Perceptual Evaluation of Color Gamut Mapping Algorithms

Fabienne Dugay, Ivar Farup,* Jon Y. Hardeberg

The Norwegian Color Research Laboratory, Gjøvik University College, Gjøvik, Norway

Received 29 June 2007; revised 7 May 2008; accepted 28 May 2008

Abstract: The recommendation of the CIE has been followed as closely as possible to evaluate the accuracy of five color gamut mapping algorithms (GMAs)—two nonspatial and three spatial algorithms—by psychophysical experiments with 20 test images, 20 observers, one test done on paper and a second one on display. Even though the results do not show any overall "winner," one GMA is definitely perceived as not accurate. The importance of a high number of test images to obtain robust evaluation is underlined by the high variability of the results depending on the test images. Significant correlations between the percentage of out-of-gamut pixels, the number of distinguishable pairs of GMAs, and the perceived difficulty to distinguish them have been found. The type of observers is also important. The experts, who prefer a spatial GMA, show a stronger consensus and look especially for a good rendering of details, whereas the nonexperts hardly make a difference between the GMAs. © 2008 Wiley Periodicals, Inc. Col Res Appl, 33, 470–476, 2008; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/col.20443

Key words: gamut mapping; *perceptual evaluation*; *psychophysical tests*; *gamut mapping algorithms*; *color reproduction*

INTRODUCTION

With the increased use of cross-media publishing, color gamut mapping has become an area of intensive research and development. The CIE[1] and Morovic[2] presented a survey of research on gamut mapping up to 2000 and Farup *et al.*[3] completed it with a review of some spatial gamut mapping algorithms (GMAs). To evaluate the performance of GMAs and to allow further comparisons, the CIE Technical Committee 8-03[4] has proposed guidelines on how to implement such tests. Evaluations with selected spatial and nonspatial GMAs have previously been done,[5–9] giving various results. Most of these evaluations compare GMAs by using pair comparison on screen with a small number of images. Morovic's paper highlights the variability of the results within different test images. For this reason, a relatively large number of images (20) have been used in this article. None of the cited papers compare pair comparison on screen and ranking on printed images to see whether there is a good correlation or not in the results. We analyze this possible difference. Moreover, we analyze the possible difference of preference of GMA between groups of observers (experts and nonexperts), as Bonnier *et al.*[8] point out that for some GMAs the opinion differs. The purpose of this article is thus to evaluate three recently developed spatial GMAs and two nonspatial in order to find out if one performs better than the others. The influence of the observers, the test images, and the paper versus display experiment are discussed. First, the experimental details are described, and then results are presented and discussed.

EXPERIMENTAL METHODS

In this section, we present the experimental setup of the evaluation in accordance with the CIE guidelines.[4]

Algorithms

According to the CIE guidelines, the following two standard (i.e., nonspatial) GMAs have to be included in the experiment.

Hue Preserving Minimum ΔE_{ab} Clipping (HPminDE).[4] This is a simple baseline algorithm that does not change in-gamut colors at all, whereas out-of-gamut colors are mapped to the closest color on the destination gamut boundary in a plane of constant hue.

SGCK.[4] This is an advanced spatially invariant sequential gamut compression algorithm. First, the lightness is

*Correspondence to: Ivar Farup (e-mail: ivar.farup@hig.no).

