From Full-Reference to No-Reference in Quality Assessment of Printed Images

Tuomas Eerola,* Joni-Kristian Kamarainen, Lasse Lensu and Heikki Kälviäinen
Machine Vision and Pattern Recognition Laboratory (MVPR)
Department of Information Technology
Lappeenranta University of Technology (LUT)
Finland

Abstract

Measuring visual quality of printed media is important as printed products play an essential role in every day life. The image quality assessment has been an active research topic in digital image processing, but adapting the developed methods to measuring visual quality of printed media has been considered rarely and is not straightforward. In this work, different methods for the quality assessment of printed images are considered. First, the so-called full-reference approach, where the original image with ideal quality is known, is presented. After that, the feasibility of using the reference-based approach for printed image quality assessments is discussed and problems related to the use of a digital reference image as the basis of the print quality analysis are shown. As a novel solution, a no-reference quality assessment approach is proposed. The solution based on a Bayesian network model of print quality is presented and its quantitative results are reported by using subjective data.

Keywords: print quality, image quality, quality assessment, full-reference, no-reference, Bayesian network, machine vision

1 Introduction

Despite the rapid development in electronic media, most people still prefer reading text printed on paper rather than reproduced on electronic displays [Imai and Omodani 2008]. Printed media can be also considered as more suitable for delivering localised news than electronic media. These, among other reasons, are why paper still has a notable role in communication, and the printed matter, such as books and newspapers, is an important part of daily life. When a customer purchases an image or printed product, one of the key factors is image quality [Engeldrum 2004]. Humans do not typically evaluate the quality of an image based on its physical parameters, but rather based on personal preferences and what they see as pleasurable [Engeldrum 2004].

The problem of how humans perceive the quality of a reproduced image is of interest to researchers of many fields related to vision science and engineering: optics and material physics, image processing (compression and transfer), printing and media technology, and psychology. The problem is especially difficult for the printed media since solving it requires understanding of paper and ink physics, visual measurements and optics and the human visual system. A measure for visual quality (print quality) cannot be defined without ambiguity because it is ultimately a subjective opinion of an “end-user” observing the result. Understandably, this evaluation has traditionally been conducted by human observers, but the recent development in computer and machine vision has made it intriguing to apply these methods to print quality evaluation. Machine vision utilises visual information reliably and may replace humans in certain laborious off-line evaluations. In addition, computational methods provide novel possibilities for on-line measurements for paper manufacturing and the printing industry.

The image quality assessment methods are usually divided into three categories: full-reference (FR), reduced-reference (RR) and no-reference (NR) quality assessment methods. In the FR methods, the reference image with ideal quality is available, whereas in the RR methods only some information of the reference image is given as input to the algorithm. In the NR methods, the reference image is unknown. FR is the main approach for evaluating and comparing the quality of digital images, especially compressed images. The digital representations of the original image and compressed image are in correspondence, i.e., there exist no spatial transformations between the images, and the compression should retain at least photometric equivalence. Therefore, the FR metrics can be computed in a straightforward manner by computing “distance metrics”. The NR image quality assessment is the most difficult task, and the majority of the proposed NR image quality assessment algorithms are designed for a single distortion type and are domain specific.

In this work, different methods for the quality assessment (QA) of printed images are presented. First, the FR assessment of printed images is considered by presenting the measurement framework proposed by the authors [Eerola et al. 2010] and by reporting the results for state-of-the-art digital image FR-QA algorithms. After that, the rationality of using the reference-based approach for printed image quality assessment and problems related to the use of a digital reference image are discussed. As a solution, a NR-QA algorithm based on the Bayesian network model and reported by the authors [Eerola et al. 2011] is presented.

2 Full-reference quality assessment of 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 by computing “distance metrics”. This is possible because digital representations are in direct 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 between the images. 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, before the other digital image for the comparison is established. In the following, the system where well-known methods are combined to form an automatic framework for analysing the full

*e-mail:tuomas.eerola@lut.fi
The framework of Fischler and Bolles [1981]. Based on these observations, the Hartley and Zisserman [2004] is applied to find the best homography for the correspondences, and finally, iv) transform one image into the coordinate system of the other image.

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 un-parametrised approach to find accurate and robust correspondence which is essential for the registration. The most popular method which combines both interest point detection and description is David Lowe’s scale-invariant feature transform (SIFT) [Lowe 2004]. The registration based on SIFT features has been utilised, for example, in mosaicking panoramic views [Brown and Lowe 2007]. The registration consists of the following stages: i) extract local features from the both images, ii) match the features (correspondence), iii) find a 2-D homography for the correspondences, and finally, iv) transform one image into the coordinate system of the other image.

