Journal of Electronic Imaging

JElectronicImaging.org

Study of no-reference image quality assessment algorithms on printed images

Tuomas Eerola
Lasse Lensu
Heikki Kälviäinen
Alan C. Bovik

SPIE. IS&T imaging.org

Study of no-reference image quality assessment algorithms on printed images

Tuomas Eerola,[a,*] Lasse Lensu,[a] Heikki Kälviäinen,[a] and Alan C. Bovik[b]
[a]Lappeenranta University of Technology, Department of Mathematics and Physics, P.O. Box 20, 53850 Lappeenranta, Finland
[b]The University of Texas at Austin, Department of Electrical and Computer Engineering, 1 University Station, Austin, Texas 78712, United States

Abstract. Measuring the visual quality of printed media is important since printed products have an important role in everyday life. Finding ways to automatically predict the image quality has been an active research topic in digital image processing, but adapting those methods to measure the visual quality of printed media has not been studied often or in depth and is not straightforward. Here, we analyze the efficacy of no-reference image quality assessment (IQA) algorithms originally developed for digital IQA with regards to predicting the perceived quality of printed natural images. We perform a comprehensive statistical comparison of the methods. The best methods are shown to accurately predict subjective opinions of the quality of printed photographs using data from a psychometric study. [DOI: 10.1117/1.JEI.23.6.061106]

Keywords: print quality; image quality assessment; no reference.

Paper 14176SS received Apr. 1, 2014; revised manuscript received Jun. 26, 2014; accepted for publication Jul. 18, 2014; published online Aug. 13, 2014.

1 Introduction

Despite rapid developments in electronic media, most people still prefer reading text printed on paper rather than reproduced on electronic displays.[1] Printed media remain more suitable for delivering local news than electronic media, and the packaging industry increasingly relies on the production of visually pleasing and personalized packages by using digital printing. In addition, an increasing number of images are captured each year, and despite the fact that the digitization has created novel ways to share and distribute images, printed images still have their users. For example, the amount of bound photobooks has grown rapidly during the recent years.[2] These, among other reasons, are why paper and other fiber-based products still play an important role in communication, and printed products, such as books, newspapers, and packages, are an important part of daily life. When a customer decides to purchase a printed book or magazine, one of the key factors is print and image quality. Rather than using technical measurements, humans do not evaluate the quality of print and images based on physical parameters, but rather based on personal preferences and what they see as pleasurable.[3]

The problem of how humans perceive the quality of a reproduced image is of interest to researchers in many fields, including optics and material physics, image processing (compression and transfer), printing and media technology, and psychology. The problem is particularly difficult for printed media, since solving it requires understanding the paper and ink physics, viewing parameters, optics, and elements of human visual perception. No measure of visual print quality can be defined without ambiguity, because it is ultimately a subjective opinion of an "end-user" observing the result. As a consequence, visual evaluations have been traditionally conducted using groups of human observers, but recent developments in perceptual models and machine vision have made it possible to develop automatic methods of print quality evaluation. The use of machine vision founded on reliable perceptual and print models promises to make it possible to replace humans in laborious off-line evaluations. In addition, such computational methods suggest the potential for quality-optimized on-line measurements during printing.

Image quality assessment (IQA) models can be divided into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR) methods. In FR methods, a reference image with presumed ideal quality is available, whereas in RR methods only a small amount of information describing the reference image is given as input. NR methods operate in the absence of any reference image. Currently, FR methods are the main approach for evaluating and comparing the quality of digital images, especially compressed ones. The digital representations of the original and compressed images are in correspondence, i.e., there exist no-spatial transformations between the images, and the compression should retain at least photometric equivalence. Therefore, FR measures can be computed in a straightforward manner by computing "distance metrics," and the actual problem is to define an appropriate metric for the task. NR-IQA is the most difficult task, and the majority of the proposed methods are designed for a single-distortion type and can be considered as domain specific.

The FR-IQA has been shown to be a good approach to predict the image quality when the reference image exists.[4] With a carefully designed measurement framework, it is possible to apply an FR approach to printed images and relatively high correlations with subjective evaluation results can be achieved, as was shown by Eerola et al.[5] However, several problems exist when the quality of printed images is evaluated by FR methods.

*Address all correspondence to: Tuomas Eerola, E-mail: tuomas.eerola@lut.fi

The first obvious weakness is the fact that the FR methods are suitable only when a digital reference image exists. This is not always the case with printed images. Even more notable problems arise from the basic assumption of the FR approach that the reference image is of ideal quality and can be used as a basis for quality evaluation. For quality assessment (QA) of compressed images, this assumption is perhaps justified; a good image compression method reduces the size of the image in such a manner that the visual appearance of the image changes as little as possible, i.e., the evaluated (compressed) image is visually similar to the reference (original) image. For printed images, however, it is not clear that such an assumption applies. First of all, the original image is in a very different form than the printed image that is being evaluated, making its use as a reference image not only difficult, but also rather questionable.[6] It is not clear how the difference between a printed photograph and a digital image should be measured. Second, visual quality may be impaired when the original image is transferred onto paper even if no printing artifacts appear, since different aspects of image quality take different degrees of importance on different media. For example, gloss is not a property of a digital image, but has a remarkable effect on the perceived quality of a printed image. Even in the hypothetical ideal situation, where the original image is of "perfect quality" (whatever that is) and the printer or paper do not cause any visible artifacts, quality may still be compromised after transferring onto paper due to the different natures of the media. Third, while making subjective evaluations of printed samples, showing a digital reference to human observers is not a simple matter, since simultaneous viewing of digital and printed images does not allow an observer to adapt both white points, whereas the memory viewing technique does not allow one to directly compare the images.[7] Often a digital reference image is not shown to the observers, and they are forced to make decisions without knowing what the printed image was supposed to look like.[5] RR-QA algorithms suffer from similar problems since a reference image is still needed.

For the aforementioned reasons, NR-QA methods are of high interest. However, NR-QA is a much more difficult task than FR-QA, and until recently, there did not exist any general NR-QA methods. All methods were either application specific or measured only a specific kind of distortion such as blur or noise. While these methods have a role in QA, no such method alone can predict the perceived quality of an image. During the last few years, significant developments have led to the creation of generic NR-QA algorithms. Thus, the objective of this study is to determine the efficacy of these new NR models for predicting the subjective quality of printed images by statistically evaluating their performances against subjective mean opinion scores (MOSs) obtained from psychometric experiments on printed samples. Since NR models have been developed for images of natural scenes, this study focuses only on the important application of printed photograph quality analysis while the quality of printed text and graphics is not considered.

The paper is organized as follows. Section 2 introduces the existing generic NR-IQA algorithms that are statistically evaluated in this study. Section 3 presents the data, a method to apply NR-IQA algorithms to printed images, and the results. The results are discussed in Sec. 4, and the conclusion is drawn in Sec. 5.

2 NR-IQA

Most existing NR-IQA algorithms fall in one of the following categories: (1) distortion-specific IQA algorithms, (2) training-based IQA algorithms, and (3) natural scene statistics (NSS)-based IQA algorithms. The first category is composed of methods that try to model the distortion such as blur[8–10] or blockiness.[11] These methods are application specific and are not in the scope of this work. However, it should be mentioned that the distortion-specific methods have also been developed for printed images.[12] The second category contains methods and models that use image-based features and require training on appropriate data.[13–15] These methods are highly dependent on the quality of the selected features and often require a large amount of training data, i.e., distorted images with subjective data that is laborious to collect. Generic image features to describe the image quality are difficult to establish, which limits the application domain. The third category contains methods that are based on the assumption that the pristine natural images form a subset of images that have different statistical properties than the distorted images. NSS methods have turned out to be a very promising approach, and most existing NR-IQA algorithms with good performance more or less rely on NSS. These methods may also require training, but the amount of training is greatly reduced relative to training-based IQA algorithms.

