image types with low noise levels and perceiving thin lines as noise for highly textured images. They propose an adaptive edge detector prior to the Laplacian convolution to improve the metric. For edge detection, Tai and Yang suggest the use of a Sobel operator

$$G = |G_x| + |G_y|,$$

(9)

where $G_x$ is defined as $I(x,y) * \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ and $G_y$ as $I(x,y) * \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}.$

The metric of Tai and Yang /69/ defines the edge map with an adaptive threshold parameter. First it computes the histogram of $G$, then it determines the threshold value as the value of $G$, when the accumulated histogram reaches $p$ % of the whole image. Even though $p$ is the same for all image contents, the threshold value adapts to each image separately. Tai and Yang as well as Olsen /51/ have used the value of 10 as $p$ for defining an adaptive threshold. After edge detection, the metric of Tai and Yang follows the procedure of Immerkær /33/.
To test the performance of their metric, Tai and Yang /69/ compared it to the metric of Immerkær /33/ by evaluating six natural image contents with several levels of added noise. They computed an estimation ratio by dividing the estimated noise with the added noise and found that while the metric maintained the good performance for high noise levels, it also made significant improvement on the low noise levels. The metric performed well also with images with fine details and textured regions. These were the most problematic situations for the metric of Immerkær /33/.

4.3.4 Contrast

For measuring contrast from test targets, two common contrast definitions exist, the contrasts of Michelson and Weber /54/. The Michelson contrast is defined as

\[ C_M = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{max}} + L_{\text{min}}}, \]  

where \( L_{\text{max}} \) is the maximum luminance and \( L_{\text{min}} \) the minimum luminance in the image /4,48,54/. The Michelson contrast value ranges from \([0, +1]\) /54/. While the Michelson contrast is an excellent metric for contrast in images with periodic luminance profiles, such as sinusoidal or square-wave gratings, it does not suit isolated stimulus elements such as single bars on a uniform background /48/. In such cases, the Weber fraction definition of contrast is commonly used. The Weber contrast is determined by

\[ C_W = \frac{\Delta L}{L}, \]  

where \( L \) is the uniform background luminance, and \( \Delta L \) the increment or decrement in the target luminance from \( L /48,54/. The Weber contrast value ranges from \([-1, +\infty]\) /54/.

For complex images with varying luminance, neither the Michelson contrast nor the Weber contrast is suitable, because they do not take the luminance distribution into account /48/. Absolute measurement of contrast with e.g. Michelson definition is easily misguided, as one or two points of extreme brightness of darkness can determine the contrast of the whole image /54/. The root-mean-square (RMS) calculation of image contrast is a common way to compute the contrast so that the contrasts of different images can be compared. Generally, \( C_{\text{rms}} \) is defined the standard deviation of the luminance values in an image

\[ C_{\text{rms}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \]  

where \( x_i \) is a normalized gray-level value such that \( 0 \leq x_i \leq 1 \), and \( \bar{x} \) the mean normalized gray level /4,48,54/. Moulden et al. /48/ tested the RMS contrast metric with five other candidates, including Michelson and Weber contrast, for the purposes of evaluating contrast in random-dot images on a monitor. Their results suggested that the standard deviation of luminance was preferable to the other candidate metrics.

Peli /54/ states that as human contrast sensitivity is a function of spatial frequency, a contrast metric attempting to predict perceived contrast should consider the spatial frequency content of an image. As this is not the case with RMS contrast or the other metrics presented above, Peli proposed a local band-limited contrast metric that assigns a contrast value for every point in the image as a function of the spatial frequency band. By definition, for every band-pass filtered image \( a(x,y) \), the
corresponding local luminance mean image $l(x,y)$ is the low-pass filtered version of the image containing all the energy below the band. The contrast is determined as a two-dimensional array $c(x,y)$ with the equation

$$c(x,y) = \frac{a(x,y)}{l(x,y)},$$

where $a(x,y)$ is the band-pass filtered image and $l(x,y)$ the corresponding local luminance mean image.

Bex and Makous /4/ compared the methods of Michelson contrast, RMS contrast and the band-limited contrast by Peli /54/ with a selection of natural images in subjective tests. Three observers took part in the test, one of which was the first author. Due to the unique luminance distributions of the images, each image had a unique contrast for each of the three metrics tested. The metric that agreed best with the observers’ opinions was the RMS contrast.

E.g. Frazor and Geisler /23/ chose to use the RMS contrast for contrast computation as it is a standard measure, and it predicts human contrast detection thresholds quite well for both natural scene patches and laboratory stimuli. The use of local RMS contrast defined as the standard deviation of luminance divided by the mean luminance in a small neighbourhood has also been reported /25/.

### 4.3.5 Colourfulness

Colourfulness is the attribute of visual sensation according to which perceived colour of an area appears to be more or less chromatic; colourfulness can also be referred to as chromaticness /18,82/. Fedorovskaya et al. /19/ proposed that colourfulness depends on the distance of image colours from neutral gray and the distances between individual image colours. The first variable is estimated with the average chroma and the second with the standard deviation of the distribution of the chroma values of all individual image colours. The colourfulness of an image is computed as the weighted sum of the average chroma and the standard deviation of chroma.

Yendrikhovskij et al. /83/ proposed a colourfulness metric that ignores lightness differences between images. The metric is based on the assumption that the average saturation of an image and its standard deviation are linearly related with the same weight to subjective colourfulness judgments. The colourfulness metric is defined by the equation

$$C_i = S_i + \sigma_i,$$

where $S_i$ is the average saturation of image $i$, and $\sigma_i$ the standard deviation of saturation of image $i$.

Yendrikhovskij et al. /83/ conducted a subjective experiment to evaluate the correlation between the colourfulness metric and subjective colourfulness evaluations. Eight participants evaluated a set of thirty digital images. The set included typical photograph categories such as portraits, landscapes, people in nature, natural objects e.g. flowers, and artificial objects e.g. planes. The metric achieved a correlation of 0.91 with the colourfulness judgments.

Hasler and Süsstrunk /28/ proposed three metrics for computing colourfulness. The first metric is defined as

$$M_{(1)} = \sigma_{ab} + 0.37 \cdot \mu_{ab},$$
where \( \sigma_{ab} \) is the trigonometric length of the standard deviation in a*b* space, and \( \mu_{ab} \) the distance of the centre of gravity in a*b* space to the neutral axis.

The second metric of Hasler and Süsstrunk /28/ is given by

\[
M^{(2)} = \sigma_{ab} + 0.94 \cdot \mu_c,
\]

where \( \sigma_{ab} \) is the trigonometric length of the standard deviation in a*b* space, and \( \mu_c \) the mean of chroma.

For the third metric, Hasler and Süsstrunk /28/ use a simple opponent colour space:

\[
rg = R - G, \quad \text{and} \\
yb = \frac{1}{2}(R + G) - B,
\]

where \( R \) is the red channel, \( G \) the green channel, and \( B \) the blue channel of the image.

The image is assumed to be coded in the sRGB colour space. The third colourfulness metric of Hasler and Süsstrunk /28/ is determined by

\[
M^{(3)} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3 \cdot \sqrt{\mu_{rg}^2 + \mu_{yb}^2},
\]

where \( \sigma_{rg} \) is the standard deviation along the \( rg \) axis, \( \sigma_{yb} \) the standard deviation along the \( yb \) axis, \( \mu_{rg} \) the mean along the \( rg \) axis, and \( \mu_{yb} \) the mean along the \( yb \) axis.

Palus /52/ analyzed the noise dependency of the third metric by Hasler and Süsstrunk /28/ with three digital images. One of the images was slightly colourful, one averagely colourful and one extremely colourful. Achromatic and chromatic impulsive noise was added to the images with several levels of degradation. Palus found that adding chromatic impulsive noise increased colourfulness, while adding achromatic impulsive noise decreased colourfulness. The increase in colourfulness was particularly noticeable in images with low colourfulness, whereas the decrease in colourfulness was visible for highly colourful images.

### 4.4 Reduced-Reference Quality Metrics

#### 4.4.1 Introduction

Reduced-reference metrics can be classified into three groups: baseline /41/, pixel block /26,79/ and image statistic /75,76/ methods. The RR feature of baseline methods is the NR metric value (baseline value) of the reference image which is used as a normalizing value for the outputs of distorted images. The RR features of pixel block methods are the local statistic values of the reference image. The RR feature of image statistic methods is the global statistic property of the reference image. This global statistic property can be parameters for probability function or some single statistic number of reference image.

The aim of the RR metric part of the project was to develop new distortion specific image content adaptive metrics for printed images. The proposed content adaptive RR metrics are dominant colour contrast, sharpness and graininess metrics. Dominant colour contrast metric can be classified into image statistic based method group. Sharpness and graininess metrics can be classified into pixel block based method group. Next motivation and function for proposed metrics are described in detail.
4.4.2 Dominant Colour Contrast

The goodness of colour reproduction is difficult to measure or describe. Colour reproduction is a preferential attribute. The higher colour reproduction accuracy of an image does not always mean higher performance. For example, saturated colours can increase the image quality compared to natural colours [12].

There are plenty of pixel-wise (FR) methods for calculations of colour quality. The simplest method is the colour error as Euclidean distance, $\Delta E$, in the CIELAB space. In reality, many factors have an effect on the perceived colour reproduction. For example, the S-CIELAB metric [84] takes into account the sensitivity of the human visual system to spatial frequencies. The FR metric of Hong and Luo [30] assigns a higher weight to dominant colours and to a colour having greater difference. Fairchild and Johnson [17] have proposed the most advanced model. Their iCAM model takes the environment parameters and the local properties of image into consideration when calculating image appearance.

When no reference image is available or it and the distorted image cannot be spatially registered, the colour quality measurement is substantially trickier. The measurement should be based on the statistics of the images and different assumptions. Hasler and Süsstrunk [28] proposed a NR metric which calculated the colourfulness of the image based on the deviation and mean values of the chromatic components in the CIELAB space. The metric assumed that the perceived colourfulness of the image correlates with mean and standard deviation of values in the chromatic plane.

We propose a RR colour metric which calculates the contrast of dominant colours in an image. We assume that natural images contain one or more objects with specific hues. The object is perceived more colourful and brighter if the contrast of its texture is high. If the contrast of the texture is low, the object is perceived pale or dim. The colours on high quality papers have higher dynamics than colours on low quality papers. The dynamics should be high enough in order, that the contrast can be high.

The proposed method calculates the first principal component from the image data in the CIELAB space [Figure 14]. The first principal component accounts for as much variability in the data as possible. Each succeeding component accounts for as much of the remaining variability as possible. With the aid of the principal component, two colour clusters are defined [Figure 15]. The first cluster includes the neighbourhood points of the principal component. The second cluster includes all the other points. The neighbourhood points of the first principal component should fulfil the conditions of equations (20) and (21):

$$\theta - \theta_{th} < \theta_{po int} < \theta + \theta_{th}$$  \hspace{1cm} (20)

$$\varphi - \varphi_{th} < \varphi_{po int} < \varphi + \varphi_{th}$$  \hspace{1cm} (21)

where $\theta$ and $\varphi$ are the angles of principal component and $\theta_{th}$ and $\varphi_{th}$ are the threshold angles. The threshold angles define the size of clusters.
When data is grouped into two clusters, the principal component is calculated also for cluster 2. The calculated principal components describe the directions of two main hues of image. For example, for the test image lake the principal component of the first cluster describes the colour contrast of the blue sea and sky and the principal component of the second cluster describes the colour contrast of the green foliage /Figure 15/.

The proposed method transforms the pixel values of distorted images to the space $S$. The axles of the space $S$ are composed of two principal components. The dominant colour contrast, $C$, is calculated by equation

$$C = \sqrt{\sigma_{p1}^2 + \sigma_{p2}^2},$$

(22)

where $\sigma_{p1}$ and $\sigma_{p2}$ are the standard deviations of the first and second dimension of space, $S$.

### 4.4.3 Graininess

Graininess of prints is often measured from simple smooth test field /40/. The measurement of graininess from natural images is a difficult task, because the graininess energy should be separated from the image structure.

Some methods have been proposed for the noise calculation of natural images. For example, Immerkær /33/ proposed a NR method which calculated the noise value using a $3\times3$ mask. One
problem of the metric is that it assumes the noise pixel to be a single and separate pixel in a smooth 3×3 pixel neighbourhood. Based on Immerkær /33/, the other problem of the metric is that it defines thin lines as noise pixels. The graininess of a printed image is also defined in the frequency and wavelet domains. Kane et al. /37/ characterized the graininess using the noise power spectrum (NPS). NPS defined the spectral graph, whose shape characterizes the graininess of image. Donohue et al. /13/ characterized the graininess in the wavelet domain. The method calculated the average wavelet coefficient energy for different image scales. The relationships between the wavelet coefficient of horizontal, vertical and diagonal directions also characterized the streaking effect. Based on Donohue et al. /13/, the method was sensitive to the structure of image and it was applicable only to the energy calculated from smooth surfaces.

We propose a graininess metric, $G$, which is based on the wavelet coefficient energy of the first scale of image. The use of the first scale was based on the assumption that the perceived graininess is high frequency energy. The method divides the image into blocks and calculates the energy value for every block. The number of low energy blocks, $n$, is calculated on the basis of the energy values of the reference image. The assumption is that the distorted image has some graininess energy and its quantity can be calculated from the $n$ blocks whose energy values are small. Figure 16 shows how low and high energy blocks locate in the images man (a), cactus (b) and lake (c). The low energy blocks are dark and high energy blocks are light. Figure 17 shows the graininess energy values (Equation 23) of the reference and the distorted images sorted in ascending order. The number of low energy blocks is located on the point where the slope of the reference image starts to increase.

![Figure 16](image.png)

*Figure 16* Graininess is calculated from the low energy pixel blocks (dark regions).
Figure 17 The number of low energy blocks is located on the point where the slope of the reference image starts to increase.

The probability distribution of wavelet coefficients of a natural image has a high peak and long tails. The proposed metric calculates the graininess values from the smooth regions of image. Based on this the wavelet coefficient values of noiseless images should be low and the probability distribution should have a high peak and short tails.

If the graininess of a noisy image is more of the gaussian type, the shape of the probability distribution is gaussian. If the graininess of a noisy image is more of the impulse type, the shape of the probability distribution is sparse and the values are mainly close to zero or large. We assume that the graininess or noise of printed images is more the impulse type than gaussian type and we use the power function for the graininess energy metric:

$$ \sum_{i=1}^{n} x_i^p, $$

where $x_i$ is a wavelet coefficient and $n$ is the number of wavelet coefficients. We assume that this kind of sparseness metric is more sensitive to graininess than the basic standard deviation. The parameter $p$ defines the weight of large coefficients compared to small coefficients. In this study value of $p$ was set to be 3. The proposed graininess metric, $G$, is calculated by Equation (12):

$$ G = 4 \left( \frac{g_{hi} + g_{dhi} + g_{vi} + g_{d2i}}{4} \right), $$

where $g_{hi}$, $g_{dhi}$, $g_{vi}$ and $g_{d2i}$ are mean graininess energy values of low energy block $i$ in horizontal, first diagonal, vertical and second diagonal directions and $n$ is the number of low energy blocks.