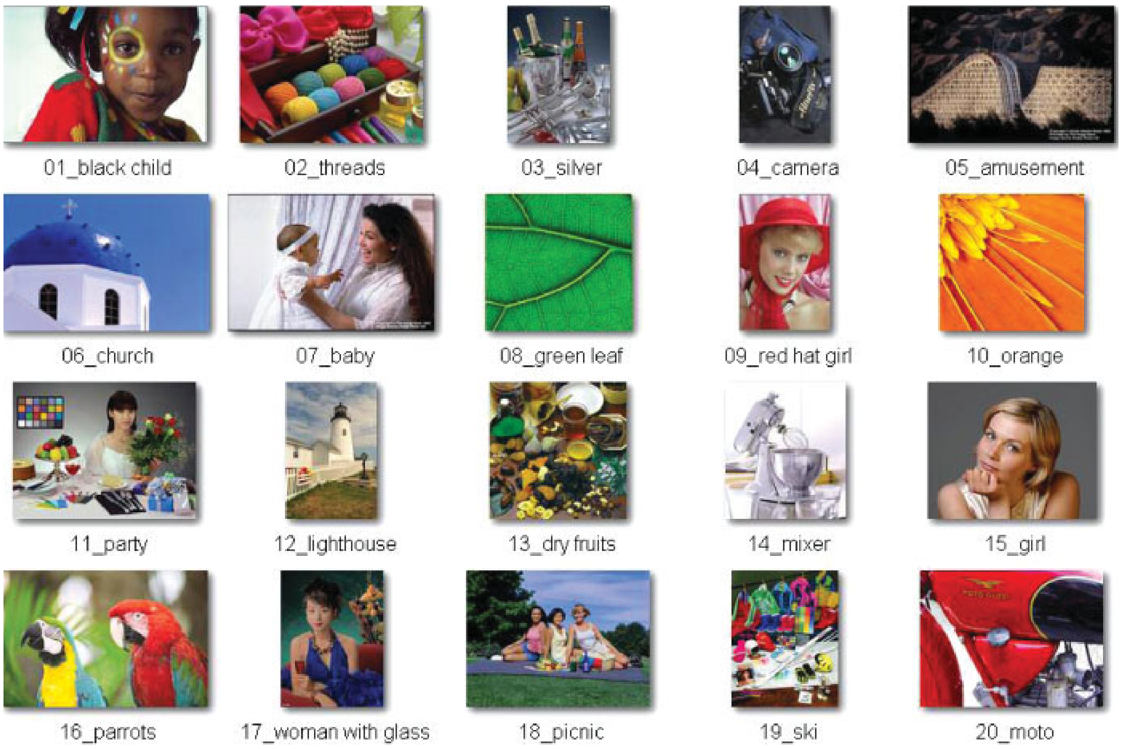

FIG. 1. The 20 test images used.

compressed using a chroma-dependent sigmoidal scaling that compresses high-chroma colors less than neutral ones. Then, the resulting colors are compressed along lines toward the cusp[2] of the destination gamut using a 90% knee scaling function. The image gamut is used as the source gamut for the final compression.

Additionally, we tested the following three recently developed spatial GMAs.

Zolliker.[5,6] This is a spatial GMA whose main goal is to recover local contrast while preserving lightness, saturation, and global contrast. First, simple gamut clipping is performed. Then, the difference between the original and the gamut-clipped images is filtered using an edge-preserving high-pass filter derived from a bilateral filter.[10] This filtered image is then added to the gamut-clipped image, resulting in an image that is mainly in-gamut and still contains most of the high-frequency information. Finally, the image is gamut clipped in order to be in-gamut. As the high-pass filtering is performed for the three color channels independently, the hue can be changed as a result of the process.

Kolås.[11] This is a new efficient hue- and edge-preserving spatial color gamut mapping algorithm. First, the image is gamut clipped along straight lines toward the center of the gamut. From the original and the clipped images, a relative compression map is constructed. Using this map, the gamut-clipped image can be constructed as a linear convex combination of the original image and neutral gray. The map is filtered using an edge-preserving decreasing filter, derived from the SNN filter.[12] Finally, the gamut-mapped image is constructed as a linear convex combination of the original image and neutral gray using the filtered map. Thus, no hues are changed.

Gatta.[2] This is a multiscale algorithm that preserves hue and local relationship between closely related pixel colors. It works by first constructing a scale-space representation of the image and then gamut clipping the lowest scale. The resulting gamut compression is then applied to the image at the next smallest scale. Various operators operating in the range are introduced to reduce haloing effects. The process is iterated until all scales are treated. To speed up the process, the filtering is performed in the Fourier domain. However, the algorithm is still $O(N(\log N)^2)$ and thus quite time consuming for large images.

For all of the GMAs, the gamut boundary is determined using the modified convex hull algorithm[13,14] with $\gamma = 0.2$, in the CIELAB color space.

Psychophysical Tests

Two methods of psychophysical tests have been chosen. For the first experiment with the printed reproductions, the rank order method was used. The five reproductions are compared simultaneously with the original displayed on a monitor. The observer is asked to rank the images from the least to the most accurate to the original. The observers were asked to mark the region(s) of the image that were the most important for their choice, and also to tell which images were difficult to distinguish. For practical reasons, this method cannot be used for the on-screen experiment, thereby the pair comparison