The presented method performs a scale and rotation invariant extraction of local features using SIFT. The SIFT method also provides descriptors which can be used for matching. As a standard procedure, the random sample consensus (RANSAC) principle presented in [Fischler and Bolles 1981] 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 thousands of points). In the framework the implemented linear homography estimation methods are Umeyama for isometry and similarity [Umeyama 1991], and the restricted direct linear transform (DLT) for affine homography [Hartley and Zisserman 2003]. The only adjustable parameters in the method are the number of random iterations and the inlier distance threshold for RANSAC. The number of iterations can be seen as a trade-off between the computation time and the probability of successful registration. However, the homography estimation forms an insignificant proportion of the total computation time, and thus, a large value for the number of iterations is advisable. A tight inlier threshold was empirically shown to cause an unstable registration result and a rather high threshold value to still produce accurate registration [Eerola et al. 2009]. Based on these observations, the number of random iterations and the inlier distance threshold can be safely set respectively to 2000 and 0.7 mm (10 pixels with 360 dpi resolution). This makes the entire registration method practically free of parameters. For the image transformation, standard remapping using bicubic interpolation is utilised.

2.2 Image quality computation

In case of printed image quality assessment, the FR-QA algorithms contain special requirements. Although the above-mentioned registration works well small errors may occur. Because of this, simple pixel-wise distance formulations, such as the root mean square error (RMSE), do not work well. In other words, a good FR-QA algorithm should not be sensitive to such small registration errors. A more notable problem emerges from the subjective tests which are carried out by using printed (hardcopy) samples while the refer-
ence (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, those FR-QA algorithms that just compute simple similarity between the reference image and the input image do not succeed.

2.3 Experiments

2.3.1 Test sets

The objective of the study 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 set of test samples consisted of natural images printed with a production-scale electrophotographic printer on 21 different paper grades. The paper grades and the printing process were selected according to the current practices, as described in detail by the authors in previous publications [Oittinen et al. 2008; Eerola et al. 2008a; Eerola et al. 2008b]. The natural images used in the study are presented in Fig. 2. The image contents were selected based on current practices 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 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.

2.3.2 Subjective evaluation

The performance of the selected FR-QA algorithms was studied against the psychometrical subjective evaluations (subjective scores). The subjective evaluation procedure is described in detail by the authors in previous publication [Oittinen et al. 2008]. In brief, the sample images were attached on neutral grey frames with the size of A5 (148 x 210 mm). Evaluators were allowed to touch the frames but not the images. Samples of a specific image content were placed in a random order on a table, covered with a grey tablecloth. Labels with the numbers from 1 to 5 were also presented on the table. The evaluators were asked to select the sample image representing the lowest quality in the set and place it on the number 1. Then, the evaluator was asked to select the highest quality sample and place it on the number 5. After that, the evaluator’s task was to place the remaining samples on the numbers so that the quality increased steadily from 1 to 5. The final subjective score was formed by computing the mean opinion scores (MOSs) over all evaluators (N=29). The level of illumination was 2200 lux with colour temperature 5000 K.

2.4 Processing of raw data

From the practical point of view, it is more interesting to properly order paper grades 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 of 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. The first option was selected since the number of samples (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 \( x_n = (x_{n,1}, \ldots, x_{n,M}) \) represent the QA algorithm scores of one FR assessment for all samples \((1, \ldots, M)\) within a single image content \(n\). Then, the linear model is

\[
\hat{x}_{n,i} = b_n \left( \frac{1}{x_{n,i}} \right),
\]

where \( b_n = (b_{n,1}, b_{n,2}) \) are selected by minimising the errors between the image contents as

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

For the first image content, \( \hat{b}_1 = (0, 1) \), and for the remaining image contents, \( \hat{b}_n \) 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.

2.4.1 Results

Two performance measures were used: the linear correlation coefficient between the MOS and QA algorithm score after nonlinear regression (NLCC) and the Spearman rank order correlation coefficient (SROCC). NLCC measures how well the QA algorithm scores correspond to each other while SROCC measures how well the QA algorithm orders the samples from the best to the worst. For NLCC, the following nonlinearity (constrained to be monotonic) was used [Sheikh et al. 2006]:

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

where \( x \) is the modified algorithm score.