Ideally, NSS features are invariant to image content, but are sensitive to distortions. Such features can be used to assess the image quality by estimating the degree of distortion or the distance of the distorted image to the pristine (natural) image, regardless of the content of the image. Several promising NSS approaches have been proposed in the literature. A typical NSS-based IQA algorithm starts with a multiscale image transform, such as the discrete cosine[16] or wavelet transform,[17] but spatial NSS methods also exist.[18] The computed transform coefficients have statistical properties that vary based on the presence of distortions. For example, the distribution of wavelet coefficients computed from natural images usually has a sharp peak near zero and long and smooth tails. Most image distortions break this regularity, which makes the shape of the coefficient distribution a good feature for IQA.

The coefficient distributions are often parametrized using the (zero mean) general Gaussian distribution (GGD),[19] which has been found to capture the broad spectrum of possible distribution shapes. The GGD is defined as

$$f(x;\alpha,\beta)=\frac{\alpha}{2\beta\Gamma\left(\frac{1}{\alpha}\right)}\exp\left[-\left(\frac{|x|}{\beta}\right)^{\alpha}\right], \tag{1}$$

where α is the shape parameter, β is the variance, and $\Gamma(\cdot)$ is the gamma function

$$\Gamma(a)=\int_0^{\infty}t^{a-1}e^{-t}\mathrm{d}t, \qquad a>0. \tag{2}$$

The related asymmetric GGD[20] is also used

$$f(x;\gamma,\beta_l,\beta_r)=\begin{cases}\frac{\gamma}{(\beta_l+\beta_r)\Gamma\left(\frac{1}{\gamma}\right)}e^{\left[-\left(\frac{x}{\beta_l}\right)^{\gamma}\right]}, & \forall\, x<0\\ \frac{\gamma}{(\beta_l+\beta_r)\Gamma\left(\frac{1}{\gamma}\right)}e^{\left[-\left(\frac{x}{\beta_r}\right)^{\gamma}\right]}, & \forall\, x\geq 0\end{cases}. \tag{3}$$

The distribution parameters (α, β) or (γ, β_l, β_r) are then used as features to either (1) classify images based on the decided distortion, then apply a distortion-specific IQA algorithm or (2) estimate the quality directly using regression techniques. Most of the methods are trained on data containing MOSs or difference MOSs (DMOSs) of the images.

The NR-IQA algorithms selected for this study and their basic information are listed in Table 1. All the selected methods are based on NSS. There are also promising training-based methods, such as the learning-based blind image quality (LBIQ)[13] measure and the visual codebook-based image quality (CBIQ) measure,[15] but these were excluded from the study due to their strong dependence on training data. As discussed later in Sec. 3.3, due to the laborious nature of preparing and subjectively evaluating printed samples, data volume remains a problem and limits the selection of the NR-IQA algorithms that can be used.

The Blind Image Quality Index (BIQI)[17] makes use of NSS features based on wavelet coefficients to first classify an image between different distortion types and to estimate distortion-specific quality scores. A support vector machine (SVM) is used for classification, and a support vector regression (SVR) is used for a distortion-specific quality score estimation. The final quality score is computed as a weighted sum of the distortion-specific quality scores. The weights are the probability estimates provided by the SVM used in the distortion classification stage.

The Distortion Identification-based Image Verity and Integrity Evaluation (DIIVINE) index[23] is an extension of BIQI with a richer set of NSS-based features. Instead of the wavelet transform, a scale-space-orientation decomposition is used.

BLIINDS-II[21,22] is an extension of the Blind Image Integrity Notator (BLIINDS).[16] It uses NSS features based on the local discrete cosine transform coefficients and a simple probabilistic model to predict the quality.

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)[25,26] differs from the methods above because it uses only spatial features, albeit over multiple scales. There is no need to map the data to a different coordinate domain, such as the wavelet or DCT domain. The NSS used is based on locally normalized luminances [mean subtracted contrast normalized (MSCN) coefficients].[18]

BRISQUE was further developed leading to the very generic, training-free Natural Image Quality Evaluator (NIQE).[27] It uses similar NSS features, but instead of establishing the quality value directly from features using an SVR, the NSS features are modeled as multivariate Gaussian. The main advantage of NIQE is that, unlike other methods, it does not require training data with subjective human evaluations. Instead, the model is constructed from features drawn from a corpus of undistorted natural images, making it the first truly distortion-independent NR-IQA algorithm.

The hybrid NR (HNR) model[24] is based on curvelet, wavelet, and discrete cosine transform coefficient statistics, and the quality is predicted using the peak coordinates of the logarithmic probability distribution of the coefficient magnitudes (LPMC).

2.1 *Previous Comparisons of NR-IQA Algorithms*

Most of the original references provide the results achieved on the well-known LIVE database.[4,28] Therefore, reliable conclusions can be made on the performance on the specific distortions present in the LIVE database, i.e., JPEG and JPEG2000 compressions, additive white Gaussian noise, Gaussian blur, and a Rayleigh fast-fading channel distortion. The results show that BRISQUE outperforms the other NR methods with a 0.94 linear correlation against the subjective scores, but LBIQ, BLIINDS-II, DIIVINE, and NIQE also attain good results with better than 0.9 correlation over all distortion types.

3 Experiments

In this section, we introduce the data used in this study, i.e., the particular types of paper used, the (printed) natural images, the psychometric subjective evaluation results (subjective scores), and the methods used to process the raw data.

3.1 *Test Sets*

The original objective guiding the data collection was to evaluate the effect of paper grade on the overall perceived visual quality of printed images. Therefore, our test sets consist of several paper grades at the cost of image contents. The first set of test samples (Test Set A) consisted of natural images printed with a prepress proofing inkjet printer on 16 different paper grades. The paper grades and the printing process were selected according to current practices, as described in detail in previous publications.[29–31] The natural images used in the study are presented in Fig. 1. The image contents were selected based on current practices and previous experience in media technology, and typical content types such as objects with details (cactus), a human portrait (man), and a landscape (landscape) were included. The fourth image content combined all the types (studio).

The second set of samples (Test Set B) consisted of images printed with a production-scale electrophotographic (EPG) printer on 21 different paper grades. The same image contents were used excluding studio [Fig. 1(d)]. The subjective evaluations, described below, were performed separately for both sets and image contents resulting in seven separate tests of 16 or 21 samples, respectively.

3.2 *Subjective Evaluation*

The performance of the selected IQA algorithms was studied against the psychometric subjective evaluations (subjective scores). The subjective evaluation procedure has been described in detail in a previous publication.[29] In brief, the sample images were attached on neutral gray frames of size A5 (148×210 mm). The observers were allowed to touch the frames, but not the images. Samples of a specific set (the same image content and printing method) were placed in random order on a table covered with a gray tablecloth. Labels with numbers ranging 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 label with number 1. Then, the observer was asked to select the highest quality sample and place it on the label with number 5. After that, the observer's task was to place the remaining samples on the labels, so that the quality increased steadily from 1 to 5. The final subjective score was formed by computing the MOSs over all observers ($N = 28$). For Test Set A, the standard deviation of the opinion scores varied from 0.26 to 0.76 depending on the sample, whereas for Test Set B, it varied from 0.18 to 1.3. The illumination

Table 1 NR-IQA algorithms used in this study.

