Perceptual Evaluation of Color Gamut Mapping Algorithms

Fabienne Dugay, Ivar Farup, and Jon Y. Hardeberg
The Norwegian Color Research Laboratory
Gjøvik University College
Gjøvik, Norway

Abstract:
The recommendation of the CIE has been followed as close as possible to evaluate the accuracy of five color gamut mapping algorithms (GMAs), two non-spatial and three spatial algorithms, by psychophysical experiments with 20 test images, 20 observers, a test done on paper and a second one on display. Even though the results do not show any overall "winner", one GMA is definitely not perceived as 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. 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 of importance. The experts, who prefer a spatial GMA, show a stronger consensus and look especially for a good rendering of details, whereas the non-experts hardly make a difference between the GMAs.

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] have presented a survey of research on gamut mapping until the end of the 90s and Farup et al.[3] has completed it with a review of some spatial gamut mapping algorithms. In order to evaluate the performance of GMAs and allow further comparisons, the CIE Technical Committee 8-03[4] has proposed guidelines on how to implement such tests. Evaluations with selected spatial and non-spatial GMAs have previously been done[5,6,7,8,9]. The purpose of this paper is to evaluate three new spatial GMAs and two non-spatial 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 also 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 that has followed as closely as possible the CIE guidelines[4].

ALGORITHMS

According to the CIE guidelines, two standard (i.e., non-spatial) GMAs have to be included in the experiment:

- HPminDE (Hue preserving minimum ΔE_{ab} clipping) [4]

This is a simple baseline algorithm which 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 compressed using a chroma-dependant sigmoidal scaling that compresses high-chroma colours less than neutral ones. Then, the resulting colours 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 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 that still contain most of the high-frequency information. Finally, the image is gamut clipped in order to be in-gamut. Since the high-pass filtering is performed for the three colour 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 towards the centre 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. 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 colours. 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. Operators operating in the range are introduced in order to reduce haloing effects. The process is iterated until all scales are treated. In order to speed up the process, the filtering is performed in the Fourier domain. However, the algorithm is still quite time consuming for large images.

For all of the GMAs, the gamut boundary is determined using the modified convex hull algorithm[12,13] with γ =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. For practical reasons, this method can not be used for the on-screen experiment, thereby the pair comparison 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

20 test images including the obligatory ski image are used (Figure 1). They have various characteristics in terms of gamut, contrast, contents, details, etc.

Figure 1: The 20 test images used

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 SpectraView Nec LCD, with a sRGB gamut, a D65 white point and a gamma set at 2.2. Their gamuts are represented Figure 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.

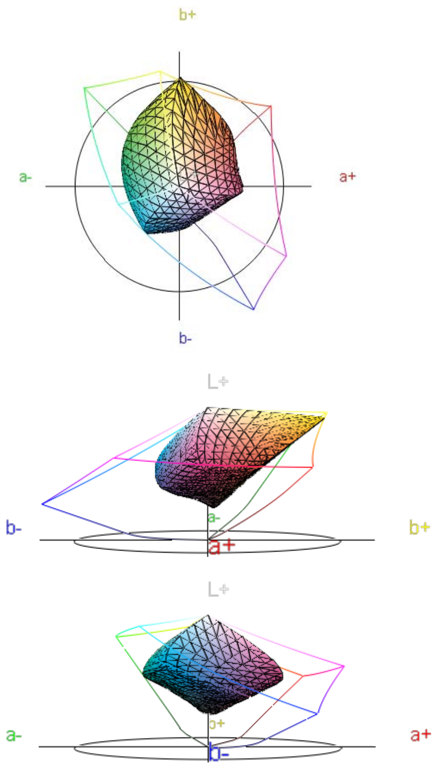

Figure 2: The Océ printer gamut (solid) and the sRGB gamut (wireframe) shown in the CIELAB color space.

VIEWING CONDITIONS

The viewing conditions were chosen as close to the ones described in 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 approximately 50 cm.