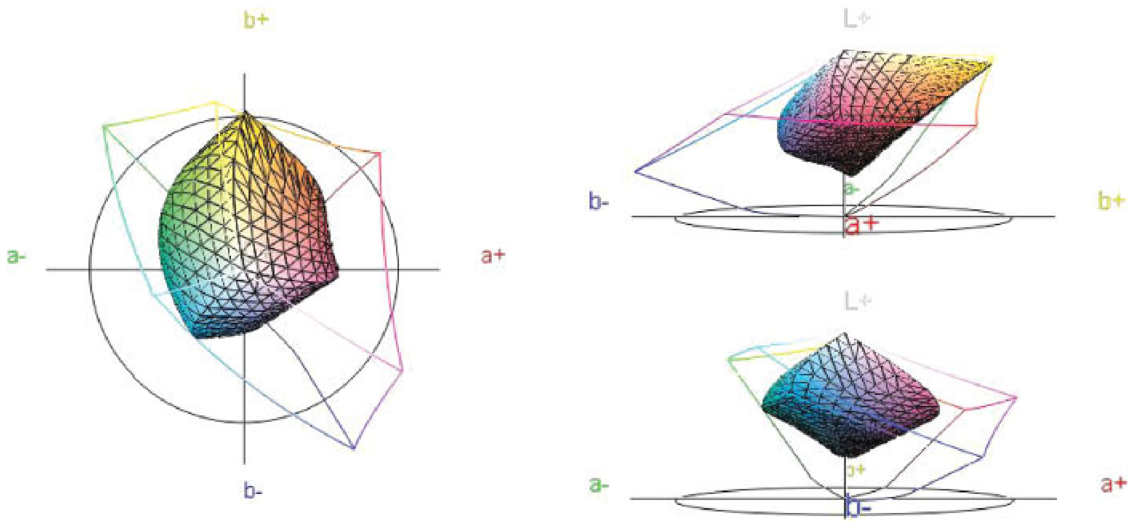

FIG. 2. The Océ printer gamut on plain paper (solid) and the sRGB gamut (wireframe) shown in the CIELAB color space.

method was used. The observer is presented with the original image along with pairs of candidate gamut-mapped images and he is asked to pick the most accurate reproduction with respect to the original image. All pairs are presented twice to avoid systematic error due to some persons who might prefer one side to the other when the images seem indistinguishable.

Images

Twenty test images including the obligatory ski image[4] are used (Fig. 1). They have various characteristics in terms of gamut, contrast, contents, details, etc. Three images are from the ISO 12640-2 standard, five from the Kodak PhotoCD, five from the ECI Visual Print Reference, and two from a local photographer.

Media

An Océ printer, the OCE TCS 500, with Océ standard paper is used. A CMYK profile was made using Profilemaker from GretagMacbeth, and the random ECI2002 CMYK test chart. The monitor where the original was displayed is a NEC SpectraView Reference 21 LCD, with a sRGB gamut, a D65 white point, and a gamma set at 2.2. Their gamuts are represented Fig. 2 in the CIELAB color space. For the pair comparison on screen, a Dell 2407WFP LCD display calibrated with a D65 white point and a 2.2 gamma was used.

Viewing Conditions

The viewing conditions were chosen in as close accordance with the CIE guidelines[4] as possible. For the ranking experiment, the printed reproductions and the original image were the same size and surrounded by, respectively, an unprinted border and a white border. The printed images were viewed in the viewing booth, the Judge II from GretagMacbeth under a D50 simulator ($x = 0.3407$, $y = 0.3601$, $L = 105$ cd/m^2) and the original on a D65 monitor (chromaticity of the white point: $x = 0.3457$, $y = 0.3585$ with a luminance of 125 cd/m^2) in a windowless room with neutral grey walls, ceiling, and floor. The level of ambient illumination on the monitor switched off was around 20 lux. The viewing booth and the display were set up side-by-side. For the pair comparison experiment, the lightning conditions were the same and the observers viewed the monitor from ~50 cm.

Observers

Twenty observers took part in the psychophysical experiment. They all passed the Ishihara color blindness test. Among them, 11 were considered as experts in terms of experience in color imaging and nine as nonexperts. The same observers did both experiments. The experi-

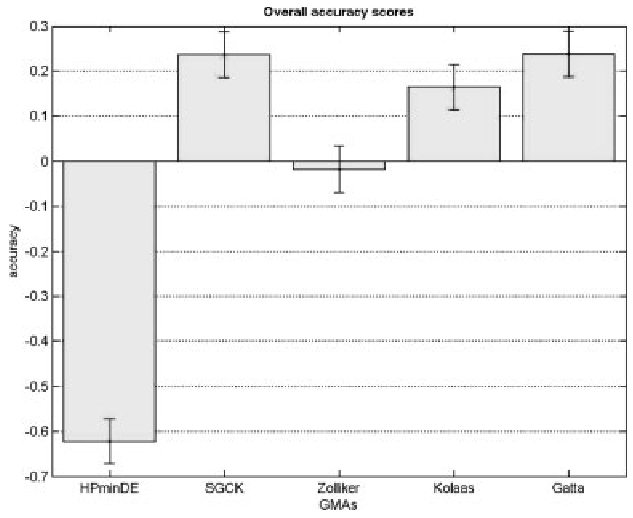

FIG. 3. Results of the experiment on paper, all images and observers.