IQA algorithm	Acronym	Features	Regression/quality estimation
Blind Image Integrity Notator—II[21,22]	BLIINDS-II	DCT coefficient statistics	Probabilistic model + support vector regression (SVR)
Blind Image Quality Index[17]	BIQI	Wavelet coefficient statistics	Support vector machine (SVM) + SVR
Distortion Identification-based Image Verity and Integrity Evaluation index[23]	DIIVINE	Scale-space-orientation decomposition coefficient statistics	SVM + SVR
Hybrid No-Reference model[24]	HNR	Curvelet, wavelet, and cosine transform coefficient statistics	Peak coordinates of logarithmic probability distribution of the coefficient magnitudes (LPMC)
Blind/Referenceless Image Spatial Quality Evaluator[25,26]	BRISQUE	Mean subtracted contrast normalized (MSCN) coefficient statistics	SVR
Natural Image Quality Evaluator[27]	NIQE	MSCN coefficient statistics	Distance between multivariate Gaussian models

used had a luminous intensity of 2200 lux and a color temperature of 5000 K.

3.3 *Applying NR Algorithms on Printed Images*

Since all the NR-IQA algorithms being studied have been developed for digital images, the printed samples need to be processed in such a manner that the algorithms can be applied. The first important consideration is related to the scanning (digitization) 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 available top-quality scanners, where subpixel accuracy of the original image can be achieved. Furthermore, to prevent photometric errors, the scanner color mapping should be adjusted to correspond to the original color information. This can be achieved by using the scanner profiling software accompanying the high-quality scanners. Second, a printed image contains halftone patterns, and, therefore, descreening is needed to remove the high-halftone frequencies and to form a continuous-tone image before the IQA algorithms developed for digital images can be applied.

To produce digitized versions of the prints, the samples were scanned using a high-quality scanner with 1250-dpi resolution and 48-bit RGB colors. A color management profile was devised for the scanner before scanning, and color correction, descreening, and other automatic settings of the scanner software were disabled. The digitized images were saved using lossless compression.

The descreening procedure was performed with a Gaussian low-pass filter which produces a continuous-tone image. To perform the descreening in a more perceptually plausible way, the images were converted to the CIE L*a*b* (Commission Internationale de l'Eclairage, L = lightness, a = chroma along red-green axis, b = chroma along yellow-blue axis) color space, in which the color channels are filtered separately. The CIE L*a*b* spans a perceptually uniform color space and does not suffer from problems related to, e.g., RGB, where color differences do not correspond to perception.[32] Moreover, the filter cut-off wavelength is limited by the halftone screen period and should not exceed 0.5 mm, which is the smallest visually disturbing detail from a viewing distance of 30 cm when the unevenness of print is evaluated.[33] In ideal conditions, the acuity limit of the human eye can be as small as 0.017 deg which corresponds to 0.1 mm.[34] It is crucial

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

Fig. 1 Image contents used in this study: (a) man; (b) lake; (c) cactus; and (d) studio.

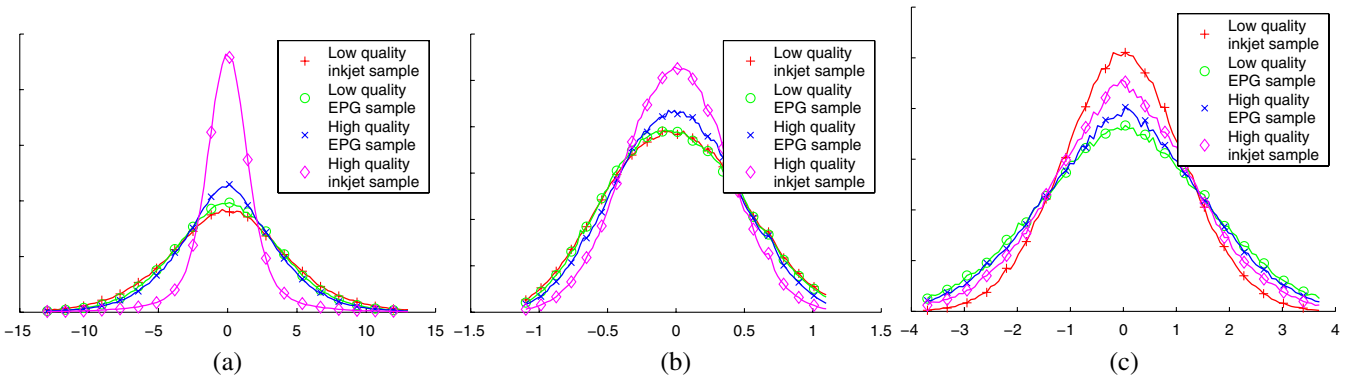


Fig. 2 Examples of coefficient histograms used to compute NSS features: (a) wavelet coefficient distribution used in BIQI; (b) MSCN coefficient distribution used in BRISQUE and NIQE; and (c) scale-space-orientation coefficient distribution used in DIIVINE.

to select a proper cut-off wavelength that corresponds to an assumed viewing distance and to the ability of a human observer to perceive distortions caused by the halftone pattern. The smaller the cut-off wavelength is, the greater the degree of distortion caused by the halftone pattern in the image after descreening. In this way, with a properly selected cut-off wavelength, the halftone pattern affects the quality predicted by the NR-IQA algorithms in a similar manner of the quality evaluation made by human observers. There also exists more sophisticated descreening methods, which are aimed at producing a visually pleasing result. However, since our goal was only to remove the halftone pattern and minimize other effects on the image, the descreening method was kept as simple as possible.

The images were downscaled to approximately match the resolution of the images that were used to develop the NR-IQA algorithms. Ideally, after descreening and scaling, all the distortions invisible to the human eye are removed. However, the optimal cut-off wavelength and scale factor need to be determined.

All the NR-IQA algorithms selected for this study require some kind of training or estimation of the NSS feature distribution. However, due to the laborious nature of preparing and subjectively evaluating printed samples, the amount of data is too small to properly train the methods. Therefore, a separate IQA database is required to train the methods. For all the selected NR-IQA algorithms, implementations available on the Internet and the provided (trained) model parameters were used. BIQI, BLIINDS-II, BRISQUE, DIIVINE, and HNR implementations were trained using the LIVE database. It contains, among other distortions, blur and noise, which have counterparts in print distortions (spreading of ink and graininess/mottling). The HNR implementation differs from the other methods by producing four different quality scores: noise, blur, JPEG2000, and JPEG quality. Only the noise measure (HNR-noise) was selected for further study, since it was shown to produce the best results with our data. The NIQE model was constructed using a larger set of only pristine images.[27]

3.4 *Results*

The descreening was performed using six different cut-off wavelengths: $0.05, 0.10, \ldots, 0.30$ mm, and seven scale factors: $0.08, 0.10, \ldots, 0.20$ corresponding to $100, 125, \ldots,$ 250 dpi. The NR-IQA algorithms were applied with every cut-off wavelength-scale pair to find the optimal parameter values for each method.

Figure 2 presents the examples of coefficient histograms that were used to compute NSS features. For visualization purposes, the histograms are presented for four samples: low-quality inkjet sample (the sample with lowest MOS in Test Set A), low-quality EPG sample, high-quality EPG sample, and high-quality inkjet sample. Although the two test sets were never combined, the quality variation between the selected samples is high. Most people would prefer the high-quality inkjet sample as the best one, followed by the high-quality EPG sample and low-quality EPG sample, the low-quality inkjet sample being the worst one. As can be seen from Fig. 2(a), the wavelet coefficient distributions used by BIQI seem to have a lower-standard deviation for high-quality samples and a higher one for low-quality samples, which indicates their potential in print QA. The same also applies for the MSCN coefficients used in BRISQUE and NIQE. However, the scale-space-orientation coefficients used in DIIVINE seem to be more suitable for distinguishing printing methods from each other than predicting the quality of printed images.

