Rank Order and Image Difference Metrics

Marius Pedersen [1] and Jon Yngve Hardeberg. Gjøvik University College, Gjøvik, Norway.

Abstract

There are a number of ways to reproduce an image, for an example gamut mapping, halftoning and compression. To find the best reproduction among a number of variants of the same reproduction algorithm, a psychophysical experiment can be carried out. Image difference metrics have been introduced to eliminate these experiments. To do this the metrics must reflect the perceived image difference. One way to evaluate the overall performance of image differnece metrics is to compute the correlation coefficient between perceived and predicted image difference. This does not always reflect the true performance of the metric, therefore we propose to use the ranking based on the predicted image difference for each scene as a data set for the rank order method. This results in a z-score similar to the overall perceived image difference, the correlation coefficient between metric z-score and perceived z-score reflects the overall performance of the image difference metrics.

Introduction

There are a number of algorithms involved in the to reproduction of an image, for instance gamut mapping, halftoning and compression. To identify the best reproduction among a number of variants of the same reproduction algorithm (e.g. JPEG compression), a psychophysical experiment can be carried out. This will result in a scale with the visual difference of the reproductions from the original. These psychophysical experiments are both time and resource demanding. Image difference metrics have been introduced to entirely or partially eliminate psychophysical experiments. One way to evaluate the overall performance of such metrics is to investigate the correlation between the visual difference and the predicted image difference for a set of test images. However, this kind of evaluation can reveal little about the performance of the image difference metrics, due to the nonlinearities in the human visual system.

The ultimate goal for image difference metrics is to be able to predict perceived image difference for all conditions and modifications. It has been shown that image difference metrics can predict perceived image difference for some scenes [1, 2], but the image difference metric should also be able to predict overall image difference. In gamut mapping the overall best gamut mapping algorithm (GMA) should be used on an image set. If the performance of the image difference metrics are reliable, and reflects their true performance, psychophysical experiments could entirely or partially eliminated.

Toet and Lucassen [3] use the ranking from the observers and ranking from their image fidelity metric to evaluate performance of the metric, this is done because we cannot expect a linear relation between the metric and perceived distortion due to the nonlinearities in the human visual system. Other researchers have also used the ranking either directly comparing metric ranking with subjective ranking or with statistical ranking methods [4, 5, 6].

One method adopted by some researchers for evaluation is the Spearman's rank order correlation [7, 8, 9]. This method assess the relationship between two variables, without making any assumption about their frequency distribution. This method will reduce the influence of extreme single observations, and can reflect more general relations.

We propose to rank the images according to their calculated distance from the original. This data is used in the rank order method, resulting in z-score directly comparable to observers z-score. We present a case based on a set of gamut mapped images, and show how ranking the results from image difference metrics reflect their performance.

State of the art of image difference metrics

In recent years, several attempts have been made to develop image difference metrics that correlate well with the perceived difference.

S-CIELAB [10] is a an extension of the CIELAB ΔE^*_{ab} metric, aiming to take into account the spatial-color sensitivity of the human eye. It was developed with two goals; to simulate the spatial blurring performed by the human visual system (HVS), and to be consistent with the basic CIELAB color difference for uniform patches. Thus it consists in transforming the image data into a perceptual opponent-color space, and blurring the channels with convolution kernels corresponding to the contrast sensitivity functions (CSF) of the HVS, before converting back to a CIEXYZ representation. Then a pixelwise color difference is calculated using the conventional CIELAB ΔE^*_{ab} equations. Practically, this means that the color difference at each pixel is weighted by the differences computed over a local neighborhood.

iCAM [11] is a framework for an image appearance model, which incorporates more sophisticated models of chromatic adaptation than S-CIELAB. It is based upon previous research in many fields such as uniform color spaces, hue linearity, the image surround importance, image difference and image quality measurement algorithms. The model uses von Kries chromatic adaptation identical to the one found in CIECAM02. The adapted signals are transformed into the IPT color space. The adapting and the surround luminance levels are taken into account, to allow for the prediction of various appearance phenomena.

