Full Reference Printed Image Quality: Measurement Framework and Statistical Evaluation

Tuomas Eerola*, Lasse Lensu, and Heikki Kälviäinen

Machine Vision and Pattern Recognition Research Group (MVPR),

Lappeenranta University of Technology,

P.O. Box 20, FIN-53851 Lappeenranta,

Finland (phone: +358 5 6212852, fax: +358 5 6212899,

e-mail: tuomas.eerola@lut.fi; lasse.lensu@lut.fi; heikki.kalviainen@lut.fi).

Joni-Kristian Kamarainen

MVPR/Computational Vision Group (Kouvola Unit),

Lappeenranta University of Technology, P.O. Box 20,

FIN-53851 Lappeenranta, Finland (phone: +358 5 6212852,

fax: +358 5 6212899, e-mail: joni.kamarainen@lut.fi).

Tuomas Leisti and Göte Nyman

Department of Psychology, University of Helsinki, P.O. Box 9, FIN-00014 Helsinki,

Finland (e-mail: tuomas.leisti@helsinki.fi; gote.nyman@helsinki.fi).

Raisa Halonen and Pirkko Oittinen

Department of Media Technology, Helsinki University of Technology, P.O. Box 5500,

FIN-02015 Espoo, Finland (e-mail: raisa.halonen@tkk.fi; pirkko.oittinen@tkk.fi).

* Corresponding author

Abstract

Full reference image quality algorithms are standard tools in digital image processing, but have not been utilised for printed images due to a "correspondence gap" between the digital domain (a reference) and physical domain (printed sample). In this work, we propose a framework for applying full reference image quality algorithms to printed images. The framework consists of accurate scanning of printed samples, and automatic registration and descreening procedures which bring the scans in correspondence with their digital originals. We complete the framework by incorporating state-of-the-art full reference algorithms to it. Using data from comprehensive psychophysical experiments of subjective quality experience, we benchmark the state-of-the-art methods and point out similar results in the digital domain: the best digital full reference measures, such as the recently introduced visual information fidelity (VIF) algorithm, perform best also for printed media.

Keywords: Print quality, image quality, quality assessment, image quality measure, full reference.

1. INTRODUCTION

Full reference (FR) image quality assessment refers to methods which evaluate visual quality based on an original image (reference) representing the "ideal quality". This is the main approach for evaluating and comparing the quality of digital images, but despite the recent digitalisation of printing technologies, it is rarely used for printed media. The reason is obvious: in the pipeline from a digital original to a print and again to a (scanned) digital image, the original undergoes various transformations which alter the image in a way that the FR assessment methods cannot be used.

In this work, we solve the problem with a framework which allows to transform the best digital full reference image quality assessments for the use with prints. We show how the effects of printing and scanning can be avoided by special image registration and descreening procedures which bring the scanned image into correspondence with the original. We incorporate state-of-the-art full reference methods to our framework and benchmark them with an extensive set of printed images and psychophysically defined ground truth. The test samples (prints) used in the study were selected by media technology experts and reflect real quality inspection problems, giving us a challenging and meaningful test set. The ground truth was collected through psychophysical experiments designed by psychologists. The subjective tests were carried out by placing the hardcopy samples on a table and giving the subjects appropriate evaluation tasks. This gave us a more intuitive and versatile "user interface" than the common computer display approach where the evaluators can see only one or a few samples at a time. As a consequence, the results are generally noteworthy for the FR image quality research. To substantiate our findings, we report an extensive statistical analysis of the the FR methods and rank them to find out the best ones. Our main contributions are the framework, where FR assessments can be applied to printed images, and the comprehensive evaluation with true images and psychophysical ground truth revealing the best performing FR methods for printed media.

The article is organised as follows. In Sect. 2, we describe the framework for computing FR image quality assessment algorithms for printed images. The theory of full reference quality assessment algorithms is presented in Sect. 3, as well as a set of state-of-the-art quality assessment algorithms selected for this study. In Sect. 4, we briefly explain the selection of our test samples and the psychophysical subjective evaluations forming the ground truth.

The results are given in Sect. 5, discussed in Sect. 6 and concluded in Sect. 7.

1.1. Related Work

This study originates from the paper and printing industry and science, and therefore, our presentation reflects the results and terminology from those fields. However, the perceived quality considered is a general phenomenon appearing in any media (electronic paper, displays, etc.). Measuring it is a challenging task since the ultimate quality evaluation can be given only by an end-user who evaluates it qualitatively, subjectively, and dependent on the context (product type, cultural factors, etc.). Understandably, industrial evaluations are conducted by human observers, but recent developments in computer vision and image processing have opened up intriguing possibilities to automate the assessments. Computational modelling of visual quality with the help of machine vision is, however, a complicated task which has not been solved. In this study, we aim to solve the problem by bringing digital image evaluation methods to the world of printed products. The work in this study can be seen as the "fusion of printed and digital visual quality assessments". In the following, we review the most important and influential works from both the fields.

Previous efforts to automate print quality evaluation using machine vision have focused on automating the current manually performed assessments, or measuring distinct quantities connected to the printing technology. For example, the KDY ImageXpert method examines selected parts of a printed test pattern and returns various indicative values, such as the roundness of a dot or edge raggedness of a line [1, 2]. The ISO 19751 standard [3], currently under development, proposes the following attributes to be measured: micro-uniformity, macro-uniformity, colour rendition, text and line quality, gloss, sharpness, and spatial adjacency. The human visual system (HVS) is partially modelled in more sophisticated methods such as the automatic evaluation of subjective unevenness in solid printed areas (e.g., [4–12]). All of these methods, however, measure only one factor of quality, possibly restricted by known physiological facts (e.g., orientation or frequency sensitivity).

Digital image quality research, on the other hand, has produced several quality assessment algorithms to measure the overall quality of images so that the result is consistent with the subjective human opinion. A good introduction and comparison has been published by Sheikh et al. [13]. In general, the image quality assessment (QA) algorithms can be

divided into full-reference (FR) and no-reference methods (NR) according to whether an input image is compared to a known reference image (FR), or the reference does not exist (NR). The FR image quality assessment algorithms are commonly used in investigations of image compression, data transmission, display optimisation, etc. However, these results are very application specific, their subjective test material is limited, and most importantly, the results are not transferable to print quality evaluation. In [14], a method to study the quality of a printed image by comparing it to its referential version has been presented. However, the method only detects discrepancies between the aligned pixels or regions of two printed and scanned images, and does not concern the overall quality of print. In [15], a colour reproduction quality metric for printed natural images based on the S-CIELAB model was presented. The method computes only colour differences between two printed samples, and the registration is done using registration marks, so it cannot be considered as a general-purpose method. The no-reference image quality assessment is a much more difficult task, and most of the proposed NR image quality assessment algorithms are designed for a single distortion type. The NR image quality assessment is not within the scope of this paper.