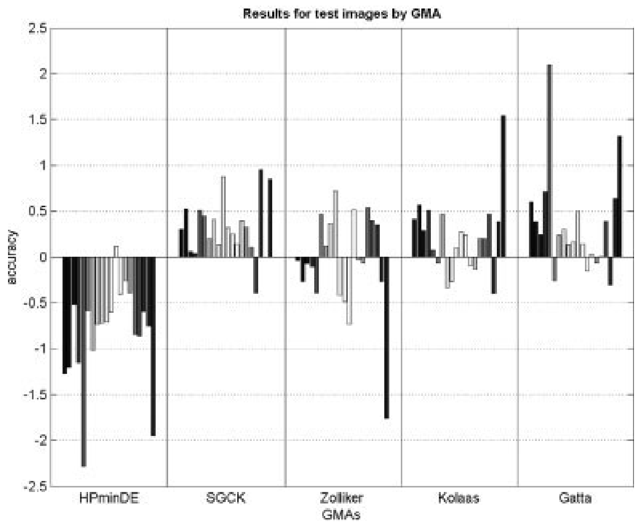

FIG. 4. Accuracy scores for the individual images in the ranking experiment with all observers. The 95% confidence interval is 0.2354.

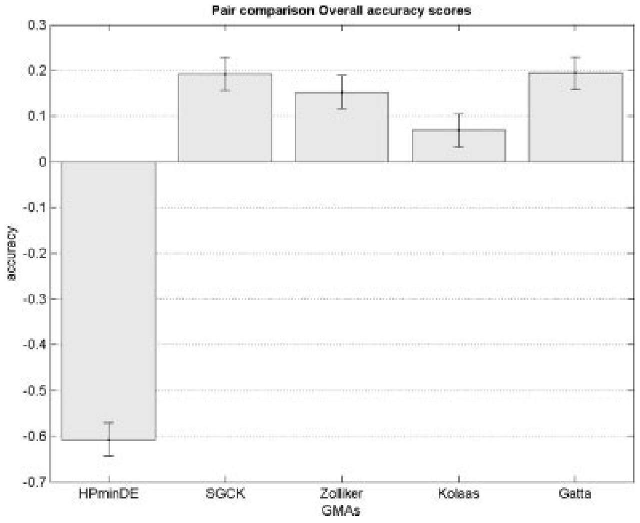

FIG. 5. Results of the experiment on display, all images and observers.

ments took in average 50 min for the experiment on paper and 39 min for the experiment on display.

Data Processing

We converted the rank data to frequency matrices,[15] and then we applied the case V of Thurstone's law of comparative judgment to obtain *z*-scores following Morovic's method.[16] For the pair comparison, software developed locally gathers the results in frequency matrices that are then processed as the other experiment to obtain the *z*-scores. The 95% confidence intervals are determined by using the empirical formula by Montag.[17]

RESULTS

The resulting *z*-scores and confidence intervals for all images and all observers with the printed reproductions are shown in Fig. 3. It is evident that HPminDE performs badly and cannot be considered as an accurate GMA. The three best algorithms do not have significantly different *z*-scores. We can mention that a spatial, Gatta, and a non-spatial GMA, SGCK, obtain the same score. Figure 4 shows the individual results per image. We notice that SGCK is stable with similar *z*-scores for each image. On the contrary, Zolliker obtains a high variability in the *z*-scores.

The results on screen (Fig. 5) also give Gatta and SGCK as the most accurate and HPminDE as the least accurate.

