Image quality metrics for the evaluation of print quality

Marius Pedersen[a], Nicolas Bonnier[b], Jon Y. Hardeberg[a] and Fritz Albregtsen[c].

[a]Gjøvik University College, P.O. Box 191, N-2802 Gjøvik, Norway;
[b]Océ Print Logic Technologies S.A., 1 rue Jean Lemoine 94015 Creteil cedex, France;
[c]Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, N-0316 Oslo, Norway.

ABSTRACT

Image quality metrics have become more and more popular in the image processing community. However, so far, no one has been able to define an image quality metric well correlated with the percept for overall image quality. One of the causes is that image quality is multi-dimensional and complex. One approach to bridge the gap between perceived and calculated image quality is to reduce the complexity of image quality, by breaking the overall quality into a set of quality attributes. In our research we have presented a set of quality attributes built on existing attributes from the literature. The six proposed quality attributes are: sharpness, color, lightness, artifacts, contrast, and physical. This set keeps the dimensionality to a minimum. An experiment validated the quality attributes as suitable for image quality evaluation.

The process of applying image quality metrics to printed images is not straightforward, because image quality metrics require a digital input. A framework has been developed for this process, which includes scanning the print to get a digital copy, image registration, and the application of image quality metrics. With quality attributes for the evaluation of image quality and a framework for applying image quality metrics, a selection of suitable image quality metrics for the different quality attributes has been carried out. Each of the quality attributes has been investigated, and an experimental analysis carried out to find the most suitable image quality metrics for the given quality attributes. For the many attributes metrics based on structural similarity are the the most suitable, while for other attributes further evaluation is required.

Keywords: Image quality, quality attributes, color printing, image quality metrics, color printing quality attributes

1. INTRODUCTION

The digital printing industry continues to grow, and a lot of effort is put into the development of new and improved technologies and products. With this increased growth the need for quality assessment also increases, for example to evaluate if new technologies outperform existing technologies, or to compare different products in order to find the best one.

There are two main classes of quality assessment; subjective and objective. The subjective assessment involves the use of human observers, while the objective assessment is without human observers, for example using measurement devices or algorithms. The use of algorithms, or rather Image Quality (IQ) metrics, are becoming more and more popular since they are time preservative, low cost, and require little competence by the user. Many IQ metrics have been proposed,[1] mostly with the goal of predicting perceived IQ. However, so far no one has been able to define an IQ metric correlated well with the percept for overall IQ. There are several reasons why, one being that IQ is multi-dimensional and very complex.

An approach to reduce the complexity, and help to bridge the gap between subjective and objective assessment of quality is to use Quality Attributes (QAs). These QAs are terms of perception,[2] such as lightness, saturation, and details. They help to reduce the complexity of IQ, and with a well-defined set of attributes also the dimensionality.

Our goal is to be able to measure perceived IQ without the involvement of human observers, more precisely using IQ metrics. In order to reach this goal we have taken the approach of using QAs, which we explain in this paper. The first step in our approach was to define a manageable a set of QAs,[3,4] and verifying them as meaningful for the evaluation of color prints.[5] Furthermore, the process of applying IQ metrics to color prints is not straightforward, since the physical printed image needs to be transformed into a digital copy. Therefore we proposed a framework to handle the transformation to a

Further author information:
Marius Pedersen: E-mail: marius.pedersen@hig.no, Telephone: +47 61 13 52 46, Fax: +47 61 13 52 40, Web: www.colorlab.no

Image Quality and System Performance VIII, edited by Susan P. Farnand, Frans Gaykema, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7867, 786702 · © 2011 SPIE-IS&T · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.876472

digital format.[6] Then we selected metrics for each QA, and evaluated them to ensure that they correlated with the percept of the QAs.[7]

This paper is organized as follows: First we investigate QAs for color prints, before proposing a set of attributes for the evaluation of color prints, which are experimentally validated. Then we propose a framework to digitize printed images, before we select and evaluate IQ metrics for each of the proposed QAs. At last we conclude and propose future work.

2. QUALITY ATTRIBUTES

QAs have been investigated in the literature, but there is no overall agreement on which attributes that are the most important. The first step towards being able to predict perceived IQ is to identify and categorize existing QAs in order to propose a refined selection of meaningful QAs for the evaluation of color prints.

2.1 State of the art

2.1.1 Quality attributes

In a study by Lindberg[8] 12 different QAs (overall quality, tone quality, detail highlights, detail shadow, gamut, sharpness, contrast, gloss level, gloss variation, color shift, patchiness, mottle, and ordered noise) were used to evaluate color prints. These QAs were reduced to two orthogonal dimensions by factor analysis based on perceptual data, one related to print mottle and one related to color gamut. These two dimensions accounted for almost all variation in the data set. Norberg et al.[9] evaluated overall quality, as well as color rendition, sharpness, contrast, detail rendition in highlight and shadow areas, color shift, gloss, mottle, and print homogeneity in a comparison of digital and traditional print technologies. Gast and Tse[10] evaluated six different QAs, banding, blur, noise, color rendition, tone reproduction and printer type, in terms of preference. Other QAs that have been investigated are: sharpness,[11] artifacts (for example noise[12] and banding[13]), naturalness,[14] contrast,[15] and color.[15–18]

Attention has not only been paid to single QAs, but also in the combined influence of them. Sawyer[19] investigated the influence of sharpness and graininess on perceived IQ separately, and their combined influence. Bartleson[20] carried out a similar study for color prints, investigating the combined influence of sharpness and graininess. The results from both these studies showed that the worst QA tended to determine the quality, and a change in other QAs would not increase quality. Natale-Hoffman et al.[21] carried out a study on the relation between color rendition and micro uniformity in terms of preference.

QAs have also been considered to be important for IQ metrics. For example, Morovic and Sun[17] proposed an IQ metric based on the lightness, hue, chromaticity QAs. Wang and Shang[22] showed that knowledge about QAs was beneficial for training IQ metrics, and thereby increasing their performance.

2.1.2 Image quality models

IQ models are intended to establish a link between subjective and objective IQ. These models are composed of QAs, and show how the QAs relate to each other and to overall IQ. The most common framework for these IQ models was proposed by Bartleson[20] in 1982. The approach of creating an IQ model was divided into three steps:

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

With these three steps Bartleson investigated the combined influence on sharpness and graininess. The advantage of using such a framework is being able to represent strengths and weaknesses of a given system by a relatively small number of QAs. Therefore, it has been used by many other researchers, such as Engeldrum,[23] Dalal et al.[18] and Keelan.[24] We also adopt this framework for the same reasons as other researchers.

Dalal et al.[18] proposed a two-sided appearance based system based on the framework of Bartleson. This system has one part for the printer and one for materials and stability. Each part of the system contains several QAs, these describe different aspects of the system, such as color rendition, uniformity, tone levels, and stability. This system has several advantages; It uses high level descriptors, which cover a wide range of IQ issues, from defects to sharpness. The printer side is also separated from materials and stability, allowing for separate analysis. The design of the model results in technology independence and the QAs are somewhat orthogonal, both being advantages. However, the system has some drawbacks, since the evaluation is mostly carried out by experts the results are influenced by the subjectivity of the observers. Also, the model might be unsuitable for non-experts due to its complexity. Since the model by Dalal et al. was designed mostly for subjective evaluation, the QAs are not adapted to IQ metrics, making it difficult to obtain a completely objective evaluation of IQ.

In addition to the models above, many others have been proposed.[17,25,26] Some of these are IQ metrics, which calculates one or several values representing IQ. Spatial-CIELAB (S-CIELAB)[25] is one of these IQ metrics, where a spatial pre-processing of the image is carried out before the CIELAB color difference formula[27] is applied to calculate IQ. S-CIELAB and other metrics are usually created to quantify overall IQ or specific QAs. It is very common that these metrics have several stages of processing, where each stage is connected to specific aspects of IQ.

2.2 Important issues regarding the selection of quality attributes

Investigation of the literature has revealed several issues that need to be taken into account when selecting QAs. First of all, the selection can based on different ideas, such as perception or technological issues. QAs based only on technological issues might not be suitable to evaluate perceived quality, while perceptual QAs might not be suitable for technological issues.

The views on how the QAs should be used greatly influence the selection of the QAs, whether they are selected for subjective or objective evaluation. If the QAs are intended for IQ metrics, they might not be directly suitable for measuring devices. The user might also influence the selection, since QAs used by experts can be different from those aimed at non-experts.

IQ models do not work with all available QAs, they usually consist of a sub-set of QAs. The number of selected QAs, or dimensionality, is a very important step. The IQ models available today have different dimensionality, mainly since there is a trade-off between the preciseness of QAs and the number of QAs. A high dimensionality in the model, will results in a more precise evaluation of IQ, where many different aspects are covered. Thus increasing the complexity, and all of the QAs might not be evaluated in a quality experiment. On the other side, having too few QAs might result in inaccurate estimation of quality.

