Estimating Print Quality Attributes by Image Quality Metrics

Marius Pedersen*, Nicolas Bonnier**, Jon Y. Hardeberg* and Fritz Albregtsen.***
**** Gjøvik University College (Norway). ** Océ Print Logic Technologies S.A. (France). *** University of Oslo (Norway).***

Abstract

Image quality assessment is a difficult and complex task due to its subjectivity and dimensionality. Attempts have been made to make image quality assessment more objective, such as the introduction of image quality metrics. However, it has been proven difficult to create an image quality metric correlated with perceived overall image quality. Because of this, and to reduce the dimensionality, quality attributes have been proposed to help linking subjective and objective image quality.

Recently, Pedersen et al. (CIC, 2009) proposed a set of meaningful quality attributes for the evaluation of color prints with the intention of being used with image quality metrics. In this paper we evaluate image quality metrics for the quality attributes, and propose a set of suitable image quality metrics for each attribute. The experimental results indicate that the Structural SIMilarity index (SSIM) by Wang et al. (2004) is the most suitable metric for measuring the sharpness quality attribute. For the other quality attributes the results are not as conclusive.

Introduction

The printing industry is continuously moving forward as new products are introduced to the market. These products are becoming more and more affordable, and the technology is constantly improved. The need to assess the quality is also increased, for example to verify that new technology advancements produce higher quality prints than the current technology.

There are two main methods to assess Image Quality (IQ), subjective and objective. Subjective assessment is carried out by human observers. Objective assessment does not involve human observers, but rather measurement devices to obtain numerical values, or alternatively IQ metrics. These IQ metrics are usually developed to take into account the human visual system, and thus with the goal of being correlated with subjective assessment.

Numerous IQ metrics have been proposed [1], but so far no one has succeeded proposing an IQ metric fully correlated with subjective IQ [2–5]. Mostly because IQ is multi-dimensional and very complex. To reduce the complexity and dimensionality, Quality Attributes (QAs) have been used in the assessment of IQ. These QAs are terms of perception [6], such as sharpness and saturation. In earlier papers [7, 8] we proposed a set of six QAs for the evaluation of color prints:

- **Color** contains aspects, such as hue, saturation, and color rendition, except lightness.
- **Lightness** will range from ”light” to ”dark”.
- **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 and definition of edges.
- **Artifacts**, like noise, contouring, and banding, contribute to degrading the quality of an image if detectable.
- The **physical** QA contains all physical parameters that affect quality, such as paper properties and gloss.

These QAs are referred to as the Color Printing Quality Attributes (CPQAs). We have created the CPQAs to help establishing a link between subjective and objective evaluation.

Our long term goal is to evaluate quality without involving human observers. In order to achieve this, with the starting point of CPQAs, we need to identify IQ metrics able to correctly measure each CPQA. Therefore, in this paper we investigate and evaluate IQ metrics in the context of CPQAs, with the goal of proposing suitable metrics for each of the CPQAs.

To achieve our goal the first step is to identify relevant IQ metrics for each of the CPQAs. Then an experiment is set up to evaluate each of the CPQAs, where both naive and expert observers are included to assure an extensive evaluation. Later, the results from the relevant metrics identified in the first step are compared against the results of the two observer groups. This enables us to refine the selection of IQ metrics for each CPQA, and to recommend a suitable set of IQ metrics able to measure each of the CPQAs.

This paper is organized as follows: First we select the relevant metrics for the different CPQAs. Then the experimental setup is explained, and the printed images are prepared for the IQ metrics. We then evaluate the metrics before we conclude and propose future work.

Selection of Image Quality Metrics for the Color Printing Quality Attributes

Numerous IQ metrics have been proposed in the literature [1], and we have selected a sub-set of these, as shown in Table 1. The selection is based on the results from previous evaluation [2–4], the criteria on which the metrics were created, and their popularity. Since many of the IQ metrics are not created to evaluate all aspects of IQ, 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.