4.4.4 Sharpness

Sharpness has often been measured by calculating the blurriness value for the image. Many blurriness metrics have been based on edge finding methods. For example, Marziliano et al. /47/ proposed a NR metric which sought the edge pixels and calculated the blurriness value from the widths of the edges. Caviedes and Oberti /8/ proposed a NR metric which calculated the sharpness value based on the kurtosis of DCT values calculated from the neighbourhood pixels of edges. Tong et al. /70/ proposed the NR metric which sought different edge types using wavelet transformation. The method assumed that the non-sharp image can not contain any sharp edge type. Crete et al. /10/ proposed the NR metric which was based on the intensity differences between the original distorted
image and the blurred distorted image. The idea was to measure the filtering degree of details when
the image was low pass filtered. The metric assumed that blurriness is due to the attenuated high
frequency energy.

The graininess or noise of printed images is high frequency energy. The above listed sharpness
metrics can interpret the graininess or noise as edges or some other image structures. Because of
this, the sharpness value of the noisy image can be too high. The perceived sharpness is also related
to the reproduction of mid frequency energy. The reproduction of high frequency energy defines
more the reproduction of image details. An image can be perceived as sharp even if the image
would not reproduce the details. Respectively, an image can be perceived as un-sharp even if the
image would reproduce the high frequency energy.

Ferzli and Karam /21/ measured the sharpness using the wavelet-based metric. This NR metric uses
the third scale of image and sharpness value was based on the edge width of Marziliano et al. /47/.
Ferzli and Karam /21/ said that this metric is noise immune because the noise energy is filtered out
before the sharpness is calculated.

The proposed sharpness metric in this study is based on the wavelet coefficients of the second scale
of image. Because the metric uses the second scale and handles only the mid-frequency energy, it is
non-sensitive to high frequency noise. The metric divides the image into blocks and calculates the
energy values for blocks. The number of high energy blocks, n, is calculated based on the β
parameters of reference image. The β parameter is calculated by fitting the wavelet coefficients
from different directions of the second scale to Generalized Gaussian distribution model (GGD):

\[
p(x; \alpha, \beta) = \frac{\beta}{2\alpha \Gamma(1/\beta)} e^{-|x|^{\alpha} \beta},
\]

where \(\Gamma\) is the gamma function and \(\alpha\) is the alfa parameter. The special cases of the GGD model are
Gaussian (\(\beta=2\)) and Laplacian (\(\beta=1\)) distributions. When \(\beta<2\) the tails are heavier than normal and
when \(\beta>2\) the tails are lighter than normal.

Figure 18 shows β values for the images man (a), cactus (b) and lake (c). The low energy blocks are
light and high energy blocks are dark. Figure 19 plots the β values of the reference and the distorted
images sorted in ascending order. The number of high energy blocks is located on the point where
the β value of the reference image is smaller than 0.5.

\(\text{Figure 18} \quad \text{Sharpness is calculated from the high energy blocks (dark regions).}\)
The number of high energy blocks is located on the point where the beta value of reference image is 0.5.

The proposed sharpness metric, $S$, is calculated as

$$S = \frac{\beta_h + \beta_{d1} + \beta_v + \beta_{d2}}{4},$$

(26)

where $\beta_h$, $\beta_{d1}$, $\beta_v$, and $\beta_{d2}$ are mean $\beta$ values of blocks in horizontal, diagonal, vertical and diagonal directions.
5 VISUAL QUALITY MODELS – FROM OBJECTIVE MEASUREMENTS TO PREDICTING HUMAN ASSESSMENT

The most important research question in DigiQ project was how to link “what is measurable” to “what was experienced (observed)”, i.e. how to connect the technical and computational low- and moderate-level objective quality measures from Sections 3.1 and 3.2 or the reference and ad-hoc based metrics from Sections 4.2, 4.3 and 4.4 to what was subjectively experienced by the subjects in psychophysical assessments described in Section 2.

To seek for the connection between subjective and objective measurements we applied the full-reference metrics /Section 5.1/, standard regression approach /Section 5.2/, content adaptive approach /Section 5.3/, cumulative match score method /Section 5.4/ and Bayesian network approach /Section 5.5/. The Bayesian networks were applied using the layered structure /Sections 5.5.2/ and the structure optimisation approaches /Section 5.5.3/.

5.1 Full-Reference Metrics

5.1.1 Introduction

In this Section we show the performance of the full-reference metrics for measuring or modelling the image quality of printed natural images. The selected full-reference metrics and their basic information, such as whether a metric works with colour or only intensity, are listed in Table 16. The basic principles of the metrics are described more detail in Section 4.2. All metrics were computed using the default parameter values proposed by the authors of the metrics.

Table 15 FR image quality metrics used in this study.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Acronym</th>
<th>Type</th>
<th>Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak signal to noise ratio</td>
<td>PSNR</td>
<td>mathematical</td>
<td>no</td>
</tr>
<tr>
<td>Universal quality index</td>
<td>UQI</td>
<td>structural</td>
<td>yes</td>
</tr>
<tr>
<td>Structural similarity index</td>
<td>SSIM</td>
<td>structural</td>
<td>no</td>
</tr>
<tr>
<td>Information fidelity criterion</td>
<td>IFC</td>
<td>information theoretic</td>
<td>no</td>
</tr>
<tr>
<td>Visual information fidelity</td>
<td>VIF</td>
<td>information theoretic</td>
<td>no</td>
</tr>
<tr>
<td>Noise quality metric</td>
<td>NQM</td>
<td>HVS error-based</td>
<td>no</td>
</tr>
<tr>
<td>(High dynamic range) Visible difference predictor</td>
<td>(HDR-)VDP</td>
<td>HVS error-based</td>
<td>no</td>
</tr>
</tbody>
</table>

5.1.2 Methodology

The first set of samples consisted of natural images /see Figure 5/ printed with a high quality inkjet printer on 16 different paper grades (CC3). The second set of samples consisted of images printed with a production-scale electrophotography printer on 21 different paper grades (CC4). The printed samples were scanned using a high quality scanner with 1250 dpi resolution and 48-bit RGB colours. A colour management profile was devised for the scanner before scanning, and colour correction, descreening and other automatic settings were disabled from the scanner software. The digitised images were saved using lossless compression. The framework to compute the image quality from printed natural image is described more detail in Section 4.2.5.

Three performance measures were chosen for the comparison of the full-reference image quality metrics. The first performance measure was the linear correlation coefficient (CC), the second was
the Spearman rank order correlation coefficient (SROCC), and the third was the root mean square error (RMSE). The performance measures were calculated after the nonlinear regression with a monotonicity constraint:

\[
Q(x) = \beta_1 \left( \frac{1}{2} - \frac{1}{1 + \exp(\beta_2 x - \beta_3)} \right) + \beta_4 x + \beta_5, 
\]

where \( \beta_i \) are fitting parameters and \( x \) is metric value.

5.1.3 Results

The results are presented in Table 16. In Figure 20 some full-reference metrics are presented with regression curves. As can be seen from Table 16 the results clearly differ between the samples of CC3 and CC4. This is because the CC3 samples (ink-jet) were considerably easier to evaluate than the CC4 samples (electrophotography). Quality variation between samples is higher, and two or three compact clusters with distinctly different quality exist in the data /see Figure 20/. In general, the methods succeed since most full-reference metrics put these clusters in the right order increasing the correlations for CC3.

Table 16  Correlation coefficient (CC), Spearman rank order correlation coefficient (SROCC) and root mean square error (RMSE) between the IQM values and MOS after nonlinear regression for both test sets.
5.1.4 Conclusions

In samples of CC3, most of the selected metrics had high correlation coefficients. UQI, SSIM, IFC, VIF, NQM and VDP were statistically indistinguishable from each other (F-test). In samples of CC4, the group of well working metrics was reduced to UQI, SSIM, IFC and VIF, which were again statistically indistinguishable from each other. The other performance measures of metrics, SROCC and RMSE, support these conclusions. If a single full-reference metric should be selected, VIF would be a safe choice since it was shown to be the best in an earlier study on digital images /64/.

5.1.5 Further Analysis

Single number as a measure of quality is problematic since quality as a subjective experience of quality is not a stable one-dimensional perceptual quantity, but often multidimensional, qualitative subjective phenomenon that, however, can be described by using a number by the observer. This number is a product of quite complicated mental processes, involving detection, attention, interpretation and decision making. How useful is a single number as a reference for objective measurements?
Figure 21 shows a method that makes a more thorough examination of the objective quality metrics possible. It is based on the attribute dimensions, such as presented in Section 2.3. For each polarity of attribute dimension, both mean quality rating and mean objective metric value is calculated. If one attribute pole has high subjective rating and another pole low subjective rating, attribute dimension can be assumed to be related to quality. This should also be case with the objective measure. If no, objective measure may not take the attribute in question into account. Figure 21 shows the abovementioned examination with two full-reference metrics used in this study; PSNR and SSIM. PSNR metric is an example of very low correlation with subjective ratings. SSIM, on the other hand, fails only with few attributes, such as yellowishness.

![Figure 21](image)

(a) (b)

Figure 21 Full-reference metric vs. quality ratings (MOS), attribute-wise; PSNR (a), SSIM (b).

5.2 Regression Approach

5.2.1 Introduction

The study approaches the concept of subjective print quality with a five-layer hierarchy shown in Table 17. In the study, the subjective quality of printed photographs (Layer 5) is predicted using regression analysis with the instrumental measurements from prints (Layer 2), computational quality attributes (Layer 3) and subjective quality attributes (Layer 4) separately as predictors. The first two sets of models are used to evaluate the feasibility of instrumental measurements from prints (Layer 2) and respective computational quality metrics (Layer 3) for the overall print quality prediction. The last set of models with subjective quality attributes (Layer 4) as predictors are utilized in analyzing the need for discrete quality models for different printing methods, papers of different quality level, and different types of image content.
Table 17  Quality hierarchy (from left to right) considered in the regression approach.

<table>
<thead>
<tr>
<th>Layer 1</th>
<th>Layer 2</th>
<th>Layer 3</th>
<th>Layer 4</th>
<th>Layer 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrumental measurements from papers</td>
<td>Instrumental measurements from prints</td>
<td>Computational quality attributes</td>
<td>Subjective quality attributes</td>
<td>Subjective print quality</td>
</tr>
</tbody>
</table>

Factors

<table>
<thead>
<tr>
<th>Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brightness, whiteness, roughness, gloss etc.</td>
</tr>
<tr>
<td>Edge sharpness, RMS noise, K100 density, colour gamut etc.</td>
</tr>
<tr>
<td>Blur, noise, contrast, colourfulness etc.</td>
</tr>
<tr>
<td>Sharpness, graininess, contrast, colourfulness etc.</td>
</tr>
<tr>
<td>MOS</td>
</tr>
</tbody>
</table>

5.2.2 Methodology

Material

The series of samples used in the study combined the electrophotographic samples (CC4) and the ink-jet samples (CC4.5). The electrophotographic and ink-jet papers were further divided into groups of plain electrophotographic (EPG) and electrophotographic photo (EPG Photo), as well as ink-jet (IJ) papers and ink-jet photo (IJ Photo) papers. Papers marketed for printing high quality colour images were considered as photo papers. EPG and EPG Photo papers as well as IJ and IJ Photo papers were printed with their respective printing methods. Multipurpose (MP) papers were printed with both methods. Table 18 presents the variation in paper characteristics in the test series. The studied image contents were cactus, man and lake /see Figure 5/.

Table 18 The variation in paper characteristics in the test series. Papers of five different types were used: multipurpose (MP), plain electrophotographic (EPG), plain ink-jet (IJ), electrophotographic photo (EPG Photo) and ink-jet photo (IJ Photo) papers.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Whiteness (%)</th>
<th>Brightness (%)</th>
<th>Gloss (GU)</th>
<th>Roughness PPS (µm)</th>
<th>Roughness Bendtsen (ml/min)</th>
<th>Permeability Bendtsen (ml/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP (n = 6)</td>
<td>79 – 157</td>
<td>89 – 100</td>
<td>4.0 – 7.4</td>
<td>4.5 – 7.2</td>
<td>91 – 247</td>
<td>392 – 706</td>
</tr>
<tr>
<td>EPG (n = 8)</td>
<td>120 – 161</td>
<td>90 – 101</td>
<td>8.2 – 70</td>
<td>0.9 – 4.1</td>
<td>6.5 – 51</td>
<td>0 – 213</td>
</tr>
<tr>
<td>IJ (n = 8)</td>
<td>111 – 150</td>
<td>92 – 98</td>
<td>2.2 – 77</td>
<td>0.7 – 5.7</td>
<td>8.1 – 270</td>
<td>26 – 193</td>
</tr>
<tr>
<td>EPG Photo (n = 7)</td>
<td>107 – 148</td>
<td>91 – 95</td>
<td>15 – 87</td>
<td>0.7 – 3.4</td>
<td>0 – 122</td>
<td>0 – 57</td>
</tr>
<tr>
<td>IJ Photo (n = 7)</td>
<td>91 – 104</td>
<td>86 – 91</td>
<td>45 – 96</td>
<td>0.6 – 4.0</td>
<td>0 – 68</td>
<td>0 – 18</td>
</tr>
</tbody>
</table>

The printed images were digitized with a calibrated camera. The digitization plane was illuminated by a filtered halogen lamp. The illumination level on the plane was 2000 lx and the color temperature was 4800 K. The samples were captured with three exposures (1/10s, 1/5s and 2/5s). From these exposures a single 16 bits linear high dynamic RGB raw image was combined. The high dynamic RGB image was transformed first to the absolute XYZ space and then to the LAB space. In addition the photometric distortion of the camera was compensated by the uniform test field of the integrating sphere.
Instrumental Measurements

A series of instrumental measurements was carried out from the samples. The instrumental measures respective to the computational quality metrics were edge sharpness, rms noise, density of full-tone black and colour gamut. Edge sharpness values for the vertical lines of the printed samples were computed in the following manner. The edge profiles of the vertical lines were first estimated with a function

\[ f(x) = \frac{b_1}{1 + e^{-b_2x}} + b_3. \]  

(28)

Edge sharpness was considered as the maximum value of the function’s derivative and was then defined as the average value over the whole edge (a minimum of 15 data points). For more details about RMS noise, density and colour gamut see Section 3.2.

Computation of Quality Attributes

The following algorithms were used for the computation of quality in the study: Crete et al. /10/ for blur, Tai and Yang /69/ for noise, rms /e.g. in 54/ for contrast, and Hasler and Süssstrunk /28/ for colourfulness. The algorithms were chosen as they are well-documented, widely used and perform well in the context of digital images. For more details about the metrics see Section 4.3.

Subjective Tests

Overall quality along with sharpness, graininess, colourfulness and contrast were evaluated from the samples in subjective tests. Observers (n = 59) were university students and naïve as regards to print and image quality. The EPG samples were evaluated separately from the IJ samples: 29 of the subjects evaluated the EPG samples (tests carried out at HU) and 30 the IJ samples (tests at TKK). Both tests followed the same procedure for the results to be comparable.

The overall quality of the samples was evaluated on a 5-point category scale (i.e. 1: bad, 5: excellent). Sharpness, graininess, colourfulness and contrast were evaluated as semantic differentials (e.g. blurry–sharp) on a 5-point category scale (e.g. 1: clearly blurrier than sample set average, 2: slightly blurrier, 3: about the average, 4: slightly sharper, 5: clearly sharper). The scale of graininess was of opposite direction, i.e. grainy – non-grainy. The subjects were asked to reach their ratings to both ends of the scale.