Comparisons of FR QA algorithms have been presented in literature. In [16], a set of simple mathematical image quality metrics, such as the average difference, and two graphical metrics known as histograms and Hosaka plots have been compared against a subjective evaluation. The test set consisted of compressed images with four different compression techniques. In [17], two well-known FR QA algorithms, the Visual Difference Predictor [18] and the Sarnoff Visual Discrimination Model [19], have been compared by using a test set composed of computer generated patterns, synthesised images and natural pictures – the distortions used were blurring, patterned noise and quantisation. In [20], three different QA algorithms have been evaluated using JPEG compressed images with different bit rates. In [21], a number of QA algorithms including mathematical distance metrics and human visual system based methods have been compared. The test set of the study consisted of compressed images, blurred images and Gaussian noise. The most notable evaluation of FR QA algorithms with a solid statistical significance analysis has been presented by Sheikh et al. [13]. They have compared the state-of-the-art image quality assessment algorithms using an extensive set of samples and subjective tests. Our statistical analysis was influenced by their work.

The main reasons for conducting this study and reporting its novel results are i) none

of the above-mentioned studies have considered the quality of printed images, and ii) the distortion types in the earlier works have been artificial distortions, such as Gaussian noise and blur, or in more realistic settings, image compression and transmission artefacts, but these do not allow extensive generalisation and are certainly not characteristic of printed media. Our approach differs from most of the earlier comparisons of image QA algorithms since we are not interested in the overall visual quality of an image within some abstract quality scale. Instead, we try to evaluate the overall visual quality of print dependent on the paper grade but independent from the image content. There does not exist a single measure to properly estimate the overall quality of paper for printed photograph applications, and thus, the scope of this paper is to test how the FR QA algorithms suit for this problem.

2. FULL REFERENCE IMAGE QUALITY FRAMEWORK FOR 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 exist no rigid, partly rigid or non-rigid (elastic) spatial shifts between the images, and the compression should retain at least 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. In the following, we describe our system where well-known methods are combined to form a novel automatic framework for analysing full reference image quality of printed products.

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 since 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 information. This can be achieved by using the 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 QA algorithm 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 discussion above, it is possible to sketch the main structure of our framework. The framework was originally presented by the authors in [22], but we will briefly describe it in the following. The framework structure and the data flow are illustrated in Fig. 1. 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 [23]. 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 [24] (in ideal conditions the acuity limit of the human eye can be as small as 0.017° which corresponds to 0.1 mm [25]). 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.

2.1. Rigid Image Registration

Rigid image registration was 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) [26]. Registration based on the SIFT features has been utilised, for example, in mosaicing panoramic

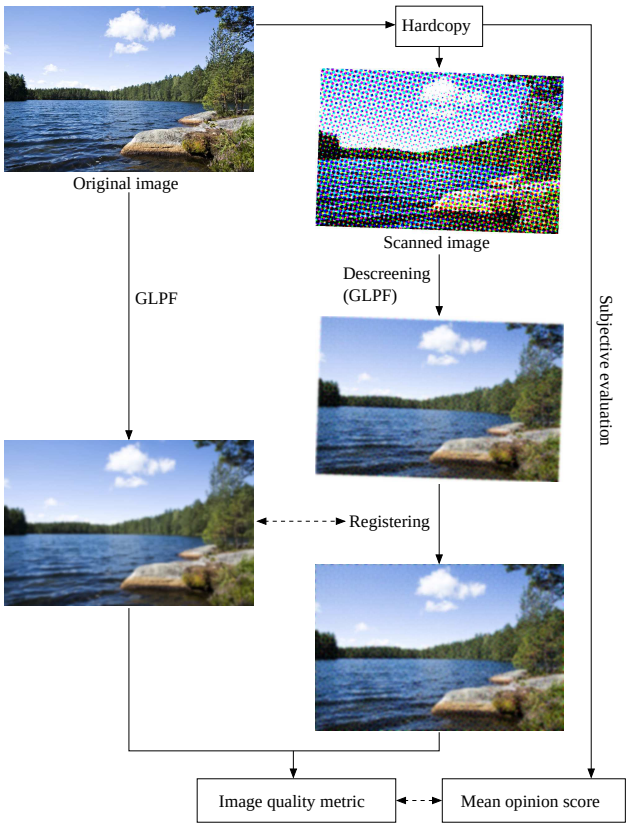


FIG. 1: The structure of the framework and data flow for computing full-reference QA algorithms for printed images.

views [27]. 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.

Our method performs a scale and rotation invariant extraction of local features using SIFT. The SIFT method also provides the descriptors which can be used for matching. As a standard procedure, the random sample consensus (RANSAC) principle presented in [28] is applied to find the best homography using exact homography estimation for the minimum number of points and linear estimation methods for all "inliers". The linear methods are robust and accurate also for the final estimation since the number of correspondences is typically quite large (several hundreds of points). In our framework the implemented linear homography estimation methods are Umeyama for isometry and similarity [29], and the restricted direct linear transform (DLT) for affine homography [30]. The only adjustable parameters in our method are the number of random iterations and the inlier distance threshold

for the RANSAC which can be safely set to 2000 and 0.7 mm, respectively. This makes the whole registration algorithm practically parameter free. For the image transformation, we utilise standard remapping using bicubic interpolation. In [22], the presented registration algorithm was shown to be accurate for printed samples when an affine transformation was used.

2.2. Image Quality Computation

In the case of printed image quality assessment, FR QA algorithms have some special requirements. Although the above-mentioned registration works well, subpixel errors do occur. Because of this, simple pixelwise distance formulations, such as the root mean square error (RMSE), do not work well. In other words, a good FR QA algorithm 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, FR QA algorithms that just compute simple similarity between the reference image and the input image, do not succeed. In the next section, we will discuss different approaches to FR QA algorithms and their differences. The QA algorithms selected for our statistical evaluation are discussed in more detail.

3. FULL REFERENCE IMAGE QUALITY ASSESSMENT ALGORITHMS

Several approaches to develop FR QA algorithm have been proposed. Generally, FR QA algorithms can be divided into three groups: arbitrary signal fidelity criteria, HVS error-based methods, methods that use natural scene statistics. 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. Methods in the third group examine the problem from an information theoretic point of view.

3.1. Arbitrary Signal Fidelity Criteria

Several mathematical distance formulations that compute similarity or dissimilarity between two matrices (images) have been evaluated in [16] and [21]. 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. MSE can be generalised to colour images by computing the Euclidean distance in the perceptually uniform CIE L*a*b* colour space as

$$LabMSE = \frac{1}{N^2}\sum_{i,j=0}^{N-1}[\Delta L^*(i,j)^2 + \Delta a^*(i,j)^2 + \Delta b^*(i,j)^2], \tag{1}$$

where $\Delta L^*(i,j)$, $\Delta a^*(i,j)$ and $\Delta b^*(i,j)$ are differences for the colour components at point (i,j). This metric is known as the L*a*b* perceptual error [21].

In [31], a method to apply fuzzy similarity measures to image quality assessment was presented. In the method, grey-level images were treated as fuzzy sets. Prior to computing the similarity, the images were further divided into regions with different weights to better adapt to the human perception. The method was extended in [32] by combining histogram similarity measures to the earlier pixelwise similarity measures, and in [33], it was generalised to colour images.