Experimental setup

In this paper, two experimental phases were carried out. In the first phase, 15 naive observers judged overall quality and the different CPQAs on a set of images. In the second phase, four expert observers judged 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. [18].

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

Metric \ CPQA	Sharpness	Color	Lightness	Contrast	Artifacts
ABF [9]		X	X		X
ΔLC [10]	X		X	X	X
Cao [11]	X				X
S-CIELAB [12]		X	X		X
S-DEE [13]		X	X		X
SHAME [14]		X	X		X
SSIM [15]	X		X	X	X
VSNR [16]	X		X		X
WLF [17]				X	X

Images

For the experiment we selected 25 images (Figure 1), which were chosen based on several image characteristics, such as lightness, saturation, and details. The images were 150 dpi uncompressed 16-bit sRGB tiff files.

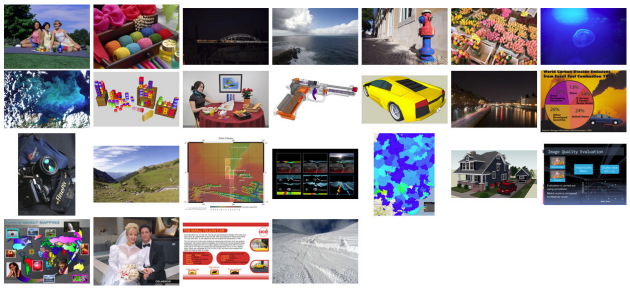

Figure 1. *The 25 images used in the experiment.*

Color workflow

The images were printed on an Océ Colorwave 600 CMYK wide format printer on Oce Red Label (LFM054) plain uncoated paper using three different rendering intents: perceptual, relative colorimetric, and relative colorimetric with Black Point Compensation (BPC). Two sets of images were printed at the same time, one set for first phase and one set for the second phase.

Viewing Conditions

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, at a color temperature of 6500K and a white luminance level of 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, an illuminance level of 450 $\pm$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. The experiment followed the CIE guidelines [19] as closely as possible.

Instructions

The instructions given to the observers focused both on the overall quality rating of the reproduction and on the QAs.

Phase 1: Naive observers

The naive observers were given the following instructions:
Judge the reproductions according to overall quality and five quality attributes (color, lightness, contrast, sharpness, and artifacts).
A description of the CPQAs, similar to the one above, were given to the observers along with the instructions. The rating of overall quality and the CPQAs was carried out as a category judgment experiment, and a seven step scale was provided to observers to assist them in their judgment.

Phase 2: Expert observers

The expert observers were given the following instructions:
Rank the reproductions according to quality.
- Elaborate on the attributes you use and quality issues you observe, i.e. all attributes you consider.
- If possible try to give an indication of the importance of the issues and attributes, and important areas.
The entire experiment was filmed, and the observers were encouraged to describe and talk about their observations.

Preparing the printed images for image quality metrics

In order to apply IQ metrics to the printed images from the experiment, they need to be digitized since the IQ metrics require a digital input. To perform this we have adopted the framework by Pedersen and Amirshahi [4]. The first step is to the scan the printed images. An Epson 10000XL was used for scanning the images for the first phase, and a HP ScanJet G4050 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. The advantage of this is that we can compare the correctness of the profiles by comparing it to the measured values of the printed test target. It also ensures that the scanner profiles are based on the same media as the printed images.

To evaluate the quality of the obtained profiles we have adopted the method by Sharma [20]. First the test target was scanned, and then used to generate a profile for the scanner. The reference data was the measured target used to generate the printer profile. The scanned test target was then opened in Adobe Photoshop CS3, and the generated scanner profile was assigned to the image. Afterwards the image was converted to CIELAB using 'convert to profile' using absolute colorimetric and the color matching module was Adobe (ACE).