All subjects rated first the overall quality of each sample in order to avoid any bias from evaluating the quality attributes. The samples of one image content at a time were placed on the table in a random order. The evaluation order of the contents was also randomized. After evaluating the overall quality of all the samples the subject was asked to rate the sharpness, graininess, colourfulness and contrast of the samples. The samples were again presented image content at a time; the subject rated the content on all scales before moving on to the next content. Again the order of the contents was randomized along with the order of the evaluated quality attributes.

Quality Modelling

Stepwise (backward) multiple linear regression analysis was used for creating the quality models relating instrumental measurements from prints (Layer 2), computational quality attributes (Layer 3) and subjective quality attributes (Layer 4) to MOS values (Layer 5). Altogether three sets of models predicting the subjective print quality were created with the Layers separately as predictors.
The goodness-of-fit of the produced models was evaluated with the coefficient of determination, $R^2$. SPSS 17.0 was used for the creation of the regression models and the assessment of their goodness. Each image content and printing method was analyzed separately.

### 5.2.3 Results

Table 19 presents the regression models predicting MOS by instrumental measurements (Layer 2) for the printing methods separately and combined. All models explained the variation in the overall quality well with coefficients of determination of over 75%. The models for the EPG and IJ samples separately had a better fit than the models for the printing methods combined. Of the attributes, edge sharpness and rms noise were of great importance in the models as they were included in six of the nine models, i.e. in all models for the EPG samples and for the samples of EPG and IJ combined. The models for the IJ samples contained only one predictor, colour gamut. This single variable explained over 94% of the variation in the overall quality for all three image contents. Overall, colour gamut appeared in five of the nine models, but had no contribution to the quality prediction of the EPG samples. The density of full-tone black appeared only in two models, both for image content lake. In all nine models, the coefficients of the variables had a logical sign, i.e. the coefficients were negative only for the rms noise. The variables included in the models seemed to depend both on the printing method and the image content. The image content with the models of the weakest explanatory power was lake.

**Table 19** The regression models predicting MOS by instrumental measurements (Layer 2) for the EPG and IJ samples separately and combined.

<table>
<thead>
<tr>
<th></th>
<th>All samples (MP + EPG + EPG Photo)</th>
<th>EPG samples (MP + EPG + EPG Photo)</th>
<th>IJ samples (MP + IJ + IJ Photo)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>$R^2$</td>
<td>Model</td>
</tr>
<tr>
<td>Cactus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.53</td>
<td>76.2 %</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>+0.41·Edge Sharpness</td>
<td></td>
<td>+1.47·Edge Sharpness</td>
</tr>
<tr>
<td></td>
<td>-0.48·RMS Noise</td>
<td></td>
<td>-0.91·RMS Noise</td>
</tr>
<tr>
<td></td>
<td>+0.36·Colour Gamut</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lake</td>
<td>1.36</td>
<td>77.3 %</td>
<td>-2.58</td>
</tr>
<tr>
<td></td>
<td>+0.36·Edge Sharpness</td>
<td></td>
<td>+1.06·Edge Sharpness</td>
</tr>
<tr>
<td></td>
<td>-0.47·RMS Noise</td>
<td></td>
<td>-0.90·RMS Noise</td>
</tr>
<tr>
<td></td>
<td>+0.43·K100 Density</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>1.45</td>
<td>79.7 %</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>+0.41·Edge Sharpness</td>
<td></td>
<td>+1.41·Edge Sharpness</td>
</tr>
<tr>
<td></td>
<td>-0.49·RMS Noise</td>
<td></td>
<td>-1.00·RMS Noise</td>
</tr>
<tr>
<td></td>
<td>+0.38·Colour Gamut</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 20 presents the regression models predicting MOS by computational attribute values (Layer 3) for the printing methods separately and combined. While the variation in the IJ print quality was explained well by the computational attribute values (coefficients of determination again over 94%), the models for the EPG samples were significantly weaker with only 29-59% of the variation explained. The models for the combined EPG and IJ samples had a reasonable explanatory power with coefficients of determination of over 70%. Of the attributes, contrast was the most significant as it appeared in eight of the nine models. Noise was included in five of the nine models and blur in four, whereas colourfulness did not contribute to the quality prediction in any of the cases. Clearly the colour metric used does not capture the right colour properties perceived as important by subjects. Contrast on the other hand was considered difficult to evaluate by the subjects. This could also be seen in the deviation of the contrast ratings between the subjects. In three of the nine models, there were coefficients with an unexpected sign. Similarly to the case of instrumental measurements as predictors, the variables of the quality models depended both on the printing method and the studied image content. With computational attribute values as predictors, image content lake had the models of the best explanatory power.

Table 20

The regression models predicting MOS by computational attribute values (Layer 3) for the EPG and IJ samples separately and combined.

<table>
<thead>
<tr>
<th>Model</th>
<th>All samples</th>
<th>R²</th>
<th>EPG samples</th>
<th>R²</th>
<th>IJ samples</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cactus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.43</td>
<td></td>
<td>-11.51·Blur</td>
<td>70.2%</td>
<td>-11.07</td>
<td>94.3%</td>
</tr>
<tr>
<td></td>
<td>-2.84·Noise</td>
<td></td>
<td>+84.10·Blur</td>
<td></td>
<td>+15.64·Contrast</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+17.22·Contrast</td>
<td></td>
<td>+21.65·Noise</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lake</strong></td>
<td></td>
<td></td>
<td>-6.61</td>
<td>80.8%</td>
<td>-21.36</td>
<td>97.6%</td>
</tr>
<tr>
<td></td>
<td>-2.35·Noise</td>
<td></td>
<td>+13.30</td>
<td></td>
<td>+1.41·Noise</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+12.83·Contrast</td>
<td></td>
<td>+17.55·Contrast</td>
<td></td>
<td>+7.10·Contrast</td>
<td></td>
</tr>
<tr>
<td><strong>Man</strong></td>
<td></td>
<td></td>
<td>-2.87</td>
<td>70.3%</td>
<td>-10.82</td>
<td>95.9%</td>
</tr>
<tr>
<td></td>
<td>-3.02·Noise</td>
<td></td>
<td>-8.74</td>
<td></td>
<td>+5.66·Blur</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+8.96·Contrast</td>
<td></td>
<td>+12.57·Contrast</td>
<td></td>
<td>+9.96·Contrast</td>
<td></td>
</tr>
</tbody>
</table>

Table 21 presents the regression models predicting MOS by respective subjective attribute values (Layer 4) for the printing methods separately and combined. All models explained the variation in the overall quality well with coefficients of determination of over 90%. Of the attributes, graininess appeared in seven of the models, colourfulness and sharpness in six. Contrast contributed to the quality prediction only in the models for the whole sample set. Colourfulness appeared to be an important predictor for the quality of the IJ samples. It was also included in all three models for the image content cactus. For comparison, the models predicting MOS by instrumental measurements /Table 19/ consisted of attributes largely congruent with those in the models created with subjective attributes as predictors. E.g. in the models for all samples, the only difference was the missing equivalent for contrast in the model by instrumental measurements for cactus. The connection between the variables in the models predicted by computational /Table 20/ and subjective attributes /Table 21/ was less clear.
Table 21  The regression models predicting MOS by subjective quality attribute values (Layer 4) for the EPG and IJ samples separately and combined.

<table>
<thead>
<tr>
<th>Model</th>
<th>All samples</th>
<th>EPG samples (MP + EPG + EPG Photo)</th>
<th>IJ samples (MP + IJ + IJ Photo)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>Model</td>
<td>$R^2$</td>
</tr>
<tr>
<td>Cactus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.44</td>
<td>+0.25·Graininess +0.37·Sharpness</td>
<td>96.1 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+0.28·Contrast +0.26·Colourfulness</td>
<td></td>
</tr>
<tr>
<td>Lake</td>
<td>-0.18</td>
<td>+0.38·Graininess +0.42·Sharpness</td>
<td>95.9 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+0.27·Contrast</td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>-0.28</td>
<td>+0.23·Graininess +0.66·Sharpness</td>
<td>97.2 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+0.21·Colourfulness</td>
<td></td>
</tr>
</tbody>
</table>

0 presents the regression models predicting MOS by subjective Layer 4 attributes for the different levels of paper quality. In the case of EPG prints, the models that best predicted the overall print quality were created for the MP papers. For the middle quality level, i.e. EPG papers, all three models contained only one predictor, graininess. Overall, graininess was the most common predictor in the EPG models as it was included in five of the nine models. When comparing image contents the models of the weakest explanatory power were for cactus with the lowest coefficient of determination being 46.4% for the EPG Photo papers.

In the case of IJ prints, the models with the best fit were for the IJ Photo papers. Again graininess was of great importance in the models, as it was included in seven of the nine models, i.e. in all models for lake and man, as well as all models for the IJ Photo papers. In general, the models for the IJ prints were better in predicting the overall quality of the samples than the models for the EPG prints. However, in all but four cases, the models for the different levels of paper quality had a weaker explanatory power compared to the models for the whole sets of EPG and IJ samples /Table 21/.
Table 22 The regression models predicting MOS by subjective quality attribute values (Layer 4) for the different levels of paper quality.

<table>
<thead>
<tr>
<th>Subject</th>
<th>EPG samples</th>
<th>IJ samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MP model</td>
<td>EPG model</td>
</tr>
<tr>
<td>Cactus</td>
<td>0.54 +0.71·graininess</td>
<td>1.40 +0.52·graininess</td>
</tr>
<tr>
<td></td>
<td>-1.55 +0.86·sharpness +0.61·colorfulness</td>
<td>0.34 +0.88·graininess</td>
</tr>
<tr>
<td>Man</td>
<td>-1.10 +0.44·sharpness +0.71·contrast +0.28·colorfulness</td>
<td>-0.12 +0.99·graininess</td>
</tr>
<tr>
<td></td>
<td>-0.37 +0.76·sharpness +0.38·contrast</td>
<td>-0.003 +0.99·colorfulness</td>
</tr>
<tr>
<td></td>
<td>0.13 +2.76·sharpness -2.03·graininess</td>
<td>-0.51 +0.68·graininess +0.52·colorfulness</td>
</tr>
<tr>
<td></td>
<td>-0.77 +1.71·graininess</td>
<td>-0.54 +0.45·graininess +0.80·contrast</td>
</tr>
</tbody>
</table>

5.2.4 Conclusions

The first objective of the study was to evaluate the feasibility of instrumental measurements from prints (Layer 2) and computational quality metrics (Layer 3) for the overall print quality prediction. The results suggest that the subjective quality of printed photographs can be predicted to a reasonable extent using instrumental measurements from objective test fields or quality attributes computed from digitized prints. Both the models with computational quality attribute values and the models with instrumental measurements as predictors explained up to 97% of the variation in the overall quality for the ink-jet samples. The models had a similar fit also for the electrophotographic and ink-jet samples combined, with coefficients of determination of about 70-80%.

The challenges of the computational models are in predicting the quality of electrophotographic prints. While the models predicting MOS by instrumental measurements explained up to 91% of the variation in the overall print quality, the models with computational quality attributes as predictors explained up to 59% of the variation at best. The model with the weakest fit accounted just for 29% of the quality variation. Of the four studied computational quality metrics, blur and contrast made the most significant contribution to the quality prediction of the samples. The instrumental measurements that contributed the most to the models were edge sharpness and rms noise.

The other objective of the study was to analyze the need for discrete quality models for different printing methods, papers of different quality level, and different types of image content. Based on
the models predicting MOS by subjective attributes, creating a single quality model seems feasible provided that the chosen set of papers covers a wide range of quality variation. The included variables, the coefficients and the goodness-of-fit of the models depended on the image content, the quality of which was predicted. This suggests that there is a real need for quality models which are adaptive to the content. The studied samples gave no indication whether the quality level of the papers affects the decision making criteria of the subjects. To address this question, papers of different quality levels should be evaluated subjectively as separate sets. Now the samples were evaluated as a single set and the data was divided afterwards based on the different quality levels.

5.3 Content Adaptive Approach

5.3.1 Introduction

The overall value of image quality is often measured by summing up all the measurable attribute values. This assumption ignores the application or content specific meanings. For example, some physical attribute can get a high distortion value but it has no significant meaning for image interpretation and quality experience with used application or image content.

There are some studies where quality dimensions have been selected for application specific ways. Mangin and Dubé /44/ proposed the quality dimensions for prints. The dimensions were based on the quality attributes that have been widely used in literature. Quintard et al. /56/ proposed that contrast, hue, saturation and texture describe the quality of display. Pastrana-Vidal et al. /53/ proposed that the quality of video is based on the spatial and temporal dimensions.

We wanted to use only the attributes which have a meaning for quality experience of printed natural images. The dimensions of used quality space were based on the attributes defined by the subjects. We used the content adaptive RR metrics developed in the project /see Section 4.4/ for quality calculations.

5.3.2 Methodology

Subjective Attributes

Subjective tests were carried out at the Department of Psychology at the University of Helsinki. Ten most used image quality attributes for describing the man, cactus and lake EPG (CC5.5) and IJ (CC5) samples are shown in Table 23.
Table 23  Ten most used subjective attributes for describing the samples.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Man EPG</th>
<th>Cactus EPG</th>
<th>Lake EPG</th>
<th>Man IJ</th>
<th>Cactus IJ</th>
<th>Lake IJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharp</td>
<td>97</td>
<td>90</td>
<td>97</td>
<td>109</td>
<td>114</td>
<td>121</td>
</tr>
<tr>
<td>Grainy</td>
<td>85</td>
<td>83</td>
<td>81</td>
<td>107</td>
<td>110</td>
<td>116</td>
</tr>
<tr>
<td>Unsharp</td>
<td>67</td>
<td>83</td>
<td>71</td>
<td>98</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td>Good colours</td>
<td>62</td>
<td>62</td>
<td>64</td>
<td>95</td>
<td>83</td>
<td>88</td>
</tr>
<tr>
<td>Faded colours</td>
<td>62</td>
<td>50</td>
<td>63</td>
<td>82</td>
<td>76</td>
<td>83</td>
</tr>
<tr>
<td>Clear</td>
<td>50</td>
<td>47</td>
<td>63</td>
<td>81</td>
<td>75</td>
<td>70</td>
</tr>
<tr>
<td>White dots</td>
<td>46</td>
<td>44</td>
<td>40</td>
<td>65</td>
<td>74</td>
<td>61</td>
</tr>
<tr>
<td>Streaking</td>
<td>35</td>
<td>41</td>
<td>39</td>
<td>63</td>
<td>73</td>
<td>56</td>
</tr>
<tr>
<td>Unclear</td>
<td>35</td>
<td>41</td>
<td>39</td>
<td>54</td>
<td>63</td>
<td>48</td>
</tr>
<tr>
<td>Faded</td>
<td>34</td>
<td>40</td>
<td>39</td>
<td>46</td>
<td>39</td>
<td>35</td>
</tr>
</tbody>
</table>

Sharp, unsharp and grainy were the three most used attributes for IJ samples regardless of image content. These attributes were also the most used for EPG samples but now there were some differences between the image contents. With the test image cactus the attribute “sharp” was the fourth most used and the attribute “unsharp” was not in the group of ten most used attributes. Whereas with the test images man and lake the sharp was the most used attribute.