The universal quality index (UQI) was introduced in [34], and a further improvement was presented in [35] 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 [35]. The HVS is assumed to be highly adapted to structural information [36], and structural distortions should be treated in a different manner than distortions arising form variations in lightning, such as brightness and contrast changes. SSIM is defined as

$$SSIM = l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma \tag{2}$$

where

$$l(x,y) = \frac{2\bar{x}\bar{y} + C_1}{\bar{x}^2 + \bar{y}^2 + C_1} \tag{3}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{4}$$

$$s(x,y) = \frac{2\sigma_{x,y} + C_3}{\sigma_x\sigma_y + C_3} \tag{5}$$

$l(x, y)$ is the luminance distortion (comparison based on mean intensities), $c(x, y)$ is the contrast distortion (comparison based on standard deviations of the intensities), and $s(x, y)$ is the structural distortion (correlation after luminance and contrast normalisation). UQI is a special case of SSIM with $\alpha = \beta = \gamma = 1$ and $C_1 = C_2 = C_3 = 0$. This is why UQI gives unstable results when the mean intensities or intensity variances of reference and input images are very close to each other. However, for printed images this is hardly ever the case. For image quality assessment, UQI and SSIM are applied locally. In UQI, this is implemented by using an 8×8 window which moves pixel-by-pixel over the image, while in SSIM, windowing is performed by using an 11×11 circular-symmetric Gaussian weighting function.

3.2. HVS Error-based Methods

A distorted image can be divided into an undistorted reference signal and an error signal. A typical HVS image QA algorithm is based on the assumption that the loss of perceptual quality is directly related to the visibility of the error signal [35]. These QA algorithms operate by weighting different aspects of the error signal based on their visibility. The approach was first introduced by Mannos and Sakrison [37]. Other popular HVS error based methods are the Visual Difference Predictor (VDP) by Daly [18] and the Sarnoff Visual Discrimination Model [19].

A typical scheme for computing HVS error-based QA algorithms consists of the following steps: preprocessing, contrast sensitivity function (CSF) filtering, channel decomposition, error-normalisation and error pooling [35]. The preprocessing 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 subbands (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 [35].

The HVS model of VDP [18] 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.

In noise quality measure (NQM) [38], a model-restored image is computed instead of the residual image. This means that both the original and degraded input image are passed through a restoring algorithm using the same parameters. The NQM is defined as

$$NQM = 10 \log_{10} \left(\frac{\sum_i \sum_j O_s^2(i,j)}{\sum_i \sum_j (O_s(i,j) - I_s(i,j))} \right), \tag{6}$$

where $O_s(i,j)$ is the simulated version of the model restored image and $I_s(i,j)$ is the restored image. The HVS model in NQM is based on Peli's contrast pyramids [39].

Perceptual Image Difference (PDiff) [40] was originally developed for the needs of image rendering. The method improved earlier HVS error-based QA algorithms by including spatial and temporal information.

Several limitations of the HVS error-based QA algorithms are listed in [35]. In brief, it is not clear that the fundamental assumptions of the HVS error-based QA algorithms, 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.

3.3. Image Quality Assessment Algorithms 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. A review on the statistical modelling of natural images and their applications can be found in [41]. 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 [42].

The information fidelity criterion (IFC) based on the NSS has been presented in [42]. 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 [43] in the wavelet domain. 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 visual information fidelity (VIF) [44], which is an extension of the IFC, the HVS model is added to include these aspects.

3.4. FR Image QA algorithms Selected for This Study

The selected FR QA algorithms and their basic information, such as whether a QA algorithm works with colour or only intensity, are listed in Table I. Most of the QA algorithms are measured using implementations available on the Internet, and all the algorithms were computed using the default parameter values proposed by the authors. If the algorithm works only with luminance, the L* component in CIE L*a*b* colour space was used. A public implementation of UQI is available at [45] and SSIM at [46]. Despite the fact that SSIM is an improved version of UQI, also UQI was tested because of the different windowing approach. The Daly VDP implementation can be found at [47], in which the original VDP is generalised to high dynamic range (HDR) images. Modifications to the original VDP are presented in [48]. A C++ implementation for PDiff is available at [49], a Matlab implementation for the NQM at [50] and for the IFC and VIF at [51]. Combined neighbourhood based and histogram similarity measures were computed as presented in [33] with a minor modification: CIE L*a*b* colours were used instead of RGB colours. Hence, differences correspond better to the human perception, and additional colour space transformations are avoided (descreening is done in the CIE L*a*b* colour space). In [31] and [32], a very large number of different fuzzy similarity measures were presented. All these similarity measures were tested, but only the highest performing ones are presented in this paper.

TABLE I: FR QA algorithms used in this study.

QA algorithm	**Acronym**	**Type**	**Colour**
Peak signal to noise ratio	PSNR	mathematical	no
L*a*b* perceptual error [21]	LabMSE	mathematical	yes
Universal quality index [34]	UQI	structural	no
Structural similarity metric [35]	SSIM	structural	no
Information fidelity criterion [42]	IFC	information theoretic	no
Visual information fidelity [44]	VIF	information theoretic	no
Noise quality measure [38]	NQM	HVS error-based	no
Perceptual image difference [40]	PDiff	HVS error-based	yes
(High dynamic range) visible difference predictor [48]	(HDR-)VDP	HVS error-based	no
Fuzzy similarity measures [31]	Fuzzy S9	mathematical	no
Combined neighbourhood-based and histogram similarity measures [33]	Fuzzy Q1,5c, Q18,9c and Q18,5c	mathematical	yes

4. DATA AND METHODS

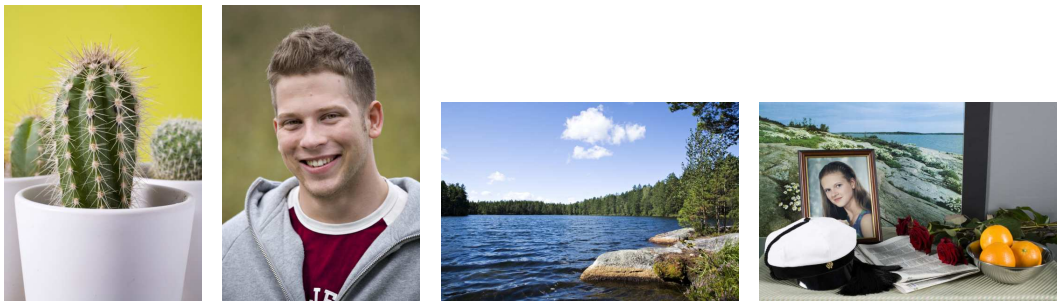
In this section, we introduce our data, i.e., the selected paper types, printed natural images, psychophysical subjective evaluations (the ground truth), and preprocessing of the raw data.

4.1. Test Sets

Our objective was to evaluate the effect of paper grade to the overall visual quality of printed images. Therefore, our test sets consisted of several paper grades at the cost of image contents. The first set of test samples consisted of natural images printed with a high quality inkjet printer on 16 different paper grades. The paper grades and the printing process were selected according to the current practises, as described in detail in [52–54]. The natural images used in the study are presented in Fig. 2. The image contents were

selected based on current practises and previous experience in media technology, and they included typical content types such as objects with details (*cactus*), a human portrait (*man*) and a landscape (*landscape*). The fourth image content combined all the types (*studio*).