The Structural Similarity Index (SSIM) proposed by Wang et al. [7] attempts to quantify the visibility of the difference between a reference and a distorted grayscale image. The algorithm defines the structural information in an image as those attributes that represent the structure of the objects in the scene, independently of the average luminance and contrast. The index is based on a combination of luminance comparison, contrast comparison and structure comparison. The comparison is done for local windows in the image and the overall image difference is computed as the mean of all these local windows.

Hong and Luo's hue angle metric [12] is based on the observed fact that systematic errors over the entire image is quite

[1] 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.

noticeable and unacceptable. Therefore the metric is constructed based on four conjectures, that pixels or areas of high significance can be identified and a suitable weight allocation can be found, larger areas of the same color should be weighted higher, larger color difference between the pixels should get higher weights, and finally that hue is an important color percept for discriminating colors within the context. The proposed algorithm creates a histogram based on the hue angle, and then sorts this ascending so that different weights can be applied to different sections of the histogram. The overall color difference is then calculated by multiplying the weighted hue angle for every pixel with the pixel-by-pixel color difference.

Psychophysical experiment

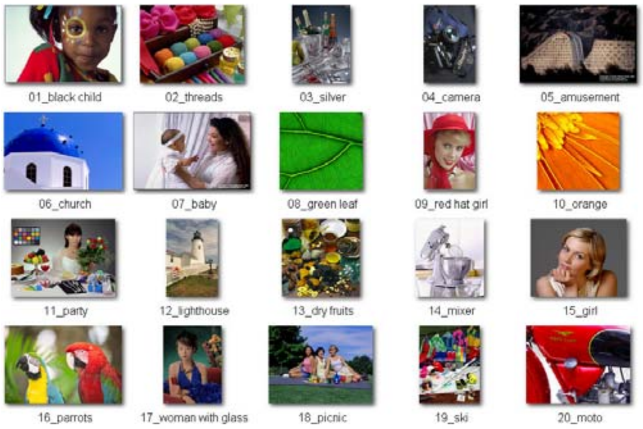


Figure 1. Images used in the experiment. They have various characteristics in terms of gamut, contrast, contents, details, etc.

20 different images (Figure 1) was chosen for a psychophysical experiment [13]. Considering the color gamut of a OCE TCS 500 printer, the amount of out-of-gamut colors ranged from 25% to 100% with a mean of 57%. The images were reproduced using 5 different gamut mapping algorithms.

- HPminDE (Hue preserving minimum ΔE^*_{ab} clipping) which is a baseline gamut mapping algorithm proposed by the CIE [14]. The algorithm does not change in-gamut colors at all, while out-of-gamut colors are mapped to the closest color on the destination gamut while preserving the hue.
- SGCK [14] is an advanced spatially invariant sequential gamut compression algorithm. The lightness is first compressed by a chroma dependent sigmoidal scaling, resulting in high chroma colors being compressed less than neutral ones. The resulting colors are then compressed along lines toward the cusp [15] of the destination gamut using a 90% knee scaling function. For the final compression the image gamut is used as the source gamut.
- Zolliker and Simon [16] proposed a spatial gamut mapping algorithm; its goal being to recover local contrast while preserving lightness, saturation and global contrast. A simple clipping is performed as first step; then by using an edge-preserving high pass filter the difference between the orginal and gamut clipped image is filtered. The filtered image is then added to the gamut clipped image. As a last step the image is clipped in-order to be in-gamut.
- Kolås and Farup [17] recently proposed a hue- and edge-preserving spatial color gamut mapping algorithm. The image is first gamut clipped along straight lines toward the center of the gamut. A relative compression map is then created from the orignal and clipped image. Using this compression map, a new image can be constructed as a linear convex combination of the original image and neutral gray image. This image is in turned filtered by a edge-preserving smoothing minimum filter. As the final step the gamut mapped image is constructed as a linear convex combination of the original image and neutral gray using the filtered map.
- Gatta et al. [18] proposed a multiscale algorithm preserving hue and local relationship between closely related pixel colours. First a scale-space representation of the image and then gamut clipping the lowest scale is constructed. The resulting gamut compression is then applied to the image at the next smallest scale. Operators are used to reduce the effect of haloing. The process is repeated until all scales are treated. The Fourier domain is used to speed up the process.