Another important issue for the selection, which is discussed in the literature,[24] is independence. With independence between the QAs they can be easily combined to get one value of IQ. However, it is very difficult to achieve complete independence between QAs, and therefore care must be taken when combining values from the QAs.

Most QAs will consist of several sub-QAs, and the number of sub-QAs of a QA is described as its size. In the cases where QAs have different size, skewness occurs, and this will influence the importance of the QAs, making it somewhat complicated to combine values from the QAs to one overall value describing quality.

The key issues that must be dealt with when selecting QAs and building IQ models are summarized as:

- Origin of QAs.
- Intended use.
- Dimensionality.
- Independence.
- QA size.

2.3 Investigation and selection of important quality attributes

In order to select the most important QAs we have taken the approach of doing a survey of the existing literature. In order not to exclude QAs in this part of the investigation, we have included QAs based on both technology and perception, and QAs used with different intentions. These QAs include, for example, lightness,[17,24] sharpness,[8,9,11,12,28] blur,[10] contrast,[8,9,15,17] noise/graininess,[10,12,19,20] banding,[10,13] details,[9,12,15–17] naturalness[14] , color,[15,16,18] hue,[17] chroma,[17] saturation,[15] color rendition,[10,18] process color gamut,[18] artifacts,[15] mottle,[8,28] gloss,[8,9] tone reproduction,[10] color shift,[9,28] ordered noise,[9] patchiness,[9] line quality,[18] text quality,[18] adjacency,[18] printer type,[10] effective resolution,[18] effective tone levels,[18] gloss uniformity,[18] skin color,[16] paper roughness,[28] paper flatness,[18] paper whiteness,[28] perceived gray value,[12] structure changes,[12] micro uniformity,[18] macro uniformity,[18] structure properties,[12] and color gamut.[28]

It is not practical to use all of these QAs in the evaluation of print quality, therefore they need to be reduced to a more manageable set. There are several issues to consider when reducing the number QAs, as mentioned previously. A long term goal of our research is to create a link between subjective and objective IQ of color prints. Based on this the QAs should be based on perception, additionally they should account for technological issues. The QAs should be suitable for novice observers, and therefore being general and straightforward to evaluate. Since we want to measure IQ using IQ metrics, the

QAs be suitable for this type of evaluation. The existing sets of QAs and models do not fulfill all of these requirements, therefore a new set of QAs is needed.

It is clear from the QAs listed above that many of them overlap and have common denominators, they can therefore be grouped within more general QAs in order to reduce the dimensionality and create a more manageable set of QA for the evaluation of IQ. When it comes to dimensionality there is usually a compromise between generality and accuracy. The accuracy is low with a small set of QAs, but the advantage is low complexity. Using a large set of QAs results in high accuracy, but at the cost of high complexity. We have linked most of above QAs to six different dimensions, considered as important for the evaluation of IQ. This results in a reasonable compromise between accuracy and complexity. We are also close to the statement by Engeldrum[23] that observers will not perceive more than five QAs simultaneously. We have reduced the QAs found in the literature to the following six:

- **Color** contains aspects related to color, such as hue, saturation, and color rendition, except lightness.
- **Lightness** is considered so perceptually important that it is beneficial to separate it from the color QA.[24] Lightness will range from ”light” to ”dark”.[2]
- **Contrast** can be described as the perceived magnitude of visually meaningful differences, global and local, in lightness and chromaticity, within the image.
- **Sharpness** is related to the clarity of details[11] and definition of edges.[29]
- In color printing some **artifacts** can be perceived in the resulting image. These artifacts, like noise, contouring, and banding, contribute to degrading the quality of an image if detectable.[30]
- The **physical** QA contains all physical parameters that affect quality, such as paper properties and gloss.

The six dimensions are concise, yet comprehensive, general high-level descriptors, being either artifactural, i.e., those which degrade the quality if detectable,[30] or preferential, i.e., those which are always visible in an image and have preferred positions.[30] These attributes are referred to as the Color Printing Quality Attributes (CPQAs).

In order to provide a simple and intuitive illustration of the CPQAs and their influence on overall IQ we have turned to Venn diagrams. Venn diagrams may be used to show possible logical relations between a set of attributes. However, it is not possible to create a simple Venn diagram with a six fold symmetry.[31] Therefore we illustrate the CPQAs using only five folds, leaving the physical QA out. This does not mean that the physical CPQA is less important than other CPQAs.

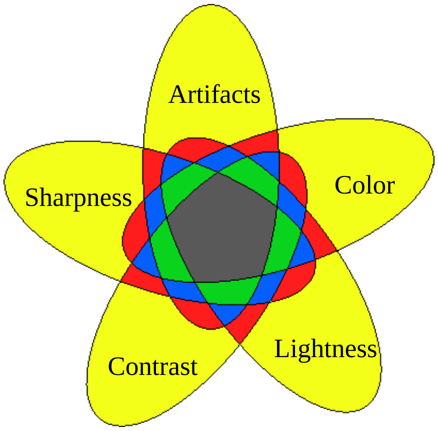


Figure 1. Simple Venn ellipse diagram with five folds used for an abstract illustration of the QAs. Five different QAs and the interaction between these are shown. Overall IQ can be influenced by one, two, three, four, or five of the QAs.

2.4 Validation of the quality attributes

2.4.1 How to validate quality attributes?

The validation should be adapted to the criteria on which the CPQAs were selected. The validation can be achieved by comparing data to the CPQAs, and analyzing the correspondence between the data and the CPQAs. Requirements need to be set to validate the CPQAs. Using the aspects on which they were selected we can derive the important requirements that the CPQAs should fulfill. For the CPQAs to be useful for the evaluation of IQ, to be perceptual, and account for technological issues, they should be able to cover the entire field of IQ. All issues encountered in the evaluation of color prints should be described using the CPQAs, making this one of the requirements to validate. As many as possible of the QAs used by the observers should be accounted for within one of the CPQAs, and not overlap several CPQAs. Minimum overlapping is considered as one of the requirements the CPQAs should fulfill. The CPQAs were selected to keep the number of QAs to a minimum, this is important for usability of the QAs, and for the CPQAs to be straightforward to use. Therefore dimensionality should be one of the requirements. For the CPQAs to be suitable for IQ metrics and straightforward to use, it is important to keep independence. Summarized, we have four different requirements the CPQAs should fulfill in order to be validated:

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

There are several ways to carry out the validation for these requirements. The validation can be carried out subjectively or objectively. In order to minimize the subjective influence, and to have an independent validation of the QAs; objective validation methods have been investigated. It is preferable to have a fully objective method, where data, for example from an experiment, can be compared to the CPQAs. This requires a database containing all QAs, categorization of them, and their relations. To our knowledge such a database does not exist, making this method inapplicable. Another possible method is to use existing definitions of QAs to create relations between the QAs, resulting in a data structure. This method is not completely objective, but it keeps the subjectivity to a minimum, being the best alternative and also the method we adopt to validate the CPQAs.

Subjective data is required for the validation since the CPQAs are perceptual. In order to validate if the CPQAs cover the entire field of IQ it is required that the observers use a wide variety QAs. Therefore, experts are the best choice, since they are more familiar with QAs and IQ issues. In addition, the color workflow on which the data is collected should guarantee many different quality issues. The image set should also include a wide variety of characteristics to ensure many different IQ issues.

One way to carry out such an experiment is to provide the CPQAs and their definitions to the observers, and ask the observers to use them in their judgment of IQ. If the observers only use the CPQAs, one could argue that they cover all aspects of IQ. However, this experimental setup can restrict the observers to the CPQAs, and prevent them from using others QAs they normally would use. Another option is to record the QAs used by the observers during the experiment, where the observers do not have prior knowledge to the CPQAs. Since the last option does not restrict the observers to the CPQAs, we will use this method.

2.4.2 Experimental setup

Images Several guidelines have been given in the literature for the selection of images, in the context of investigating IQ issues.[32] We have followed these guidelines in the selection of 25 images (Figure 2) for the experiment.

To address the customer segment of Océ, we have also included 3D models, maps, posters, presentations, and pdf-like documents. The images have been collected from different sources. One image from ISO,[33] two from CIE,[32] ten images from the authors, one image from MapTube,[34] three images from ESA,[35] four images from Google 3D Warehouse,[36] one image reproduced with permission from Ole Jakob Skattum, and one image from Halonen et al.[37] The images were 150 dpi 16-bit sRGB, saved as tiff files without compression.

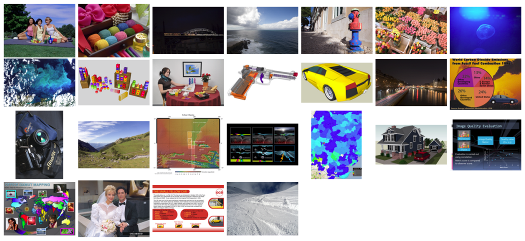


Figure 2. The 25 images used in the experiment to validate the CPQAs.