The second set of samples consisted of images printed with a production-scale electrophotographic printer on 21 different paper grades. The same image contents were used excluding *studio* (Fig. 2(d)). The subjective evaluations, described later, were performed separately for both sets and image contents resulting in 7 separate tests of 16 or 21 samples, respectively.



(a) (b) (c) (d)

FIG. 2: Natural images used in the study: (a) cactus; (b) man; (c) landscape; (d) studio.

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 of the scanner software were disabled. The digitised images were saved using lossless compression.

The descreening was performed using 4 different cut-off wavelengths: 0.1 mm, 0.2 mm, 0.3 mm and 0.5 mm. The cut-off wavelength approximately corresponds to the the smallest detail remaining after the descreening. The smallest cut-off wavelength was selected to correspond to the printing resolution (360dpi), i.e., the filter removes the halftone pattern. The effect of the cut-off wavelength was also studied. The registration was made using the affine transformation.

4.2. Subjective Evaluation

The performance of the selected FR QA algorithms was studied against the psychophysical subjective evaluations (the ground truth). The subjective evaluation procedure is described in detail in [54]. In brief, samples of a specific set (the same image content and

printing method) were placed on a table in random order. Labels with the numbers from 1 to 5 were also presented on the table. The observer was asked to select the sample image representing the lowest quality in the set and place it on the number 1. Then, the observer was asked to select the highest quality sample and place it on the number 5. After that, the observer's task was to place the remaining samples on the numbers so that the quality increased steadily from 1 to 5. The final ground truth was formed by computing the mean opinion scores (MOSs) over all observers (N=28).

4.3. Processing of Raw Data

From the practical point of view, it is more interesting to put paper grades in a proper order than to find the overall quality of a single printed image on some abstract quality scale. Therefore, the subjective evaluation as well as QA algorithm scores should be similar over different image contents for the same paper grade. The subjective evaluation results were always scaled to the interval 1–5, but the image quality QA algorithm scores may differ significantly between the image contents. Therefore, either the QA algorithm scores need to be scaled to a common scale or the analysis needs to be done separately for different image contents. We selected the first option since the number of samples (16 or 21) was not enough to find statistically significant differences between the QA algorithms. Therefore, different image contents were combined to form a larger test set by scaling the QA algorithm scores. Here, the scaling was performed linearly. Let $\boldsymbol{x_n} = (x_{n,1}, ..., x_{n,M})$ represent the QA algorithm scores of one FR assessment for all samples (1-M) within a single image content n. Then, in the linear model we have

$$\hat{x}_{n,i} = \hat{\boldsymbol{b}}_{\boldsymbol{n}} \begin{pmatrix} 1 \\ x_{n,i} \end{pmatrix}, \tag{7}$$

where $\hat{\boldsymbol{b}}_{\boldsymbol{n}} = (b_{n,1}, b_{n,2})$ are selected by minimising the errors between the image contents:

$$\hat{\boldsymbol{b}}_{\boldsymbol{n}} = \arg\min_{\boldsymbol{b_n}} \sum_i [x_{1,i} - (b_{n,1} + b_{n,2} x_{n,i})]^2. \tag{8}$$

For the first image content, $\hat{\boldsymbol{b}}_{\boldsymbol{1}} = (0, 1)$, and for the remaining image contents, $\hat{\boldsymbol{b}}_{\boldsymbol{n}}$ are such that the QA algorithm scores are converted to values similar to the values of the first image content with the same paper grade. The above-mentioned method does not allow combining

results from the different printing methods (different paper grades), and therefore, we finally had two test sets: Test set A containing 64 inkjet samples and Test set B containing 63 electrophotography samples.

5. RESULTS

5.1. Performance Measures

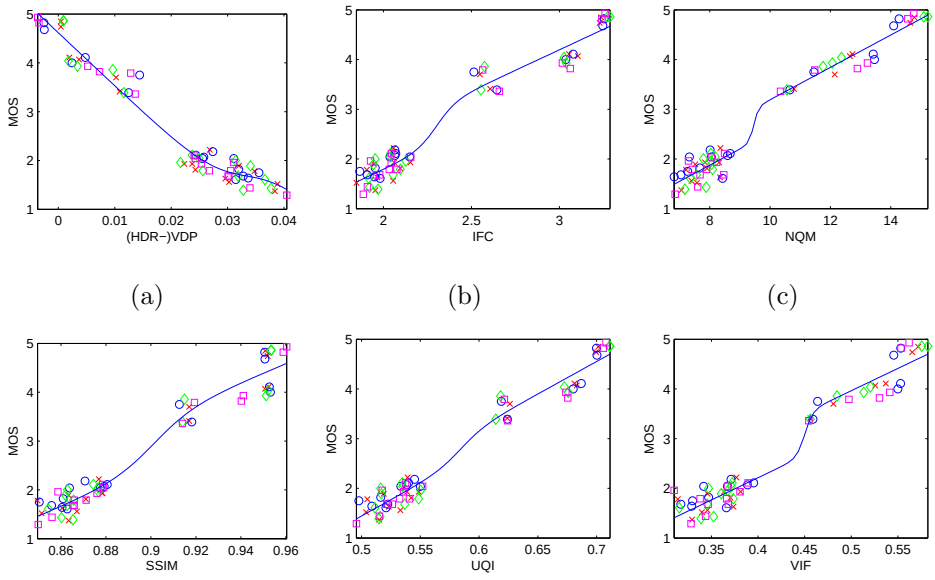
Three performance measures were chosen for the comparison of the QA algorithms. The selected measures are similar to those presented in [13]. The first one is the linear correlation coefficient between MOS and the QA algorithm score after nonlinear regression. The following nonlinearity (constrained to be monotonic) was used [13]:

$$Q(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2 x - \beta_3)} \right) + \beta_4 x + \beta_5, \tag{9}$$

where x is the modified algorithm score. The second performance measure is the Spearman rank order correlation coefficient (SROCC), and the third measure is the root mean squared error (RMSE) between MOS and QA algorithm score after the nonlinear regression. The results are presented in Tables II, III and IV. In Fig. 3 and 4 some QA algorithms are presented with regression curves. In the plots, the best cut-off wavelengths were used for each QA algorithm.