A script was written to calculate the mean CIELAB values for each patch in the scanned image. The color difference between the measured CIELAB values and scanned CIELAB values was calculated to get a measure of the correctness of the profile. The ΔE^*_{ab} has been used as a measure for the differences between them. For the Epson 10000XL a mean ΔE^*_{ab} of 2.06 was found, and for the HP ScanJet G4050 a mean ΔE^*_{ab} of 1.63. These values are only slightly higher than the ones found by Sharma [20], and therefore considered to be acceptable.

Since both experimental phases were carried out under mixed illumination, the CIECAM02 chromatic adaptation transform [21] 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 [21].

Evaluation of Image Quality Metrics

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 this phase each observer judged overall quality and the five CPQAs for each image, which enabled us to compute z-scores for each of these. For the first phase 24 of the 25 images (Figure 1) were used in the experiment.

For calculating the performance of the IQ metrics we have adopted several different methods. We will investigate the performance of the IQ metrics both image by image, and the overall performance over the entire set of images. For the image wise evaluation, the Pearson correlation [22] between the calculated quality and the observed quality is used. 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 [23], 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

Table 2 shows the results of the evaluation for the IQ metrics selected for the sharpness CPQA. SSIM performs quite well with a correlation above 0.6 in 50% of the images, but the mean correlation over the 24 images is only 0.29. 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 a performance measure, and SSIM shows an excellent correlation (1.00) with a low p-value (0.03). Investigation of the z-scores from SSIM and the observers shows a striking resemblance as seen in Figure 2, therefore SSIM seems to be the best metric for the sharpness CPQA.

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 [23] 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
CAO	-0.15	17	0.99 (0.07)
ΔLC	-0.04	21	0.84 (0.37)
SSIM	0.29	50	1.00 (0.03)
VSNR	0.04	33	-0.68 (0.52)

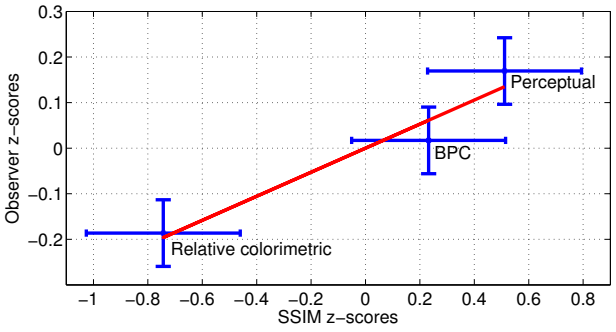

Figure 2. *The rank order z-scores for SSIM plotted against the observer z-scores for the sharpness CPQA, both with a 95% confidence interval. The red line showing the linear regression between the z-scores indicates that SSIM is a suitable metric to measure the sharpness CPQA.*

Color

For the color CPQA none of the evaluated IQ metrics stand out in terms of the mean correlation or percentage above 0.6 (Table 3). For the rank order correlation some better results are found, indicating that the ranking for each image on an overall basis agrees to a certain extent with the naive observers. This applies especially for S-DEE$_{\text{Color}}$ and SHAME, but these also have a high p-value. An investigation of the z-scores, similar to what was done for SSIM regarding sharpness in Figure 2, reveals that SHAME gives a similar ranking of the z-scores, while S-DEE does not. It is also interesting to notice that SHAME$_{\text{Color}}$ has a low performance, which indicates that the lightness differences contribute to SHAME's high performance. The results for the color CPQA does not give any concrete results on which metric that should be used to measure it, but for an overall indication of color quality SHAME might be adequate.

Table 3: Performance of the metrics for the color CPQA.For further explanation see Table 2. The subscript $_{\text{Color}}$ indicates that only the color part of the metric has been evaluated.