Observers

20 observers took part in the test. They all passed the Ishihara color blindness test. Among them, 11 were considered as experts in terms of experience in color imaging and 9 as non-experts. The same observers did both tests. The tests took in average 50 min for the test on paper and 39 min for the test on display. 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.

Data processing

We converted the rank data to frequency matrices[14], and then we applied the case V of Thurstone law of comparative judgment to obtain z-scores with the method of Morovic[15]. 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 per cent confidence intervals are determined by using the empirical formula by Montag [16].

RESULTS

The resulting z-scores and confidence intervals for all images and all observers with the printed reproductions are shown in Figure 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 undivided 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 (Figure 5) also give Gatta and SGCK as the most accurate and HPminDE as the least accurate.

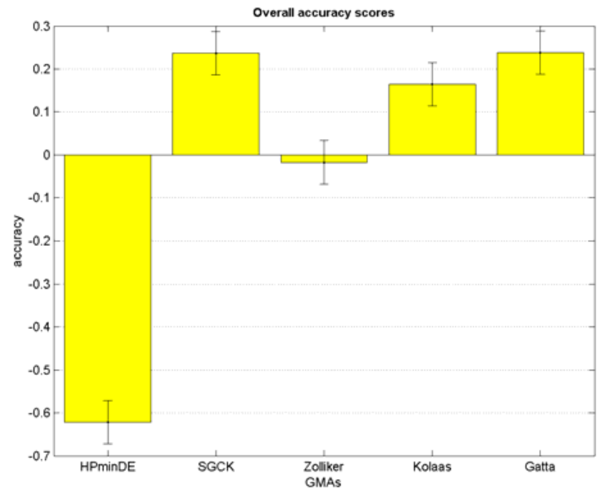

Figure 3: Results of the experiment on paper, all images and observers

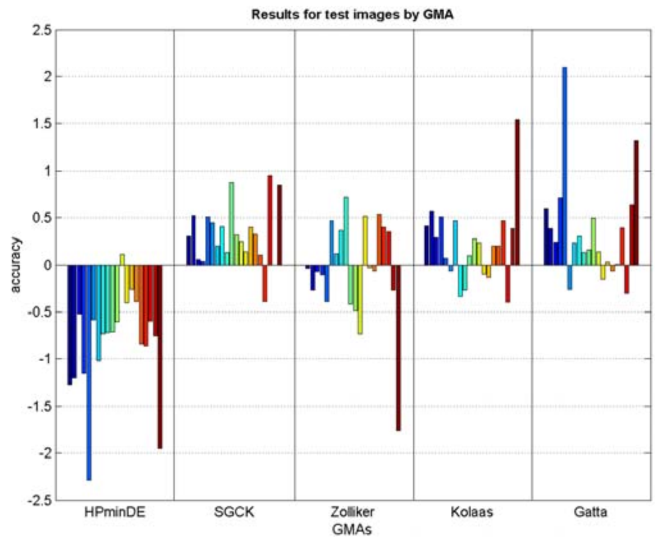

Figure 4: Accuracy scores for the individual images in the ranking experiment with all observers. The 95 percent confidence interval is 0.2354.