5.2. Statistical Significance

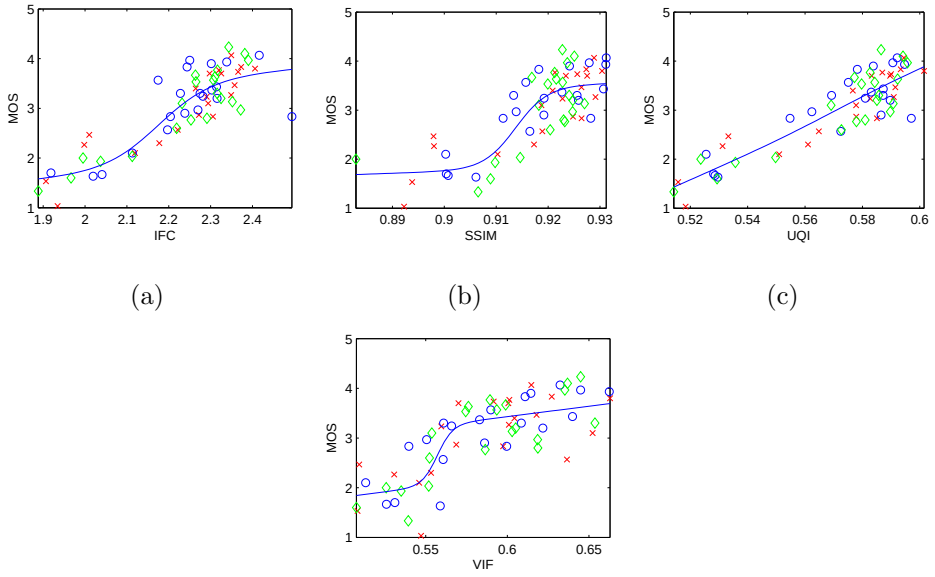
The statistical significance of the previous results was studied using the variance test. It expresses the trust in the superiority or inferiority of one QA algorithm over another based on the performance measures. The test is based on the assumption that the residuals (difference between MOS and the QA algorithm score) are normally distributed. The normality of the residuals was tested using Lilliefors test [55] with the 5% significance level and the results are shown in Table VII. The F-test was used to test whether the variances of the residuals of two QA algorithms are identical, i.e., the QA algorithm residuals are randomly drawn from the same distribution. The null hypothesis is that the residuals of both QA algorithms come from the same distribution and are statistically indistinguishable with 95% confidence.



(a) (b) (c)

(d) (e) (f)

FIG. 3: Data and nonlinear regression curves (Test set A). The symbols represent the image contents: red x-mark - man, blue circle - lake, green diamond - cactus, magenta square - studio. (a) (HDR-)VDP; (b) IFC; (c) NQM; (d) SSIM; (e) UQI; (f) VIF.



(a) (b) (c)

(d)

FIG. 4: Data and nonlinear regression curves (Test set B). The symbols represent the image contents: red x-mark - man, blue circle - lake, green diamond - cactus. (a) IFC; (b) SSIM; (c) UQI; (d) VIF.

TABLE II: Correlation between the QA algorithm scores and MOS after nonlinear regression for both test sets and all cut-off wavelengths

	Test set A				Test set B			
QA algorithm	0.1mm	0.2mm	0.3mm	0.5mm	0.1mm	0.2mm	0.3mm	0.5mm
PSNR	0.440	0.424	0.404	0.349	0.536	0.547	0.522	0.524
LabMSE	0.289	0.385	0.364	0.215	0.605	0.601	0.598	0.601
UQI	0.982	0.983	0.984	0.984	0.849	0.871	0.882	0.887
SSIM	0.978	0.979	0.978	0.970	0.776	0.819	0.802	0.721
IFC	0.983	0.982	0.982	0.982	0.842	0.851	0.860	0.873
VIF	0.982	0.982	0.982	0.982	0.818	0.807	0.811	0.793
NQM	0.988	0.988	0.988	0.988	0.702	0.700	0.695	0.697
PDiff	0.928	0.948	0.957	0.955	0.302	0.178	0.081	0.207
(HDR-)VDP	0.982	0.980	0.978	0.975	0.675	0.608	0.556	0.381
Fuzzy S9	0.686	0.687	0.686	0.639	0.628	0.581	0.580	0.595
Fuzzy Q1,5c	0.951	0.962	0.970	0.966	0.396	0.358	0.586	0.570
Fuzzy Q18,9c	0.881	0.776	0.657	0.581	0.670	0.672	0.679	0.691
Fuzzy Q18,5c	0.887	0.828	0.742	0.699	0.634	0.740	0.716	0.726

The significance test results are shown in Tables V and VI for both test sets and between all QA algorithms.

5.3. Sensitivity to Image Content

Subjective image quality depends more on how the distortion is perceived than how much there is distortion. For example, if the distortion is not visible in all contents, their perceived quality may significantly differ. Our research originates from the viewpoint of paper, not the printing process, and therefore, we seek out QA algorithms which are independent of the image content. The sensitivity of the tested QA algorithms to image content is illustrated in Fig. 5. The lines represent the variance of the QA algorithm scores, and the dots the mean value for each content (order: *man*, *landscape*, *cactus* and *studio*). To fit all QA algorithms

TABLE III: SROCC between QA algorithm scores and MOS for both test sets and all cut-off wavelengths.

	Test set A				Test set B			
QA algorithm	0.1mm	0.2mm	0.3mm	0.5mm	0.1mm	0.2mm	0.3mm	0.5mm
PSNR	0.294	0.269	0.254	0.230	0.432	0.428	0.419	0.412
LabMSE	0.347	0.315	0.301	0.270	0.470	0.471	0.470	0.466
UQI	0.912	0.914	0.914	0.915	0.678	0.763	0.790	0.788
SSIM	0.899	0.915	0.921	0.906	0.509	0.689	0.697	0.660
IFC	0.901	0.900	0.903	0.891	0.719	0.731	0.744	0.761
VIF	0.901	0.893	0.884	0.887	0.738	0.729	0.730	0.726
NQM	0.885	0.885	0.885	0.886	0.633	0.633	0.631	0.625
PDiff	0.800	0.798	0.803	0.803	0.316	0.206	0.171	0.241
(HDR-)VDP	0.924	0.923	0.899	0.827	0.608	0.537	0.487	0.387
Fuzzy S	0.606	0.578	0.561	0.550	0.543	0.537	0.532	0.519
Fuzzy Q1,5c	0.821	0.884	0.902	0.872	0.441	0.101	0.317	0.395
Fuzzy Q18,9c	0.790	0.740	0.677	0.597	0.588	0.618	0.615	0.620
Fuzzy Q18,5c	0.805	0.780	0.737	0.706	0.563	0.623	0.634	0.632

into a single image, the QA algorithm scores were normalised to zero mean and unit variance over all image contents.

6. DISCUSSION

As can be seen from Tables II, III and IV, the results clearly differ between the test sets A (inkjet) and B (electrophotography). For almost every QA algorithm, the correlations are higher and the errors smaller for Test set A. The reason for this is the fact that Test set A is considerably easier than Test set B, both subjectively and computationally. Quality variation between the samples is higher, and two or three compact clusters with distinctly different quality exist in the data (see Fig. 3). In general, the methods succeed since most QA algorithms put these clusters in the right order, increasing the correlations using Test set

TABLE IV: RMSE between the QA algorithm scores and MOS after nonlinear regression for both test sets and all cut-off wavelengths.