The 20 different images have been evaluated by 20 observers in a pair comparison experiment [13]. All observers had normal or corrected to normal color vision. The observers were presented with the original image in the middle of the screen, with 2 different reproductions on each side. The observers were asked to pick the image with the most accurate reproduction with respect to the original image. When the observer had picked one image, a new pair of reproductions was shown until all combinations were evaluated. All pairs were also shown twice in opposite order for consistency. The monitor was a Dell 2407WFP LCD display calibrated with a D65 white point and a 2.2 gamma. The viewing conditions were chosen as close to the ones described in the CIE guidelines [14] as possible. The level of ambient illumination was measured to approximately 20 lux. The observer was seated approximately 50 cm from the screen.

Z-scores

The Z-scores are based on Thurstone's law of comparative judgement [19, 20]. Data collected are transformed into interval scale data where scores represent the distance of a given image from the mean score of a set of images in the scene [21], and therefore being relative. The 95% confidence intervals are calculated in the same way as proposed by Morovic [21]. The error bars are then computed as

$$\bar{X} \pm \frac{\sigma}{\sqrt{N}}$$

where $\bar{X}$ is the Z-score, σ is the the standard deviation and N is the size of the sample set. For these experiments this is the number of observers multiplied with 2, because each image pair was shown twice for consistency. With this confidence interval there is a 95% estimate that the value will be within the interval, and if the confidence interval of another GMA is outside this interval the difference is significant.

Image difference metrics

5 image difference metrics have been choosen, ΔE^*_{ab}, S-CIELAB [10], iCAM [11], SSIM [7] and the hue angle algorithm [12]. All metrics except SSIM has a scale where closer to 0 indicate a reproduction closer to the original, while SSIM has a scale between 0 and 1, where 1 indicate an identical reproduction.

Results

From the pair comparison experiment z-scores were calculated, indicating the performance of the different GMAs. From Figure 2 we can see that the Gatta algorithm gets the highest

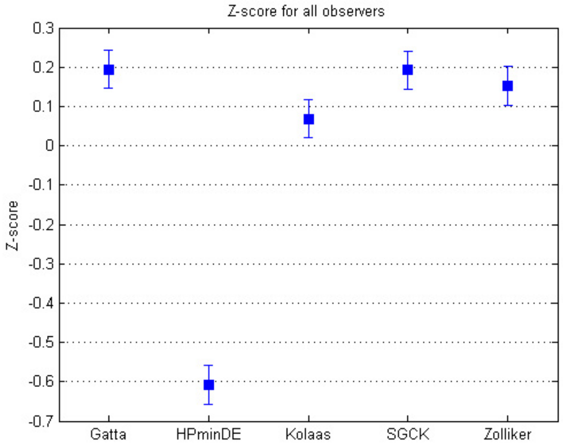


Figure 2. Z-score from pair comparison experiment I. Gatta cannot be differeniated from SGCK and Zolliker. Kolås is rated barely lower than the Gatta, SGCK but cannot be differeniated from Zolliker. HPminDE receives the lowest z-score from the observers.

score from the observers, indicating the lowest visual difference from the original. This algorithm gives statistically the same visual difference from the original as the SGCK and Zolliker algorithm. Kolås has the fourth best score, but has the same visual difference as Zolliker. HPminDE clearly gives the highest visual difference from the original, the low score here indicating a large consensus among the observers about the low performance of this algorithm.