Color workflow The images were printed on an Océ Colorwave 600 CMYK wide format printer on Oce Red Label (LFM054) plain uncoated paper. The profile of the printer was generated using a GretagMacbeth TC3.5 CMYK + Calibration test chart in ProfileMaker Pro 5.0.8. A round trip test was carried out to ensure a correct profile as suggested by Sharma,[38] and we performed a visual inspection of color gradients to verify that no artifacts occurred. The images were printed with three different rendering intents: perceptual, relative colorimetric, and relative colorimetric with black point compensation.

Viewing conditions The experiment took place in a controlled viewing room under a color temperature of 5200K, and an illuminance level of 450 ±75 lux and a color rendering index of 96. The observers were presented with a reference image on an an EIZO ColorEdge CG221 display at a color temperature of 6500K and a white luminance level of 80 cd/m^2, following the specifications of the sRGB. The observers viewed the reference image and the printed image simultaneously from a distance of approximately 60 cm. The experiment followed the CIE guidelines[32] as closely as possible.

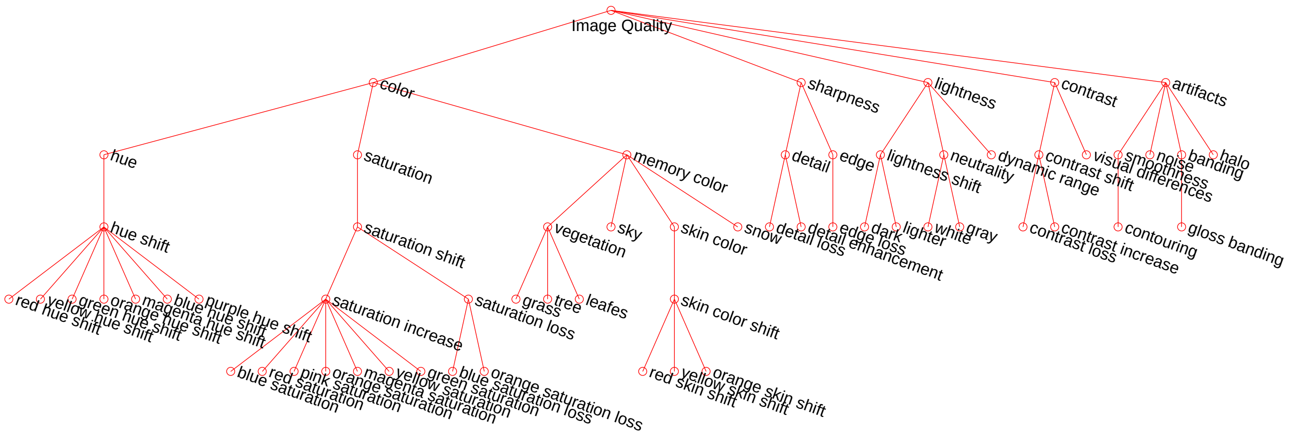


Figure 4. The QA tree generated from the attributes used by four expert observers. Each level of a sub-tree has been sorted from left to right based on frequency (high to low).

Instructions The instructions given to the observers focused both on the overall quality rating and on the QAs, where they were asked to rank the reproduction according to quality. The whole experiment was filmed, and the observers were encouraged to describe and talk about their observations. The video enabled the authors to extensively analyze the results, and it resulted in a more free experiment than if they were to write down their observations.

2.4.3 Fitting the quality attributes data to the color printing quality attributes

Four experts observers judged the 25 images, which resulted in a total of 100 observations, and more than six hours of video were recorded. The video was transcribed by the authors with focus on the QAs used by the observers. Numerous QAs, more than 750 in total and more than 350 different QAs, were used by the expert observers. This data constitutes the basis for the validation of the CPQAs.

Many of the words and phrases from the observers are similar, and some even synonyms. Therefore, the QAs from the experiment need to be categorized. Similar words and phrases should be grouped together, and relations between terms found. We have chosen to use existing definitions for this, which results in two different approaches; top-down or bottom-up. The top-down approach builds relations from the most general QAs and downwards to the most specific QAs. This method requires building a full tree structure with all relations, further it can be compared to the QAs used by the observers. In the bottom-up approach, the QAs from the observers are the starting points. The QAs are grouped into more general attributes till the most general QA is reached. This method has the advantage that it does not require building a full tree structure prior to the analysis. Therefore, it is the most suitable method to validate the CPQAs.

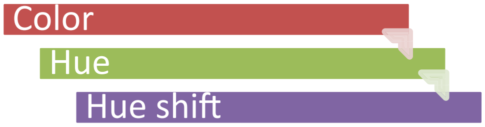


Figure 3. Bottom-up procedure for the QA hue shift, which belongs to the more general hue attribute, which in turn belongs to the color attribute.

The analysis is carried out as follows; given that the observer has used the QA hue shift, this QA belongs to the more general QA hue. Using the relations of Pedersen et al.[3,4] and the definition by Wyszecki and Styles,[2] hue is considered a part of the more general color QA, which is one of the CPQAs (Figure 3).

The bottom-up approach described above has been used to generate a tree for all the images and observers in the experiment (Figure 4). Since the physical CPQA was not changed in the experiment, we limit the discussion to five of the six CPQAs, excluding the physical CPQA.

Discussion on the fitting of quality attributes Several issues were encountered while fitting the QAs, that are discussed below.
Overlapping QAs regards the issues that some of the QAs are difficult to group within only one of the CPQAs. Naturalness is one of these attributes. We have previously argued that naturalness can be accounted for by using several of the main or sub-attributes.[3,4] In this experiment the observers used several QAs together with naturalness, the analysis carried out reveals that a change in one or several of the other QAs very often lead to the impression of an unnatural image. Naturalness

was used in five observations, and in all of these observations the term color was used, contrast was used in three of the five, while memory colors in four of the five observations. Also in the literature it has been shown that naturalness depends on chroma and colorfulness,[39] contrast,[39] and memory colors.[40] Because of this, naturalness is most likely accounted for if these QAs are of reasonable quality.

Gamut is another overlapping QA, which cannot be listed as a sub-QA under one of the CPQAs, since it is dependent on both the color CPQA and the lightness CPQA. The observers used both color and lightness in the three observations where gamut was used. Therefore, gamut can be accounted for using the color and lightness CPQAs.

Readability and legibility were also used in the experiment, which are often used to describe textual information. Zuffi et al.[41] found these two to be related, and they are also influenced by contrast[42,43] and sharpness. In five of the eleven observations from the experiment where legibility and readability were used, the observers used both contrast and sharpness, in the remaining six observations either sharpness or contrast was used. Therefore legibility and readability are most likely accounted for with the CPQAs.

For the QA memory colors the observers only specified changes in color (saturation and hue), not changes in lightness, and therefore memory colors are located under color in Figure 4. Nevertheless, changes in lightness might occur and in these cases memory color might also be placed under the lightness QA as well.

Independence is another issue encountered in the fitting of the QAs. Dynamic range is considered as a sub-QA of lightness, but it has also been shown to influence contrast.[44] The observers indicated a relation between dynamic range and visibility of details, but it should be mentioned that dynamic range was only used in two observations.

Contrast is one of the QAs where the experimental data indicates independence. Contrast is influenced by saturation and lightness, but also linked to detail. Since the definition of contrast contains both color and lightness it is perhaps the least independent QA. The observers often used contrast separated from color and lightness, which make contrast a very complex QAs. As mentioned above contrast is important to account for naturalness and readability. Without the contrast CPQA we would not cover the whole field of quality, and it is therefore required in order to fulfill the criteria on which the CPQAs were selected, even at the expense of independence.

To analyze the independence between the QAs used by the observers we have carried out a cross-tabulation and chi-square analysis. The input data to the analysis was whether or not one of the five CPQAs was used by the observers for each of the 25 images, and we use a 5% significance level. This analysis has the disadvantage that it does not give any information on the nature of the relationship between the CPQAs, it only gives information about when two CPQAs are used together. The results show a dependence between artifacts and lightness, but also dependence between artifacts and sharpness, and contrast and lightness. It was mentioned above that the observers indicated a relation between contrast and dynamic range, one of the sub-QAs of lightness. However, the dependence analysis between these does not reveal a relation, neither for detail visibility and dynamic range. It should be noted that for all of these the amount of data is too small to conclude for these specific QAs (Figure 6).

Global and local issues have also been commented on by the observers. The QAs above can be divided into global and local attributes, this differentiation can be important for the assessment of IQ, but also in the method used to combine results from different QAs.