Figures 3 and 4 present the results of the best NR-IQA algorithm with the optimal cut-off wavelength and scale factor for each image content. As can be seen, the correlations are very high. However, selecting the optimal parameter values for each method produces overly optimistic results. Since the image content is often unknown beforehand in a practical application, the parameters should be fixed in such a manner that the NR-IQA algorithm is not sensitive to image content. Figure 5 shows the optimal parameter values for each method for different image contents, while Table 2 presents the corresponding correlations [linear correlation coefficient (LCC) and Spearman's rank correlation coefficient (SRCC)] between the MOS and algorithm scores. Also, the parameter values that maximize the mean correlation over all image contents and the corresponding correlation coefficients are presented.

As it can be seen from Fig. 3, Test Set A contains two relatively distinct clusters of different qualities: high-quality photopapers and lower-quality multipurpose papers. Any NR-IQA algorithm that distinguishes the two clusters and places them into the right order gets a high LCC value, while a method that fails to correct select the better cluster gets a high negative correlation coefficient. Moreover, the

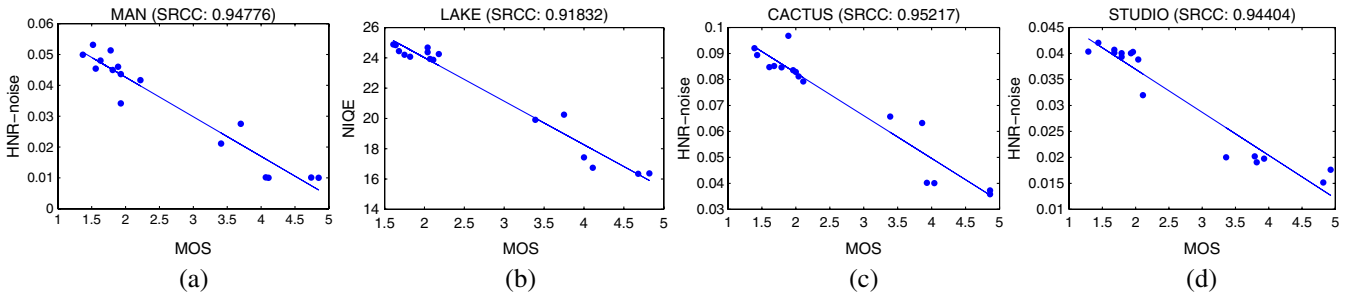


Fig. 3 The best method, cut-off wavelength, and scale factor selected for each image content of Test Set A: (a) man; (b) lake; (c) cactus; and (d) studio.

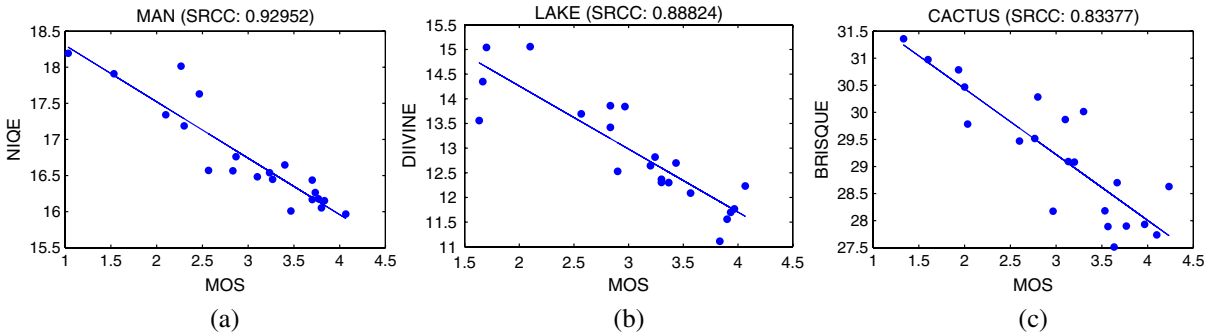


Fig. 4 The best method, cut-off wavelength, and scale factor selected for each image content of Test Set B: (a) man; (b) lake; and (d) cactus.

overall correlation coefficient increases if an algorithm places the cluster further away even if the correlations inside the clusters do not change. Therefore, based on the LCCs obtained using Test Set A, one can only determine whether a method works adequately at a coarse level or not, but it does not reveal the relative performances of the method studied in detail. Hence, when selecting the optimal combination parameter values, SRCC appears to be a more suitable measure.

To further study the performances of the various NR-IQA models on Test Set A using LCC, the samples representing the two clusters were divided into two and an additional test

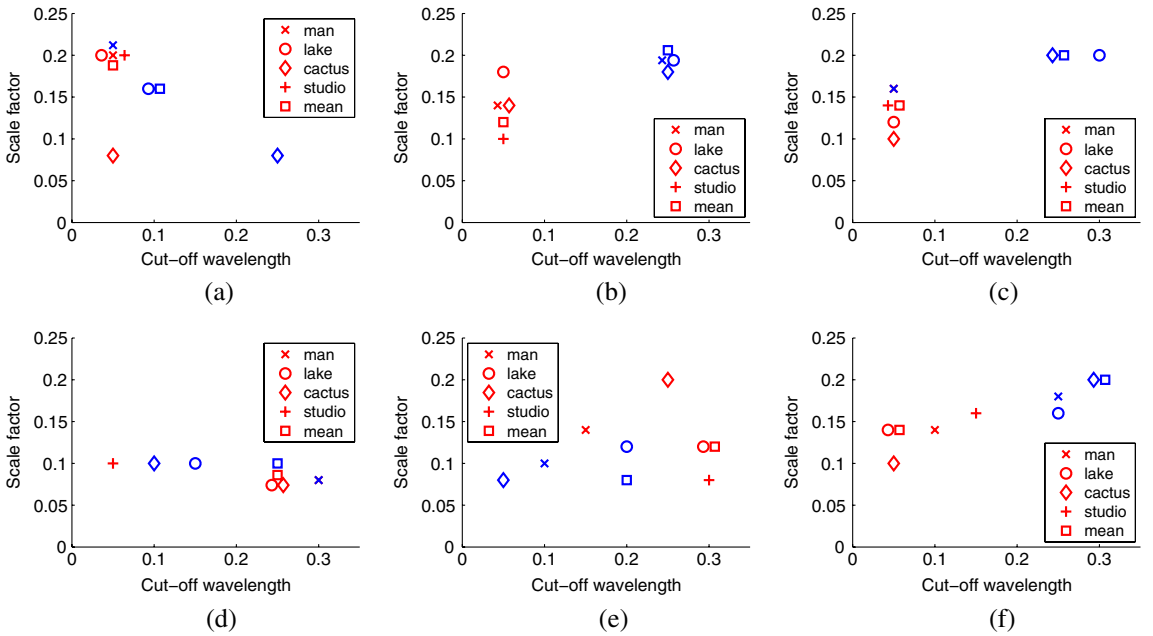


Fig. 5 Optimal parameter values for different image contents in Test Set A (red) and Test Set B (blue): (a) BIQI; (b) BLIINDS-II; (c) BRISQUE; (d) DIIVINE; (e) HNR-noise; and (f) NIQE. If for two or more image contents of the optimal parameter values were the same, then the markers were slightly shifted to make the figures more readable.