Metric	Mean correlation	Percentage above 0.6	Rank order correlation
ABF	-0.06	25	0.28 (0.82)
ABF$_{\text{Color}}$	0.02	21	0.28 (0.82)
S-CIELAB	-0.11	21	0.59(0.60)
S-CIELAB$_{\text{Color}}$	-0.01	33	0.43 (0.72)
S-DEE	-0.04	21	0.72 (0.49)
S-DEE$_{\text{Color}}$	0.04	25	0.84 (0.36)
SHAME	0.00	25	0.84(0.36)
SHAME$_{\text{Color}}$	-0.09	17	0.28 (0.82)

Lightness

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. The rank order correlation is, as for the sharpness CPQA, very high together with a low p-value. The reason for this is that the rating of lightness by the observers is very similar to the rating of sharpness, and therefore SSIM has a similar performance. Some metrics, such as SHAME and S-CIELAB$_{\text{Lightness}}$ have a fairly high percentage of images above 0.6 in correlation, however, they have a low correlation for the

rank order method. This indicates that they are not able to correctly rank the reproductions as the observers, but that they have a similar frequency distribution as the observers' z-scores.

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
ABF	-0.20	25	-0.94 (0.22)
$\text{ABF}_{\text{Lightness}}$	-0.03	33	-0.94 (0.22)
ΔLC	0.14	33	0.89 (0.30)
S-CIELAB	-0.20	17	-0.77 (0.44)
$\text{S-CIELAB}_{\text{Lightness}}$	0.03	38	-0.77 (0.44)
S-DEE	-0.22	13	-0.65 (0.55)
$\text{S-DEE}_{\text{Lightness}}$	-0.05	29	-0.83 (0.38)
SHAME	0.01	38	-0.49 (0.68)
SSIM	0.32	58	0.99 (0.05)
$\text{SSIM}_{\text{Lightness}}$	-0.22	33	-0.98 (0.12)
VSNR	-0.18	21	-0.76 (0.45)

Contrast

SSIM also performs quite well for the contrast CPQA. From Table 5 we can see that it has a mean correlation of 0.32, but for half the images in the dataset it has a correlation above 0.6. The rank order method confirms a good performance by indicating a similar ranking by the metric as by the observers. It is also interesting to notice that by extracting the contrast calculation in SSIM we are able to slightly increase the percentage of images above 0.6 in correlation. The WLF, calculated as the difference in contrast between the original and reproduction, has the same performance as $\text{SSIM}_{\text{Contrast}}$, but with a higher rank order correlation. Visual inspection of the z-scores shows that SSIM provides a more correct z-score with larger differences between the rendering intents, even though WLF has a higher correlation.

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	0.27	38	0.48 (0.68)
SSIM	0.32	50	0.86 (0.34)
$\text{SSIM}_{\text{Contrast}}$	0.32	54	0.84 (0.37)
WLF	0.32	54	0.97 (0.17)

Artifacts

The results from the observers for the artifacts CPQA showed small differences between the different rendering intents, making it a very difficult task for the IQ metrics. The best metric was ΔLC (Table 6). The rank order correlation gives a high correlation, but also a fairly high p-value. The visual inspection of the z-scores show that ΔLC correctly ranks the rendering intents as the observers, but it is difficult to give a conclusive result due to the small visual differences.

It is also worth noticing that the Cao metric performs second best in terms of mean correlation and rank order correlation, being specifically designed for artifacts.

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
ABF	0.01	25	-0.18 (0.88)
Cao	0.15	38	0.50 (0.67)
ΔLC	0.25	42	0.84 (0.37)
S-CIELAB	-0.05	25	0.16 (0.90)
S-DEE	-0.01	17	0.33 (0.78)
SHAME	-0.03	33	0.51 (0.66)
SSIM	0.09	29	0.44 (0.71)
VSNR	-0.20	25	-0.94 (0.22)
WLF	0.03	21	0.18 (0.88)

Phase 2: Expert Observers

During the second experimental phase observers stated and pointed out regions where different quality issues were perceived. Since the experiment was filmed, it enabled the authors to go through the video and mark the issues and regions found by the observers. 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 3) to show how the metrics perform regarding the different CPQAs. 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:

(a) Picnic image (b) Tree mask

Figure 3. *The picnic image has been used to show the differences of the IQ metrics. On the right side one of the masks used to evaluate the IQ metrics, where the mean has been calculated within the black region*

- Tree: in this region the observers commented mainly on details, but also on lightness and contrast.
- 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.