One child with several own children is another issue seen in Figure 4, where some QAs have only one child, and this child has several own children. It could then be argued that these QAs could be discarded, and the QA below could be linked to the parent of the removed QA. For example, saturation could be removed and replaced with saturation shift. However, observers have used these terms, as saturation alone, without further specification, which indicates that these levels are important and should be kept. Another argument for keeping these QAs, is that there might be several children, such as for the edge QA, where one could suggest having two children as for the detail QA, one for edge loss and another for edge enhancement.

Skewness among the QAs have been found in the experimental data. The color CPQA has significantly more sub-QAs than the other CPQAs. This can be used as an additional argument for separating lightness from the color CPQA in order to reduce skewness. Additionally, separation of these CPQAs leads to a simple adaptation for evaluation of grayscale images. Another very important issue strongly influenced by skewness is how to combine IQ values for the different CPQAs to one IQ value representing IQ. When the CPQAs are skewed, this combination might not be straightforward, and thereby complex.

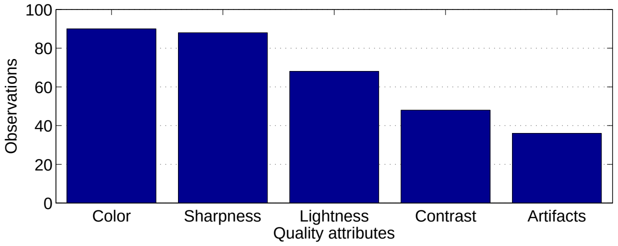


Figure 5. Frequency of use for the CPQAs for the 100 observations. Color is the CPQA used the most by the observers.

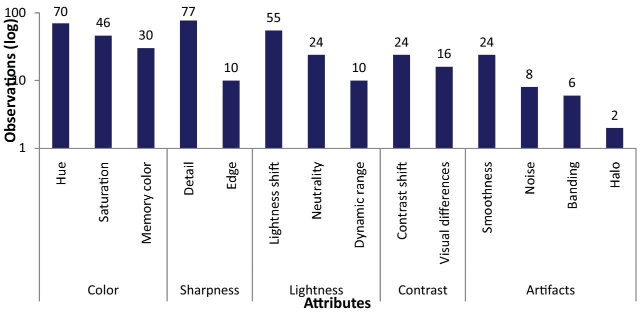


Figure 6. Distribution of sub-QAs for the CPQAs. Detail and hue are the most used, while the sub-QAs of artifacts are the least used.

Dimensionality is important for the complexity of the CPQAs. In the experiment the observers used all CPQAs, and because of this none of the CPQAs can be directly removed to reduce the number of QAs. However, the color and lightness CPQAs could be merged, but at the cost of increased skewness. Analysis of the use of the color and lightness CPQAs show that there were 39 observations where both CPQAs were used, and nine observations where lightness was addressed without color. These results reflect the printing workflow and images on which the analysis was carried out. Therefore, it is not unlikely that for specific workflows with specific documents, the dimensionality can be reduced.

2.4.4 Observations on the color printing quality attributes

The experimental data also leads to different observations on the CPQAs, which can be valuable in the assessment of IQ. One of these observations is how often the different CPQAs have been used, which is shown in Figure 5. Color is the CPQA used most frequently by the observers, closely followed by sharpness. Lightness the third most used CPQA, contrast the fourth, and artifacts is the least used CPQA by the experts. These observations indicate that the color and sharpness CPQAs should be evaluated in all images, and that the artifact and contrast CPQA only needs to be evaluated in some images.

The distribution of the sub-QAs for the CPQAs is also interesting to look at (Figure 6). Detail is the most frequently used sub-QA, closely followed by hue. It is not surprising that detail is the most used CPQA since the rendering intents reproduce details differently, especially visible in the shadow regions. The perceptual rendering intent gave a slight, but noticeable, hue shift in some images, resulting in the frequent use of hue sub-QA. The overview shows that the sub-QAs of the artifacts CPQA are the least used, this was expected since these attributes are very specific.

2.4.5 Validation summary

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

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

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

The final and last requirement regarded dependence. We found some dependence between the CPQAs, but the CPQAs are not fully independent,[3,4] because it is very difficult, if not impossible, to account for all quality issues while maintaining a low dimensionality. The experimental results indicate that contrast is the least independent CPQA. However, contrast cannot be removed since it is often used by the observer, and it is also used without relations to other CPQAs. For that reason we consider the dependence found to be acceptable.

3. APPLYING IMAGE QUALITY METRICS TO PRINTS

The CPQAs proposed above are selected with the intention of being used with IQ metrics. However, applying IQ metrics to printed images is not straightforward since the metrics require a digital input. In order to accomplish this, the printed reproduction needs to be transformed into a digital copy. We will first give an overview of existing framework, before we introduce new a framework for this process.

3.1 Existing frameworks for using image quality metrics on printed images

A few frameworks have been proposed for using IQ metrics on printed images. All these frameworks follow the same procedure; as a first step the printed image is scanned, which can be followed by a descreening procedure to remove halftoning patterns. Then image registration is performed to match the scanned image with the original. Finally, an IQ metric can be applied.

In 1997 Zhang et al.[45] proposed a framework, where the image is scanned, then three additional scans are performed, each with a different color filter. This results in enough information to transform the images correctly to CIEXYZ. The scanning resolution was set to 600 dpi, but little information is given on the registration and the descreening.

Another framework was proposed by Yanfang et al.[46] in 2008. Two control points were applied to the image before printing, one point to the upper left corner and one to the upper center, to assist in the registration. Resolution for the scanning was 300 dpi, and descreening was performed by the scanner.

In 2009 Eerola et al.[47] proposed a new framework, which follows the same steps of the previous frameworks. First the printed reproduction is scanned, then both the original and the reproduction go through a descreening procedure, which is performed using a Gaussian low-pass filter. Further, image registration is carried out, where local features are used with a Scale-Invariant Feature Transform (SIFT). A RAnd SAmple Consensus principle (RANSAC) was used to find the best homography. This is different from the previous framework, since it uses local features instead of control points. The scanner resolution was 1250 dpi, and scaling was performed using using bicubic interpolation.

Recently, Vans et al.[48] proposed a framework for defect detection. It uses a scanner for acquisition of the print. Image registration is performed using a binary version of the print, where regions are extracted to de-skew and align the printed image. The purpose of the framework was to detect defects, which is done with a modified version of SSIM.[26] The framework has the advantage of being able to do real-time defect detection.

3.2 A new framework based on control points

We modify and propose a framework similar to the one by Yanfang et al.,[46] where image registration is carried out using control points. Prior to printing these control points are added to the image. These points are black squares placed just outside the four corners of the image. Scanning is performed after printing, followed by assigning the scanner profile. The next step in the framework is to find the coordinates of the center of the control points, in both the original image and the scanned image. A simple automatic routine for finding the coordinates was used. The scanned image can be affected by several geometric distortions, such as translation, scaling, and rotation. Because of this, image registration must be carried out. The coordinates of the control points, found by the automatic routine, are used used to create a transformation for the registration. Since there are several options available for the registration and interpolation methods for the scaling, we have investigated these. The results showed that a simple transformation correcting for translation, rotation, and scaling and bilinear interpolation gave the smallest error.[6] After successful registration, a simple procedure is applied to remove the the control points. Finally, an IQ metric can be used to calculate the quality of the printed image. An overview of different stages of the framework is shown in Figure 7. This framework differs from the one from Eerola et al.,[47] not only in the registration method, but also in the descreening. In our modified framework we do not perform any direct descreening, but we leave this to the IQ metrics in order to avoid a double filtering of the image.

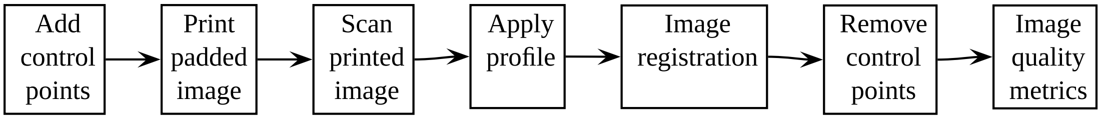


Figure 7. Overview of the proposed framework for using IQ metrics with printed images.

3.2.1 Comparison against another framework

The most important aspect to consider when selecting a framework for color prints is that the errors introduced are as small as possible. In order to test the performance of our framework we compared it against the framework by Eerola et al.[47] Three images with different characteristics were selected for the comparison, where they were rotated and scaled. The best framework should have the least difference between the original image and the registered image. The results show that the proposed framework introduces less error than the framework by Eerola et al.[47] The biggest difference is found in an image with uniform areas, which is a problem for frameworks based on local features since no registrations points can be found. In images with a lot of details the difference in error is small, but the proposed framework performs slightly better. Investigation of the computational time shows that the proposed method is more than 20 times faster.