Rank order based on metric order

Aiming to develop an universal image difference metric, this metric should work across mulitple scenes and in different conditions. One way of evaluating the performance of image difference metrics is to check the correlation between the perceived image difference and the calculated image difference [22, 23] as seen in Figure 3. We can see from Figure 3 that the data points are very spread, and there is very little correlation found. Thus it is not possible in this way to use the image difference results as a of evaluating the performance of the best gamut mapping algorithm. This is the case for all the metrics, where the correlation is generally low for all scenes as seen in Table 1 for both Pearson's correlation and Spearman's rank order correlation. The Spearman's rank order correlation does not take into account the frequency distribution of the variables, and should therefore be less sensitive to extreme values, but as seen in Table 1 this is not a good measure for overall prediction of performance. The z-score for each scene does not say anything about the difference between scenes, it is based on how preferred each image is within each scene. We propose to process the image difference metric results in the same way as we commonly do with the results from a rank order perceptual evaluation.

The rank for each metric in the 20 scenes have been used as a basis for the overall performance of the GMAs, this correspond to 20 observers in a ranking experiment. If the result from the metrics match the overall results from the psychophysical experiment (Figure 2), the metrics predict perceived image difference. This method will only provide information about the order of the image samples, not information of the distance between the samples.

In principle the rank order and pair comparison approaches provide the same information [24]. The rank order data has been used to generate corresponding pair comparison data [24, 25], and the z-scores were computed as for a pair comparison experiment [21, 20]. Babcock [26] got similar score for pair comparison, rank order and graphical rating, this implicate that scale values from one type of experiment can be directly compared to scale values from another type of experiment. The rank order z-scores in this experiment have been calculated by using the Colour Engineering Toolbox [27].

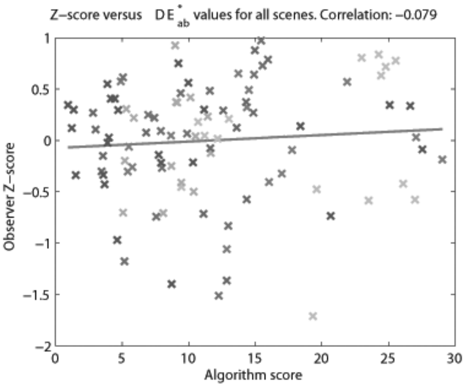


Figure 3. Z-score from observers against ΔE^*_{ab} values. The data points are very spread,and we get a very low correlation between the z-scores and ΔE^*_{ab} values. The HPminDE algorithm is rated as the best by ΔE^*_{ab} the opposite of the observers.

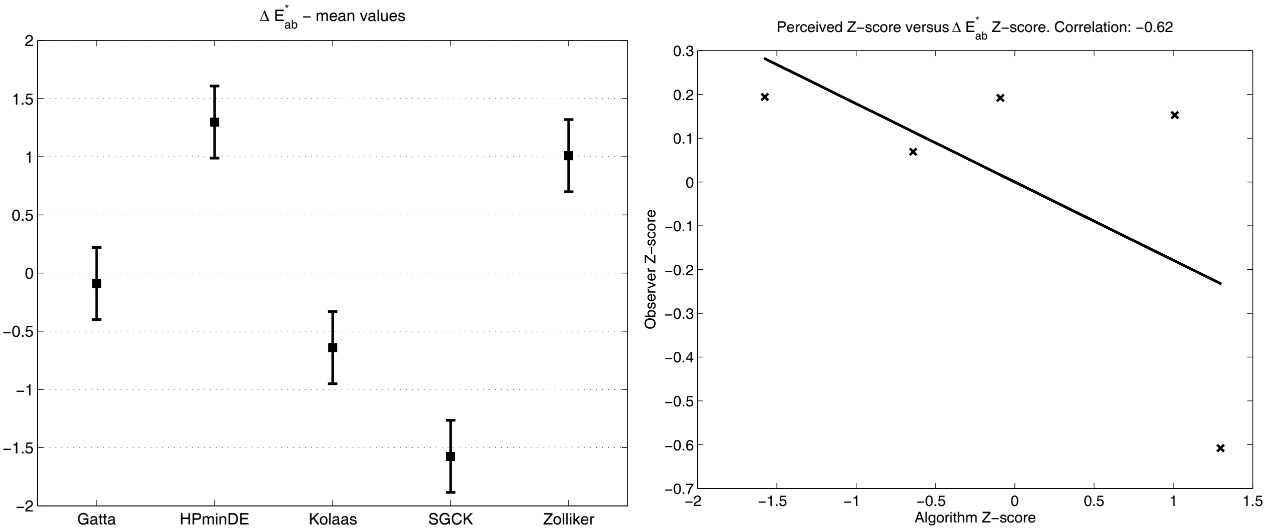
Table 1. Correlation between all z-scores and all algorithms. The correlation here is low for all metrics both for Pearson and Spearman. SSIM has the highest Pearon's correlation, but this is not good, indicating a low performance. For the Spearman's correlation the hue angle algorithm has the highest correlation, but still very low. The plot for ΔE^*_{ab} with a linear fitted line and calculated Pearson's correlation is found in Figure 3.