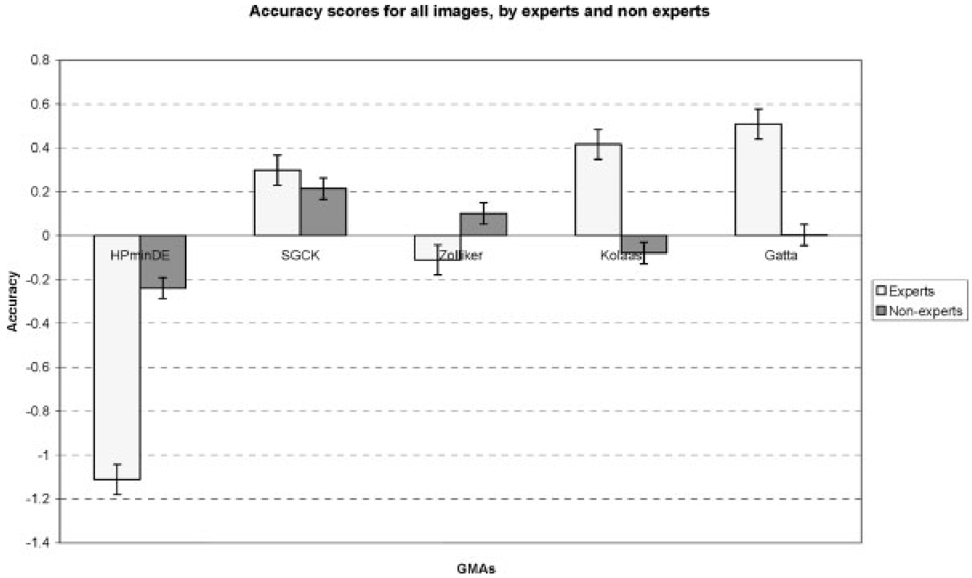

FIG. 6. Accuracy scores for the experiment on paper, all images, expert (light gray) and nonexpert (dark gray) observers.

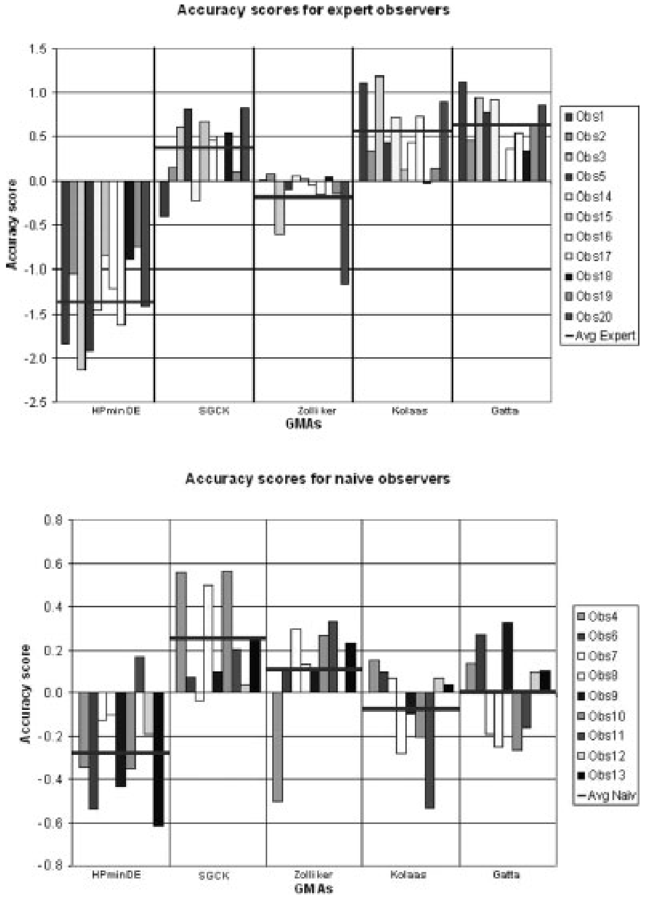

FIG. 7. Accuracy scores of each observer for each GMA compared to the group's average scores, experts (top) and nonexperts (bottom), in the experiment on paper.

Observers

Two groups of observers performed the experiments, the experts and nonexperts. We obtain different results for those two groups (Fig. 6). The experts distinguish more the different GMAs, with a difference of 1.62 points of *z*-scores between the least and most accurate, compared to only 0.45 points of difference for the nonexperts. It means that experts see greater differences between the alternatives than nonexperts. When comparing the individuals' responses with the group responses for each experts and nonexperts, we notice by calculating the relative standard deviation that there is a stronger consensus among the opinions in the expert than nonexpert groups (Fig. 7).