4. IMAGE QUALITY METRICS FOR THE QUALITY ATTRIBUTES

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

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

Numerous IQ metrics have been proposed in the literature.[1] These take into account different aspects of IQ, and therefore a single metric might not be suitable to measure all CPQAs. Because of this we will first discuss the properties that the metrics should have for each CPQA, before a subset of the existing metrics is selected for further evaluation.

4.1.1 Sharpness

Sharpness is related to edges and details, and IQ metrics for sharpness should account for these two factors. Most of the metrics for sharpness found in the literature work by the principle to find edges in the image, calculating quality in a local region at the edges, and then combining the values from the local regions into an overall value representing sharpness quality.[49] However, these methods do not directly take into account details. Another approach, which better takes into account details, is based on structural similarity, where the metrics usually work on local neighbourhoods, such as the Structural SIMimilarity (SSIM) index.[26] There are also other metrics that are based on the visibility of salient objects to assess detail preservation, see for example Cao et al.[50] Previous analysis of sharpness indicates a relation to contrast,[3] therefore metrics accounting for local contrast could be suitable to assess sharpness.

4.1.2 Color

There are many potential IQ metrics for the color CPQA. First of all the metrics for this CPQA should account for color information, making all metrics based on color differences possible candidates. However, applying color difference formulas directly, such as the ΔE^*_{ab}, will most likely not predict perceived quality since the impression of quality is influenced by the viewing conditions. Therefore, the metrics should incorporate a simulation of the human visual system, such as the S-CIELAB[25] that performs a spatial filtering before applying the CIELAB color difference formula. These color difference formulas, which many IQ metrics use, usually consist of three parts, one for lightness and two for chromaticity. Since lightness and color are separate CPQAs, they should be evaluated separately. As a result the chromaticity part of these metrics will most likely be more suitable than using all three parts. Additionally, it has also been argued in the literature that some regions are more important than others. Therefore, metrics such as the Spatial Hue Angle MEtric (SHAME),[51] which use different weighting functions are potentially suitable metrics.

4.1.3 Lightness

Metrics for the lightness CPQA should mainly follow the same principles as for the color CPQA. IQ metrics based on color differences commonly calculates lightness quality separated from color, such as S-CIELAB.[25] Metrics working only on grayscale can also be suitable if they analyze lightness, such as SSIM[26] that performs a comparison of lightness between the original and the reproduction. However, metrics analyzing specific aspects in the lightness channel are most likely not suitable.

4.1.4 Contrast

The definition of the contrast CPQA states that contrast is both global and local, as well as dependent on lightness and chromaticity. Therefore metrics for the contrast CPQA should account for these aspects. Metrics computing contrast over a local neighbourhood for further combining the local values into a global value for contrast are potentially suitable metrics, such as the ΔLocal Contrast (ΔLC)[52] or SSIM.[26] Nevertheless, most contrast metrics do not account for chromaticity, and therefore they might not be optimal to measure contrast. One of the metrics that uses chromoticity to calculate contrast is the Weighted multi-Level Framework (WLF),[53] which also takes into account locality and globality.

4.1.5 Artifacts

The artifacts CPQA contains many different and specific QAs. Because of this it might be difficult to find an overall metric to evaluate all aspects of this CPQA. However, artifacts have some common denominators, if they are not detectable, they do not influence image quality. Therefore, the metrics used to measure artifacts should account for the sensitivity of the human visual system, such as the Visual Signal to Noise Ratio (VSNR)[54] that has a threshold for when artifacts are perceptible. The characteristics of artifacts is also an important issue, for noise metrics based on local contrast might be suitable, as ΔLC,[52] since noise affects local contrast. Artifacts like banding can be detected by metrics using edge-preserving filters, for example the Adaptive Bilateral Filter (ABF) metric,[55] opposite of metrics performing non-edge preserving filtering, as the S-CIELAB.[25]

4.1.6 Selected image quality metrics

We cannot evaluate all of the IQ metrics in the literature for all the CPQAs, and because of this a sub-set of metrics were selected, as shown in Table 1. The selection is based on the results from previous evaluations,[6,16] the criteria on which the metrics were created, the discussion above, and their popularity. Many of the metrics have been created to judge some aspects of IQ, and therefore only the suitable metrics for each CPQA will be evaluated. Furthermore, for specific CPQAs we also evaluate parts of the metrics. For example, S-CIELAB combines the lightness and color differences to obtain an overall value. When suitable, we will evaluate these separately in addition to the full metric.

Table 1. Selected IQ metrics for the evaluation of CPQAs.

CPQA Metric	Sharpness	Color	Lightness	Contrast	Artifacts
ABF[55]		X	X		X
Busyness[56]	X		X		
blurMetric[57]	X				
Cao[50]	X				X
CW-SSIM[58]	X		X	X	X
ΔLC[52]	X		X	X	X
LinLab[59]		X	X		X
MS-SSIM[60]	X		X	X	X
M-SVD[61]	X		X		X
RFSIM[62]	X		X	X	X
S-CIELAB[63]		X	X		X
S-DEE[64]		X	X		X
SHAME[51]		X	X		X
SSIM[26]	X		X	X	X
VSNR[54]	X		X		X
WLF[53]				X	X
YCXCzLab[65]		X	X		X

4.2 Evaluation of selected image quality metrics

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

4.2.1 Experimental setup

Two experimental phases were carried out for the evaluation of the metrics. 15 naive observers participated in the first experimental phase, where they judged overall quality and the different CPQAs on a seven step scale for a set of images. In the second phase, four expert observers ranked the quality of a set of images and elaborated on different quality issues. We will give a brief introduction of the experimental setup, for more information see Pedersen et al.[5,7]

We selected the same 25 images as in Figure 2, identically processed and printed. Two sets of images were printed at the same time, one set for first phase and one set for the second phase.

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

4.2.2 Evaluation of image quality metrics

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

Since both experimental phases were carried out under mixed illumination, the CIECAM02 chromatic adaptation transform[66] was used to ensure consistency in the calculations for the metrics. The measured reference white point of the monitor and the media were used as input to the adaptation transform, following the CIE guidelines.[66]

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

Phase 1: naive observers In the first experimental phase each observer judged overall quality and the five CPQAs for each image, which enabled us to compute z-scores[67] for each of these. For the first phase 24 of the 25 images (Figure 2) were used in the experiment.

To assess the performance of the evaluated metrics we have adopted several different methods. In order to achieve extensive evaluation we will investigate the performance of the IQ metrics both image by image, and the overall performance over the entire set of images. The Pearson correlation[68] is used for the image-wise evaluation, comparing the calculated quality and observed quality. The mean of the correlation for each image in the dataset and the percentage of images with a correlation above 0.6 is used as a measure of performance. While for the overall performance, we will use the rank order method,[69] where the correlation between the z-scores from the observers and the z-scores of the metric is the indication of performance. However, for the rank order correlation one should consider that we only have three data points, and therefore it is also important to perform a visual comparison of the z-scores.

Sharpness: The results of the selected metrics for the sharpness CPQA is shown in Table 2. SSIM has a mean correlation of 0.29, but it has a correlation above 0.6 in 50% of the images. The rank order method used to evaluate the overall performance calculates z-scores for the metric, which can be compared against the z-scores from the observers. A metric capable of correctly measuring the CPQA will have z-scores similar to the z-scores from the observers. The correlation between the z-scores is used as a performance measure, and SSIM shows an excellent correlation (1.00) with a low p-value (0.03). Visual investigation of the z-scores from SSIM and the observers shows a striking resemblance, therefore SSIM seems to be a suitable metric for the sharpness CPQA. Other versions of the SSIM, as the Multi Scale-SSIM (MS-SSIM) and Complex Wavelet-SSIM (CW-SSIM) increases the performance. Since these account better for the viewing conditions they might be more robust than the single scale SSIM. Other metrics as the ΔLC and the Riesz-transform based Feature SIMilarity metric (RFSIM) also perform very well for this CPQA. We can also see from Table 2 that the metrics perform similar in terms of rank order correlation, and that for an overall evaluation of sharpness these metrics produce similar results.

Table 2. Performance of the metrics for the sharpness CPQA. Mean correlation implies that the correlation has been calculated for each image in the dataset, and then averaged over the 24 images. Percentage above 0.6 is the percentage of images where the correlation is higher than 0.6. The rank order correlation indicates the correlation between the metric's z-scores computed with the rank order method[69] and the observer's z-scores for the CPQA, in addition the p-value for the correlation is found in the parenthesis.

Metric	Mean correlation	Percentage above 0.6	Rank order correlation
CW-SSIM	0.36	63	1.00 (0.05)
ΔLC	0.26	50	0.97 (0.14)
MS-SSIM	0.29	58	0.97 (0.15)
RFSIM	0.34	63	0.99 (0.09)
SSIM	0.29	50	1.00 (0.03)

Color: None of the evaluated IQ metrics perform significantly better than the others, in terms of the mean correlation, percentage above 0.6 and rank order correlation (Table 3). The results here indicates that none of the evaluated metrics can accurately measure the color CPQA. It is interesting to notice that all of these metrics are based on color differences, which might indicate that to find a suitable metric for color one should look of metrics based another principle than color differences.