	Correlation	
Metric	Pearson	Spearman
ΔE^*_{ab}	-0.08	0.11
SSIM	0.16	0.05
S-CIELAB	-0.06	0.10
iCAM	0.01	0.07
Hue angle	-0.11	0.13

One disadvantage by using rank order is the number of scenes needed, in order to provide useful results the number of scenes must be high. The more scenes used, the more accurate results. The more data provided the more accurate the results will become, and the confidence intervals in the metric's z-score will become smaller. The confidence intervals can be used as a measure of when a reproduction method is significantly better than another, and this provides more information than a Pearson's or Spearman's rank order correlation calculation.

From Figure 4(a) we can see that results from the rank order for ΔE^*_{ab}. The HPminDE have the highest z-score, but this has the lowest z-score from the observers. This GMA will clip the color to the minimum ΔE^*_{ab} and will always be rated as the best by the ΔE^*_{ab} formulae. The SGCK gets a very low score in the ranking, while the observers rated this as one of the best gamut mapping algorithms. Figure 4(b) shows z-score from the observers plotted against the rank order z-score from ΔE^*_{ab}. The Pearson's correlation here is -0.62, indicating different scores for the observers and ΔE^*_{ab}.

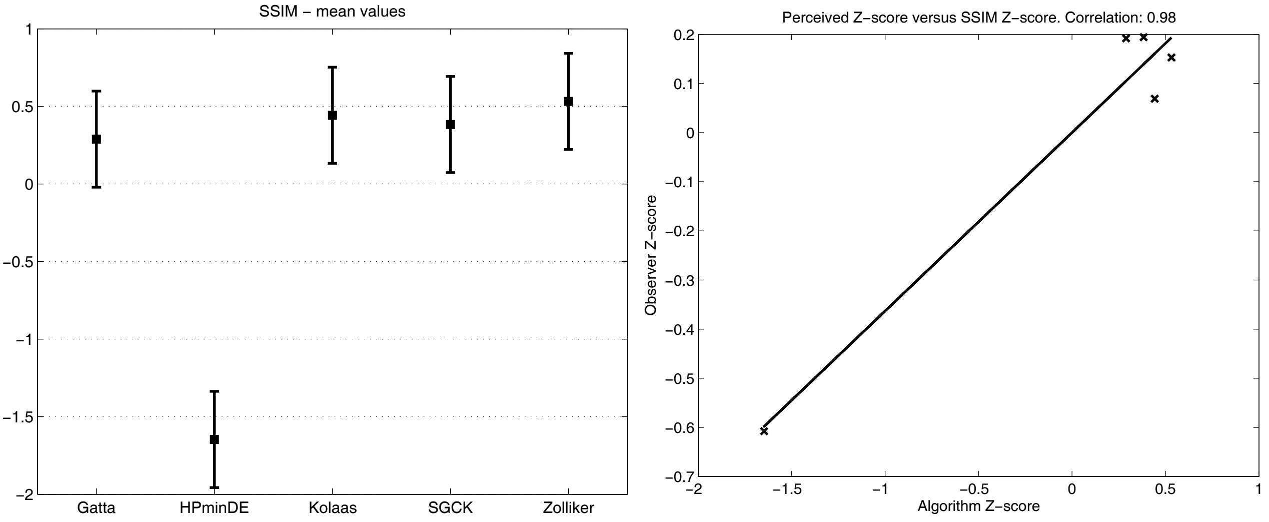
Figure 5(a) shows the results for SSIM, as we can see the HPminDE gets the lowest score by SSIM. This is the same as