In Fig. 3 and Table 1, regression curves and performance measure values are presented for six different QA algorithms: Information fidelity criterion (IFC) [Sheikh et al. 2005], Multi-scale structural similarity metric (MS-SSIM) [Wang et al. 2003], Noise quality measure (NQM) [Damera-Venkata et al. 2000], Peak signal-to-noise ratio (PSNR), Visual information fidelity (VIF) [Sheikh and Bovik 2006], and Visual signal-to-noise ratio (VSNR) [Chandler...
Even more notable problems arise from the basic assumption of the FR approach: the reference image contain the ideal quality and it can be used as a basis for the quality evaluation. For the quality assessment of compressed images, this assumption is justified. A good image compression method reduces the size of the image in such 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 the assumption is correct. First of all, the original image is in very different form than the printed image that is evaluated, making its use as the reference image not only difficult, but also rather questionable. It is not unambiguous how the difference between a printed photograph and a digital image should be measured. Secondly, it is not clear that the best quality is achieved when the original image is transferred unchanged to the paper. Even in a hypothetical ideal situation where the original image is of the “perfect quality” (what ever that is) on the display, the quality is not necessarily “perfect” after transferring the image visually unchanged on paper due to the different nature of the media. Thirdly, while making subjective evaluations with print samples for method development purposes, it is not often possible to show the digital reference image to the evaluators and the evaluators are forced to make the decisions without knowing what the printed image was supposed to look like.

3 No-reference quality assessment

Because of the aforementioned reasons, it is justified to investigate NR quality assessment methods. Due to the obvious reasons, the NR quality assessment is much more difficult task than FR, and until recently there did not exist any general NR quality assessment methods. All of the proposed methods are either very application specific or measure only specific kind of distortion such as blur or noise. However, during the last few years, great developments have occurred and also more universal NR quality assessment methods have been established. These methods include curvelet, wavelet, and cosine transform based Hybrid No-Reference (HNR) model [Shen et al. 2011], natural scene statistics based BLind Image Integrity Notator (BLIINDS) [Saad et al. 2010], Distortion Identification-based Image Verity and INtegrity Evaluation (DIVINE) index [Moorthy and Bovik 2011] and a (No-reference) Free-Energy-based Quality Metric (NFEQM) [Zhai et al. 2012].

None of the proposed methods is an universal NR image quality assessment method because they consider mainly the quality of digital images and the test sets for validation have been rather limited. The proposed methods produce a single scalar value that is intended to describe the visual overall quality. However, in realistic and challenging visual evaluation tasks involving aesthetic or even personal attributes, it is highly unlikely that the overall visual print quality can be measured with a single measurement and represented by a single scalar value [Keelan 2002]. Even in the restricted settings with artefactual or preferential attributes, human evaluators are likely to give different ratings for the same samples. As a possible resolution, a statistical Bayesian network model for quality assessment for printed images is presented.

4 Bayesian network model of print quality

The stochastic nature of perception and interpretation of visual information motivates to treat the overall quality and its attributes
as probability distributions. For this purpose, the Bayesian theory provides a natural tool for modelling and analysis. A Bayesian network is an attractive choice since it is a probabilistic model that represents a set of random variables (instrumental measurements and subjective attributes) and their conditional independences with a directed graph. The idea of using Bayesian networks for modelling visual quality is not completely new. In [de Freitas Zampolo and Seara 2004] and [Pulla et al. 2008], Bayesian networks were used to describe the overall image quality. However, these studies were not complete. In [de Freitas Zampolo and Seara 2004], a network was used to combine noise [Damera-Venkata et al. 2000] and distortion measures [de Freitas Zampolo and Seara 2003]. The work reported in [Pulla et al. 2008] was more similar to this work since the authors used the network to combine objective and subjective assessment data. The objective measurements were given as input values, and the overall image quality was viewed as a probability distribution of ratings. The preceding works did not consider the problem of how to establish the network structure automatically based on true data. Instead, they showed the potential of Bayesian networks to model image quality and similar phenomena.

The Bayesian network model presented here was originally reported by the authors in [Eerola et al. 2011]. The idea in [Pulla et al. 2008] was advanced by proposing a method which automatically optimises the structure of the Bayesian network for using it as a model of visual print quality. This was done by making elementary hypotheses about the behaviour of the overall quality with respect to the objective (instrumental) measures (prior) and by computing the model fitness through simulation. The structure optimisation method was a genetic algorithm mainly due to the complexity of the optimisation problem and to the need for simulation to evaluate the fitness of a solution. The main contribution was an evolve-estimate-simulate optimisation loop where the structure and connections are evolved using an evolutionary approach, network parameters estimated using the maximum likelihood rule, and the network performance evaluated using simulation. The final network forms statistical dependencies between the psychometric data and instrumental measurements. The network can be used as a unified model representing and explaining the phenomenon and as a more practical tool producing a single visual quality index (VQI) for any printed product.