Table 3. Performance of the metrics for the color CPQA. For further explanation see Table 2.

Metric	Mean correlation	Percentage above 0.6	Rank order correlation
LINLAB	-0.25	17	-0.93 (0.24)
S-CIELAB	-0.29	13	-0.95 (0.19)
S-DEE	-0.34	13	-0.92 (0.25)
SHAME	-0.04	21	-0.25 (0.84)

Lightness: MS-SSIM shows the highest mean correlation for the evaluated IQ metrics (Table 4), it also has the highest percentage of images above 0.6 in correlation, and it has an excellent rank order correlation with a low p-value. However, the single scale SSIM also performs well. The other metrics in Table 4 also have a high rank order correlation, indicating that many metrics have an overall similarity to the observers rating.

Table 4. Performance of the metrics for the lightness CPQA. For further explanation see Table 2. The subscript $_{\text{Lightness}}$ indicates only the lightness part of the metric.

Metric	Mean correlation	Percentage above 0.6	Rank order correlation
ΔLC	0.31	50	0.94 (0.22)
MS-SSIM	0.50	63	0.99 (0.08)
S-CIELAB$_{\text{Lightness}}$	0.14	46	1.00 (0.01)
SSIM	0.32	58	1.00 (0.05)
VIF	0.34	54	0.99 (0.09)

Contrast: SSIM also performs well for this CPQA as seen in Table 5. It has a mean correlation of 0.32, 50% of the images have a correlation above 0.6, and the rank order correlation is fairly high. It is worth noticing that using just the contrast calculation in SSIM the number of images with a correlation above 0.6 is increased. Also by using MS-SSIM the number of images with a correlation above 0.6 is slightly increased compared to the single scale SSIM. ΔLC has the highest mean correlation, and in 58% of the images a correlation above 0.6, and the rank order correlation is high. It should be noticed that all metrics have fairly high p-values for the rank order correlation. The difference between the metrics are is small for all performance measures, and therefore the results do not give a clear indication of which metric that is the best.

Artifacts: The observers gave similar results for the different rendering intents in terms of artifacts, because of this it is a very demanding task for the IQ metrics. The evaluation shows that RFSIM has the most images with a correlation above 0.6 together with MS-SSIM, but MS-SSIM has the highest mean correlation (Table 6). MS-SSIM and WLF have the highest rank order correlation, but they do also have high p-values. Therefore the results do not give a clear indication of a suitable metric. One should also consider that the artifacts CPQA contains many different sub-QAs, therefore it could be difficult to find just one IQ metric to measure overall artifact quality, and several metrics might be required.

Table 5. Performance of the metrics for the contrast CPQA. The subscript $_{\text{Contrast}}$ indicates only the contrast part of the metric. For further explanation see Table 2.

Metric	Mean correlation	Percentage above 0.6	Rank order correlation
ΔLC	0.34	58	0.97 (0.17)
MS-SSIM	0.30	58	0.74 (0.47)
RFSIM	0.32	54	0.94 (0.22)
SSIM	0.32	50	0.86 (0.34)
$\text{SSIM}_{\text{Contrast}}$	0.32	54	0.84 (0.37)

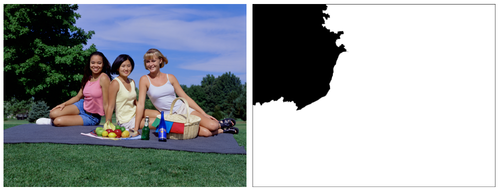
Table 6. Performance of the metrics for the artifacts CPQA. For further explanation see Table 2.

Metric	Mean correlation	Percentage above 0.6	Rank order correlation
MS-SSIM	0.23	46	0.61 (0.59)
RFSIM	0.14	46	0.26 (0.83)
SSIM	0.09	29	0.44 (0.71)
WLF	0.02	33	0.76 (0.45)

Phase 2: expert observers In the second experimental phase a group of expert observers commented on quality issues in the reproductions. A video of the experiment was used by the authors to extract regions were the observers perceived quality issues. This enabled us to perform an in-depth evaluation of the IQ metrics, which ensures that the metrics are capable of measuring the different CPQAs. We will only include the metrics that performed well in the first evaluation phase, since these are the ones most likely to be suitable for the CPQAs.

We will use the picnic image (Figure 8) to evaluate the IQ metrics. The observers indicated that this image contained a wide variety of QAs and different quality issues. These quality issues are the important issues for the IQ metrics to detect. Based on the comments from the observers important regions have been found, each containing different quality issues:

- Tree: mainly details, but also lightness and contrast issues.
- Shoe: loss of details perceived in one of the reproductions.
- White shirt: a hue shift in one of the reproductions.
- Hair: a hue shift in the hair of the asian girl in the middle.
- Pink shirt: one reproduction was too saturated.
- Grass: detail and saturation issues.
- Skin: a hue shift found in some reproductions.
- Cloth: a reproduction had a lighter red cloth than the others.
- Blanket: lightness issues.
- Sky: saturation and detail issues.



(a) Picnic image (b) Tree mask

Figure 8. The picnic image has been used to show the differences of the IQ metrics. On the right side one of the masks, where the mean value from the IQ metrics has been calculated within the black region.

To evaluate the IQ metrics we compare the rank of the metrics, based on the mean value of each region, to the rank of the observers for each region. A mask for each region was created based on the comments from the observers (example shown in Figure 8(b)), and this mask was used to obtain the ranking from the metrics. The observers did not rank all reproductions for all regions or quality issues, but instead they indicated which one was the best or the worst. We consider it to be important for the IQ metrics to predict which reproduction that is clearly better or worse. In addition to the ranking of the metrics, a visual inspection of the quality maps from each IQ metric has been carried out by the authors. This visual inspection will reveal more information about the performance of the metrics than the mean value.

SSIM: For the first experimental phase SSIM had a high performance for the sharpness, lightness, and contrast CPQAs, and fairly well for the artifact CPQA. In the second phase SSIM was able to detect the correct order regarding details, and gives results similar to the observers, as seen from the tree, grass, and shoe regions in Table 7(a). The visual inspection supported this, and revealed that SSIM is able to detect even small loss of details. These findings corresponds well with the results from the first experimental phase where SSIM was on of the best performing metrics for the sharpness CPQA. SSIM also correctly detected an area with a hue shift (hair), since this area in addition had a lightness shift. In the cloth

Table 7. Ranking for the different regions in the image where observers commented on quality issues. P = perceptual rendering intent, R = relative colorimetric rendering intent, and B = relative colorimetric rendering intent with BPC. If (R,P) > B, then B was ranked as the worst, but the observers did not rank the two other reproductions. () for the metric side indicates that the mean values are not significantly different with a 95% confidence level. A mask was created based on the comments from the observers, and the mean of the results from the IQ metric was used a basis for the ranking.

(a) SSIM

Region	Observers	SSIM	Correct rank
Tree	P > R (B)	P > B > R	Yes
Shoe	P > R (B)	P > B > R	Yes
White shirt	P > B (R)	P > B > R	Yes
Hair	(P,B) > R	P > B > R	Yes
Pink shirt	(P,B) > R	P > B > R	Yes
Grass	P > (R,B)	P > B > R	Yes
Skin	R > B > P	P > B > R	No
Cloth	(B,R) > P	P > B > R	No
Blanket	(R,B) > P	P > B > R	No
Sky	P > (R,B)	P > B > R	Yes

(b) ΔLC

Region	Observers	ΔLC	Correct rank
Tree	P > R (B)	P > B > R	Yes
Shoe	P > R (B)	B > R> P	No
White shirt	P > B (R)	P > B > R	Yes
Hair	(P,B) > R	P > B (B,R)	No
Pink shirt	(P,B) > R	(B,P) > R	Yes
Grass	P > (R,B)	P >B > R	Yes
Skin	R > B > P	P > B > R	No
Cloth	(B,R) > P	P > B > R	No
Blanket	(R,B) > P	P > B > R	No
Sky	P > (R,B)	P > B > R	Yes

region, where lightness differences were perceived by the observers, SSIM gave the correct ranking. SSIM also gave the same ranking as the observers in the tree region, where lightness and contrast were used by the observers. This shows that SSIM can be suitable to measure both lightness and contrast, but further analysis is required to ensure that SSIM is able to measure these CPQAs. We also notice that SSIM gives similar ranking for all regions in the image.