With the IJ samples the fourth most used attribute was “faded colours” regardless of image content. With the EPG samples “faded colours” was the fifth or sixth most used attribute. With the EPG samples “good colours” was the third or fourth most used attribute. With the IJ samples “good colours” was in the group of ten most used attribute. The third colour attribute that the subjects used was the deep colours. In special, with the IJ samples it was pretty often used.

The attribute “good colours” relates to the overall colour quality. The other colour attributes characterize more on specific colour properties. We assume that the attributes “deep colours” and “faded colours” compose a bipolar dimension. The attribute “deep colours” is the positive and “faded colours” is the negative polar of this dimension. In addition we assume that the attribute “faded” has an effect on this dimension.

The attributes clear and unclear were also in the group of ten most used attributes. Clarity (clear-unclear) is a high-level attribute /see Section 2.3.1/; brightness, sharpness and higher contrast make images seem clearer.
Based on the subjective frequencies of the attributes we formed the human quality space for prints. The dimensions were sharpness, graininess and colour contrast. The dimension of sharpness was composed of the attributes sharp and unsharp. The dimension of graininess was composed of the attributes grainy and white dots. The dimension of colour contrast was composed of the attributes deep colours, faded colours and faded. A quality space depends on the context. The context of this space is a photograph printed with the size of 10 cm × 15 cm. We selected only the low-level attributes for the space. The high-level attribute clarity was restricted out of the space. Anyway, the dimensions of sharpness and colour contrast have an effect on the clarity.

**Image Quality Calculation Method**

Based on the human quality space defined above, we propose the content adaptive image quality calculation method for printed natural images. The proposed method digitizes the printed image, decomposes it into scales and orientations by the steerable pyramid technology /67/ and calculates the image quality value using content adaptive RR metrics.

The printed images were digitized by the calibrated camera /see Section 5.1.2/. The different image attributes were calculated using different image scales /Figure 22/. For example, graininess was calculated from the first scale and sharpness and dominant colour contrast were calculated from the second scale. The use of higher scale for sharpness and colour contrast calculation filtered the noise energy and its disruptive effect was lower. Graininess and sharpness metric used only the L channel information of Lab space. Dominant colour contrast metric used all three Lab space channels.

![Figure 22](image-quality-framework.png)  
*Figure 22* Image quality framework calculates the different attributes using different scales.

The performance of the proposed method was studied against the subjective MOS values. Because the MOS values of the image content were always scaled to the interval 1-5, we normalized the RR metric values to a common scale [0-1]. The image quality, $IQ$, was calculated by Equation 2:

$$IQ = \frac{(1 - DC) + G + S}{3}, \quad (29)$$

where $DC$ is dominant colour contrast, $G$ is graininess and $S$ is sharpness dimension.
The performance of the proposed IQ method was measured with linear correlation (CC), rank order correlation (SROCC), and root mean square error (RMSE) metrics. In addition we compared our method to the general RR image quality assessment method (RRIQA) proposed by Wang and Simoncelli [76].

5.3.3 Results

MOS values for EPG (CC5.5) and IJ (CC5) samples as a function of IQ values of the proposed method are shown in Figure 23(a) and Figure 23(c). Figure 23(b) and Figure 23(d) show MOS values for EPG and IJ samples as a function of the RRIQA values. The lines added on the figures define the ideal values of the predicted IQ if the relation between predicted and subjective MOS values would be a linear. The added lines also show if the IQ model over- or underemphasizes the values. Table 24 shows CC, SROCC and RMSE values for the proposed IQ method and RRIQA. The CC and SROCC values of the proposed IQ method were statistically significant (p<0.0001) for all the image contents excluding the EPG image lake. SROCC value for the EPG image lake was statistically significant (at p<0.01 level).

Figure 23 Predicted MOS as a function of subjective MOS: proposed IQ method for EPG (a) and for IJ (c), RRIQA for EPG (b) and for IJ (d).
Table 24  Linear correlation (CC), rank-order correlation (SROCC) and root mean square error (RMSE) values for EPG and IJ samples.

<table>
<thead>
<tr>
<th></th>
<th>EPG</th>
<th></th>
<th></th>
<th>IJ</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>SROCC</td>
<td>RMSE</td>
<td>CC</td>
<td>SROCC</td>
<td>RMSE</td>
</tr>
<tr>
<td>Proposed IQ</td>
<td>-0.883</td>
<td>-0.727</td>
<td>0.405</td>
<td>-0.979</td>
<td>-0.962</td>
<td>0.245</td>
</tr>
<tr>
<td>RRIQA</td>
<td>-0.786</td>
<td>-0.631</td>
<td>0.535</td>
<td>-0.888</td>
<td>-0.830</td>
<td>0.560</td>
</tr>
<tr>
<td>Proposed IQ (man)</td>
<td>-0.879</td>
<td>-0.748</td>
<td>0.409</td>
<td>-0.979</td>
<td>-0.969</td>
<td>0.253</td>
</tr>
<tr>
<td>RRIQA (man)</td>
<td>-0.864</td>
<td>-0.680</td>
<td>0.432</td>
<td>-0.926</td>
<td>-0.871</td>
<td>0.463</td>
</tr>
<tr>
<td>Proposed IQ (cactus)</td>
<td>-0.881</td>
<td>-0.772</td>
<td>0.429</td>
<td>-0.980</td>
<td>-0.945</td>
<td>0.236</td>
</tr>
<tr>
<td>RRIQA (cactus)</td>
<td>-0.783</td>
<td>-0.617</td>
<td>0.565</td>
<td>-0.946</td>
<td>-0.932</td>
<td>0.386</td>
</tr>
<tr>
<td>Proposed IQ (lake)</td>
<td>-0.927</td>
<td>-0.636</td>
<td>0.310</td>
<td>-0.980</td>
<td>-0.973</td>
<td>0.240</td>
</tr>
<tr>
<td>RRIQA (lake)</td>
<td>-0.845</td>
<td>-0.721</td>
<td>0.441</td>
<td>-0.813</td>
<td>-0.762</td>
<td>0.712</td>
</tr>
</tbody>
</table>

5.3.4 Conclusions

Based on the subjective data, the attributes of sharpness, graininess and colour contrast had an effect on the image quality of printed natural images. This was an expected result. The image quality has often been evaluated by the attributes related to sharpness, noise and colour reproduction. One goal of the study was to find if any image content specific quality spaces were evident. Based on the subjective data, all the contents were evaluated mainly with the same dimensions. One small exception was that with the IJ samples the deep colours attribute was slightly more used than with the EPG samples. With the EPG samples the attribute ‘good colours’ was slightly more used than with the IJ samples. For the final quality space we selected the dimension of colour contrast instead of good colours or deep colours. This dimension of colour contrast had some features from both these attributes.

The proposed method was based on an application specific approach. It was designed for natural images printed with EPG or IJ printers. In the study, the performance of the proposed method was compared to a general application-independent method. The output of application-independent methods is often a single number related to the image quality. One benefit of the proposed method is that the output includes individual attribute values behind the overall quality. This can be valuable information for research and development work. The second benefit of the application specific approach is the ease of tuning it to take into account the application specific features. For example the proposed method eliminates the influence of graininess for the calculations of other attributes. The character of graininess or noise energy is a highly application specific feature. The graininess energy of a print differs from the noise energy of a digital image. Respectively, the noise energy of a digital or a printed image depends on the imaging device. The noise characteristics of a low-end camera differ from the characteristics of a high-end camera. Also the graininess characteristics of an ink jet printer differ from the characteristics of an electrophotographic printer.
5.4 Cumulative Match Score Approach

5.4.1 Introduction
The main goal was to find evidence that the human visual quality experience can be estimated using computational and physical features. The research problem was to devise a method to reveal the possible connections between these very different measurements. The first obvious measure of correspondence would be simple linear correlation. This, however, includes several severe problems. Correlation measures only linear relationship which cannot be the case between computational print quality measures and subjective visual quality experience. In addition, due to the laborious nature of carrying out visual experiments, the amount of data remains small, and therefore, the correlation coefficients very unreliable. The correlation could be computed for a single computational feature, but its explanatory power would remain insufficient.

On the other hand, studying combinations of several features would be numerically infeasible. Most importantly, the human evaluation results could not be ranked because it was possible to realise only pair-wise comparisons, i.e., for each sample it is possible to decide which of the remaining samples is the most similar, but it is impossible to order all samples to some monotonously increasing or decreasing scale. Therefore, any comparison procedure should be based on pair-wise similarities, not ranking.

5.4.2 Methodology
For the investigation of the similarity (explanatory power) between the computational features and human evaluation features, the use of cumulative match score is proposed. The idea for the method originates from the biometric authentication research community where an important problem is how to compare different methods for face recognition or verification. A generally accepted measure is the match score introduced in the FERET protocol /55/. The cumulative match score prefers methods where similar samples (different images of the same person in FERET) are close to each other.

In FERET, the ground truth information is the known identities of persons in the images. In our case, the ground truth is the human evaluation: if a computational measure always prefers the same sample as the human does, there must be a relationship between the computational features and the “features” humans use in their judgement. It is also noteworthy that this evaluation method is inherently based on topological properties (which samples are close to each other in the feature space), not the base (manifold) structure from the problem point of view. The tailored procedure to compute the cumulative match score is presented in Algorithm 1. Note that the algorithm is not restricted to any particular distance measure.
Algorithm 1: Cumulative match score for comparing distances in any vector space to a space where only pair-wise similarity information is available
1. Set pair-wise data as the ground truth (human evaluation)
2. for each sample do
3. Find the best match in the ground truth space
4. In the evaluation space (computational features) compute all pair-wise distances to the sample
5. Sort samples based on the distances
6. Find the best ground truth match in the sorted list
7. Store index number
8. end for
9. Compute the cumulative histogram of the match indices

The $n$th bin of the cumulative histogram tells "how many times (percent of all samples) the ground truth match (human evaluation) has been within the set of $n$ closest samples in the second evaluation space (computational)". The advantage of the method is that information about the type of the relation (linear or non-linear) is not required and neither is the optimal value for the measure. The comparison of any evaluation spaces is straightforward. However, since the scale and type of the relation are not taken into account, the method does not tell how the quality experience should be modelled. Only evidence of the measures which should be used is available.

The cumulative match score is not symmetric, i.e., it may give different results with respect to which of the evaluation spaces is used as the ground truth. In our case, for example, if the computational measures were used as the ground truth, the cumulative match score would answer the question: "Are technically close samples (based on a single measure) visually close to each other?" Thus, the cumulative match score would give good results only when the examined computational measure explains the whole quality experience. Usually, no matter how important a factor a computational measure is for the quality experience, there can be another quantity that causes large visual differences to the samples that are similar based on the selected computational measure. This is why the human evaluation is used as the ground truth: "Are visually close samples technically close to each other?"

The cumulative match score can utilise any distance function, but the Euclidean distance is the most straightforward. To prevent domination of large scale measures, the normalised Euclidean distance can be used (each coordinate in the sum of squares is inverse weighted by the sample variance of that coordinate).

5.4.3 Results

In Figure 24, the cumulative match scores over 16 IJ samples (CC3) for different physical properties and print quality measures are shown. The dotted lines represent the mean and 95% confidence intervals (2 times standard deviation) for randomly generated vectors (estimated by simulation). The confidence intervals of the random vectors cannot be generalised to the cumulative match scores of the metrics. For the random vectors, (random) variation determines the whole order, and therefore, causes very large variation to the cumulative match score histograms. For the measures, the variation (inaccuracies caused by a small test set, measuring errors, etc.) causes only
small changes to the orders (samples close to each other may change places, but not samples far from each other), and therefore, the actual variation in the cumulative match scores is smaller.

Figure 24  Cumulative match scores over 16 IJ samples (CC3): (a) physical properties of paper, (b) print quality measures computed from the technical fields, (c) print quality measures computed from the natural images, (d) combinations of the physical properties and print quality measures.

From Figure 24(a), it can be seen that paper gloss is the most important physical property in the sense of the quality experience: for 50% of the samples, the subjectively closest sample is also the closest sample based on the gloss value, and for 81%, the closest sample is within the set of 4 closest samples. It is also evident in Figure 24(a) that paper opacity does not explain the quality experience (its importance would increase if the images were printed on both sides).

From Figure 24(b), it can be seen that the mean colour difference, mottling, density of black and edge raggedness all have a significant influence on the quality experience. Figure 24(c) shows that the computational measures from natural images (used in the human evaluation) have, unexpectedly, an inferior ability to explain the quality experience when compared to measures from the technical fields. Especially, the noise measure from natural images seems to have a very insignificant role in the quality experience.
In Figure 24(d), the cumulative match scores for several combinations of physical properties and computational print quality measures are shown. It can be seen that for 55% of the samples, the subjectively most similar sample is also the most similar based on the combination of the physical properties and computational print quality measures, and for 90%, the most similar sample is within 3 samples. This presents clear evidence of the explanatory power of the combined measures.

Figure 25 presents the cumulative match score histograms for 21 EPG samples (CC4). From Figure 25(a) it can be seen that the roughness of the paper is the most important paper physical property when electrophotographic printing is used. For almost 60% of the samples, the subjectively closest sample is in the set of 3 closest samples (20 possibilities). Other important single physical properties are permeability, density and gloss.

Figure 25(b) shows that the gloss of printed regions, variance of printing dot area (evenness of the print) and edge blurriness are the most important print quality measures. As an interesting result, colour properties, such as the colour gamut, have a relatively small contribution to the experienced quality when compared to inkjet printing. The reason for this is that the colour reproduction in electrophotographic printing is rather independent of the paper grade, and therefore, humans cannot see differences in the reproduced colours.

From Figure 25(c) it can be seen that, similarly to inkjet printing, the used natural image computational measures have an inferior ability to explain the quality experience. Natural image noise is the only measure that seems to have some influence on experienced quality.

Figure 25(d) shows the cumulative match score histograms for the combinations of physical properties and computational print quality measures. The cumulative match score of the combinations of physical properties seems to be slightly better than for the combination of all print quality measures. However, the number of print quality measures is large (dot quality measures added) and some of them add more noise instead of new information, resulting in a lower cumulative match score histogram.
Figure 25  Cumulative match scores over 21 EPG samples (CC4): (a) physical properties of paper, (b) print quality measures computed from the technical fields, (c) print quality measures computed from the natural images, (d) combinations of the physical properties and print quality measures.

5.4.4 Conclusions

Based on our experiments, we conclude that the optimal combination is highly dependent on the used printing method. For inkjet printing, the most important computational print quality measures are mottling, colour gamut, contrast, blurriness and the densities of primary colours, and the most important physical properties are paper gloss, brightness and whiteness.

For electrophotographic printing, the most important computational print quality measures are the unevenness of print (mottling or variance of dot area), edge blurriness, print gloss and print densities, and the most important physical properties are paper roughness, brightness and opacity.
5.5 Bayesian Network Approach

5.5.1 Introduction

The Bayesian network (BN) approach for modelling the interconnections between instrumentally measured properties and subjective quality perception was introduced in Section 1.6. The nodes in the BN model represent instrumental measurements, computational quality characteristics and perceived quality attributes, and the directed edges of BN describe the probabilistic relations between the nodes.

The model parameters are identified as conditional state probabilities according to the identification data. For example, a conditional state probability distribution might express that if a person assessed a sample’s usefulness as low and naturalness as low then he/she would most likely assess the evaluated overall quality as low. The conditional probability distribution of a node is estimated at all combinations of states of parent nodes by counting the corresponding observations in the identification data. If no observations of a combination of parent’s states exist, the state of the node is assumed evenly distributed according to the maximum entropy principle.