4.1 Main structure and variables

The first step to construct a Bayesian network is to select the nodes, that is, the random variables. In this work, the basic structure depicted in Fig. 4 is used for the Bayesian network. On the left, the nodes represent the objective measures that are the essential external model inputs. In the figure, the rightmost node represents the overall quality, the network output in the form of a probability distribution within the range of possible values. The middle portion consists of the subjective attributes which represent the abstract quality concepts shared and used by the individuals to form the basis of overall quality. These attributes were identified in the human experiments using well-defined psychometric tests. The subjective attributes form an intermediate layer which transforms the objective measures into the probabilistic overall quality.

The arrows indicate the causality of the model. In the network, the subjective attributes are interpreted as “the reality” that is desired to be measured. On the one hand, the subjective attributes induce a certain measuring result (the objective measurements), and additionally, their combination forms the perceived overall quality. This is why the direction of the causality is from the subjective attributes to the objective measures as well as to the overall quality. However, constraining the direction of the causality towards the objective measures does not prevent the inference of subjective attributes and the overall quality based on the objective measures.

It is important to notice that the objective measures on the left in Fig. 4 can be accurately and repeatedly measured from printed photographs and test fields. The overall quality, or more precisely, its distributions from experiments with a jury of evaluators, can be estimated by carrying out psychometric experiments. From this viewpoint, the model produces the most likely distribution of the overall quality opinions of the evaluators, if the same material is physically presented to a number of them.

The best objective measures for the left portion in Fig. 4 were selected according to the results of the prior works by the authors [Eerola et al. 2008a; Eerola et al. 2008b] where the most important instrumental and computational measures were surveyed and their relevance for explaining the overall quality was evaluated. The task was not possible by using the standard linear correlation; instead, non-linear relationships were evaluated and ranked using the proposed cumulative match score histogram (CMSH) [Eerola et al. 2008a]. The main idea behind the CMSH is the assumption that if two samples are visually perceived as being close to each other, they should be close to each other also based on the objective measures. If a measure fails to meet the criterion, it was classified as irrelevant for subjective overall quality. Using the method, it was possible to rank the existing measures, and even exhaustively search for the optimal combinations of \( N = 1, 2, \ldots, 6 \) best measures. For digital printing (inkjet and electrophotography), the following six measures were selected: computational motting [Sadovnikov et al. 2007], colour gamut, mean colour density, print gloss, edge blurriness and edge raggedness. This result is well in accordance with the current practises: these measures are commonly used in paper mill laboratories as well.

The selection of subjective attributes was based on systematic interviews of evaluators during the far-reaching subjective experiments. As a standard psychological interview technique [Radun et al. 2008] the evaluators were asked to describe visual factors that affected their ratings after they had given a rating for overall quality for each image. Later, a common vocabulary was established from the factors by using manual search, frequency analysis, and term mappings, and it was revised in the next independent experiments. Specifically, the most common subjective attributes were as follows: naturalness, clarity, colourfulness, subjective gloss, graininess, lightness, contrast, and sharpness. It should be noted that these subjective attributes do not necessarily correspond to their physical analogues since the semantic meaning of a term varied between the evaluators. This is typical for the natural, fuzzy concepts that naïve evaluators use in their everyday speech in contrast to the well-defined concepts used by the professionals. This difference
is not necessarily caused by the incorrect use of the terms, but by the fact that visual impressions of contrast, sharpness, naturalness etc. are not unambiguously related to any physical property of an image. For example, higher colour saturation may make the image look subjectively sharper, although the use of these concepts is separated among professionals. For this reason, the graph edges cannot be formed manually, but the relationships need to be learned.

4.2 Structure learning

Learning the optimal structure for a Bayesian network has been shown to be NP-complete [Chickering 1996]. As a consequence, full search methods are infeasible. Moreover, the laborious nature of collecting subjective data severely limits the available amount of training data. Therefore, also most heuristic methods, such as the PC algorithm [Spirtes and Glymour 1991], are not applicable. The structure optimisation is, however, essential for solving the problem and needs to be implemented into the learning process.