ΔLC: In the first experimental phase ΔLC had one of the best performances for the sharpness CPQA. It is able to detect the detail issues in the grass and tree regions as seen in Table 7(b). This is also the reason why it performs quite well for the rank order correlation for the sharpness CPQA in the first experiment phase. ΔLC also gave good results for the lightness CPQA, but it does not predict the lightness issues commented on by the expert observers. For the contrast CPQA ΔLC correlated with the observers from the first experimental phase. The experts did not directly comment on contrast, but in the pink shirt one of the reproduction was too saturated, which is one of the aspects looked at when evaluating contrast. ΔLC correctly detects this issue, even though it does account for chromaticty. Additionally, it correctly detects the detail issues as discussed above, which is also one of the aspects of contrast. However, since ΔLC does not account for chromaticity, it might not be suitable to evaluate all aspects of the contrast CPQA.

MS-SSIM: MS-SSIM is one of the best performing metrics for the artifact CPQA. The analysis for the second experimental phase is difficult, since MS-SSIM is a multilevel approach it is difficult to compute a value for each region. However, investigation of the SSIM maps for the different levels reveals that MS-SSIM is able to detect banding in several of the levels. The reason for the higher performance for the artifact CPQA and other CPQAs compared to the single scale SSIM might be that the multi-scale version takes better into account the viewing conditions, due to the subsampling of the image. The analysis for SSIM above is also partly valid for MS-SSIM, since the first level of MS-SSIM is the same as SSIM.

CW-SSIM: CW-SSIM does not compute a map like SSIM and some of the other metrics. But it is based on some principles that makes it suitable for some aspects. Linear and uniform changes correspond to lightness and contrast differences to which CW-SSIM is not sensitive because the structure is not changed, this is the reason why CW-SSIM most likely does not work well for contrast and lightness. However, in the sharpness CPQA where details are lost in several regions, the structure is changed and therefore it is suitable for sharpness.

RFSIM: This metric does not produce a quality map either, and therefore it is difficult to assess its performance in the second experimental phase. However, RFSIM uses an edge map to detect key locations in the image. This results in changes that occur at edges are the main component for the quality calculation, which explains why RFSIM gives good results for sharpness, since sharpness is related to the definition of edges.[29] This would also make RFSIM suitable for certain artifacts, such as banding, and could also explain why it is one of the best metrics for the artifacts CPQA. However, the artifacts CPQA contains many different artifacts, and therefore one cannot conclude that RFSIM is the best metric to measure the artifacts CPQA.

5. CONCLUSION

Our long term goal is to be able to measure image quality without human observers, more specifically using image quality metrics. Our approach to reach this goal is through quality attributes. First we presented a set of six quality attributes (sharpness, color, lightness, contrast, artifacts, and physical) for the evaluation of color prints, originated from the existing literature and with the intention of being used with image quality metrics. These quality attributes have been experimentally validated as suitable and meaningful for the evaluation of color prints. Furthermore, we introduced a framework for using image quality metrics with color prints. As the next step, we selected suitable image quality metrics for each quality attribute. These image quality are then evaluated against the percept for each attribute, before a final selection was made. For the sharpness quality attribute the Structural SIMilarity (SSIM) based metrics are suitable. None of the evaluated metrics can predict perceived color quality. SSIM based metrics also perform well for the lightness CPQA, which is also the case for the contrast CPQA. While inconclusive results are found for the artifacts attribute.

Future work includes additional evaluation of the metrics for the quality attributes, but also how to combine the values from the image quality metrics to obtain one quality value representing overall image quality.

6. ACKNOWLEDGEMENTS

The author hereof has been enabled by Océ-Technologies B.V. to perform research activities which underlies this document. This document has been written in a personal capacity. Océ-Technologies B.V. disclaims any liability for the correctness of the data, considerations and conclusions contained in this document.

REFERENCES

[1] Pedersen, M. and Hardeberg, J., "Survey of full-reference image quality metrics," Høgskolen i Gjøviks rapportserie 5, The Norwegian Color Research Laboratory (Gjøvik University College) (Jun 2009). ISSN: 1890-520X.

[2] Wyszecki, G. and Styles, W., [*Color Science, Concepts and Methods, Quantitative Data and Formulae*], Wiley Interscience, Derby, UK, 2nd ed. (Aug 2000).

[3] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F., "Attributes of image quality for color prints," *Journal of Electronic Imaging* **19**, 011016–1–13 (Jan 2010).

[4] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F., "Attributes of a new image quality model for color prints," in [*Color Imaging Conference*], 204–209, IS&T, Albuquerque, NM, USA (Nov 2009).

[5] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F., "Validation of quality attributes for evaluation of color prints," in [*Color and Imaging Conference*], 74–79, IS&T/SID, San Antonio, TX, USA (Nov 2010).

[6] Pedersen, M. and Amirshahi, S., "A modified framework the evaluation of color prints using image quality metrics," in [*5th European Conference on Colour in Graphics, Imaging, and Vision (CGIV)*], 75–82, IS&T, Joensuu, Finland (Jun. 2010).

[7] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F., "Estimating print quality attributes by image quality metrics," in [*Color and Imaging Conference*], 68–73, IS&T/SID, San Antonio, TX, USA (Nov 2010).

[8] Lindberg, S., *Perceptual determinants of print quality*, PhD thesis, Stockholm University (2004).

[9] Norberg, O., Westin, P., Lindberg, S., Klaman, M., and Eidenvall, L., "A comparison of print quality between digital, offset and flexographic printing presses performed on different paper qualities," in [*International Conference on Digital Production Printing and Industrial Applications*], 380–385, IS&Ts (May 2001).

[10] Gast, G. and Tse, M.-K., "A report on a subjective print quality survey conducted at NIP16," in [*NIP17: International conference on Digital Printing Technologies*], 723–727 (Oct 2001).

[11] Bouzit, S. and MacDonald, L., "Colour difference metrics and image sharpness," in [*Color Imaging Conference*], 262–267, IS&T/SID (2000).

[12] Nilsson, F. and Kruse, B., "Objective quality measures of halftones images," in [*IS&T's NIP 13: International Conference on Digital Printing Conference*], 353–357 (Mar 1997).

[13] Cui, C., Cao, D., and Love, S., "Measuring visual threshold of inkjet banding," in [*Image Processing, Image Quality, Image Capture, Systems Conference (PICS)*], 84–89, IS&T, Montreal, Quebec, Canada (2001).

[14] Fedorovskaya, E. A., Blommaert, F., and de Ridder, H., "Perceptual quality of color images of natural scenes transformed in CIELUV color space," in [*Color Imaging Conference*], 37–40, IS&T/SID (1993).

[15] Bonnier, N., Schmitt, F., Brettel, H., and Berche, S., "Evaluation of spatial gamut mapping algorithms," in [*Color Imaging Conference*], **14**, 56–61, IS&T/SID (Nov 2006).

[16] Hardeberg, J., Bando, E., and Pedersen, M., "Evaluating colour image difference metrics for gamut-mapped images," *Coloration Technology* **124**, 243–253 (Aug 2008).

[17] Morovic, J. and Sun, P., "Visual differences in colour reproduction and their colorimetric correlates," in [*Color Imaging Conference*], 292–297, IS&T/SID, Scottsdale, AZ (2002).

[18] Dalal, E. N., Rasmussen, D. R., Nakaya, F., Crean, P. A., and Sato, M., "Evaluating the overall image quality of hardcopy output," in [*Image Processing, Image Quality, Image Capture, Systems Conference*], 169–173, IS&T, Portland, OR (May 1998).

[19] Sawyer, J., "Effect of graininess and sharpness on perceived print quality," in [*Photographic Image Quality Symposium*], 222–231, Royal Photographic Society (Sep 1980).

[20] Bartleson, C., "The combined influence of sharpness and graininess on the quality of color prints," *J. Photogr. Sci.* **30**, 33–38 (1982).

[21] Natale-Hoffman, K., Dalal, E., Rasmussen, R., and Sato, M., "Effect of selected image quality on overall preference," in [*Image Processing, Image Quality, Image Capture, Systems Conference (PICS)*], 266–269 (Apr 1999).

[22] Wang, Z. and Shang, X., "Spatial pooling strategies for perceptial image quality assessment," in [*IEEE International Conference on Image Processing*], 2945–2948 (Oct 2006).

[23] Engeldrum, P. G., "Image quality modeling: Where are we?," in [*Image Processing, Image Quality, Image Capture, Systems Conference (PICS)*], 251–255, IS&T (1999).

[24] Keelan, B. W., [*Handbook of Image Quality: Characterization and Prediction*], Marcel Dekker, New York (2002).

[25] Zhang, X. and Wandell, B., "A spatial extension of CIELAB for digital color image reproduction," in [*Soc. Inform. Display 96 Digest*], *Proc. Soc. Inform. Display 96 Digest* , 731–734 (1996).

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

[27] Commission Internationale de l'Eclairage, "Colorimetry, 2nd ed. publication CIE 15.2, Bureau Central de la CIE," (1986).