(a) ΔE^*_{ab} rank order score.

(b) Observer z-score against ΔE^*_{ab} rank order score.

Figure 4. *Rank order score for ΔE^*_{ab} values, and these values plotted against z-scores from observers.*



(a) SSIM rank order score.

(b) Observer z-score against SSIM rank order score.

Figure 5. *Rank order score for SSIM values, and these values plotted against z-scores from observers. The results here indicate a very high performance by the SSIM, the correct ranking of the HPminDE is the main reason for this, but also the similar ranking of the remaining gamut mapping algorithms.*

the observers. The four other gamut mapping algorithms cannot be differentiated with a 95% confidence interval. The observers also have very small differences between these algorithms, the Kolås algorithm has a score just lower than the SGCK, Gatta and Zolliker. The score from SSIM is very similar to the score from the observers, this is also verified with a correlation between the scores of 0.98 (Figure 5(b)). The correct ranking of the HPminDE gamut mapping algorithm is the basis for the excellent correlation here, and this ranking (Figure 5(a)) also reflect the observers ranking (Figure 2). The overall Pearson's correlation of only 0.16 and Spearman's rank order correlation of 0.05 (Table 1) between the z-scores and SSIM scores, these are therefore not a good measure of overall performance, even though high correlation can be found within each scene in both measures. The correlation within a scene can be average, but the ranking can be correct as seen on Figure 6. Here the Spearman's correlation will perform well, while the Person's only perform average. In other cases the Spearman's correlation will perform average, while the Pearson's correlation will perform excellent. Spearman's rank order correlation can provide low correlation of clusters of data are found, but the ranking within the cluster is not necessarily correct. When the ranking is used in the rank order method, the normal distribution is taken into account and will therefore better handle extreme values. Because of this the ranking of the results within each scene and using these as a basis for the rank order z-scores, SSIM will reflect perceived image difference.

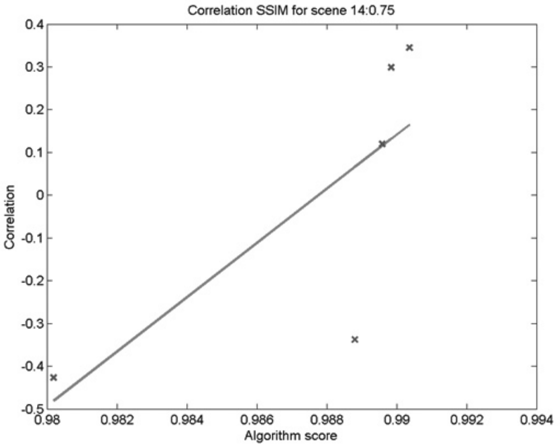


Figure 6. Correlation between z-score and calculated image difference for SSIM on scene 14. The visual difference between the images is not correctly predicted by SSIM, but the ranking is. Based on this the rank order method on the SSIM data will be a better measure.

The are large differences between the score from iCAM and the observers, mostly due to the calculated values for HPminDE in iCAM. iCAM has a spreading of the 4 best algorithms from the observers, and the HPminDE is rated as the best by iCAM opposite of the observer rating. The Pearson's correlation between the observer z-score and iCAM z-score is -0.68, resulting in a low performance by iCAM.