We have taken the approach of comparing the rank of the IQ metrics to the rank of the observers in these regions. A mask for each region was created (Figure 3(b)), and the mean value for the IQ metrics was used to calculate the rank. 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 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 authors. This visual inspection will reveal more information about the performance of the metrics than the mean value would.

SSIM

SSIM performed quite well for the sharpness, lightness, and contrast CPQA, and fairly well for the artifact CPQA in the first phase. In the second phase SSIM was able to detect the correct order regarding detail preservation, and corresponds well with the results from the observers, as seen from the tree, grass, and shoe regions in Table 7. The visual inspection of the quality map revealed that SSIM is able to detect even small loss of details. These results support the findings in the first phase, where SSIM performed well for the sharpness CPQAs. SSIM also correctly detected an area with a hue shift (hair), since this area in addition had a lightness shift. In the cloth 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. One should also notice that SSIM gave the same ranking for all regions in the image.

Table 7: Ranking from SSIM 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.

Region	Observers	SSIM	Correct ranking
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

SHAME

SHAME produced the best results for the color CPQA in the first experimental phase. However, since this metric does not produce a value for each pixel, it is difficult to evaluate against the results of the experts. Nevertheless, SHAME uses a weighting map based on hue angle, where the more pixels of the same hue, the higher importance they are given. The highest weights are given to the blue and green regions of the image. The sky region, containing most of the blue hue pixels, is given the highest importance. It is also a region where observers noticed saturation issues. The grass and tree regions are most likely what caused SHAME's higher performance compared to the other metrics, since these also have been commented on by the observers.

The reason for SHAME's varying results for the different images in the dataset most like stems from the weighting based on the hue angles. In some images large uniform areas are found, but these are not necessarily the regions-of-interest for the observers. In these cases the large uniform areas are given a high weight, while smaller regions, as the cloth and hair, are given a lower weight. However, these smaller region might draw attention and have high importance in the quality evaluation, and thereby resulting in a disagreement with the observers. Because of this SHAME should be used with care when measuring the color CPQA.

ΔLC

ΔLC has the best performance for the artifact CPQA. The expert observers did not specifically comment on artifacts, but in one of the reproductions banding is visible in some regions. The ΔLC is not capable of detecting this, while other metrics such as the ABF that has an edge preserving filter is able to detect this artifact. However, the ΔLC is able to detect the detail issues in the shoe, grass, and tree region as seen in Table 8. This is also the reason why it performs quite well for the rank order correlation for the sharpness CPQA in the first experiment phase. The evaluation based on the second phase does not confirm whether ΔLC is suitable to measure the artifact CPQA, but it does show that artifacts vary, and that it might be necessary to have different metrics for different sub-artifacts.

Table 8: Ranking from ΔLC for regions in the image where observers commented on quality issues. See Table 7 for more information.

Region	Observers	ΔLC	Correct ranking
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

Conclusion and Future Work

In this study we evaluated a group of IQ metrics against the perceptual results from an experiment, with the intention of proposing suitable IQ metrics for a set of five CPQAs. The analysis carried out show that the SSIM from Wang et al. [15] is the most suitable IQ metric for the sharpness CPQA. This metric also performs well for the lightness and contrast, but further investigation of these CPQAs should be carried out to ensure that SSIM is the most suitable metric. For the color CPQA, SHAME from Pedersen and Hardeberg [14] shows good results, but further eval-

uation is required for this CPQA as well. For the artifact CPQA the results are inconclusive because it contains many different artifacts.