	Test set A				Test set B			
QA algorithm	0.1mm	0.2mm	0.3mm	0.5mm	0.1mm	0.2mm	0.3mm	0.5mm
PSNR	8.542	8.614	8.700	8.923	5.282	5.237	5.335	5.329
LabMSE	9.111	8.882	8.922	9.386	4.981	4.997	5.012	5.000
UQI	1.791	1.753	1.711	1.683	3.307	3.072	2.948	2.884
SSIM	1.977	1.947	1.969	2.310	3.943	3.590	3.738	4.335
IFC	1.737	1.776	1.782	1.773	3.372	3.281	3.189	3.051
VIF	1.795	1.790	1.789	1.773	3.594	3.692	3.663	3.810
NQM	1.460	1.461	1.461	1.459	4.456	4.465	4.500	4.483
PDiff	3.533	3.017	2.752	2.813	5.962	6.155	6.235	6.120
(HDR-)VDP	1.783	1.881	1.990	2.121	4.613	4.966	5.202	5.784
Fuzzy S9	6.917	6.908	6.920	7.322	4.868	5.090	5.096	5.030
Fuzzy Q1,5c	2.938	2.592	2.319	2.468	5.743	5.843	5.076	5.138
Fuzzy Q18,9c	4.516	6.000	7.169	7.738	4.651	4.634	4.594	4.538
Fuzzy Q18,5c	4.399	5.328	6.380	6.805	4.842	4.210	4.367	4.304

A. Another practical reason for the different results between Test set A and Test set B arises from the two different kinds of printing methods; their underlying artifacts and distortions affect the subjective quality experience differently. Respectively, the QA algorithms differ by being sensitive to different kinds of distortions, which explains why some QA algorithms are better suited for inkjet (Test set A) and some for electrophotographic (Test set B) prints. It is noteworthy that the best QA algorithms for both sets model the phenomenon reasonably well and consistently, as is evident in Figs. 3 and 4 (the shapes are similar).

As expected, the simple pixelwise metrics, such as the PSNR and LabMSE, do not work well. With Test set A, even the subjectively distinguishable clusters were not assorted correctly. However, most of the advanced methods showed high correlation coefficients. UQI, SSIM, IFC, VIF, NQM and HDR-VDP are the best ones and statistically indistinguishable from each other. With Test set B, the group of well working QA algorithms is reduced to

TABLE V: F-test results (Test set A). Each entry contains five binary numbers. The first four numbers represent the different cut-off wavelengths used in the descreening (from smallest to highest), and the fifth number represents the best result for each QA algorithm (the cut-off wavelength is treated as a parameter of the QA algorithm). 0 means that QA algorithms are statistically indistinguishable from each other and 1 means that QA algorithms have a statistically significant difference.

	PSNR	LabMSE	UQI	SSIM	IFC	VIF	NQM	PDiff	VDP	FS9	FQ1,5c	FQ18,9c	FQ18,5c
PSNR	0000	00000	11111	11111	11111	11111	11111	11111	11111	00000	11111	11001	11110
LabMSE	00000	00000	11111	11111	11111	11111	11111	11111	11111	10001	11111	11001	11111
UQI	11111	11111	00000	00010	00000	00000	00000	11111	00000	11111	11111	11111	11111
SSIM	11111	11111	00010	00000	00010	00010	11111	11101	00000	11111	11000	11111	11111
IFC	11111	11111	00000	00010	00000	00000	00000	11111	00000	11111	11111	11111	11111
VIF	11111	11111	00000	00010	00000	00000	00000	11111	00000	11111	11111	11111	11111
NQM	11111	11111	00000	11111	00000	00000	00000	11111	01110	11111	11111	11111	11111
PDiff	11111	11111	11111	11101	11111	11111	11111	00000	11111	11111	00000	01111	01111
(HDR-)VDP	11111	11111	00000	00000	00000	00000	01110	11111	00000	11111	11001	11111	11111
Fuzzy S9	00000	10001	11111	11111	11111	11111	11111	11111	11111	00000	11111	10001	11000
Fuzzy Q1,5c	11111	11111	11111	11000	11111	11111	11111	00000	11001	11111	00000	11111	11111
Fuzzy Q18,9c	11001	11001	11111	11111	11111	11111	11111	01111	11111	10001	11111	00000	00001
Fuzzy Q18,5c	11110	11111	11111	11111	11111	11111	11111	01111	11111	11000	11111	00001	00000

TABLE VI: F-test results (Test set B).

	PSNR	LabMSE	UQI	SSIM	IFC	VIF	NQM	PDiff	VDP	FS9	FQ1,5c	FQ18,9c	FQ18,5c
PSNR	00000	00000	11111	11101	11111	11111	00000	00000	00000	00000	00000	00000	00000
LabMSE	00000	00000	11111	01101	11111	11111	00000	00000	00000	00000	00000	00000	00000
UQI	11111	11111	00000	00010	00000	00010	11111	11111	11111	11111	11111	11111	11111
SSIM	11101	01101	00010	00000	00010	00000	00000	11111	01110	01101	11101	01001	00000
IFC	11111	11111	00000	00010	00000	00000	11111	11111	11111	11111	11111	11111	10111
VIF	11111	11111	00010	00000	00000	00000	00000	11111	01110	11111	11111	10001	10000
NQM	00000	00000	11111	00000	11111	00000	00000	11111	00010	00000	11000	00000	00000
PDiff	00000	00000	11111	11111	11111	11111	11111	00000	10001	00000	00000	01111	01111
(HDR-)VDP	00000	00000	11111	01110	11111	01110	00010	10001	00000	00000	00000	00000	00010
Fuzzy S9	00000	00000	11111	01101	11111	11111	00000	00000	00000	00000	00000	00000	00000
Fuzzy Q1,5c	00000	00000	11111	11101	11111	11111	11000	00000	00000	00000	00000	00000	01000
Fuzzy Q18,9c	00000	00000	11111	01001	11111	10001	00000	01111	00000	00000	00000	00000	00000
Fuzzy Q18,5c	00000	00000	11111	00000	10111	10000	00000	01111	00010	00000	01000	00000	00000

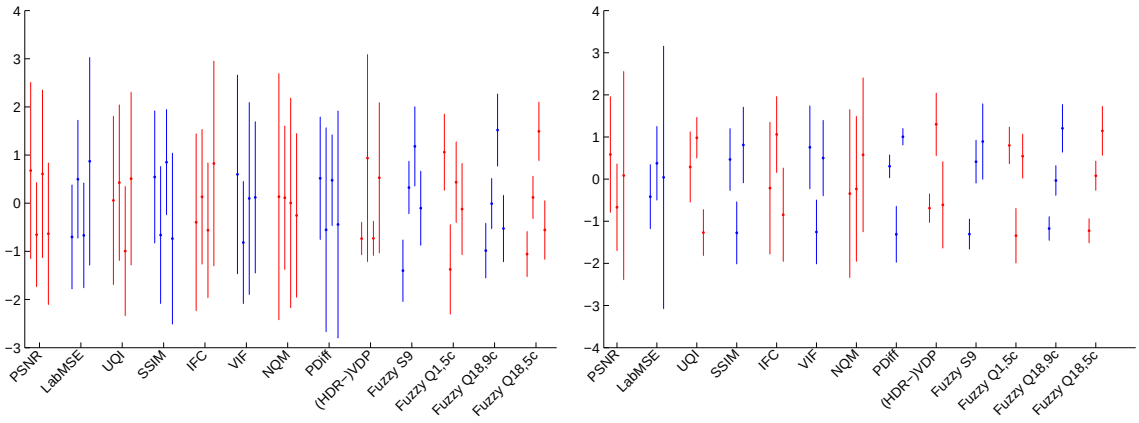
TABLE VII: Gaussianity of residuals. 1 means that the data are normally distributed according the Lilliefors' composite goodness-of-fit test.