The HPminDE receives the highest score by S-CIELAB, both the SGCK and Kolås get low scores. The results for S-CIELAB is very similar to the ones found with ΔE^*_{ab}. This results in a low Pearson's correlation of -0.62.

The hue angle algorithm have similar results as ΔE^*_{ab} and S-CIELAB, this is not surprising due to the familiarity between these metrics. The ranking of the GMAs are the same with only minor differences in the z-score values, this results in almost the same Pearson's correlation between observer z-score and metric z-score as S-CIELAB, with -0.72.

Table 2. Rank indicate the pearson's correlation between rank order z-score and observer z-score. Mean Pearson indicate correlation between ranked metric score and observer z-score calculated as Pearson's correlation, where the correlation for each scene has been averaged. Mean Spearman is similar to Mean Pearson but for Spearman's rank order correlation.

	Correlation		
Metric	Rank Pearson	Mean Pearson	Mean Spearman
ΔE^*_{ab}	-0.62	0.28	0.17
SSIM	0.98	0.84	0.78
S-CIELAB	-0.62	0.28	0.20
iCAM	-0.68	-0.03	-0.07
Hue angle	-0.72	0.27	0.15

Overall observations

SSIM is the image difference metric with the best fit between the observers and the algorithm z-score (Table 2), indicating a good prediction of perceived image difference. All other metrics have a correlation below 0, indicating that these metrics do not predict perceived image difference well. In all metrics except SSIM the HPminDE gamut mapping algorithm has been miscalculated, i.e. given a too high rank by the metrics. These metrics are based on ΔE^*_{ab} and therefore the HPminDE will be given a high rank.

The results here indicate that a ranking of the algorithms value within each scene and using this ranking to calculate the rank order z-score gives a better prediction of perceived image difference than calculating the correlation between algorithm score and observer z-score from a pair comparison experiment.

By using the rank order method the results are also less sensitive to extreme values as long as a reasonable number of scenes are used.

Conclusions and perspectives

We propose a new way of uning image difference metrics to evaluate the performance of image reproduction algorithms, ranking the score from the metrics and calculating z-scores based on this. This gives an overall score for the performance, and we avoid the problem of scale differences between scenes and the results are less sensitive to single extreme values. The results show that image difference metrics can predict overall perceived image difference, when we look at the ranking and discard information about the distance between the reproductions. This method could be applied to different reproduction algorithms, such as compression, halftoning and gamut mapping.

It has been shown that image difference metrics still have problems with predicting perceived image difference, even though the ranking might be correct more work must be carried out to improve the metrics. A larger dataset with an increased number of image reproduction methods should carried for a better testing of the proposed method.

References

[1] E. Bando, J. Y. Hardeberg, and D. Connah. Can gamut mapping quality be predicted by color image difference formulae. In *Human Vision and Electronic Imaging X, ed. B. Rogowitz, T. Pappas, S. Daly, Proc. of SPIE - IST Electronic Imaging,SPIE*, volume 5666, pages 180 – 191, 2005.

[2] M. Pedersen, J. Y. Hardeberg, and P. Nussbaum. Using gaze information to improve image difference metrics. In B. Rogowitz and T. Pappas, editors, *Human Vision and Electronic Imaging VIII (HVEI-08)*, volume 6806 of *SPIE proceedings*, San Jose, USA, Jan 2008. SPIE.

[3] A. Toet and M.P. Lucassen. A new universal colour image fidelity metric. *Displays*, 24:197–204, 2003.

[4] N. Chaddha and T.H.Y. Meng. Psycho-visual based distortion measures for monochrome image and video compression. In *Proc. og the Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 841–845, 1993.

[5] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for blocking artifacts in images based on human visual sensitivity. *IEEE TIP*, 4:713–724, 1995.

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

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

[8] P. Le Callet and D. Barba. A robust quality metric for color image quality assessment. In *ICIP*, volume 1, pages 437–440, 2003.