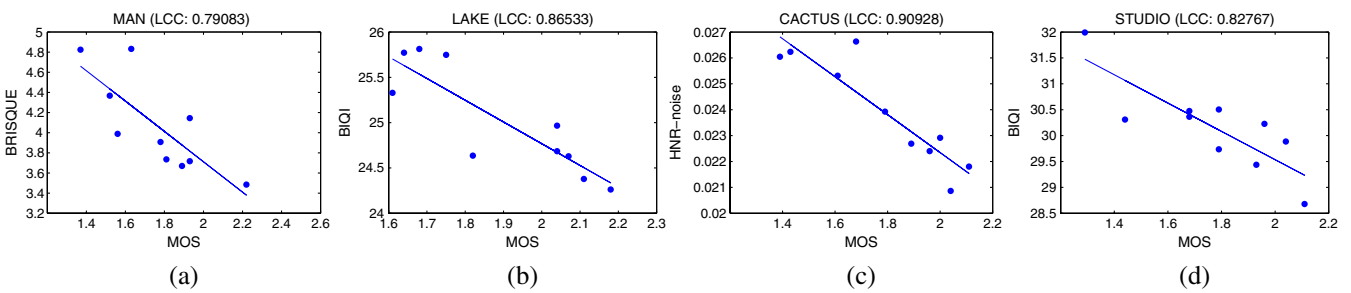
Table 2 Correlations with optimal parameter values for each image content. The first number in each entry is LCC value, and the second is SRCC. The numbers shown in bold indicate the best results.

		Method					
	Image content	BIQI	BLIINDS-II	BRISQUE	DIIVINE	HNR-noise	NIQE
Test Set A	Man	**0.97/0.92**	0.87/0.77	0.79/0.82	0.74/0.64	**0.98/0.95**	**0.97/0.92**
	Lake	**0.98/0.91**	0.69/0.56	0.88/0.60	0.95/0.83	**0.98/0.91**	**0.98/0.92**
	Cactus	**0.97/0.93**	0.72/0.62	0.93/0.85	0.92/0.86	**0.98/0.95**	**0.97/0.92**
	Studio	0.94/0.93	0.76/0.75	0.94/0.77	0.91/0.91	**0.98/0.94**	**0.97/0.84**
	Mean	**0.97/0.89**	0.33/0.32	0.84/0.69	0.85/0.77	**0.97/0.90**	**0.97/0.85**
Test Set A (multipurpose papers only)	Man	0.68/0.78	0.56/0.54	**0.79/0.75**	0.23/0.27	0.64/0.81	0.57/0.71
	Lake	**0.87/0.88**	0.60/0.54	0.69/0.70	0.39/0.40	0.76/0.76	0.59/0.69
	Cactus	0.77/0.87	0.85/0.86	0.76/0.82	0.62/0.68	**0.91/0.87**	0.78/0.78
	Studio	**0.83/0.79**	0.58/0.55	0.42/0.35	0.81/0.70	0.75/0.80	0.43/0.38
	Mean	**0.64/0.72**	0.39/0.39	0.50/0.43	0.27/0.25	**0.64/0.66**	0.34/0.44
Test Set B	Man	0.86/0.88	0.82/0.81	0.87/0.87	0.57/0.56	0.89/0.85	**0.92/0.93**
	Lake	**0.89/0.81**	0.65/0.60	0.77/0.70	**0.88/0.89**	0.70/0.54	0.85/0.73
	Cactus	**0.85/0.76**	0.75/0.75	**0.87/0.83**	0.71/0.72	0.78/0.67	0.70/0.67
	Mean	**0.82/0.79**	0.67/0.64	**0.80/0.76**	0.59/0.64	0.67/0.53	0.73/0.66

was carried out using the challenging lower-quality multipurpose papers. The result is shown in Fig. 6 and Table 2, and as expected, the correlations are much lower than on the full Test Set A. However, although the quality variation inside the set is very low, clear correlations between the MOS values and the algorithm scores can be observed.

Figures 7 and 8 show the results with optimal parameter values (mean SRCC over all image contents). Based on the figures, it is clear that the image content significantly affects the algorithm scores. Although the within-content correlations are high, different image quality score values for different contents cause a low overall correlation over all contents. However, it should be noted that the subjective evaluation results were separately scaled to the interval 1 to 5 for each content, and the MOS values are not directly comparable between the image contents. Therefore, combining the plots without preprocessing is not a well-grounded approach.

One option to avoid the above problem is to scale the IQA algorithm scores for each image content, as was done in Ref. 5. From a practical viewpoint, it is more interesting to put the 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 the IQA 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 to 5, but the IQA algorithm scores may differ significantly across image contents. Therefore, either the IQA algorithm scores need to be



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

Fig. 6 The best method, cut-off wavelength, and scale factor selected for each image content of Test Set A (only multipurpose papers): (a) man; (b) lake; (c) cactus; and (d) studio.

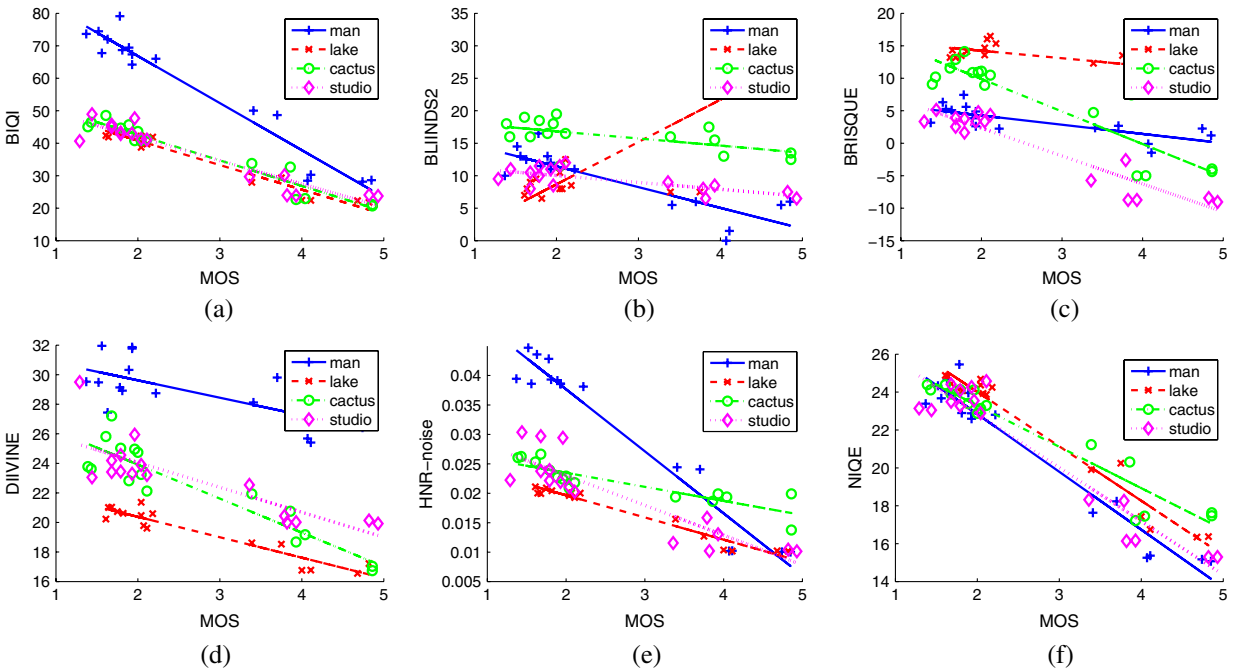


Fig. 7 Unaligned algorithm scores with the best cut-off wavelengths and scale factors selected for each method on Test Set A: (a) BIQI; (b) BLIINDS-II; (c) BRISQUE; (d) DIIVINE; (e) HNR-noise; and (f) NIQE.