	Test set A				Test set B			
QA algorithm	0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5
PSNR	0	0	0	0	1	1	1	1
LabMSE	0	0	0	0	1	1	1	1
UQI	1	1	1	0	1	1	1	1
SSIM	1	1	1	1	1	1	1	1
IFC	1	1	1	0	1	1	1	1
VIF	1	1	1	1	0	1	1	1
NQM	0	0	0	0	1	1	1	1
PDiff	0	1	1	1	0	0	0	0
(HDR-)VDP	1	1	0	1	1	1	1	1
Fuzzy S9	0	0	0	0	1	1	1	1
Fuzzy Q1,5c	1	1	1	1	0	1	1	1
Fuzzy Q18,9c	0	0	0	0	1	1	1	1
Fuzzy Q18,5c	0	0	0	0	1	1	1	1

UQI, SSIM, IFC and VIF, which are again statistically indistinguishable from each other. The other performance measures, SROCC and RMSE, support these conclusions. If a single optimal QA algorithm should be selected, VIF would be a safe choice since it was shown to be the best in an earlier study on digital images [13].

As a secondary task, the selection of the cut-off wavelength of GLPF in the descreening was studied. A notable result concerning it is that the optimal selection depends on the QA algorithm (see Tables II, III and IV), and a single universal best cut-off wavelength cannot be defined. For example, for UQI the best cut-off wavelength seems to be 0.5 mm, while for VIF it is 0.1 mm. However, the effect is not too dramatical, and it seems that selection of the cut-off wavelength is not a crucial step, as long it is large enough to filtered out the halftone pattern and small enough not to be visually disturbing to the human eye.

The requirement for the QA algorithm to be independent of the image content leads to



(a) (b)

FIG. 5: Illustration of the sensitivity to image content. The lines represent the variance of the QA algorithm scores and the dots the mean value for each content (order: *man*, *landscape*, *cactus* and *studio*). (a) Test set A; (b) Test set B.

a new problem. The original objective to automatically estimate the printability of paper using printed natural images is divided into two subproblems: i) how to compute the quality of a natural image so that it is consistent with the subjective human evaluation, and ii) how to measure the paper or print quality in a way that it does not depend on the content of the examined image. Unfortunately, there is a conflict between the solutions for these two. For example, the unevenness of a print (noise) is a serious problem in a landscape image with a large solid colour region of sky, while it is almost imperceptible in an image with a great deal of small details. A QA algorithm that models the human perception produces a different result than an algorithm that computes the amount of distortion independent of the image content. This problem is apparent in Table VIII, where the correlation coefficients between the subjective evaluations with different image contents are shown. The images printed on the same paper grade using the same printer should have the same amount of signal-level distortion, and therefore, if the image content did not affect the quality, the correlations would be near to unity. This holds quite well for Test set A, where the quality variation is large, but for Test set B, the image contents had a notable influence on the perceived quality despite the similar level of distortion. In summary, both conditions, a good correspondence to human perception and independence of the image content, cannot be achieved simultaneously in a sufficient manner, and therefore, choosing the correct QA algorithm depends on the circumstances. From Fig. 5, it can be seen that the NQM is

the best QA algorithm based on its insensitivity to image content. However, no significant differences in the earlier selected group of the best QA algorithms, UQI, SSIM, IFC and VIF, can be revealed in terms of sensitivity to content.

TABLE VIII: Correlation coefficients between the subjective evaluations with different image contents.

	Test set A				Test set B		
Cont.	man	lake	cactus	studio	man	lake	cactus
man	1.000	0.990	0.995	0.992	1.000	0.800	0.920
lake	0.990	1.000	0.988	0.989	0.800	1.000	0.889
cactus	0.995	0.988	1.000	0.993	0.920	0.889	1.000
studio	0.992	0.989	0.993	1.000			

7. CONCLUSIONS

In this work, we presented a novel and complete framework to compute full reference (FR) quality assessment algorithms for printed natural images. FR QA algorithms are popular and well studied in digital image quality research, but we report the first conclusive results obtained with printed media. Using our framework, we evaluated and statistically verified the performance of several state-of-the-art FR QA algorithms for an extensive set of printed products. As the main conclusion, we found out that the UQI, SSIM, IFC and VIF algorithms outperformed other QA algorithms, while the NQM was the least sensitive to image content. In the experimental part of the work, we reported remarkable correlations between the FR QA algorithms and subjective visual quality evaluations, promoting their use in printing and media technology. Since the tested methods have been mainly developed for digital image quality analysis, we have established a new research direction, i.e., how the digital FR QA algorithms can be further developed towards FR print quality evaluation. Whether or not they could replace the existing complicated and ad hoc metrics in the printing industry will remain a challenge to be dealt with in future studies.

Acknowledgement

The authors would like to thank the Finnish Funding Agency for Technology and Innovation (TEKES decision 40176/06) and the partners of the DigiQ project for their support.

---

[1] D. Wolin, K. Johnson, and Y. Kipman. Automating image quality analysis. In *IS&T's NIP 14: International Conference on Digital Printing Technologies*, pages 627–630, Toronto, Ontario, Canada, 1998.

[2] D. Wolin, K. Johnson, and Y. Kipman. The importance of objective analysis in image quality evaluation. In *IS&T's NIP 14: International Conference on Digital Printing Technologies*, pages 603–606, Toronto, Ontario, Canada, 1998.

[3] T. Bouk, E. N. Dalal, K. D. Donohue, S. Farnand, F. Gaykema, D. Dmitri, A. Haley, P. L. Jeran, D. Kozak, W. C. Kress, O. Martinezb, D. Mashtare, A. McCarthy, Y. S. Ng, D. R. Rasmussen, M. Robb, H. Shin, S. M. Quiroga, E. H. Barney Smith, M.-K. Tse, D. Williams, E. Zeise, and S. Zoltner. Recent progress in the development of incits w1.1, appearance-based image quality standards for printers. In *Proc. of SPIE IS&T Electronic Imaging*, volume 6494, 2008.

[4] D. Armel and J. Wise. An analytic method for quantifying mottle - part 1. *Flexo*, pages 70–79, December 1998.

[5] D. Armel and J. Wise. An analytic method for quantifying mottle - part 2. *Flexo*, pages 38–43, January 1999.

[6] J. Briggs, D. Forrest, A. Klein, and M.-K. Tse. Living with ISO-13660: Pleasures and perils. In *IS&Ts NIP 15: 1999 International Conference on Digital Printing Technologies*, pages 421–425, IS&T, Springfield VA, 1999.

[7] ISO/IEC. 13660:2001(e) standard. information technology - office equipment - measurement of image quality attributes for hardcopy output - binary monochrome text and graphic images. ISO/IEC, 2001.

[8] P.-Å. Johansson. *Optical Homogeniety of Prints*. PhD thesis, Kunglika Tekniska Högskolan, Stockholm, 1999.