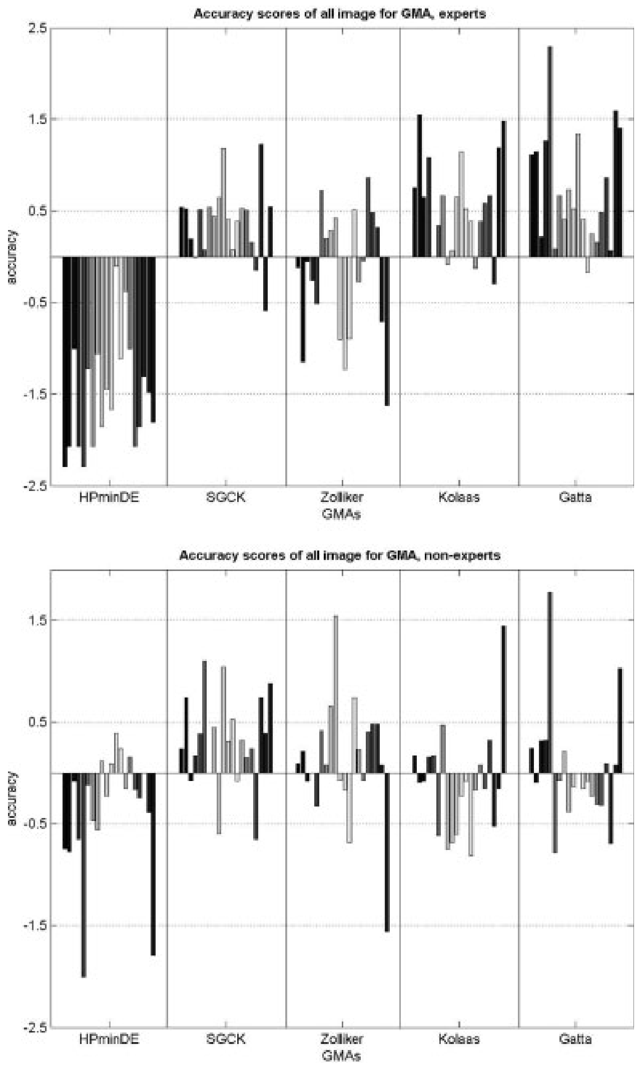

FIG. 8. Accuracy scores for each image in the experiment on paper, viewed by the experts (top) and the nonexperts (bottom).

Observers were asked to circle the regions they were looking at to make their ranking. From these data, we also notice that experts look at more regions of smaller sizes in the image to make their choice. The experts ranked the Gatta and Kolås GMAs as the most accurate and those two GMAs rendered the best details. Thus for the experts, a good rendering of details is an important criterion of accuracy. For the nonexperts, the nonspatial GMA, SGCK, is globally preferred.

TABLE I. Correlation coefficients, *r*, and *P* values between the percentage of out-of-gamut pixels, the perceived difficulty, and the number of distinguishable pairs of GMAs.

Correlation coefficients and (*P* value)	% of out-of-gamut pixels	Perceived difficulty	Number of distinguishable pairs of GMAs on paper
% of out-of-gamut pixels		−0.6113 (0.0042)	0.6798 (0.0010)
Perceived difficulty	−0.6113 (0.0042)		−0.7573 (0.0001)
Number of distinguishable pairs of GMAs on paper	0.6798 (0.0010)	−0.7573 (0.0001)	

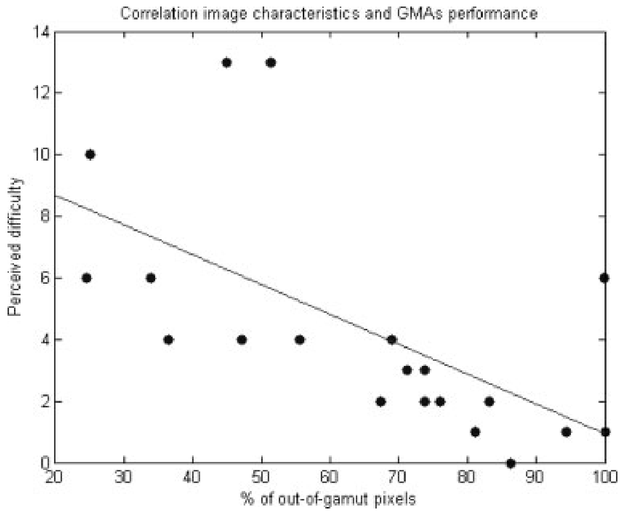

FIG. 9. Correlation between the perceived difficulty and the percentage of out-of-gamut pixels.

When looking at the results per image (Fig. 8), we notice that the results for the nonexperts are really image dependent. The low average scores are due to a high variability of the results and not to a low score for each image. The nonexperts cannot really distinguish the GMAs.

On the contrary, for the experts the results are quite consistent, except for the Zolliker algorithm which is highly image dependent.

Images

We have performed the tests with a high number of test images. As we have already seen, the results obtained show the variability depending on the images. We look for correlation between image characteristics and GMA performance. There is a correlation, *r*, between the perceived difficulty and the number of distinguishable pairs of GMAs (Table I, Fig. 11). The perceived difficulty is estimated by the number of times an image was said to be very difficult to rank by the observers. The number of distinguishable pairs of GMAs is the number of times GMAs are significantly different from the others. The percentage of out-of-gamut pixels is linked to both the perceived difficulty and the number of distinguishable pairs of GMAs (Table I, Figs. 9 and 10). So, the more an image is out-of-gamut, the more important is the choice of the GMA.