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 IQA algorithms. Therefore, the different image contents were combined to form a larger test set by scaling the IQA algorithm scores. Here, the scaling was performed linearly. Let $\boldsymbol{x_n} = (x_{n,1}, \ldots, x_{n,M})$ represent the IQA algorithm scores of one assessment for all samples $(1, \ldots, M)$ within a single-image content n. Then, in the linear model, we have

$$\hat{x}_{n,i} = \hat{\boldsymbol{b}}_{\boldsymbol{n}} \begin{pmatrix} 1 \\ x_{n,i} \end{pmatrix}, \tag{4}$$

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

$$\hat{\boldsymbol{b}}_{\boldsymbol{n}} = \underset{\hat{\boldsymbol{b}}_{\boldsymbol{n}}}{\operatorname{argmin}} \sum_{i} [x_{1,i} - (b_{n,1} + b_{n,2} x_{n,i})]^2. \tag{5}$$

For the first image content $\hat{\boldsymbol{b}}_{\boldsymbol{1}} = (0,1)$ and for the remaining image contents $\hat{\boldsymbol{b}}_{\boldsymbol{n}}$ is such that the IQA algorithm scores

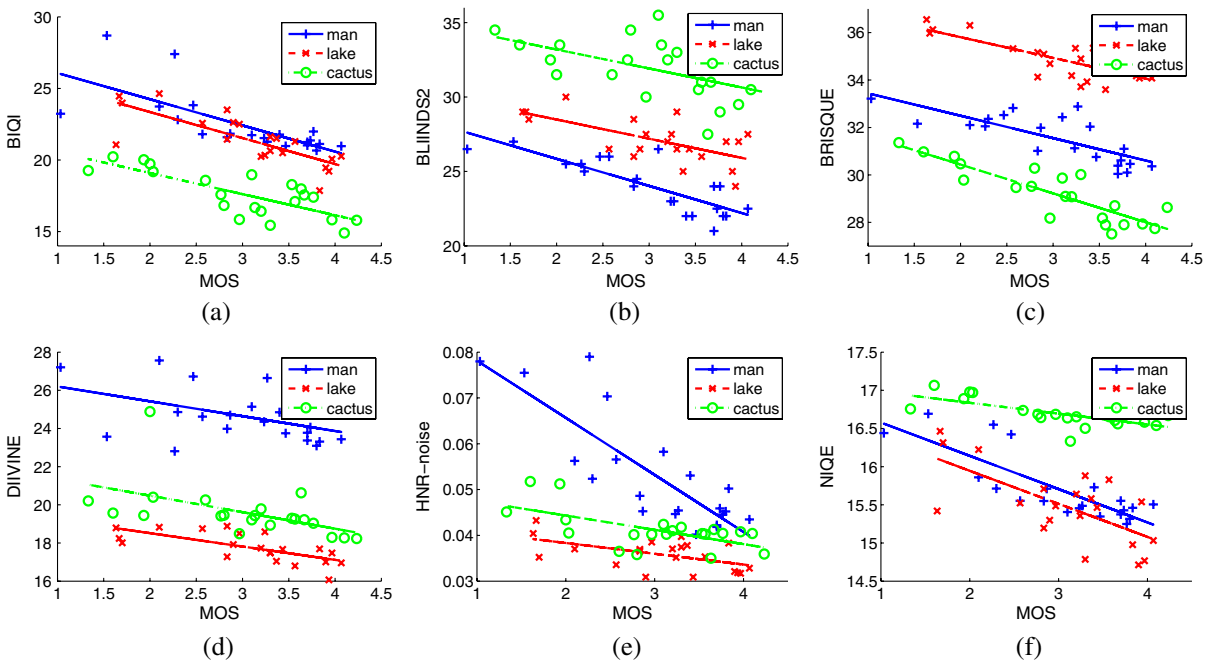


Fig. 8 Unaligned algorithm scores with the best cut-off wavelengths and scale factors selected for each method on Test Set B: (a) BIQI; (b) BLIINDS-II; (c) BRISQUE; (d) DIIVINE; (e) HNR-noise; and (f) NIQE.

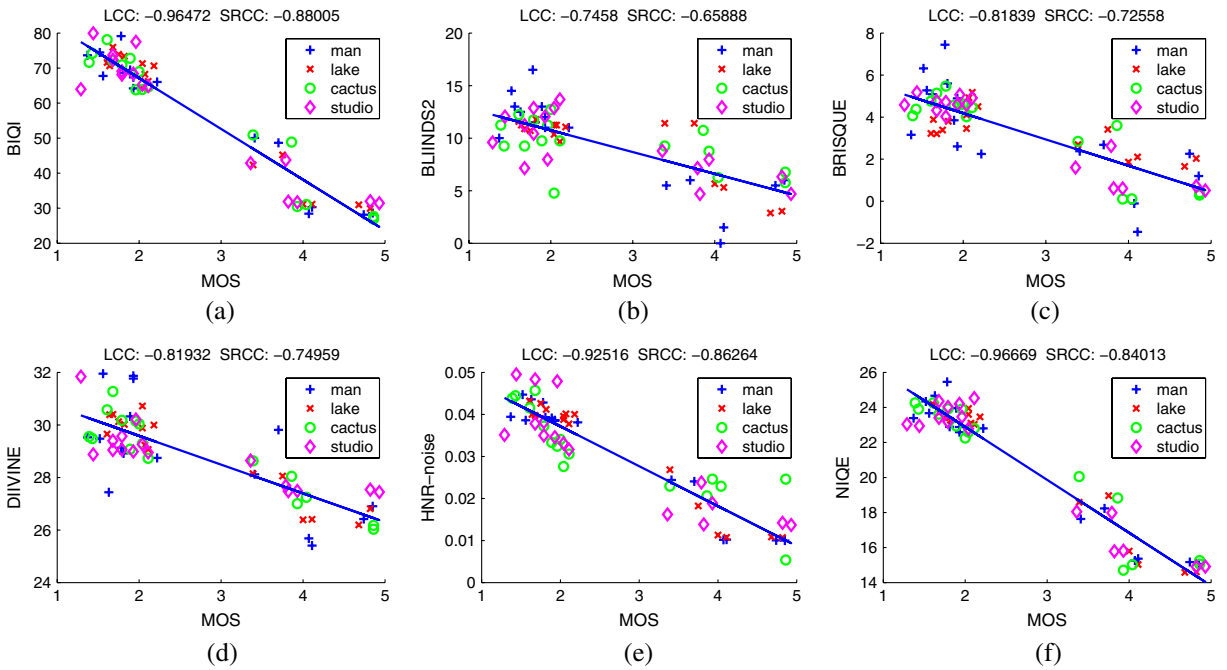


Fig. 9 Aligned algorithm scores with the best cut-off wavelengths and scale factors selected for each method on Test Set A: (a) BIQI; (b) BLIINDS-II; (c) BRISQUE; (d) DIIVINE; (e) HNR-noise; and (f) NIQE.

are converted to values similar to the values of the first image content on the same paper grade. The above-mentioned method does not allow combining Test Sets A and B, since there were no samples with comparable quality in different sets that could have been used to scale the MOS values to the same scale. Figures 9 and 10 and Table 3 present the results with aligned algorithm scores. Table 3 also lists the results for selected FR-IQA algorithms obtained using similar procedures:[5] peak signal-to-noise ratio (PSNR), structural similarity metric (SSIM),[35] multiscale SSIM (MS-SSIM),[36] and visual information fidelity (VIF).[37]

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 performance measures. The test is based on the assumption that the residuals (difference between MOS and the IQA algorithm score linearly fitted to MOS) are normally distributed. The normality of the residuals was tested using