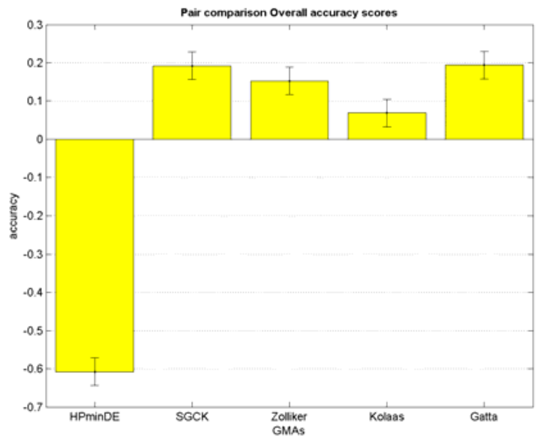

Figure 5: Results of the experiment on display, all images and observers

OBSERVERS

Two groups of observers did the tests, the experts and non-experts. We obtain different results for those two groups (Figure 6). The experts distinguish more the different GMAs, with a difference of 1.62 points of z-scores between the least and more accurate, compared to only 0.45 points of difference for the non-experts. It means that there is a stronger consensus among the opinions in the expert than non-expert groups. Observers were asked to encircle the regions they were looking at to make their ranking. From theses data, we notice also that experts look at more regions of smaller sizes in the image to make their choice. The experts rank the Gatta and Kolås GMAs as the most accurate and those two GMAs render the best the details. Thus for the experts, a good rendering of details is an important criterion of accuracy. For the non-experts, the non-spatial GMA, SGCK, is globally preferred.

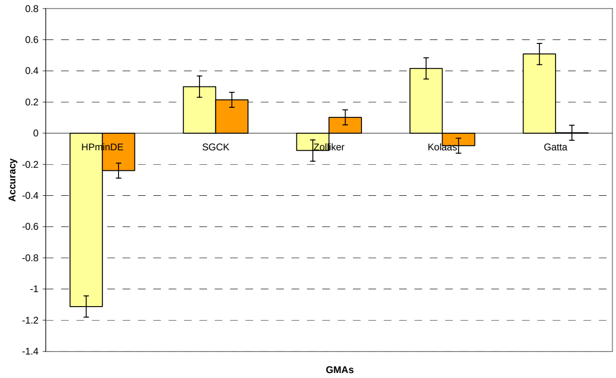

Figure 6: Accuracy scores for the experiment on paper, all images, expert (yellow) and non-expert (orange) observers

When looking at the results per image (Figure 7), we notice that the results for the non-experts are really image dependant. The low average scores are due to a high variability of the results and not to a low score for each image. The non-experts cannot really distinguish the GMAs. On the contrary, for the experts the results are quite consistent, except for the Zolliker algorithm that is highly image dependant.

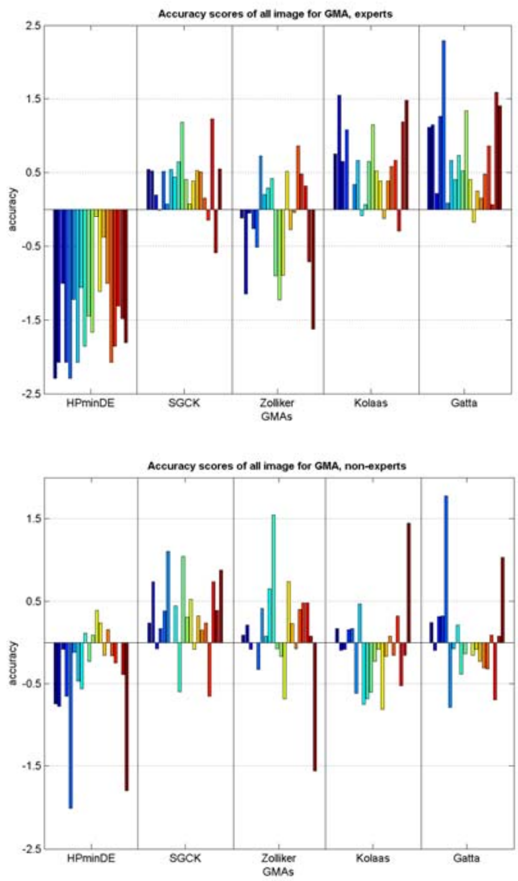

Figure 7: Accuracy scores for each image in the experiment on paper, viewed by the experts (up) and the non-experts (down)

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 images characteristics and GMA performance. There is a high correlation between the perceived difficulty and the number of distinguishable pairs of GMAs (Table 1). 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 link to both the perceived difficulty and the number of distinguishable pairs of GMAs (Figures 8,9,10). So, the more an image is out of gamut, the more important is the choice of the GMA.