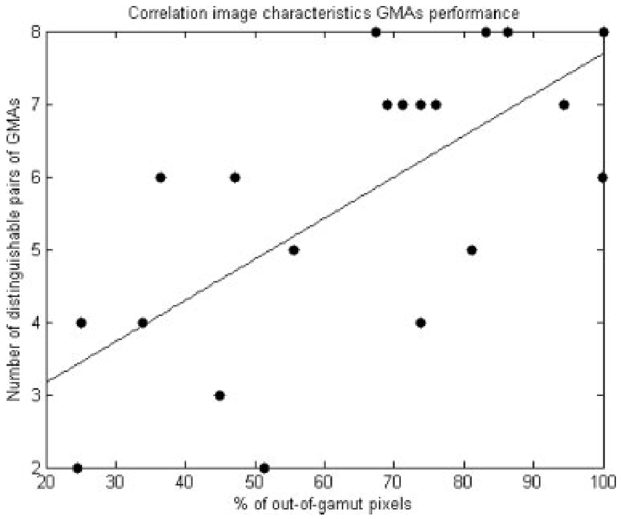

FIG. 10. Correlation between the number of distinguishable GMAs and the percentage of out-of-gamut pixels.

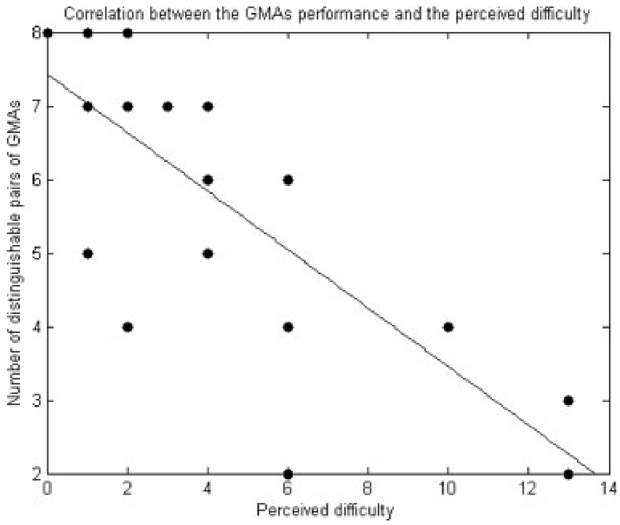

FIG. 11. Correlation between the number of distinguishable pairs of GMAs and the perceived difficulty.

By looking at each image, we can find some common trends. The images with saturated colors are better rendered by the Zolliker GMA. Those with details in dark area are much better rendered with the Gatta GMA. The color range of the image is not the only parameter that drives the performance of a GMA. For example, the Zolliker GMA performs differently on two images with red content. For one image (image 9 on Fig. 1) it is ranked the first, whereas for another red image (image 20 on Fig. 1) it has a very low negative score. Some artifacts appear in that image with the red and pink.

Experiments

It is not uncommon to perform the evaluation of GMAs on display.[6,8] It is thus natural to ask whether the results on screen are comparable to the ones obtained with the printed reproductions. The results with all observers for the two experiments are given in Fig. 12. For three of five GMAs, the *z*-scores are really close. The slightly lower scores for the screen experiment may be due to the fact that each pair is compared twice. When two images are almost indistinguishable, the observer may have chosen one time one image and the second time the other, thus no algorithm is preferred. On the contrary, in the ranking experiment, each pair is virtually compared only once and the observer is forced to make a choice. For the Zolliker algorithm, a minor mistake was discovered after finishing the experiments. While for the printed images the

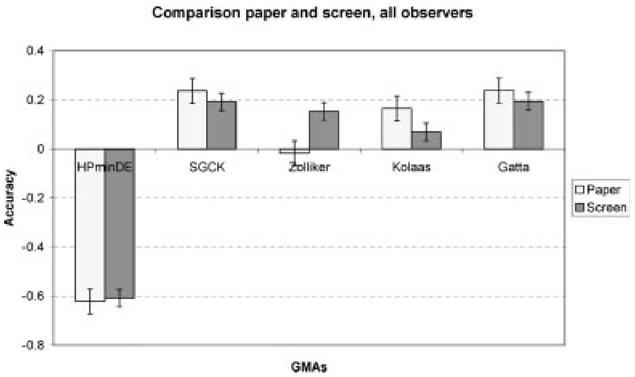

FIG. 12. Accuracy scores for the test on paper (light gray) and the test on display (dark gray) with all images and all observers.

algorithm was implemented correctly, the last step of the algorithm was not performed for the images that were displayed on screen. This means that these images were slightly out of gamut. Thus, the Zolliker algorithm should perform slightly better in the experiments on screen than it would in the correct case. This is indeed seen in the results as well. Zolliker is the algorithm that has the highest difference in *z*-score between the two experiments. In addition, this may also come from the fact that this algorithm, to a large extent, preserves local contrast and lightness at the expense of color accuracy.