The BN is intended to be identified with full-reference and reduced-reference instrumental measurements. Once identified, the model operates as a no-reference measurement system. Given instrumental measurements as input, the network produces a probability distribution on visual quality assessments. Sensitivity analysis of the identified model reveals how much the output probability distribution changes if small changes are made in the parameters or in the input evidence values. This gives insight into the capability of the network to predict the visual quality and points out the parameters that the output is most (and least) sensitive to. For instance, if there are such combinations of parent node states in the identification data set that only one or very few observations exist, the output of the BN will be sensitive to the corresponding state changes.

This Section is divided into two parts: Subsection 5.5.2 presents the methodology to construct a layered Bayesian network in the field of image quality and Subsection 5.5.3 describes the novel structure optimisation approach and show the result of the method.

5.5.2 The Layered Structure Approach

In this Subsection we describe a four-layered Bayesian network that consists of the overall image quality (OQ), high-level image quality attributes (HL), low-level image quality attributes (LL) and instrumental measurements (inst). Figure 26 shows schematically the structure of the model. The arrows indicate the statistical causality of the model. The construction of the model consists of selecting the structure, i.e. how the arrows between the blocks are connected, and identifying the statistical dependence between the blocks as conditional probabilities. As there will be only a single path between any two nodes, the Bayesian model is considered as a polytree graph.
Figure 26 An example of the image quality measurement system modelled as a Bayesian network.

The foundation of the model is the LL level. The low-level quality attributes are assumed to be the attributes of the image that the humans can directly assess and that their assessments – in spite of the non-physical nature – are objective, i.e. the jury would largely have consistent opinions about them. Hence low-level quality attributes represent the “ground truth” about perceived quality and thus determine statistically both the instrumental measurement values and the high-levels of subjective image quality.

The overall quality is its own layer as a single node. High level quality consists of the attributes naturalness and usefulness which are abstract aspects of the image and together compose the overall quality. The low-level quality attributes and instrumental measurements included in the model may be changed according to the image evaluation task in question. The model can be identified when observations from each layer are available: instrumental measurement values together with the subjective evaluations of low-level quality attributes, high level quality attributes and overall quality.

The key idea in modelling the quality assessment with a Bayesian network is that once the model has been identified, it is possible to consider the overall image quality and other attributes as joint probability distributions. In particular, any evidence about a node state can be propagated through the network so that the probabilities of the states of each node are updated. Obtaining evidence of instrumental values through measurements and then inferring about the overall quality is the main intended use. Our choice of using a probabilistic modelling framework for this purpose is in accordance with describing measurement uncertainty in more traditional measurement systems.

The joint probability of the states in the directed Bayesian model is:

$$p(x) = \prod_{k=1}^{K} p(x_k | pa_k),$$

(30)

where $K$ is the number of nodes in the model, $x_k$ is the examined node, and $pa_k$ are the parent nodes of $x_k$. For nodes without parents, a priori information is used for the distribution of states of the examined node. In our model, we assume that the model is an acyclic graph. The graph of Figure 26 consists of four layers and can be written – neglecting the detailed structure within layers – as:
When the model is being used and experimental evidence on instrumental measurements is obtained, i.e. \( x_{\text{inst}} = x_{\text{inst}}^{(e)} \), the LL level probability is derived through Bayesian formula as:

\[
p^{(e)}(x_{LL} \mid x_{\text{inst}}^{(e)}) = \frac{p(x_{\text{inst}}^{(e)} \mid x_{LL}) p(x_{LL})}{\sum_{x_{\text{LL}}} p(x_{\text{inst}}^{(e)} \mid x_{LL}) p(x_{LL})}.
\]  

Then information about the overall image quality, \( x_{OQ} \), is the corresponding conditional probability distribution:

\[
p(x_{OQ} \mid x_{\text{inst}}^{(e)}) = \sum_{x_{\text{LL}}} \sum_{x_{\text{HLL}}} p(x_{OQ} \mid x_{\text{HLL}}) \cdot p(x_{\text{HLL}} \mid x_{LL}) \cdot p^{(e)}(x_{LL} \mid x_{\text{inst}}^{(e)}),
\]  

For instance, if we had evidence on attributes in the LL level, the child nodes in the instrumental level would not give any extra information about high level quality.

In general, the real technical measurements give continuous values as a result. The relation between continuous values in instrumental level and discrete LL is

\[
p(x_{LL} \mid x_{\text{inst}}^{(e)}) = \frac{f(x_{\text{inst}}^{(e)} \mid x_{LL}) \cdot p(x_{LL})}{\sum_{x_{LL}} f(x_{\text{inst}}^{(e)} \mid x_{LL}) \cdot p(x_{LL})},
\]  

where \( f(x_{\text{inst}} \mid x_{LL}) \) describes the measurement, i.e. the probability density of obtaining measurement value \( x_{\text{inst}} \), when LLs take the values \( x_{LL} \).

The directed edges of a Bayesian network describe the probabilistic relations between the nodes. The edge structure in a Bayesian model is identified from the relations between attributes in an identification data set. Mutual information and Pearson correlation are examples of methods for choosing edges between nodes in a Bayesian model.

Using either of the methods, the dependency between each pair of attributes is examined and this information is used to establish the relevant interconnections as edges. The model parameters are identified as conditional state probabilities according to the identification data. For instance, it is typical in an identification data set that the overall quality is judged as low, if both usefulness and naturalness are judged as low. The conditional probability distribution of a node is estimated at all combinations of states of parent nodes by simply counting the observations in the identification data.

**Experimental Results**

To demonstrate the idea of Bayesian network modelling in image quality assessment, we conducted a small visual evaluation test with one image content. The model presented in Figure 26 was identified through jury assessments that consisted of overall quality, high level quality attributes and low-level quality attributes for modified images. At first, the instrumental measurements layer consisted of image manipulations that simulated the results obtained with real measuring devices.
For testing the validity of continuous measurement, the simulated measurements were replaced by algorithmic measurements.

The edges in the model between high-level quality attributes and overall quality are based on earlier research that suggested the overall image quality to be explained with naturalness and usefulness attributes. The rest of the edges were chosen according to the mutual information between the attributes on successive quality levels, i.e., between simulated instrumental measurements and low-level quality attributes, and between the low-level quality attributes and high level quality attributes. The five low-level quality attributes applied in our test were sharpness, brightness, colourfulness, graininess and clarity. The meaning of each attribute was shortly explained to each evaluator so that the variance of subjective understanding could be decreased.

In the test case, a digital image shot in a studio was modified by three methods: HSV saturation adjustment, low-pass filtering and noise addition. The degree of these modifications simulated the instrumental measurements in the Bayesian model. We used three distinct degrees of modification for each method: no modification, mild, and moderate level. The combination of all modified images was used in the evaluation test, that is, the number of images was 27. Each image was assessed with respect to eight attributes (5+2+1) on a scale from 1 to 5, one attribute at a time. In practice this means that each subject was asked to label the 27 images with the grades 1-5 eight times. The image used in the test is shown in Figure 5 (studio image). The Studio image had been designed to cover the subjective and objective factors as much as possible. Consequently, it contained several detailed objects that required evaluator’s attention. Each evaluator weights differently the components of the image while assessing certain attributes, which is a reason that may cause irrationality in jury decision. The irrationality is a source of uncertainty in the model identification and therefore it may cause modelling error. The human assessments were done using a monitor display and a graphical user interface built for MATLAB.

The edges between the instrumental level and low-level quality attributes were defined by the largest mutual information values between a node in the instrumental measurement level and two nodes in low-level quality attributes, thus the number of edges in the model was constrained to two. The edges leaving the low-level and directing to either high-level node were also allocated according to the largest mutual information. The attributes were combined to a model shown in Figure 27.
Figure 27 The Bayesian model constructed according to the mutual information results computed from the test case data.

Figure 27 shows that as the edge decision rule is the mutual information, according to the human evaluations, the model is divided into two branches. The mutual information and Pearson correlation values between the attributes related in the model are shown in Table 25.

Table 25 Mutual information and Pearson correlation values for nodes in Bayesian model identification.

<table>
<thead>
<tr>
<th>Related attributes in the model</th>
<th>$I$ as nats</th>
<th>$r_{x,y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSV saturation</td>
<td>$\text{Brightness}$</td>
<td>0.0755</td>
</tr>
<tr>
<td></td>
<td>$\text{Colourfulness}$</td>
<td>0.1825</td>
</tr>
<tr>
<td>Low-pass filtering</td>
<td>$\text{Sharpness}$</td>
<td>0.5633</td>
</tr>
<tr>
<td></td>
<td>$\text{Clarity}$</td>
<td>0.2101</td>
</tr>
<tr>
<td></td>
<td>$\text{Graininess}$</td>
<td>0.3343</td>
</tr>
<tr>
<td></td>
<td>$\text{Clarity}$</td>
<td>0.1989</td>
</tr>
<tr>
<td>Noise</td>
<td>$\text{Sharpness}$</td>
<td>0.5300</td>
</tr>
<tr>
<td></td>
<td>$\text{Graininess}$</td>
<td>0.1134</td>
</tr>
<tr>
<td></td>
<td>$\text{Clarity}$</td>
<td>0.4340</td>
</tr>
<tr>
<td>Usefulness</td>
<td>$\text{Brightness}$</td>
<td>0.1550</td>
</tr>
<tr>
<td></td>
<td>$\text{Colourfulness}$</td>
<td>0.1907</td>
</tr>
<tr>
<td>Naturalness</td>
<td>$\text{Usefulness}$</td>
<td>0.2662</td>
</tr>
<tr>
<td>Overall quality</td>
<td>$\text{Naturalness}$</td>
<td>0.1649</td>
</tr>
</tbody>
</table>

The values of Pearson correlation vary between -1 and 1. The theoretical maximum value of mutual information is the entropy of a uniform distribution of a node with fewer states, as instrumental measurement nodes have three states and all other nodes have five states. In the model identified
according the Pearson correlations, the structure is very similar to model identified with mutual information. The only difference is that the edge from brightness to HSV saturation is changed to edge from clarity to HSV saturation. In that case, the brightness attribute would not be directly utilized in inference from instrumental level to high level quality.

To better model the instrumental measurements, algorithmic measurements were conducted for all 27 images. Algorithms for evaluating blur, contrast and noise as no-reference measurements were utilized. For defining the relations between low-level quality attributes and algorithmic measurements, the continuous results were first quantized to five states. The relations of these measurements to the low-level quality attributes as mutual information are shown in Table 26.

<table>
<thead>
<tr>
<th>Table 26</th>
<th>Mutual information as nats between algorithmic measurements and low-level quality attributes.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noise</td>
</tr>
<tr>
<td>Sharpness</td>
<td>0.4303</td>
</tr>
<tr>
<td>Brightness</td>
<td>0.0950</td>
</tr>
<tr>
<td>Colourfulness</td>
<td>0.0476</td>
</tr>
<tr>
<td>Graininess</td>
<td>0.3919</td>
</tr>
<tr>
<td>Clarity</td>
<td>0.3496</td>
</tr>
</tbody>
</table>

Table 26 implies that if the number of edges directed to each instrumental measurement node is constrained to two, there would be no edges leaving brightness or colourfulness nodes, because they have the smallest amounts of mutual information in relation to noise, contrast and blur measurements. This could be solved by further developing the measurement algorithm or increasing the number of edges.

A simple example of the efficiency of inference with the continuous measurements is shown in Figure 28.

![Figure 28](image.png)  
*Figure 28*  The pdfs of algorithmic noise at each state of sharpness attribute.

According to Equation 34, we may evaluate the maximum likelihood estimate for the continuous result given the observations from the low-level quality attribute level. The probability distributions
of the amount of noise at each state of sharpness are presumed to be Gaussian distributions whose mean values and standard deviations have been calculated from the known noise levels combined with the human assessment data.

The pdf curves in Figure 28 can now be used to deduce the most likely state of sharpness, given that the noise level has been measured. For example, if the noise was algorithmically measured to be 14 (black vertical line in Figure 28), the probabilities of the states 1, 2, 3, 4 and 5 of sharpness are 0.1806, 0.2588, 0.2224, 0.3375 and 0.0007, respectively, and thus the most likely state of sharpness is 4.

5.5.3 The Structure Optimisation Approach

In this Subsection we describe a full framework of Bayesian network which is optimised using the collected objective measurements and psychophysical subjective data. The optimisation strategy for the network is novel and is therefore discussed first. Then we will present the results using our optimisation strategy.

Optimising the Model Structure

Learning the optimal structure for a Bayesian network has been shown to be NP-complete, and thus, methods applying full search for structure learning are out of the question. In addition, due to the laborious nature of subjective test, the amount of training data remains considerably limited. This makes also most of the heuristic methods, such as the PC algorithm, unreliable. However, optimising the structure is essential for our problem and we need to establish it in our learning process.

In the print quality modelling case, it is possible to form a number of hypotheses how the model should behave. For example, if the undesired unevenness of print (Mottling) increases while the other inputs (objective measures) do not change, the overall quality should decline. Similarly, if the colour gamut (a subset of colours a paper can reproduce with the available inks) expands, the overall quality should improve. Now, it is possible to test a hypothesis with a model, and produce a scalar value representing how logically the model works (how often the model behaviour follows the hypotheses). This leads to a complicated optimisation task: find such a structure for Bayesian network that after it is trained with the existing data, the obtained model behaves as logically as possible. Here, the prior knowledge of behaviour acts as a regularisation term which enables the optimisation process with a small number of data points.

The hypotheses about the behaviour of a logical model are presented in Table 27. The hypotheses can be tested by computing the marginal distributions for overall quality with some objective measure values set as the evidence, and then change the value of a single objective measure and examine how the marginal distribution of overall quality changes. Instead of comparing two distributions, the expected values of the overall quality (estimation of the mean opinion score of a jury with similar distribution in opinions) can be used.
Table 27  Prior rules for the optimization (hypotheses).

<table>
<thead>
<tr>
<th></th>
<th>Change on input</th>
<th>Effect on overall quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Mottling increases</td>
<td>Decrease</td>
</tr>
<tr>
<td>H2</td>
<td>Colour densities increase</td>
<td>Increase</td>
</tr>
<tr>
<td>H3</td>
<td>Colour gamut increases</td>
<td>Increase</td>
</tr>
<tr>
<td>H4</td>
<td>Gloss increases</td>
<td>Increase</td>
</tr>
<tr>
<td>H5</td>
<td>Edge blurriness increases</td>
<td>Decrease</td>
</tr>
<tr>
<td>H6</td>
<td>Edge raggedness increases</td>
<td>Decrease</td>
</tr>
</tbody>
</table>

If the number of edges increases too much in the graph, situations occur where the number of parents of some node is too large to properly estimate the conditional probability distribution with the limited training data. This causes hypothesis tests to fail and the fitness function value to decline. This, in part, keeps optimal networks compact, and only the most important edges remain. Therefore, the presented fitness function also prevents over-fitting to the existing data. In following experiments, weights were set to 1 for all nodes, i.e., all hypotheses were treated as equally important.