Table 1: Correlation coefficients and p-values between percentage of out-of-gamuts pixels, perceived difficulty and number of distinguishable pairs of GMAs

Correlation coefficient 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	

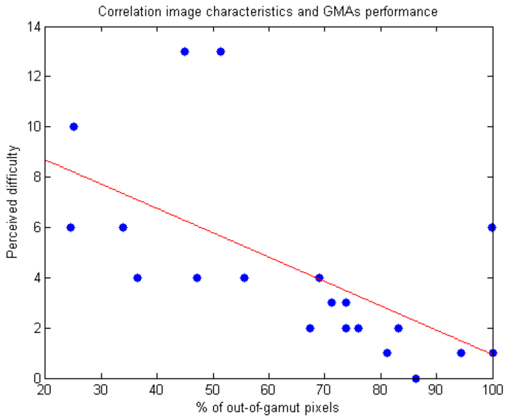

Figure 8: Correlation between the perceived difficulty and the % of out-of-gamut pixels

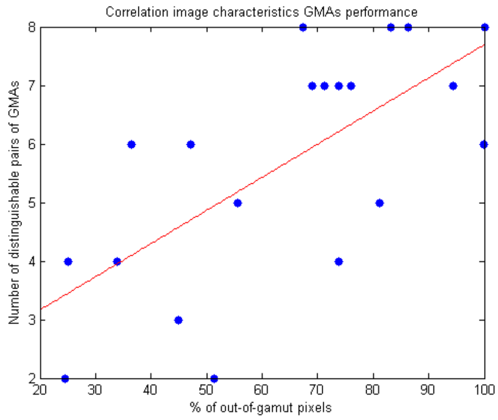

Figure 9: Correlation between the number of distinguishable GMAs and the % of out-of-gamut pixels

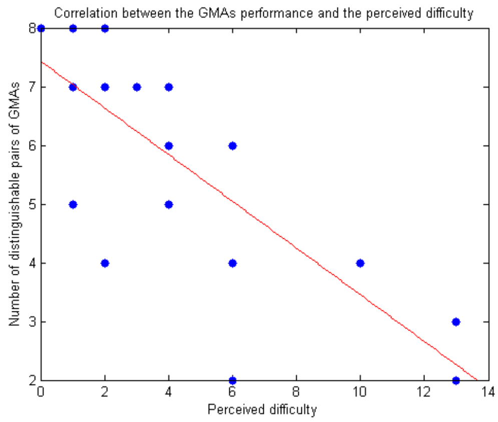

Figure 10: Correlation between the number of distinguishable pairs of GMAs and the perceived difficulty

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 it is ranked the first whereas for another red image it has a very low negative score. Some artifacts appear in that image with the red and pink.

EXPERIMENTS:
It is common 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 in the graph Figure 11. For three of the 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. The media used may also have an influence, as the printer has a low resolution where we could see the halftoning. The quality was better on the screen, but the observers mentioned that the pair comparison test was more wearing and boring than the ranking.

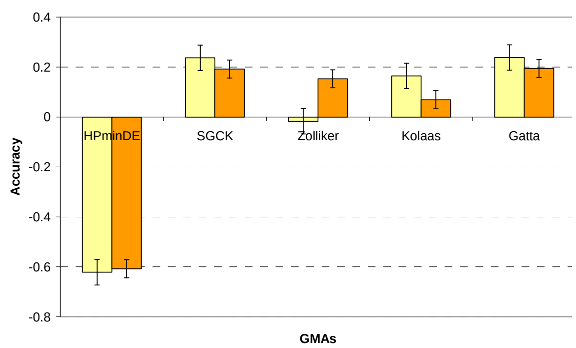

Figure 11: Accuracy scores for the test on paper (yellow) and the test on display (orange) with all images and all observers

CONCLUSIONS

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

- 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 and Zolliker GMAs in the evaluation on display.
- SGCK is the algorithm that performs the most steadily.
- Experts and non-experts have different opinions.
- Experts have a stronger consensus, look especially at a good rendering of details and rank first two spatial GMAs.
- Non-expert observers do not really distinguish the different algorithms.
- The dependency on the test images is high, 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 and the perceived difficulty and 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.