Thus, as the printer has a low resolution where we could see the halftoning and the screen reproduces much more contrast than printed media, this could also explain this difference between the two experiments.

The media used may also have an influence. The quality was better on the screen, but the observers mentioned that the pair comparison test was more wearing and boring than the ranking.

CONCLUSIONS

This study has evaluated five selected spatial and nonspatial color gamut mapping algorithms by psychophysical experiments following the CIE guidelines. The conclusions and observations from this evaluation are summarized as follows:

- HPminDE is definitely not perceived as an accurate GMA.
- The Gatta GMA obtains the highest *z*-score, but not significantly different from SGCK and Kolås GMAs in the evaluation on paper and from SGCK GMA in the evaluation on display.
- SGCK is the algorithm that performs most steadily.
- Experts and nonexperts have different opinions.
- Experts have a stronger consensus. They look especially at the good rendering of details. The two highest accuracy scores for this group of observers are for spacial GMAs.
- Nonexpert observers do not really distinguish the different algorithms: experts see greater differences between the alternatives than nonexperts.
- The dependency on the test images is high, and thus it is important to have a high number of test images to obtain a robust evaluation.
- There are correlations between the percentage of out-of-gamut pixels, the perceived difficulty, and the number of distinguishable pairs of GMAs.
- Paper and display evaluations show similar but not identical results.
- Observers found the pair comparison more wearing and boring than the ranking experiment.

ACKNOWLEDGMENTS

The authors thank Nicolas Bonnier and Océ Print Logic Technologies for implementing the Zolliker algorithm and printing the gamut-mapped images on their OCE TCS 500. The authors also thank all the 20 observers for their contribution to the psychophysical evaluation.

1. CIE T.C. 8-03. Survey of gamut mapping papers. 1999. Available at www.colour.org/tc803/survey/survey_index.html (accessed March 29, 2007).
2. Morovič J, Luo MR. The fundamentals of gamut mapping: A survey. J Imaging Sci Technol 2001;45:283–290.
3. Farup I, Gatta C, Rizzi A. A multiscale framework for spatial gamut mapping. IEEE Trans Image Process 2007;16:2423–2435.
4. CIE 156:2004. Guidelines for the evaluation of Gamut Mapping Algorithms. Vienna: CIE 2004.
5. Zolliker P, Simon K. Adding local contrast to global gamut mapping algorithms. CGIV 2006 Final Program and Proceedings of Society for Imaging Science and Technology, 2006. p 257–261.
6. Zolliker P, Simon K. Retaining local image information in gamut mapping algorithms. IEEE Trans Image Process 2007;16:664–672.
7. Morovic J, Yang Y. Influence of test image choice on experimental results, Proceedings of 11th Color Imaging Conference, Springfield, VA: IS&T 2003, p 143–148.
8. Bonnier N, Schmitt F, Brettel H, Berche S. Evaluation of spatial gamut mapping algorithms. Proceedings of 14th Color Imaging Conference: Springfield, VA: IS&T, 2006. p 56–61.
9. Balasubramanian R, deQueiroz R, Eschbach R, Wu W. Gamut mapping to preserve spatial luminance variations. J Imaging Sci Technol 2001;45:436–443.
10. Tomasi C. Bilateral filtering for gray and color images. 1998 IEEE International Conference on Computer Vision, IEEE Computer Society, 1998. p 839–846.
11. Kolås Ø, Farup I. Efficient hue- preserving and edge-preserving spatial gamut mapping. Proceedings of 15th Color Imaging Conference, Springfield, VA: IS&T, 2007. p 207–212.
12. Harwood D, Subbarao M, Hakalathi H, Davis LS. A new class of edgepreserving smoothing filters. Pattern Recogn Lett 1987;6:155–162.
13. Balasubramanian R, Dalal E. A method of quantifying the color gamut of an output device. Proceedings of SPIE, Color Imaging: Device-Independent Color, Color Hard copy, and Graphic Arts II, Vol. 3018, San Jose, CA, 1997. p 110–116.
14. Bakke AM, Hardeberg JY, Farup I. Evaluation of gamut boundary descriptors. Proceedings of 14th Color Imaging Conference, Springfield, VA: IS&T, 2006. p 50–55.
15. Engeldrum PG. Psychometric Scaling: A Toolkit for Imaging System Development. Winchester: Imcotek Press; 2000. ISBN: 0-9678706-0-7-34.
16. Morovič J. To develop a universal gamut mapping algorithm, Condensed format edition, PhD Thesis, University of Derby, Derby, UK. 215 p.
17. Montag ED. Empirical formula for creating error bars for the method of paired comparison. J Electron Imaging 2006;15:222–230.