[9] M. Carnec, P. Le Callet, and D. Barba. Full reference and reduced reference metrics for image quality assessment. In *Seventh International Symposium on Signal Processing and Its Applications*, volume 1, pages 477– 480, Jul 2003.

[10] X. Zhang and B. A. Wandell. A spatial extension of CIELAB for digital color image reproduction. In *Soc. Inform. Display 96 Digest*, pages 731–734, San Diego, 1996.

[11] M. D. Fairchild and G. M. Johnson. The iCAM framework for image appearance, image differences, and image quality. *Journal of Electronic Imaging*, 13:126–138, 2004.

[12] G. Hong and M.R. Luo. Perceptually based colour difference for complex images. In R. Chung and A. Rodrigues, editors, *Proceedings of SPIE: 9th Congress of the International Colour Association*, volume 4421, pages 618–621, 2002.

[13] Fabienne Dugay. Perceptual evaluation of colour gamut mapping algorithms. Master's thesis, Gjøvik University College, 2007.

[14] CIE. Guidelines for the evaluation of gamut mapping algorithms. Technical Report ISBN: 3-901-906-26-6, CIE TC8-08, 156:2004.

[15] J. Morovic and M.R. Luo. The fundamentals of gamut mapping: A survey. *Journal of Imaging Science and Technology*, 45(3):283–290, 2001.

[16] P. Zolliker and K. Simon. Adding local contrast to global gamut mapping algorithms. In *CGIV 2006 Final program and Proceedings, Society for Imaging Science and Technology*, 2006.

[17] Ø. Kolås and I. Farup. Efficient hue-preserving and edge-preserving spatial gamut mapping. In *Fifteenth Color Imaging Conference*, 2007.

[18] I. Farup, C. Gatta, and A. Rizzi. A multiscale framework for spatial gamut mapping. *IEEE Transactions on Image Processing*, 16(10):2423–2435, 2007.

[19] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:237, 1927.

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

[21] J. Morovic. *To Develop a Universal Gamut Mapping Algorithm*. PhD thesis, University of Derby, 1998.

[22] E. Bando, J. Y. Hardeberg, D. Connah, and I. Farup. Prediction visible image degradation by colour image difference fomulae. In *The 5th International Conference on Imaging Science and Hardcopy*, pages 121–124, 2004.

[23] J. Morovic and P. Sun. Predicting image differences in color reproduction from their colorimetric correlates. *The Journal of imaging science and technology*, 47:509–516, 2003.

[24] C. Cui. Comparision of two psychophysical methods for image color quality measurements: Paired comparision and rank order. In *IS&T/SID Eighth Color Imaging Conference*, pages 222–227, 2000.

[25] P. J. Green. *Gamut mapping and appearance models in colour management*. PhD thesis, University of Derby, UK., 2003.

[26] J. S. Babcock, J. B. Pelz, and M. D. Fairchild. Eye tracking observers during rank order, paired comparison, and graphical rating tasks. In *Image Processing, Image Quality, Image Capture Systems Conference*, 2003.

[27] P. Green and L. MacDonald, editors. *Colour Engineering: Achieving Device Independent Colour*. John Wiley & Sons, 2002.

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 currently pursuing a PhD in Color Imaging. 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.*

Jon Yngve Hardeberg *is a Professor of Color Imaging at Gjøvik University College. He received his Ph.D from Ecole Nationale Suprieure des Télécommunications in Paris, France in 1999,with a dissertation on colour image acquisition and reproduction, using both colorimetric and multispectral approaches. He has more than 10 years experience with industrial and academic colour imaging research and development, and has coauthored over 100 research papers within the field. His research interests include various topics of colour imaging science and technology, such as device characterisation, gamut visualisation and mapping, image quality, and multispectral image acquisition and reproduction. He is a member of IS&T, SPIE, and the Norwegian representative to CIE Division 8. He has been with Gjøvik University College since 2001 and is currently head of the Norwegian Color Research Laboratory.*