In the case of print quality modelling, it is possible to form a number of hypotheses on how the model should behave. For example, if the undesired solid printed area unevenness (mottling) increases while the other objective measures remain the same, the overall quality should decline. Similarly, if the colour gamut (a subset of colours a paper grade can reproduce with the available inks) expands, then the overall quality should improve. Using these heuristic and intuitively correct regulation rules, it is possible to produce a scalar value representing how logically correct a model is, that is, by randomly pruning how well model behaviour follows the hypotheses. This leads to a complex optimisation task: finding such a Bayesian network structure that the model behaves as consistently as possible after its parameters have been estimated using the training data. In this learning scheme, a network is not evaluated according to how well it represents the training data, but how well it represents the prior knowledge after the estimation with the training data. Therefore, the prior knowledge of behaviour acts as a regularisation term which enables the optimisation process with a small number of data points. The structure optimisation method is presented in more detail in [Eerola et al. 2011].

4.3 Experiments

To test the Bayesian network approach, the same test set was used as in the FR experiments (Sec 2.3). The objective measures were computed from the technical test fields (the lower row in Fig 2), and the subjective evaluation was carried out using the three natural image contents (human portrait, landscape and cactus). The subjective evaluations were performed using the procedure presented in Sec 2.3.2. In addition to the overall quality, the numerical scale was revealed also for the subjective attributes by asking human evaluators to label the samples based on a single attribute, such as sharpness or graininess. The subjective evaluation was conducted separately for each image content, and the number of evaluators was 29, so the number of training samples was $21 \times 3 \times 29 = 1827$. However, it should be noted that the objective measures were constant for each paper grade, and thus, the training data is extensive only for the subjective part of the model (only 21 different combinations of objective measures).

The initial population for the genetic algorithm consisted of 20 educated guesses, 20 fully random networks, and 20 partly educated guesses (some of the edges manually selected). Due to the long computation time, the number of simulations needed to evaluate the fitness of network structure, was set to small, and thus, the error margin for the fitness function value was relatively large. Therefore, a list of the 100 best structures was maintained during the optimisation process. Moreover, due to the stochastic nature of genetic algorithm, the structure learning was repeated 10 times resulting 1000 network structures. All the 1000 networks were evaluated against the subjective MOS using leave-one-out cross-validation. The best structure according to the correlation coefficient between the model output and MOS is shown in Fig. 5. In Fig. 6, the correlation against the subjective evaluation is plotted. The expectation values of the overall quality were used as the visual quality index (VQI). For a comparison, a tree-structured network learned using Chow-Liu algorithm [Pearl 1988] was also tested. With a correlation coefficient of 0.75 against MOS, the tree-structured network was outperformed by the network found using proposed structure optimisation method. A more extensive analysis of the found models can be found in [Eerola et al. 2011].

5 Conclusions and future work

In this work, different approaches to estimate the overall quality of printed images were presented. First, the full-reference approach was presented and analysed. The problems related to the full-reference approach, most notably the unclear definition of the reference (image), was discussed and investigation towards no-reference approaches justified. Different methods to no-reference quality assessment were shortly presented. A Bayesian network model of printed image quality was introduced and was shown to predict well the subjective human evaluations.

The main advantage of the Bayesian network approach is its versatility. The factors of low quality are easier to establish, since numerical values for the instrumental measurements are known. In comparison, the FR quality measures usually return only one scalar
value (some measures return also a dissimilarity map) that tells whether the overall quality is high or low. The Bayesian network helps us also to understand the perceived quality as a phenomenon. The structure of the network gives information about the relations of the objective measurements to the subjective attributes, and about the relations of the subjective attributes to each other. In addition, the Bayesian network works with incomplete measurements, i.e., only a portion of the objective measurements can be used as the evidence for predicting the overall quality. To be more specific, in the Bayesian network any nodes can be used as evidence to predict the value of any other node. This enables, for example, to fix the desired overall quality and examine the distribution of one instrumental measure with a certain combination of other instrumental measures.

The presented Bayesian network was a proof-of-concept, and more work is needed to make the approach an indispensable tool for the quality assessment of digital images.

Acknowledgements

The authors would like to thank Raisa Halonen and Prof. Pirkko Ottinen from the Department of Media Technology in Helsinki University of Technology for providing the test material, Tuomas Leisti and Prof. Göte Nyman from the Department of Psychology in the University of Helsinki for providing the subjective evaluation data, and Prof. Risto Ritala from the Department of Automation Science and Engineering in Tampere University of Technology for contributions related to Bayesian networks. The authors would like to thank also the Finnish Funding Agency for Technology and Innovation (TEKEs) and partners of the DigiQ project (TEKES Project No. 40176/06) for support.

References


Pulla, J., Mettänen, M., Konkarikoski, K., and Ritala, R. 2008. Bayesian network model as an overall image qual-