**Experimental Results**

The applied optimisation method was a genetic algorithm mainly due to the complexity of the optimisation problem and the need for simulation to evaluate the fitness of a solution. CC4 samples (EPG) were used. As a single result the most optimal network is shown in Figure 29 where the connections describe which elements are important to which elements in the final network. Despite of its limitations this Bayesian network can be considered as a unified model which explains human observer quality experience based on instrumental measurements (and vice versa). The best structures of Bayesian network found by the optimisation process are shown in Figure 29, Figure 31 and Figure 33. A list of 100 best structures was maintained during the optimisation process. After the optimization step, the fitness function was computed for the 100 networks with a larger sample size.

The best structure according to the fitness function is shown in Figure 29. All 100 networks were evaluated against the subjective MOS using leave-one-out cross-validation. The best structure according to the correlation coefficient between the model output and MOS is shown in Figure 31. In Figure 30, Figure 32 and Figure 34 the correlations against the subjective evaluation are plotted. The expectation values of the overall quality were used as the VQI. In Figure 33 and Figure 34, results are shown when, instead of optimising the fitness function, the correlation coefficient to subjective evaluation was optimised. As it can be seen, the number of edges increases dramatically, and due to the small amount of training data, the generalisability of the model becomes weak.

Significance of the edges in the network structure was studied by removing each edge and by computing the RMSE between the output of the original model and reduced model with large amount of random inputs. The results are shown in network images. The line width represents the significance of an edge (RMSE). Similarly, the significance of the inputs (objective measures) was tested, and the RMSEs between the output of the original model and reduced model (the objective
measure removed from the model) are shown in network figures. In this case, larger numbers imply more important objective measures for the model.

Generalisability of the models was tested as follows: (i) train the model with real subjective data, (ii) generate new data based on the model, (iii) train the model again with the generated data, and (iv) validate the simulated model with real subjective data. To generate data based on a model, its inputs (objective measures) were set by using random sampling from uniform distribution. The inputs were set as the evidence, and the marginal distributions for all other nodes (subjective attributes and the overall quality) were computed. New data (an integer value for each node) was determined by sampling the values from the computed marginal distributions. This way, the generalisability of the models was tested in two stages: how well the generated data correspond to the real world data, and how well the simulated model predicts the real subjective data. The results are shown in Figure 30, Figure 32 and Figure 34. It can be seen that the correlation drops significantly for the model that was found by optimising the correlation to subjective evaluation with the small training set /Figure 34/, but the correlations even increase for the models where the behaviour of the model was optimised using hypothesis testing /Figure 30 and Figure 32/. This confirms the previously mentioned assumption that generalisability becomes weak if only the correlation coefficient is optimised to subjective evaluation by using the small training set.

![Figure 29](image)

**Figure 29** The best Bayesian network found: best simulation result.

![Figure 30](image)

**Figure 30** Correlations between the visual quality index and subjective MOS: (a) Leave-one-out method, correlation 0.884; (b) simulation model, correlation 0.957.
Figure 31  The best Bayesian network found: best correlation against the subjective evaluation (leave-one-out).

Figure 32  Correlations between the visual quality index and subjective MOS: (a) leave-one-out method, correlation 0.935; (b) simulation model, correlation 0.955.

Figure 33  The best Bayesian network found: correlation coefficients optimised instead of the fitness function.
Figure 34  Correlations between the visual quality index and subjective MOS: (a) leave-one-out method, correlation 0.994; (b) simulation model, correlation 0.704.
6 DEVELOPING TEST IMAGES FOR PRINT QUALITY EVALUATION

6.1 Introduction

The choice of test images for quality evaluation – whether for camera, display or print quality – is a challenging task. This is true of objective as well as subjective quality evaluation. A known fact is that image content exerts an influence on quality assessment. On the low (syntactic, i.e. conforming to the rules of syntax) level of quality attributes, this is due to masking and facilitation phenomena. On a higher (semantic, i.e. reflecting intended meaning) level, it can be associated with cognitive effects, such as variation of the attention value of image objects and also memory effects. On the highest level, emotional responses may vary, depending on the quality level.

Several arguments can be used in composing test images. First, they may be related to technological considerations; images are chosen so that they contain objects which test the limits of performance of a particular technology. Second, images may be selected so that they represent typical cases of use. This is exemplified by the photo space concept used in mobile imaging contexts. Photo space is a statistical description of the frequency of picture taking as a function of subject illumination level and subject-to-camera distance /32/. Third, images may be chosen so that they include objects with memory colours and memory shapes.

In the literature, several single test images or test image collages have been presented /Figure 35-Figure 39/.

![Figure 35 Roman16 Reference Images /7/](image-url)
Figure 36  ISO SCID Images (a part of the set) /34/.

Figure 37  Altona Test Suite 1.2a Visual /14/.

Figure 38  Kodak PhotoCD PCD0992 (16 of 24 images) /22/.
While a range of test images have been presented (and even standardized) mainly from the standpoint of criticality in printing, aspects such as computation of quality from the images or visual salience of objects, or visual appeal have been largely ignored. This section deals with the requirements to be set on test images and outlines the process by which test images have been developed at TKK.

### 6.2 General Principles

The issues in designing the composition for a test image are related on one hand to content and on the other to technical characteristics. Table 28 presents a summary of aspects to be considered from the content standpoint.

**Table 28** Content related requirements for test images.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme</strong></td>
<td>A neutral, consistent and recognizable theme makes it easier to evaluate the image and does not confuse or mislead the observer. The theme should not evoke difficult feelings, such as disgust or fear.</td>
</tr>
<tr>
<td><strong>Memory colours</strong></td>
<td>The presence of versatile memory colours (the most important ones being sky and skin) is important because they offer an efficient mental comparison basis for colour reproduction.</td>
</tr>
<tr>
<td><strong>Memory shapes</strong></td>
<td>The presence of human face and other familiar objects are useful for evaluating sharpness and distortions.</td>
</tr>
<tr>
<td><strong>Surface materials</strong></td>
<td>The surface materials should include smooth surfaces, shiny surfaces, patterned/geometrical surfaces and detailed/textured surfaces.</td>
</tr>
<tr>
<td><strong>Details</strong></td>
<td>There should be a sufficient number of details in the image, but not too much so that the image will not appear confusing.</td>
</tr>
<tr>
<td><strong>Salience</strong></td>
<td>There should be no features that are excessively salient and demand a major share of the observer’s total attention. Features that easily stand out are exaggerated colours and sizes, for example.</td>
</tr>
<tr>
<td><strong>Perceptual aspects</strong></td>
<td>For example, depth cues make it easier to address the dimensions and proportions in the image, making it more effortless to look at the image. The layout also has a big role in the balance of the image.</td>
</tr>
<tr>
<td><strong>Aesthetical aspects</strong></td>
<td>The image should be visually pleasing, because “ugly” images are systematically judged to be of weaker image quality than “beautiful” images. One of the components that make the image pleasing is layout. This viewpoint is also related to the marketing and breakthrough of the final product. The aim, however, is not to create an artistically valid image.</td>
</tr>
</tbody>
</table>

First, the theme should be easily recognizable and understandable in different cultures. Second the image should contain semantically meaningful objects and surfaces with memory shapes, texture and colour. Moreover, requirements are set on the layout and aesthetics.

To be able to evaluate the chosen visual properties, sharpness, noise, contrast, colourfulness and gloss, further aspects need to be considered. These are listed in Table 29.
Table 29  Requirements on test image arising from assessment of given quality attributes.

<table>
<thead>
<tr>
<th>Quality attribute</th>
<th>Important perceptual characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharpness</td>
<td>Small details, edges with high contrast</td>
</tr>
<tr>
<td>Noise</td>
<td>Smoothly coloured surfaces, dark colours, colour slides</td>
</tr>
<tr>
<td>Contrast</td>
<td>Memory colours, same hue colours next to each other, colours with only small differences next to each other</td>
</tr>
<tr>
<td>Colourfulness</td>
<td>Missing no single colour; memory colours, bright colours, shades of brown, pastel colours, neutral gray</td>
</tr>
<tr>
<td>Gloss</td>
<td>Edges with high contrast, smoothly coloured surfaces</td>
</tr>
</tbody>
</table>

More requirements arise from consideration of the viewing situation. Many studies have shown that the type of task in image viewing tends to influence gaze paths /2,50,62,71/: task-free viewing may differ from viewing for quality assessment. To make the results from quality testing experiments representative of real life situations, it seems logical to require that the test image is such that the differences in eye paths are minor. To support full reference based quality computation and assessment, the test image should also have maximal quality with respect to relevant quality attributes.

6.3  Gaze-Driven Approach for Test Image Development

A total of three reference images were created in the project. The procedure used in developing the reference images is described in the following.

The development of the first reference image from a scratch progressed in two steps. The first step involved designing studio compositions of the selected theme with different kinds of objects and finding out how feasible the objects were for assessing the critical attributes, sharpness, noise, contrast, colourfulness and gloss. Based on subjective quality assessment from prints, the first test image was constructed. The second step consisted of eye-tracking experiments, in which the balance of the test image was evaluated and the subjects’ gaze paths during quality assessment were analyzed.

6.3.1  Quality Assessment from Prints

A breakfast table with several objects was chosen as the theme. The theme was selected because it is well known in many cultures and also the objects associated with it are familiar. The idea of objects on a table is not new since it has been used in many test images before such as for instance the ISO SCID images /Figure 36/ and Sony sRGB standard images /Figure 39/.

Variations in the objects on the table were used to find the most useful for visual quality assessment by tests. In the scenes photographed as candidate images the table and the person were always the same. The visual tests were made using ink jet prints of the captured images. The digital images had been systematically manipulated in Photoshop to achieve controlled variations in the four quality attributes of interest. The manipulation was done at eleven levels for each attribute. Printing was accomplished with an Epson Stylus Pro 3800 printer in the size 100×150 mm. In order to create variation in gloss eleven grades of paper were used.

The number of naive observers assessing the suitability of different objects for quality evaluation of the selected quality attributes was 20. The test persons performed quality assessment for five sets of
prints in which one quality attribute was varied at a time (sharpness, noise, contrast, colourfulness or gloss). The test persons were not told which of the quality attributes had been varied. They were asked to circle the regions in the image that were most useful in making the quality ranking. ‘Useful’ meant that it was easy and effective to rate image quality from the region. The test room illumination in all tests was 300–400 lux.

6.3.2 Quality Assessment from a Display

Based on the data obtained by the visual tests described above, a hypothetically optimal composition of the breakfast scene was set up. The scene was photographed and minor processing performed to correct for defects. Eye-tracking experiments were performed to evaluate the balance of the test image and to analyze the subjects’ gaze paths during quality assessment. The task given to the subjects composed of two parts, a free-viewing task and a quality rating task. In the free-viewing task, the subjects were given 20 seconds to freely view the final test image. The purpose of this was to find out how balanced the test image is with respect to salience of different objects.

In the quality rating task, the subjects rated manipulated versions of the final test image on a scale from 1 (bad) to 5 (excellent) based on image quality. For this purpose a new set of manipulated images was created. The set consisted of the original image and 16 manipulated images (sharpness, noise, contrast and colourfulness manipulated one at a time at four levels). The purpose of the quality rating task was to find out how the subjects viewed the image when conducting quality assessment.

The images were shown on a 17-inch TFT, LG Flatron L1717S screen. At the pixel resolution of 1280×1024 the resolution on the display surface was about 100 ppi. The size of the images per angle of view was the same in the print and display conditions. The images were shown for ten seconds one at a time. The subjects’ eye movements were tracked using the SMI (SensoMotoric Instruments GmbH) iView X RED P/T eye tracking device /Figure 40/. Ten naive subjects participated in the tests, but due to experimental problems, the data of eight was accepted.

Figure 40 The set-up of the eye-tracking experiment.
6.4 Test Images Developed in the Project

6.4.1 Test Image I

The photographs of the different breakfast compositions were taken with a Canon EOS 5D camera fitted with a Canon 24–70 mm f/2.8L USM lens. The exposure settings were: exposure time 1/6 sec, aperture F/13, ISO 200 and focal length 40 mm. The images were saved in RAW format. After standard processing with the Adobe Photoshop software, the images were saved in TIFF format in the size 2126×1417.

The first single test image developed in the project (Test Image I) is presented in Figure 41. The content of the image can be roughly divided in four categories: woman, table with objects, landscape picture and gray wall. With reference to Table 28 the image of the person contains memory colours (skin) and shapes (face) as well as details (eyes). The landscape picture includes outdoor memory colours (sky, foliage). The table and the objects on it are of different surface materials (teddy bear, table cloth), memory shapes and colours (fruit), details (teddy bear, clock, newspaper, spectacles), and of neutral (teapot, teacup, plate, passe-partout of the landscape picture) and metallic (plate, cutlery) tones. To meet the requirements of colour assessment /cf. Table 29/, the image contains additional coloured objects (brown shirt of the woman, napkin, flowers, water glass, and picture of a dog) and a uniform surface for perception of noise (gray wall).

While Test Image I met the expectations in several aspects, it was considered to have four main areas of improvement: the busyness of the image, the control of the colour world, the salience of the object categories, and the naturalness of the event and the setting. To address the shortcomings, the image was further developed.

Figure 41 The first reference image developed in the project /111/.

6.4.2 Test Image II

The image was photographed with a professional digital SLR camera, Nikon D300, supplemented with a Sigma 18-50mm f/2.8 EX DC Macro lens. The camera was placed on a robust Manfrotto tripod. Two Nikon SB-600 flash units provided the lighting for the scene. The image was captured in 14-bit raw format. The following exposure settings were used: ISO 200, shutter speed 1/250s,
aperture f/7.1 and focal length of 31 mm. ISO 200 provided the lowest noise levels and the widest attainable dynamic range for the camera. The shutter speed of 1/250 s presented the fastest possible flash sync speed for the camera. The aperture provided a sufficient depth of field to achieve high sharpness over the whole image and was large enough to allow sufficient light for a proper exposure.

A lot of effort was put on achieving a uniform lighting with stable white balance over the whole image. Light of the left flash was bounced from the ceiling, providing a large area of soft light over the whole scene. The right flash was shot through a diffuser with lower power, resulting in a quite small area of harder light. The hard light effectively softened the shadows caused by the left flash, and provided extra light for improved illumination of the human model. In addition, the side flash effectively increased the visibility of surface textures of the objects on the table. For additional lighting, two reflectors were constructed from aluminum foil. The left reflector provided extra light for the human model, while the right reflector softened the shadows on the right corner of the table cloth. Some faint shadows were intentionally left on the image to increase image naturalness.

The captured image was edited with two different image processing software, Nikon Capture NX and Adobe Photoshop CS3. A calibrated and profiled LCD display, 24” Eizo ColorEdge CG242W, was used in editing the image. The image was processed in two phases. During the first phase, basic adjustments for saturation, contrast, curves and sharpness were applied in Nikon Capture NX software. The white balance of the image was adjusted by first photographing a 24-color GretagMacbeth ColorChecker chart in the same lighting as the test image, and later correcting the white balance with Nikon Capture NX software by choosing a neutral color field from the photographed chart as a white balance gray point. All the adjustments were done for a 14-bit RAW image.

After the corrections for the whole image were completed, the image was converted to 16-bit Adobe RGB and further processed in Adobe Photoshop CS3. The second phase adjustments included minor corrections on specific areas of the image (e.g. reflections on the landscape picture), as well as hiding specific colour targets in the image. Ten specific colour patches were chosen from the 24-color GretagMacbeth ColorChecker. The implementation of the colour targets was done by choosing ten objects in the image with colours closely matching to the desired colour targets. The average colour of each object was then slightly adjusted on a specific area to match the corresponding colour reference. Finally the test image was saved as a 16-bit TIFF image in Adobe RGB format, with a resolution of 360 dpi and a print size of 100 × 150 mm.