ACKNOWLEDGEMENTS

The authors would like to thank Nicolas Bonnier and Océ Paris for implementing Zolliker algorithm and printing the gamut mapped images on their OCE TCS 500.

We would also like to thank all the 20 observers for their contribution to the psychophysical evaluation.

REFERENCES

1. CIE T.C. 8-03 (1999), *Survey of gamut mapping papers*, available on http://www.colour.org/tc8-03/survey/survey_index.html (consulted the 29th March 2007)

2. Morovič J. and Luo M.R. (2001), The fundamentals of gamut mapping: A survey. *Journal of Imaging Science and Technology* vol. 45, no. 3, pp. 283–290, 2001.

3. Farup I., Gatta C. and Rizzi A. (2007), A general multiscale framework for spatial gamut mapping. Accepted for publication in *IEEE Transactions on Image Processing*

4. CIE 156:2004, *Guidelines for the evaluation of Gamut Mapping Algorithms*. Technical report, ISBN: 3 901 906 26 6.

5. Zolliker P. and Simon K. (2006), Adding Local Contrast to Global Gamut Mapping Algorithms. *CGIV 2006 Final program and Proceedings, Society for Imaging Science and Technology.*

6. Zolliker P. and Simon K. (2007), Retaining Local Image Information in Gamut Mapping Algorithms. *IEEE Transactions on Image processing, Vol.16, NO.3, March 2007.*

7. Morovič, J. and Yang, Y. (2003), Influence of test image choice on experimental results, *Proceedings of IS&T/SID 11th Color Imaging Conference.*

8. Bonnier, N. Schmitt, F., Brettel, H. and Berche, S. (2006), Evaluation of Spatial Gamut Mapping Algorithms, *Proceedings of IS&T and SID's 14th Color Imaging Conference: Color Science and Engineering: Systems, Technologies, Applications*, Scottsdale, Arizona, 2006. pp 56-61

9. Balasubramanian, R., deQueiroz, R., Eschbach, R. And Wu , W. (2001), Gamut mapping to preserve spatial luminance variations. *Journal of Imaging Science and Technology*, vol. 45, NO. 5, pp. 436–443, September/October 2001.

10. Tomasi C. (1998), Bilateral Filtering for Gray and Color Images, *1998 IEEE International Conference on Computer Vision*

11. Kolås Ø. and Farup I. (2007), Efficient hue-preserving and edge-preserving spatial gamut mapping, submitted for publication in *Color Imaging Conference*

12. Balasubramanian R. and Dalal E. (1997), A method of quantifying the color gamut of an output device. *Color Imaging: Device-Independent Color, Color Hard copy, and Graphic Arts II,* volume 3018 of Proceedings SPIE, San Jose, CA.

13. Bakke A.M., Hardeberg J.Y. and Farup I. (2006), Evaluation of gamut boundary descriptors, *Proceedings of IS&T and SID's 14th Color Imaging Conference: Color Science and Engineering: Systems, Technologies, Applications*, Scottsdale, Arizona, 2006. pp 50-55

14. Engeldrum P. G. (2000), *Psychometric Scaling: A Toolkit for Imaging System Development*. Imcotek Press, Winchester. ISBN 0-9678706-0-7

15. Morovič J. (1998), *To Develop a Universal Gamut Mapping Algorithm*. Condensed format edition. PhD Thesis. Derby, UK. University of Derby. 215p.

16. Montag E.D. (2006), Empirical formula for creating error bars for the method of paired comparison, *Journal of Electronic Imaging*,1 Jan–Mar 2006/Vol. 15(1)