[28] Klaman, M., *Aspects on Colour Rendering, Colour Prediction and Colour Control in Printed Media*, PhD thesis, Stockholm University (2002).

[29] Caviedes, J. and Oberti, F., "A new sharpness metric based on local kurtosis, edge and energy information," *Signal Processing: Image Communication* **19**, 147–161 (2004).

[30] Topfer, K., Keelan, B., O'Dell, S., and Cookingham, R., "Preference in image quality modelling," in [*Image Processing, Image Quality, Image Capture, Systems Conference (PICS)*], 60–64, IS&T, Apr (Portland, OR 2002).

[31] Grunbaum, B., "The search for symmetric Venn diagrams," *Geombinatorics* **8**, 104 – 109 (1999).

[32] CIE, "Guidelines for the evaluation of gamut mapping algorithms," Tech. Rep. ISBN: 3-901-906-26-6, CIE TC8-03 (156:2004).

[33] ISO, "ISO 20462-3 photography - psychophysical experimental methods to estimate image quality - part 2: Quality ruler method," (jul 2004).

[34] MapTube. http://maptube.org/ (2010).

[35] European Space Agency. www.esa.int (2010).

[36] Google 3D Warehouse. `http://sketchup.google.com/3dwarehouse/` (2010).

[37] Halonen, R., Nuutinen, M., Asikainen, R., and Oittinen, P., "Development and measurement of the goodness of test images for visual print quality evaluation," in [*Image Quality and System Performance VII*], Farnand, S. P. and Gaykema, F., eds., **7529**, 752909–1–10, SPIE, San Jose, CA, USA (Jan 2010).

[38] Sharma, A., "Measuring the quality of ICC profiles and color management software," *The Seybold Report* **4**, 10–16 (Jan 2005).

[39] Janssen, T., *Computational Image Quality*, PhD thesis, Technische Universiteit Eindhoven (1999).

[40] Yendrikhovskij, S., *Color reproduction and the naturalness constraint*, PhD thesis, Technische Universiteit Eindhoven (1998).

[41] Zuffi, S., Scala, P., Brambilla, C., and Beretta, G., "Web-based versus controlled environment psychophysics experiments," in [*Image Quality and System Performance IV*], Cui, L. C. and Miyake, Y., eds., *SPIE proceedings* **6494**, 649407 (Jan 2007).

[42] Knoblauch, K., Arditi, A., and Szlyk, J., "Effects of chromatic and luminance contrast on reading," *Journal of the Optical Society of America A* **8**, 428–439 (Feb 1991).

[43] Legge, G. E., Parish, D. H., Luebker, A., and Wurm, L. H., "Psychophysics of reading. XI. Comparing color contrast and luminance contrast," *Journal of the Optical Society of America A* **7**, 2002–2010 (Oct 1990).

[44] Capra, A., Castrorina, A., Corchs, S., Gasparini, F., and Schettini, R., "Dynamic range optimization by local contrast correction and histogram image analysis," in [*International Conference on Consumer Electronics (ICCE '06)*], 309 – 310 (Jan 2006).

[45] Zhang, X., Silverstein, D., Farrell, J., and Wandell, B., "Color image quality metric S-CIELAB and its application on halftone texture visibility," in [*COMPCON97 Digest of Papers*], 44–48, IEEE, Washington, DC, USA (1997).

[46] Yanfang, X., Yu, W., and Ming, Z., "Color reproduction quality metric on printing images based on the s-cielab model," in [*2008 International Conference on Computer Science and Software Engineering*], 294–297 (2008).

[47] Eerola, T., Kamarainen, J.-K., Lensu, L., and Kalviainen, H., "Framework for applying full reference digital image quality measures to printed images," in [*Scandinavian Conference on Image Analysis*], Salberg, A.-B., Hardeberg, J. Y., and Jenssen, R., eds., *Lecture Notes in Computer Science* **5575**, 99–108, Springer Berlin / Heidelberg, Oslo, Norway (June 2009).

[48] Vans, M., Schein, S., Staelin, C., Kisilev, P., Simske, S., Dagan, R., and Harush, S., "Automatic visual inspection and defect detection on variable data prints," Tech. Rep. HPL-2008-163R1, HP Laboratories (June 2010).

[49] Katajamaki, J. and Saarelma, H., "Objective quality potential measures of natural color images," *Journal of Imaging Science and Technology* **42**, 250–263 (may/june 1998).

[50] Cao, G., Pedersen, M., and Baranczuk, Z., "Saliency models as gamut-mapping artifact detectors," in [*5th European Conference on Colour in Graphics, Imaging, and Vision (CGIV)*], 437–443, IS&T, Joensuu, Finland (Jun 2010).

[51] Pedersen, M. and Hardeberg, J. Y., "A new spatial hue angle metric for perceptual image difference," in [*Computational Color Imaging*], *Lecture Notes in Computer Science* **5646**, 81–90, Springer Berlin / Heidelberg, Saint Etienne, France (Mar 2009). ISBN: 978-3-642-03264-6.

[52] Baranczuk, Z., Zolliker, P., and Giesen, J., "Image quality measures for evaluating gamut mapping," in [*Color Imaging Conference*], 21–26, IS&T/SID, Albuquerque, NM, USA (Nov 2009).

[53] Simone, G., Pedersen, M., Hardeberg, J. Y., and Rizzi, A., "Measuring perceptual contrast in a multilevel framework," in [*Human Vision and Electronic Imaging XIV*], Rogowitz, B. E. and Pappas, T. N., eds., **7240**, SPIE (Jan 2009).

[54] Chandler, D. and Hemami, S., "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing* **16**, 2284–2298 (Sep 2007).

[55] Wang, Z. and Hardeberg, J. Y., "An adaptive bilateral filter for predicting color image difference," in [*Color Imaging Conference*], 27–31, IS&T/SID, Albuquerque, NM, USA (Nov 2009).

[56] Orfanidou, M., Triantaphillidou, S., and Allen, E., "Predicting image quality using a modular image difference model," in [*Image Quality and System Performance V*], Farnand, S. P. and Gaykema, F., eds., *SPIE Proceedings* **6808**, 68080F–68080F–12, SPIE/IS&T, San Jose, USA (Jan 2008).

[57] Crete, F., Dolmiere, T., Ladret, P., and Nicolas, M., "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., *Proceedings of SPIE* **6492**, 64920I (Mar. 2007).

[58] Wang, Z. and Simoncelli, E., "Translation insensitive image similarity in complex wavelet domain," in [*IEEE international conference on acoustics, speech and signal processing*], *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on* **2**, 573–576 (March 2005).

[59] Kolpatzik, B. and Bouman, C., "Optimized error diffusion for high-quality image display," *Journal of Electronic Imaging* **1**, 277–292 (Jul 1992).

[60] Wang, Z., Simoncelli, E. P., and Bovik, A. C., "Multi-scale structural similarity for image quality assessment," in [*Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers*], 1398–1402 (Nov 2003).

[61] Shnayderman, A., Gusev, A., and Eskicioglu, A. M., "An SVD-based grayscale image quality measure for local and global assessment," *IEEE Transactions On Image Processing* **15**(2), 422–429 (2006).

[62] Zhang, L., Zhang, L., and Mou, X., "RFSIM: A feature based image quality assessment metric using riesz transforms," in [*Internatonal Conference on Image Processing*], 321–324 (Sep 2010).

[63] Zhang, X., Farrell, J., and Wandell, B., "Application of a spatial extension to CIELAB," in [*Very high resolution and quality imaging II*], *SPIE proceedings* **3025**, 154–157 (Feb 1997).

[64] Simone, G., Oleari, C., and Farup, I., “Performance of the euclidean color-difference formula in log-compressed OSA-UCS space applied to modified-image-difference metrics,” in [*11th Congress of the International Colour Association (AIC)*], (Oct 2009).

[65] Kolpatzik, B. and Bouman, C., “Optimal universal color palette design for error diffusion,” *Journal of Electronic Imaging* **4**, 131–143 (Apr 1995).

[66] CIE, “Chromatic adaptation under mixed illumination condition when comparing softcopy and hardcopy images,” Tech. Rep. ISBN: 3-901-906-34-7, CIE TC8-04 (162:2004).

[67] Engeldrum, P. G., [*Psychometric Scaling, a toolkit for imaging systems development*], Imcotek Press Winchester USA (2000).

[68] Kendall, M. G., Stuart, A., and Ord, J. K., [*Kendall’s Advanced Theory of Statistics: Classical inference and relationship*], vol. 2, A Hodder Arnold Publication, 5 ed. (1991).

[69] Pedersen, M. and Hardeberg, J. Y., “Rank order and image difference metrics,” in [*4th European Conference on Colour in Graphics, Imaging, and Vision (CGIV)*], 120–125, IS&T, Terrassa, Spain (Jun 2008).