Figure 42 presents the second version of the test image for print quality evaluation, Test Image II. In designing the image, the following changes were made with regard to the first image. The busyness of the image was tackled by cropping the image looser and by reducing the number of objects in the image. The salience of the object categories was balanced by redirecting the woman’s gaze away from the camera towards the table and giving the outdoor image a larger role in the image. To make the event and the setting more natural, a lighter shade of gray was used for the wall. The objects on the table were carefully chosen to support a consistent theme in the image.
Figure 42  Test Image II shown in its original size (360 dpi, print size 100 × 150 mm).

Ten specific colour targets were placed in Test Image II to have a better control of the colour world. Figure 43 presents the locations of the colour targets and their corresponding colour patches on the 24-color GretagMacbeth ColorChecker. The chosen colours were blue, green, red, yellow, magenta, cyan, orange and three neutral greys.

Figure 43  Ten specific colour targets were hidden in Test Image II. The locations of the colour targets are shown in the left, and the corresponding ten patches of the GretagMacbeth ColorChecker are shown on the right.

By participant self-report, it was found that while some aspects in the image had been improved, the shortcomings of Test Image II were in the same four areas as in the first image. The image was now considered too spacious both on the sides of the image due to the loose cropping, and on the table due to too organized object placement. The role of the woman became too insignificant and her reading posture was considered unnatural. The colour world was reported unbalanced due to a large area of red in the form of the table cloth. Furthermore, the environment and the lighting were considered to be office-like instead of home-like.
6.4.3 Test Image III

The image was photographed with the same professional-level digital camera, Nikon D300, as the previous image but using different lens, exposure settings and lighting. The used lens was a Nikkor 35mm f/1.8G AF-S DX. The exposure settings used were: ISO 200, shutter speed 1/8s and aperture f/11. The sensor in the camera is Nikon APS-C format and therefore the corresponding focal length with the used fixed lens in the 35 mm format is around 52 mm. The aperture was selected so that the depth of field was large enough and all the objects appeared sharp.

Natural daylight was utilized in taking the image. The base illumination of the studio environment was adjusted with the windows covered. The base illumination was created with two Arri 650 Plus halogen lamps with 650 W and 3200 K. A filter was used to convert the color temperature of the lamps from 3200K to 5000K. The lights were reflected via the ceiling. To have a more natural lighting, two windows at the right side of composition were uncovered. Diffusers were placed in front of the windows to slightly diffuse the harsh sunlight. The spectrum of light was measured with a PhotoResearch PR-670 spectroradiometer right after capturing the image. The spectrum was measured from a LabSphere 99% Spectralon Reflectance Standard, which was placed on the table facing towards the camera. The light intensity of the spectrum was around 2700 lux and color temperature around 5700 K. The illumination was not uniform.

The post-processing of the image was also done with a 24” Eizo display. The raw image (14-bit NEF) was converted with Adobe Camera Raw 4.6 and edited as 16-bit ProPhoto RGB image in Photoshop CS3. Some editing was necessary to remove dust spots and other artefacts. The overall image contrast and sharpness were increased slightly. In addition, due to the natural light source, there was too much contrast between the left and right side of the image. To balance this, brightness was increased on the left side of the image with a gradient mask. Finally the image was saved, similarly to Test Image II, as a 16-bit TIFF in Adobe RGB format, with a resolution of 360 dpi and a print size of 100 × 150 mm.

To combine the insights of the first two images and to try to avoid their pitfalls, a third attempt, Test Image III, was made /Figure 44/. Special attention was paid to the naturalness of the image. The idea was to create a test image similar to what could be printed at home by the end-user e.g. on a commercial ink-jet paper. The naturalness of the image was improved by changing the reading posture of the woman and using daylight as a part of the lighting. The tableware was set for two people with the implication that the other person is behind the camera.
In designing the image, cropping was taken into significant consideration to address the busyness vs. spaciousness of the image. The image was cropped considerably closer than the previous image, but still retaining a sufficient distance from the setting. The cropping also balanced the salience of the object categories. In order to avoid too large salience of the woman, her gaze was kept towards the newspaper. Suitable places were sought for the objects on the table with the attempt of not creating too organized an impression. Furthermore, no specific colour targets were included in the image, but the balance of the colour world was improved by considering the hue angles of the included colours and the sizes of the colour targets. The table cloth was removed from the setting revealing the wooden surface of the table.

6.5 Goodness of Test Images Developed

6.5.1 Colour World of the Test Images

Figure 45 illustrates the colour values of the three test images plotted on the chromaticity coordinates a* (redness-greenness) and b* (yellowness-blueness) of the CIELAB colour space. From the figures it can be seen that including a vast range of colours with positive yellowness-blueness values is considerably easier than the insertion of colour values with a negative b*. In all three test images the hues of green, yellow, orange and red were well covered, but particularly in the first test image, the spread of the hues of blue and magenta was narrow. Some improvement was seen in Test Image II for the hues of blue, but overall the best spread of chromaticity was achieved with Test Image III with an improved variation also in the hues of magenta.
Figure 45  The colour values of the photographs plotted on the chromaticity coordinates a* and b* of the CIELAB colour space, from left to right Test Image I, II and III.

6.5.2 Perceived Usefulness of Object Categories and Image Regions

Figure 46 shows the perceived usefulness of the four object categories in quality evaluation of the five quality attributes. The three test images largely followed the same trends, i.e. the participants considered the usefulness of the object categories similarly, independent of the image at hand. The larger proportions of the woman in Test Image III e.g. for quality and sharpness were explained by the fact that the woman was holding the newspaper in her hands, thus receiving more hits.

Figure 46  Perceived usefulness of the four object categories for rating overall quality, sharpness, graininess, contrast and colourfulness.

Figure 47 visualizes the perceived usefulness of the object categories as heat maps based on the participants’ self-report done by circling the useful regions by pen in the questionnaires. The scale on the heat map goes from blue (considered useful by few participants) via cyan, green, yellow and orange to red (considered useful by many participants). From top down the evaluated quality attributes are the overall quality, sharpness, graininess, contrast and colourfulness.
Clearly the most important single object for print quality evaluation in the test images was the human face. On the three upmost rows, quality, sharpness and graininess, it received the most hits as being useful for evaluating the particular aspect of quality. For contrast, the objects considered most useful were the outdoor picture and its frame, along with the plant on the table. For the evaluation of colourfulness, the participants were rather unanimous in circling the fruit selection as useful. Also the plant was considered useful for colourfulness evaluation.
Figure 47  The regions in the images the participants found useful in rating print quality. From top down: overall quality, sharpness, graininess, contrast and colourfulness, from left to right: Test Images I, II and III. The scale indicates how many percent of the participants found the area useful in the evaluation. Areas circled by only one participant are not shown. To maximize the use of the scale, the dark reds at end of the scale are used already at 80 %.
6.5.3 **Subjective Appearance of the Test Images**

Figure 48 presents the perceived naturalness, balance, pleasantness and visual engagement of the three test images.

![Figure 48](image)

**Figure 48**  Perceived naturalness, balance, pleasantness and visual engagement of the test images.

Test Image I was considered as the most pleasant and visually engaging of the images, whereas Test Image III had the highest mean value for naturalness and balance. The trends of bars seem to suggest that the more natural the test image is, the less engaging it is visually. Of the test images, Test Image III had the smallest standard deviation in all four cases, which indicates that the participants were the most unanimous in evaluating its properties. One-way repeated measures ANOVAs were performed to analyze whether the differences in the mean values were statistically significant. In the case of perceived naturalness, Test Image III had the highest value with statistical significance $F(1.52, 44.0) = 3.79, p < .001$ (Greenhouse-Geisser, $\epsilon = .76$), pair-wise comparisons I-III: $p < .001$, II-III: $p < .001$. The differences between the mean values were not statistically significant for balance, $F(2, 58) = 2.89, p = .06$ (sphericity not violated), and pleasantness, $F(1.64, 47.6) = 3.26, p = .06$ (Greenhouse-Geisser, $\epsilon = .82$). For visual engagement, Test Image I had the highest value with statistical significance, $F(2, 58) = 25.5, p < .001$ (sphericity not violated), pair-wise comparisons I-II: $p < .001$, I-III: $p < .001$.

Figure 49 presents the perceived ease and certainty of evaluation. Perceived ease was measured by asking how quick the quality evaluation was using the particular test image.

![Figure 49](image)

**Figure 49**  Perceived ease and certainty of quality evaluation of the test images.

There seemed to be no significant differences between the test images in the perceived ease of evaluation for overall quality, sharpness or graininess. For the evaluation of contrast and colourfulness, Test Image II was considered to be the quickest. The evaluation of the attributes was apparently faster due to the presence of the red table cloth and the high contrast generated by the dark-coloured plant against the light gray wall. According to the performed one-way repeated
measures ANOVAs, the best image could be distinguished statistically only in the case of
colourfulness evaluation, for which Test Image II had the highest value of perceived ease, $F(2, 58) = 10.1, p < .001$ (sphericity not violated), pair-wise comparisons I-II: $p < .01$, II-III: $p = .001$. For contrast evaluation, only Test Images I and II differed with a statistical significance, $F(2, 58) = 4.61, p < .05$ (sphericity not violated), pair-wise comparisons I-II: $p < .05$, II-III: $p = 0.10$.

The perceived certainty of evaluation was of the same magnitude for all other quality attributes but
colourfulness. Test Image II had a slightly higher value than the other two images. Based on the
one-way repeated measures ANOVA, the mean value was not, however, highest with a statistical
significance, as it only differed statistically from Test Image I, $F(2, 58) = 4.17, p = < .05$ (sphericity
not violated), pair-wise comparisons I-II, $p < .05$, II-III, $p = .18$.

6.5.4 Conclusions

Each of the three test images has its advances. Test Image I was considered visually highly
engaging (mean 5.5, std 1.0), with a statistical significance (I-II: $p < .001$, I-III: $p < .001$). Test
Image II, on the other hand, stood out in the perceived ease of colourfulness evaluation. While no
differences were found in the evaluation of other quality attributes, colourfulness evaluation from
Test Image II was considered the easiest with a statistical significance (I-II: $p < .01$, II-III: $p = .001$). Test Image II can also easily be used for the objective evaluation of colourfulness, as there
are ten specific colour targets hidden in the image.

Test Image III was considered very natural (mean 5.9, std 0.8) by the participants; the difference
with the two other test images was statistically significant (I-III: $p < .001$, II-III: $p < .001$). As for
the balance and pleasantness of the test images, no statistically significant differences were found
between the images. The self-report data from the participants suggested that the more natural the
test image is the less engaging it is perceived visually.

The participants agreed best on the properties of Test Image III, as it had the smallest standard
deviation for naturalness, balance, pleasantness and visual engagement. It also covered the widest
range of chromaticity of the three test images. In addition, the cropping of the image was found to
solve the spaciousness vs. busyness problematic of the image as the participants did not bring up the
issue in their comments on Test Image III.

To conclude, three of the four improvement areas of Test Images I and II were addressed efficiently
in Test Image III. The salience of the objects left a bit to be desired as the placement of the
newspaper in the woman’s hands re-increased the salience of the woman. However, as heavy
emphasis was placed on the naturalness of the event and the setting in the test image, the final
version of the test image, Test Image III, met its expectations perfectly.
7 VISUAL QUALITY INDEX FOR DIGITALLY PRINTED IMAGES – DREAMS AND REALITY

This chapter is devoted to discussion of the results of the study and beyond from the standpoints of the different research approaches. The tone is even critical and tries to pinpoint the gaps which still exist in the research area of the project.

7.1 Influence of Context and Image Content

The purpose of the quality index is to give an estimation of the quality that potentially can be achieved by using certain paper. How this potential is used is dependent on both technical context (e.g. printing) and use context. A static quality index cannot predict people’s experience of quality in all situations; the question of the experience of quality in finished products is far more complicated and beyond the scope of this project.

The other issue concerns the number of images and their presentation. The mean quality rating that is given to, e.g. a single paper grade, is dependent on the image contents used in the evaluation. What is a representative set of contents in this case? Most subjective tests were made with only three contents. However, it was found out that the subjective ratings for these images, which represented quite dissimilar subjects, were surprisingly similar. It seems that in case of paper with electrophotographic or ink-jet printing, image content seems to have quite little moderating influence on the perception of subjective quality. Most likely this is due to the context of quality variation arising from paper; even in cases where image content has an effect on the relevant quality attributes, the different attributes correlate strongly and suppress the influence of image content.

Figure 51 shows the mean quality ratings for different contents in CC4 (EPG) and Figure 51 for CC5 (IJ). There are some papers that have larger variation between contents (e.g. with EPG samples 6, 12, 13, 14, 15, 16), but difference is less than one on a five-point scale. CC4 had the largest variation between contents.

![Graph showing mean quality ratings and 95% confidence intervals for EPG samples (numbers 1-21).](image)

*Figure 50* Mean quality ratings and 95% confidence intervals for EPG samples (numbers 1-21).
Image content is likely to influence quality assessment at all levels of image information. Jaimes and Chang /35/ define ten levels. This suggests that algorithmic control of content effects is highly challenging and is largely an open question.

In this project three approaches were employed to tackle the role of content. These were all low level approaches. The reduced reference approach uses computed knowledge about the distribution of image energy and colour, and hypotheses about the visibility of quality characteristics. The methodology can be further enhanced for computation of quality attributes irrespective of the model through which the attributes are combined to predict overall quality. In the full reference approach, content is taken into consideration on pixel and local levels. Also this methodology shows promise. Its areas of application are different from the reduced reference approach. The RR methodology does not require the existence of a digital original, but quality can be computed from any printed image.

The third approach follows fully different argumentation than the previous two, as it consisted of the systematic development of image content to produce a universally usable reference image, as discussed in Chapter 6. To compose a balanced – in the sense of content and quality characteristics – setting for image capture proved to be demanding, but appears to be doable.

7.2 Visual Quality Index Model

The subjective visual quality experience is without doubt multidimensional, but yet observers are capable of assessing overall quality. There is variation between observers and they may score samples differently. Also when differences between samples are small, one observer may be somewhat inconsistent over time. This may mean that VQI needs to be defined as a mean and some range of variation.

The situation motivates treating overall quality and its elements as probability distributions, and the Bayesian theory provides the necessary tool for analysis. Bayesian network provides an attractive model as it is a probabilistic model that represents a set of random variables (objective measurements and visual experiences) and their conditional independencies via directed graph. This was discussed in the context of Bayesian networks in Sections 1.6 and 5.5.

Figure 51 Mean quality ratings and 95% confidence intervals for IJ samples (numbers 16-36).
As the ultimate result of this work, we have advanced the preliminary idea and developed a method which automatically optimises the structure of the Bayesian network used as overall print quality model. This is done by making hypotheses about the behaviour of the overall quality with respect to objective measures (prior) and by computing the model fitness by simulation. The purpose of making hypotheses is to make the network obey prior known heuristics (e.g. if the mottling increases the quality should decrease).