Future work will include additional evaluation to ensure that the correct metrics are used in the evaluation of the CPQA, especially for the contrast, lightness, and artifact CPQA. When quality values for each CPQA are obtained, the question of how to combine them into a single number representing overall quality constitutes an important future step. The results for the color CPQA also reveal the importance of the characteristics of the image. How to use these characteristics to select IQ metrics is also an interesting possibility for future work.

Acknowledgments

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] M. Pedersen and J.Y. Hardeberg. 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] S. A. Ajagamelle, M. Pedersen, and G. Simone. Analysis of the difference of gaussians model in image difference metrics. In *5th European Conference on Colour in Graphics, Imaging, and Vision (CGIV)*, pages 489–496, Joensuu, Finland, Jun 2010. IS&T.

[3] J.Y. Hardeberg, E. Bando, and M. Pedersen. Evaluating colour image difference metrics for gamut-mapped images. *Coloration Technology*, 124(4):243–253, Aug 2008.

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

[5] N. Bonnier, F. Schmitt, H. Brettel, and S. Berche. Evaluation of spatial gamut mapping algorithms. In *14th Color Imaging Conference*, volume 14, pages 56–61. IS&T/SID, Nov 2006.

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

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

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

[9] Z. Wang and J. Y. Hardeberg. An adaptive bilateral filter for predicting color image difference. In *Color Imaging Conference*, pages 27–31, Albuquerque, NM, USA, Nov 2009.

[10] Z. Baranczuk, P. Zolliker, and J. Giesen. Image quality measures for evaluating gamut mapping. In *Color Imaging Conference*, pages 21–26, Albuquerque, NM, USA, Nov 2009.

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

[12] X. Zhang, J.E. Farrell, and B.A. Wandell. Application of a spatial extension to CIELAB. In *Very high resolution and quality imaging II*, volume 3025 of *SPIE proceedings*, pages 154–157, San Jose, CA, USA, Feb 1997.

[13] G. Simone, C. Oleari, and I. Farup. 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)*, Sydney, Australia, Oct 2009.

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

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

[16] D.M. Chandler and S.S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, Sep 2007.

[17] G. Simone, M. Pedersen, J. Y. Hardeberg, and A. Rizzi. Measuring perceptual contrast in a multilevel framework. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging XIV*, volume 7240. SPIE, Jan 2009.

[18] M. Pedersen, N. Bonnier, J. Y. Hardeberg, and F. Albregtsen. Validation of quality attributes for evaluation of color prints. In *Color Imaging Conference*, San Antonio, TX, USA, Nov 2010.

[19] CIE. Guidelines for the evaluation of gamut mapping algorithms. Technical Report ISBN: 3-901-906-26-6, CIE TC8-03, 156:2004.

[20] A. Sharma. Measuring the quality of ICC profiles and color management software. *The Seybold Report*, 4(20):10–16, Jan 2005.

[21] CIE. Chromatic adaptation under mixed illumination condition when comparing softcopy and hardcopy images. Technical Report ISBN: 3-901-906-34-7, CIE TC8-04, 162:2004.

[22] M. G. Kendall, A. Stuart, and J. K. Ord. *Kendall's Advanced Theory of Statistics: Classical inference and relationship*, volume 2. A Hodder Arnold Publication, 5 edition, 1991.

[23] M. Pedersen and J. Y. Hardeberg. Rank order and image difference metrics. In *CGIV 2008 Fourth European Conference on Color in Graphics, Imaging and Vision*, pages 120–125, Terrassa, Spain, Jun 2008. IS&T.

Author Biography

Marius Pedersen *received his BsC in Computer Engineering in 2006, and MiT in Media Technology in 2007, both from Gjøvik University College, Norway. He is pursuing a PhD in Color Imaging, under the supervision of Pr. Hardeberg and Pr. Albregtsen, sponsored by Océ. He is also a member of the Norwegian Color Research Laboratory at Gjøvik University College. His work is centered on image quality metrics for color prints.*