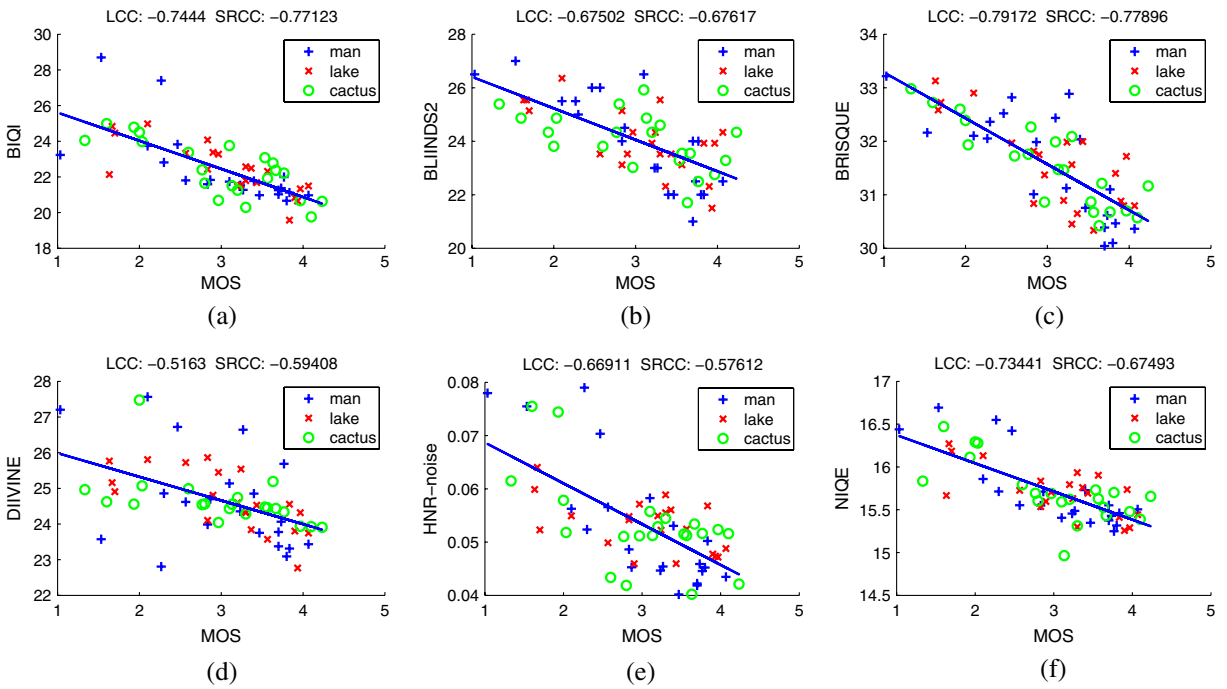


Fig. 10 Aligned algorithm scores with the best cut-off wavelengths and scale factors selected for each method on Test Set B: (a) BIQI; (b) BLIINDS-II; (c) BRISQUE; (d) DIIVINE; (e) HNR-noise; and (f) NIQE.

Table 3 Correlations between the MOS and the aligned algorithm scores from NR-IQA algorithms and from selected FR-IQA algorithms.[5] The numbers shown in bold indicate the best results.

	Test Set A		Test Set B	
IQA algorithm	LCC	SRCC	LCC	SRCC
PSNR (FR)	0.44	0.29	0.54	0.43
SSIM (FR)	0.98	0.92	0.80	0.70
MS-SSIM (FR)	0.98	0.91	0.89	0.79
VIF (FR)	0.98	0.90	0.82	0.74
BIQI (NR)	**0.96**	**0.88**	**0.74**	**0.77**
BLIINDS-II (NR)	0.75	0.66	0.68	0.68
BRISQUE (NR)	0.82	0.73	**0.79**	**0.78**
DIIVINE (NR)	0.82	0.75	0.52	0.59
HNR-noise (NR)	**0.93**	**0.86**	0.67	0.58
NIQE (NR)	**0.97**	**0.84**	**0.73**	0.67

the Lilliefors test[38] at a 5% significance level, and the residuals were shown to follow a normal distribution for all methods in Test Set A and for all methods except DIIVINE in Test Set B. Moreover, since DIIVINE had the lowest LCC in Test Set B, the non-normality has no significant effect on our conclusions. 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 90% confidence. The significance test results for the aligned algorithm scores are shown in Tables 4 and 5 for both test sets and for all possible pairings of QA algorithms.

Table 4 F-test results for Test Set A: 0 means that the QA algorithms are statistically indistinguishable from each other, 1 means that the IQA algorithm for the row is statistically better than the IQA algorithm in the column, and −1 means that the IQA algorithm in the row is statistically worse than the IQA algorithm for the column.

	BIQI	BLIINDS-II	BRISQUE	DIIVINE	HNR-noise	NIQE
BIQI	—	1	1	1	1	0
BLIINDS-II	−1	—	0	0	−1	−1
BRISQUE	−1	0	—	0	−1	−1
DIIVINE	−1	0	0	—	−1	−1
HNR-noise	−1	1	1	1	—	−1
NIQE	0	1	1	1	1	—

Table 5 F-test results for Test Set B.

	BIQI	BLIINDS-II	BRISQUE	DIIVINE	HNR-noise	NIQE
BIQI	—	0	0	1	0	0
BLIINDS-II	0	—	0	0	0	0
BRISQUE	0	0	—	1	0	0
DIIVINE	−1	0	−1	—	0	−1
HNR-noise	0	0	0	0	—	0
NIQE	0	0	0	1	0	—

4 Discussion

As described earlier, the training of IQA algorithms was performed using a separate set of distorted or pristine digital images, and the printed and scanned images were used only for testing. Therefore, the training and testing data were very different from each other and the training data did not contain the printing distortions. It is understandable that this reduced the performance of the learned models and might favor certain IQA algorithms. However, as the results showed, most IQA algorithms were still able to achieve high correlations with MOS, suggesting that the distortions in the training data were close to the distortions in the testing data. Moreover, since the ultimate goal is to develop an NR-IQA algorithm that needs to be trained only once, after which it should be possible to apply it to any appropriate image quality application, it is worthwhile to also evalate the algorithms' dependency on the quality of the training data. This justifies the use of different training and testing data as it reveals not only the performance of the methods, but also their sensitivity to imperfect training data.

On both test sets, most of the NR methods outperformed PSNR and the best methods were shown to produce almost as good results as state-of-the-art FR-IQA algorithms. As mentioned earlier, Test Set A contains two relatively distinct clusters, making SRCC a more reliable metric for comparing the performance of IQA algorithms. Based on the SRCC results, the best methods are BIQI, NIQE, and HNR-noise and these are also statistically significantly better than the rest of the methods. Test Set B does not contain similar distinguishable clusters, and therefore, LCC also provides useful information. Unlike SRCC, it also measures the linear dependence, i.e., whether a change in the quality at the high-quality end of the scale corresponds to a similar change at the low-quality end. Based on the LCC results, the best methods are BIQI, BRISQUE, and NIQE. The SRCC results support this conclusion. However, it should be noted that no statistically significant differences in performance were found between these methods and BLIINDS-II or HNR-noise.

The notably lower performance of DIIVINE compared with BIQI on both test sets is a surprising result, since DIIVINE is an extension of BIQI and it outperformed BIQI in experiments made using the LIVE database. The results shown in Fig. 2 suggest that the scale-space-orientation decomposition used by DIIVINE is more sensitive to the different printing methods and, therefore, a less suitable

approach for the QAs of printed images. Moreover, since it is the more complex of the two (88 NSS features in DIIVINE compared with 18 in BIQI), there is a higher chance of overtuning, making it more vulnerable to the large differences between training and testing data.

As can be seen in Figs. 7 and 8, image content has a noticeable effect on the algorithm scores. The main reasons for this are the fact that the NR-IQA algorithms were trained using a separate database containing different image contents, and the scaling of the MOS values for each type of image content was performed separately. Scaling the algorithm scores similarly to the MOS values does not really solve the problem, since we do not know if the effect of image content will be eliminated even if the above problems did not exist. Therefore, the results presented in Figs. 9 and 10 can be seen as an upper limit of method performance and to achieve such results, a manual effort is needed. However, if the parameters a and b that define the scaling of the algorithm scores could be predicted based on the image content, then the results presented in Figs. 9 and 10 could be achieved using a fully automatic method.