The optimisation algorithm used is a genetic algorithm mainly due to the complex nature of the optimisation problem and the need for simulation for fitness evaluation of the solution. For all members of each population the network weights are trained using the standard expectation maximization (EM) algorithm. As a complete novel contribution we have introduced a method which automatically learns a full Bayesian network from psychophysical data and instrumental measurements described in the previous sections. The network can be used as a unified model explaining the phenomenon and as a more practical tool it acts as a model producing a single visual quality index (VQI) for any printed product.
8 CONCLUSIONS

The study was initiated with a vision of a “transparent” imaging pipeline from paper through printing to final viewing in terms of the build-up of quality. The vision reflected the general change of thought from reasoning that product quality is determined by production quality to reasoning that quality arises when end users interact with the products and do assessment and get impressions and experiences. In the study the users were in the consumer role. The product context was limited to separate printed images, and did not include some recognizable product context.

The vision was achieved in two senses. First, a quality framework and the most important quality attributes were identified and presented as a hierarchical model. Second, different modelling approaches were analyzed to find the quantitative relations between the layers of the model and across the model from the bottom layer to the topmost layer – representing overall image quality. Modelling of overall subjective visual quality with objectively determined quality attributes as variables proved to be feasible using different approaches. The Bayesian network approach was the most comprehensive showing a lot of promise, but requiring a lot of data for establishing the model. Computation of a visual quality index for specifying the quality potential of paper in a given context either as a single number or as an average and a range for the confidence interval is one of the applications of the models.

In the study image content exerted a strong influence on quality assessments of different papers printed using optimized settings on a given device. Approaches to manage the influence of image content included the reduced-reference and full-reference approaches. Both were found to have potential to come up with quality measures which separately or as simple combinations showed statistically highly significant linear correlations with visual quality. A fully different way of thought was behind development of one single test image to be a representative of the whole quality content space. The success of this approach needs further testing.
9 REFERENCES

## APPENDICES

### Appendix 1  Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>BN</td>
<td>Bayesian Network</td>
</tr>
<tr>
<td>CC</td>
<td>Concept Case / Correlation Coefficient</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>CIE</td>
<td>Commission Internationale de l'Eclairage</td>
</tr>
<tr>
<td>CMS</td>
<td>Cumulative Match Score</td>
</tr>
<tr>
<td>CSF</td>
<td>Contrast Sensitivity Function</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DF</td>
<td>Degrees of Freedom</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>EPG</td>
<td>Electrophotography</td>
</tr>
<tr>
<td>FERET</td>
<td>Facial Recognition Technology</td>
</tr>
<tr>
<td>FR</td>
<td>Full-Reference</td>
</tr>
<tr>
<td>GGD</td>
<td>General Gaussian Distribution</td>
</tr>
<tr>
<td>GLPF</td>
<td>Gaussian Low-Pass Filter</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>HSD</td>
<td>Honestly Significant Difference</td>
</tr>
<tr>
<td>HSV</td>
<td>Hue Saturation Value</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>iCAM</td>
<td>Color Appearance Model</td>
</tr>
<tr>
<td>ICC</td>
<td>International Colour Consortium</td>
</tr>
<tr>
<td>IFC</td>
<td>Information Fidelity Criterion</td>
</tr>
<tr>
<td>IJ</td>
<td>Inkjet</td>
</tr>
<tr>
<td>IQ</td>
<td>Image Quality</td>
</tr>
<tr>
<td>IQC</td>
<td>Image Quality Circle</td>
</tr>
<tr>
<td>IQM</td>
<td>Image Quality Metric</td>
</tr>
<tr>
<td>JND</td>
<td>Just Noticeable Difference</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MP</td>
<td>Multipurpose</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>NPS</td>
<td>Noise Power Spectrum</td>
</tr>
<tr>
<td>NR</td>
<td>No-Reference</td>
</tr>
<tr>
<td>NSRPJ</td>
<td>Normalized Sum of Rational Partial Juries</td>
</tr>
<tr>
<td>NSS</td>
<td>Natural Scene Statistics</td>
</tr>
<tr>
<td>PSNR</td>
<td>Peak Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Square</td>
</tr>
<tr>
<td>RR</td>
<td>Reduced-Reference</td>
</tr>
<tr>
<td>RRIQA</td>
<td>Reduced-Reference Image Quality Assessment</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale-Invariant Feature Transform</td>
</tr>
<tr>
<td>SROCC</td>
<td>Spearman Rank Order Correlation Coefficient</td>
</tr>
<tr>
<td>SSIM</td>
<td>Structural Similarity Metric</td>
</tr>
<tr>
<td>UQI</td>
<td>Universal Quality Index</td>
</tr>
<tr>
<td>VDP</td>
<td>Visual Difference Predictor</td>
</tr>
<tr>
<td>VIF</td>
<td>Visual Information Fidelity</td>
</tr>
<tr>
<td>VQI</td>
<td>Visual Quality Index</td>
</tr>
</tbody>
</table>
Appendix 2  Publications and Theses

**Journal Articles**


**Conference Papers**


MSc Theses


MSc Dissertation

## Appendix 3  Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha rationality</td>
<td>A measure of the coherence of the assessment made by an evaluator or a jury. Alpha can have values between zero and one. The larger the value of alpha rationality, the more rational decisions the person/jury is making.</td>
</tr>
<tr>
<td>Band-pass filtering method</td>
<td>A method that passes frequencies within a certain range and rejects frequencies outside that range.</td>
</tr>
<tr>
<td>Bayesian network (BN)</td>
<td>A directed graphical model in which each node represents a random variable and the links express probabilistic relationships (conditional dependences) between these variables. In this project, Bayesian network has been proposed as an image quality measurement model in which the nodes represent image quality attributes.</td>
</tr>
<tr>
<td>Category scaling</td>
<td>Psychometric method of placing the stimuli in categories according to a magnitude of a certain attribute.</td>
</tr>
<tr>
<td>CIELAB</td>
<td>Device independent colour space with dimension L for lightness, and dimensions a (redness-greenness) and b (yellowness-blueness) for chroma.</td>
</tr>
<tr>
<td>Clarity</td>
<td>The degree of how easy it is to discriminate objects from an image.</td>
</tr>
<tr>
<td>Computational method</td>
<td>A method used for computing quality from a digital original or a digitized printed photograph.</td>
</tr>
<tr>
<td>Context</td>
<td>The circumstances in which the image evaluation is performed and which affect the interpretation of the problem setting thus affecting the outcome of the evaluation.</td>
</tr>
<tr>
<td>Contrast sensitivity function (CSF)</td>
<td>Description of contrast sensitivity (in HVS) with respect to intensity and/or frequency.</td>
</tr>
<tr>
<td>Correspondence analysis</td>
<td>A descriptive, exploratory statistical method representing the rows and columns of a two way contingency table in a joint plot.</td>
</tr>
<tr>
<td>Cumulative match score (CMS)</td>
<td>A method for the investigation of the similarity (explanatory power) between two evaluation spaces, here, the instrumental measures and human evaluation results. The CMS curve is the rank n versus percentage of samples for which the ground truth match (human evaluation) is within the set of n closest samples in the space of instrumental measures.</td>
</tr>
<tr>
<td>Cumulative match score histogram</td>
<td>Histogram of CMS values, describes how many “guesses” are required from an objective method until if converges to the same result to as ground truth (monotonically increasing).</td>
</tr>
<tr>
<td>Dominant colour contrast</td>
<td>Statistically most frequent colour tone (in the project).</td>
</tr>
<tr>
<td>Factor analysis</td>
<td>A multivariate statistical method for analyzing whether the correlations between a set of observed variables stem from their relationship to one or more latent variables in the data, each of which take the form of a linear model.</td>
</tr>
<tr>
<td>First scale of image</td>
<td>High-resolution representation of an image.</td>
</tr>
<tr>
<td>Full-reference metric</td>
<td>A metric that requires a reference image for the quality computation.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Graininess (objective)</td>
<td>High to mid frequency unwanted random or fixed pattern intensity distortion on image</td>
</tr>
<tr>
<td>Graininess (subjective)</td>
<td>Subjective experience of the degree of granularity in an image, synonymous with noise</td>
</tr>
<tr>
<td>Ground truth</td>
<td>Subjectively assessed quality (in this project)</td>
</tr>
<tr>
<td>Grouping</td>
<td>Method of dividing a set of stimuli, such as image samples, into distinct categories</td>
</tr>
<tr>
<td>High-level quality attribute</td>
<td>An abstract subjective attribute which cannot be directly related with an objective property of an image, e.g. naturalness</td>
</tr>
<tr>
<td>Human visual system (HVS) models</td>
<td>Computational models used in computer vision and image processing to deal with biological and psychological processes of the human visual system that are not yet fully understood. Such models are used to simplify the behaviour of the complex system</td>
</tr>
<tr>
<td>Image content adaptive</td>
<td>Method takes local (and/or global) information of the studied image into account</td>
</tr>
<tr>
<td>Image content</td>
<td>The subject matter of an image, all the objects and the setting of an image</td>
</tr>
<tr>
<td>Image quality attribute</td>
<td>See high/low-level quality attribute</td>
</tr>
<tr>
<td>Image quality metric</td>
<td>The specific measurement of a given attribute or overall quality</td>
</tr>
<tr>
<td>Image quality model</td>
<td>A mathematical model which predicts the image quality rating using the values of physical parameters, instrumental measurements, image quality metrics or image attributes</td>
</tr>
<tr>
<td>Instrumental measurement</td>
<td>A measurement made from a simple test target with an instrument</td>
</tr>
<tr>
<td>Jury</td>
<td>A group of human observers</td>
</tr>
<tr>
<td>Just noticeable difference (JND)</td>
<td>The smallest difference in the sensory input that can be noticed by an observer. The JND usually reported is the difference that an observer notices on 50% of trials or 75% of a group notices in one trial</td>
</tr>
<tr>
<td>Low-level quality attribute</td>
<td>A concrete attribute that is associated with an objective property of an image, such as graininess</td>
</tr>
<tr>
<td>Memory colour</td>
<td>A colour that has a special association for a human observer, examples of such colours that have a memory reference are grass green, sky blue and skin tone</td>
</tr>
<tr>
<td>Minkowski method</td>
<td>Principle is to ensure that overall IQ value cannot be better than the lowest quality contributor of quality model</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean opinion score, average rating of quality given to an image by a group of observers</td>
</tr>
<tr>
<td>Multidimensional scaling</td>
<td>Statistical methods for revealing an underlying multidimensional space for a set of distance or similarity data</td>
</tr>
<tr>
<td>Mutual information (MI)</td>
<td>A statistical measure of interdependence that indicates how much the uncertainty of one variable is reduced by knowing the value of the other variable. MI is not restricted to linear dependence. MI is always non-negative</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>----------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Natural image</td>
<td>Photograph</td>
</tr>
<tr>
<td>Natural scene statistics (NSS)</td>
<td>Refers to computational models which compute statistics typical to natural scenes</td>
</tr>
<tr>
<td>Naturalness</td>
<td>How well an image represents the real world according to an observer</td>
</tr>
<tr>
<td>No-reference metric</td>
<td>A metric that operates without a reference image, attempts to imitate the evaluation of a human observer</td>
</tr>
<tr>
<td>Normalized sum of rational partial juries NSRPJ</td>
<td>A measure of the coherence of the assessment made by a jury. NSRPJ is computed as the ratio of rational partial juries to the total number of partial juries that can be formed from the original jury. Partial juries are considered rational if they form no circular preferences.</td>
</tr>
<tr>
<td>Objective measurement</td>
<td>Measurement made with an instrument or a computational method, is not affected by subjective bias</td>
</tr>
<tr>
<td>Overall quality</td>
<td>The overall rating of quality for (or an overall value of) an image that is dependent of all quality-related attributes of that image</td>
</tr>
<tr>
<td>Pair-wise comparison</td>
<td>Psychometric method of evaluating differences in magnitude between two stimuli and placing them on a scale</td>
</tr>
<tr>
<td>Pixel-block method</td>
<td>Method uses pixel block (e.g. 8 × 8 pixel area) specific data for computations</td>
</tr>
<tr>
<td>Pixel-wise method</td>
<td>Method uses pixel specific data for computations</td>
</tr>
<tr>
<td>Printed image</td>
<td>Printed photograph</td>
</tr>
<tr>
<td>Qualitative methodology</td>
<td>Methods for finding qualitative differences and relations between phenomena, not measuring predefined variables (in this project)</td>
</tr>
<tr>
<td>Reduced-reference metric</td>
<td>Metric extracts features from the reference image and employs them as side information in the assessment</td>
</tr>
<tr>
<td>Reference image</td>
<td>Original image that is considered to be of perfect quality or free of distortions</td>
</tr>
<tr>
<td>Relative MOS</td>
<td>MOS value that is given in relation with the best and the weakest sample of an image set</td>
</tr>
<tr>
<td>RMS contrast</td>
<td>Root-mean-square value of the luminance channel of a digitized photograph (in the project)</td>
</tr>
<tr>
<td>RMS noise</td>
<td>Root-mean-square value of the luminance channel of a digitized 50% gray field (in the project)</td>
</tr>
<tr>
<td>Second scale of image</td>
<td>Reduced-resolution approximation of an image</td>
</tr>
<tr>
<td>Semi-structured interview</td>
<td>Interview that has a structure, which can be altered during the interview depending on the answers of the interviewee</td>
</tr>
<tr>
<td>Sensitivity of Bayesian Network</td>
<td>Sensitivity refers to how much the output probability distribution changes if small change is made in the parameters or in the observed values of input variables of the model</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Sub-bands</td>
<td>A frequency axis can be divided into a number of different frequency bands</td>
</tr>
<tr>
<td>Subjective evaluation</td>
<td>Evaluation done by a human observer, is affected by subjective bias</td>
</tr>
<tr>
<td>Subjective experience of visual quality</td>
<td>Visual quality as it is individually experienced, including all the aspects that are subjectively related to quality</td>
</tr>
<tr>
<td>Test image</td>
<td>Digital or printed natural image used for subjective evaluation or objective measurement</td>
</tr>
<tr>
<td>Test target</td>
<td>Contains one or several technical test fields</td>
</tr>
<tr>
<td>Tukey’s HSD</td>
<td>A measure for determining statistically significant differences (at a chosen significance level)</td>
</tr>
<tr>
<td></td>
<td>between sample groups. One group can consist of, for instance, the ranking points given by several jurors to one printed image, and the group is represented by the mean of these ranking points. If two groups are further away from each other than HSD, the difference between their average ranking points is statistically significant</td>
</tr>
<tr>
<td>Universal quality index (UQI)</td>
<td>A method for assessing quality of any images and providing a quality value in correspondence to average human evaluation</td>
</tr>
<tr>
<td>Visual quality index</td>
<td>Index for visual quality calculated from the objectively measured properties of an image or paper in a given context</td>
</tr>
<tr>
<td>Wavelet</td>
<td>Mathematical functions similar to sinusoidal waves, but delay (i.e. localized in space)</td>
</tr>
<tr>
<td>Wavelet domain</td>
<td>Domain where signal is decomposed into different scale and orientation components using a mathematical function called wavelet</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

The coordinator of the project, the Department of Media Technology at Aalto University extends its thanks to all collaborators in the project for a very productive and fruitful working atmosphere. The role of the industrial partners, KCL, Mondi Business Paper, M-Real, Stora Enso and LabVision Technologies, and funding partners, Tekes and GTTS, is gratefully acknowledged.