[9] R. R. Rosenberger. Stochastic frequency distribution analysis as applied to ink jet print

mottle measurement. In *IS&Ts NIP 17: 2001 International Conference on Digital Printing Technologies*, pages 808–812, IS&T, Springfield VA, 2001.

[10] A. Sadovnikov, L. Lensu, J. Kamarainen, and H. Kälviäinen. Quantifed and perceived uneveness of solid printed areas. In *Xth Ibero-American Congress on Pattern Recognition*, pages 710–719, 2005.

[11] B. Streckel, B. Steuernagel, E. Falkenhagen, and E. Jung. Objective print quality measurements using a scanner and a digital camera. In *DPP 2003: IS&T Int. Conf. on Digital Production Printing and Industrial Applications*, pages 145–147, 2003.

[12] D. Wolin. Enhanced mottle measurement. In *PICS 2002: IS&T's PICS conference*, pages 148–151. IS&T, 2002.

[13] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions On Image Processing*, 15(11):3440–3451, November 2006.

[14] J. Rakun. The computer-aided detection of inferior printing quality and errors. In *Proc. of the IEEE Mediterranean Electrotechnical Conference*, pages 1236–1240, Malaga, Spain, 2006.

[15] X. Yanfang, W. Yu, and Z. Ming. Color reproduction quality metric on printing images based on the s-cielab model. In *Proc. of the 2008 International Conference on Computer Science and Software Engineering*, pages 294–297, Wuhan, China, 2008.

[16] A. M. Eskicioglu and P.S. Fisher. Image quality measures and their performance. *IEEE Transactions On Communications*, 43(12):2959–2965, December 1995.

[17] B. Li, G. W. Meyer, and R. V. Klassen. A comparison of two image quality models. In *Proc. SPIE Vol. 3299, Human vision and electronic imaging III*, pages 98–109, San Jose, USA, 1998.

[18] S. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Proc. SPIE Vol. 1666, Human Vision, Visual Processing, and Digital Display III*, pages 2–15, San Jose, USA, 1992.

[19] J. Lubin and D. Fibush. Contribution to the IEEE standards subcommittee: Sarnoff JND vision model, August 1997.

[20] A. Mayache, T. Eude, and H. Cherifi. A comparison of image quality models and metrics based on human visual sensitivity. In *Proc. IEEE International Conference on Image Processing*, pages 409–413, Chicago, IL, USA, 1998.

[21] I. Avcibaş, B. Sankur, and K. Sayood. Statistical evaluation of image quality measures.

Journal of Electronic Imaging, 11(2):206–223, April 2002.

[22] T. Eerola, J.-K. Kamarainen, L. Lensu, and H. Kälviäinen. Framework for applying full reference digital image quality measures to printed images. In *16th Scandinavian Conference on Image Analysis*, pages 99–108, Oslo, Norway, 2009.

[23] G. Wyszecki and W. S. Stiles. *Color science : concepts and methods, quantitative data and formulae*. Wiley, second edition, 2000.

[24] A. Sadovnikov, P. Salmela, L. Lensu, J. Kamarainen, and H. Kälviäinen. Mottling assessment of solid printed areas and its correlation to perceived uniformity. In *14th Scandinavian Conference of Image Processing*, pages 411–418, Joensuu, Finland, 2005.

[25] J. M. Wolfe, K. R. Kluender, and D. M. Levi. *Sensation & Perception*. Sinauer Associates, 2006.

[26] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[27] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.

[28] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing*, 24(6), 1981.

[29] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE-TPAMI*, 13(4):376–380, 1991.

[30] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge Universiy Press, second edition, 2003.

[31] D. van der Weken, M. Nachtegael, and E. E. Kerre. Using similarity measures and homogeneity for the comparison of images. *Image and Vision Computing*, 22(9):695–702, August 2004.

[32] D. van der Weken, M. Nachtegael, and E. E. Kerre. Combining neighbourhood-based and histogram similarity measures for the design of image quality measures. *Image and Vision Computing*, 25(2):184–195, February 2007.

[33] D. van der Weken, V. D. Witte, M. Nachtegael, S Schulte, and E. E. Kerre. Fuzzy similarity measures for colour images. In *IEEE Conference on Cybernetics and Intelligent Systems*, 2006.

[34] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, March 2002.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[36] Z. Wang, A. C. Bovik, and L. Lu. Why is image quality assessment so difficult? In *IEEE International Conference on Acoustics, Speech, & Signal Processing*, 2002.

[37] J. L. Mannos and D. J. Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions on Information Theory*, 20(4):525–536, July 1974.

[38] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions On Image Processing*, 9(4):636–650, April 2000.

[39] E. Peli. Contrast in complex images. *Journal of the Optical Society of America*, 7(10):2032–2040, 1990.

[40] H. Yee, S. Pattanaik, and D. P. Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics*, 20(1):39–65, January 2001.

[41] A. Srivastava, A.B. Lee, E.P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, January 2003.

[42] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions On Image Processing*, 14(12):2117–2128, December 2005.

[43] M.J. Wainwright, E.P. Simoncelli E.P., and A.S. Willsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Computational Harmonic Analysis*, 11(1):89–123, July 2001.

[44] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions On Image Processing*, 15(2):430–444, February 2006.

[45] Z. Wang. Universal quality index [online]. Available: http://www.cns.nyu.edu/∼zwang/files/research/quality_index/ demo.html.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Structural similarity index [online]. Available: http://www.ece.uwaterloo.ca/∼z70wang/research/ssim/.

[47] Visible difference metric for high dynamic range images [online]. Available: http://www.mpi-inf.mpg.de/resources/hdr/vdp/index.html.

[48] R. Mantiuk, S. Daly, K. Myszkowski, and H.-P. Seidel. Predicting visible differences in high dynamic range images - model and its calibration. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, pages 2763–2769, 2004.

[49] H. Yee. Perceptual image diff [online]. Available: http://pdiff.sourceforge.net/.

[50] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model [online]. Available: http://users.ece.utexas.edu/∼bevans/papers/2000/ imageQuality/index.html.

[51] Image & video quality assessment at LIVE [online]. Available: http://live.ece.utexas.edu/research/Quality/index.htm.

[52] T. Eerola, J.-K. Kamarainen, T. Leisti, R. Halonen, L. Lensu, H. Kälviäinen, G. Nyman, and P. Oittinen. Is there hope for predicting human visual quality experience? In *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*, Singapore, 2008.

[53] T. Eerola, J.-K. Kamarainen, T. Leisti, R. Halonen, L. Lensu, H. Kälviäinen, P. Oittinen, and G. Nyman. Finding best measurable quantities for predicting human visual quality experience. In *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*, Singapore, 2008.

[54] P. Oittinen, R. Halonen, A. Kokkonen, T. Leisti, G. Nyman, T. Eerola, L. Lensu, H. Kälviäinen, R. Ritala, J. Pulla, and M. Mettänen. Framework for modelling visual printed image quality from paper perspective. In *Image Quality and System Performance V*, San Jose, USA, 2008.

[55] H. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, June 1967.