BIQI and NIQE perform well on both the test sets. While BIQI achieves slightly higher correlations, NIQE contains one significant benefit: it requires only pristine images to be trained. This enables training with a much larger number of images, making it less vulnerable to new image content. The results show that the NIQE is less sensitive to different image contents than the other tested NR-IQA algorithms [see Fig. 7(f)].

5 Conclusion

We applied the leading general-purpose NR-IQA algorithms to printed images and evaluated the performance of several state-of-the-art NR-IQA algorithms on an extensive set of printed photographs. We found that the BIQI and NIQE algorithms outperformed the other QA algorithms. Of these two, NIQE was less sensitive to image content, making it the most promising current method for NR-printed IQA.

Acknowledgments

The authors would like to thank the Finnish Funding Agency for Technology and Innovation (TEKES) and the partners of the DigiQ project (No. 40176/06) and the EffNet project (No. 913/10) for support. The research groups of Prof. Göte Nyman from University of Helsinki, Prof. Pirkko Oittinen from Aalto University, and Prof. Risto Ritala from Tampere University of Technology are especially acknowledged. The work was also supported by the Finnish Cultural Foundation. Alan Bovik's research was supported in part by the National Science Foundation under Grant IIS-1116656.

References

1. J. Imai and M. Omodani, "Reasons why we prefer reading on paper rather than displays: studies for seeking paper-like readability on electronic paper," *J. Imaging Sci. Technol.* **52**(5), 1–5 (2008).
2. Prophoto, "Trends in the Photo and Imaging Market—Photokina," 2012, http://www.prophoto-online.de/img/ftp/broschueren/Trends-in-the-photo-and-imaging-market-photokina-2012.pdf (3 June 2014).
3. P. G. Engeldrum, "A theory of image quality: the image quality circle," *J. Imaging Sci. Technol.* **48**(5), 446–456 (2004).
4. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.* **15**(11), 3440–3451 (2006).
5. T. Eerola et al., "Full reference printed image quality: measurement framework and statistical evaluation," *J. Imaging Sci. Technol.* **54**(1), 1–13 (2010).
6. U. Rajashekar et al., "Performance evaluation of mail-scanning cameras," *J. Electron. Imaging* **19**(2), 023008 (2010).
7. K. M. Braun, M. D. Fairchild, and P. J. Alessi, "Viewing techniques for cross-media image comparisons," *Color Res. Appl.* **21**(1), 6–17 (1996).
8. C. Li et al., "No-reference blur index using blur comparisons," *Electron. Lett.* **47**(17), 962–963 (2011).
9. P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Process. Lett.* **19**(7), 423–426 (2012).
10. C. T. Vu, T. D. Phan, and D. M. Chandler, "S3: a spectral and spatial measure of local perceived sharpness in natural images," *IEEE Trans. Image Process* **21**(3), 934–945 (2012).
11. Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. of the Int. Conf. on Image Processing*, pp. 477–480, IEEE, Rochester, New York (2002).
12. T. Eerola et al., "Bayesian network model of overall print quality: construction and structural optimisation," *Pattern Recogn. Lett.* **32**(11), 1558–1566 (2011).
13. H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. of the IEEE Computer Vision and Pattern Recognition*, pp. 305–312, IEEE, Colorado Springs (2011).
14. C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Trans. Neural Networks* **22**(5), 793–799 (2011).
15. P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Trans. Image Process.* **21**(7), 3129–3138 (2012).
16. M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.* **17**(6), 583–586 (2010).
17. A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.* **17**(5), 513–516 (2010).
18. D. L. Ruderman, "The statistics of natural images," *Network: Comput. Neural Syst.* **5**(4), 517–548 (1994).
19. K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.* **5**(1), 52–56 (1995).
20. N.-E. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *Proc. of the 16th IEEE Int. Conf. on Image Processing*, pp. 2281–2284, IEEE, Cairo, Egypt (2009).
21. M. A. Saad, A. C. Bovik, and C. Charrier, "DCT statistics model-based blind image quality assessment," in *Proc. of the IEEE Int. Conf. on image Processing*, pp. 3093–3096, IEEE, Brussels, Belgium (2011).
22. M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: a natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.* **21**(8), 3339–3352 (2012).
23. A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: from scene statistics to perceptual quality," *IEEE Trans. Image Process.* **20**(12), 3350–3364 (2011).
24. J. Shen, Q. Li, and G. Erlebacher, "Hybrid no-reference natural image quality assessment of noisy, blurry, JPEG2000, and JPEG images," *IEEE Trans. Image Process.* **20**(8), 2089–2098 (2011).
25. A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Proc. of the 45th Asilomar Conf. on Signals, Systems and Computers*, pp. 723–727, IEEE, Pacific Grove, California (2011).
26. A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.* **21**(12), 4695–4708 (2012).
27. A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.* **20**(3), 209–212 (2013).
28. H. R. Sheikh et al., "Live image quality assessment database release 2," 2005, http://live.ece.utexas.edu/research/quality (12 September 2012).
29. P. Oittinen et al., "Framework for modelling visual printed image quality from paper perspective," *Proc. SPIE* **6808**, 68080L (2008).
30. T. Eerola et al., "Is there hope for predicting human visual quality experience?," in *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics*, IEEE, Singapore (2008).
31. T. Eerola et al., "Finding best measurable quantities for predicting human visual quality experience," in *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics*, IEEE, Singapore (2008).
32. G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd ed., Wiley, New York (2000).
33. A. Sadovnikov et al., "Mottling assessment of solid printed areas and its correlation to perceived uniformity," in *14th Scandinavian Conf. of Image Processing*, pp. 411–418, Springer Berlin Heidelberg, Joensuu, Finland (2005).
34. J. M. Wolfe, K. R. Kluender, and D. M. Levi, *Sensation & Perception*, Sinauer Associates, Sunderland, Massachusetts (2006).
35. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
36. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. of the Thirty-Seventh Asilomar Conf. on Signals, Systems and Computers*, Vol. 2, pp. 1398–1402, IEEE, Pacific Grove, California (2003).

37. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* **15**(2), 430–444 (2006).
38. H. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *J. Am. Stat. Assoc.* **62**(318), 399–402 (1967).

Tuomas Eerola received his MSc and PhD degrees (doctor of technology) in information processing (computer science) in 2006 and 2010, respectively, from the Department of Information Technology in Lappeenranta University of Technology, Lappeenranta, Finland. He is currently a postdoctoral researcher with the Machine Vision and Pattern Recognition Laboratory, Lappeenranta University of Technology. His research interests include digital image processing, pattern recognition, and image quality assessment.

Lasse Lensu is a professor of computer science and engineering at Lappeenranta University of Technology (LUT), Finland. He received his MSc degree in data communications in 1991 and LicSc and DSc degrees in computer science in 2001 and 2002 from the Department of Information Technology of LUT. His research interests include digital imaging and image processing applications, computational vision, and biosensing. He has also been involved in the technology transfer to companies.

Heikki Kälviäinen has been a professor of computer science and engineering in the Machine Vision and Pattern Recognition Laboratory (MVPR) at Lappeenranta University of Technology (LUT) in Finland since 1999, currently located in the LUT Department of Mathematics and Physics. Besides LUT, he has been a visiting professor at Brno University of Technology, Czech Technical University, and University of Surrey. His primary research interests include machine vision, pattern recognition, and digital image processing and analysis.

Alan C. Bovik is the Curry/Cullen Trust Endowed chair professor at the University of Texas at Austin and director of the Laboratory for Image and Video Engineering (LIVE). His research interests include image and video processing, computational vision, and visual perception. He has published more than 700 technical articles and holds four US patents. His several books include the recent companion volumes *The Essential Guides to Image and Video Processing* (Academic Press, 2009).