Influence of Test Image Choice on Experimental Results

Jan Morovic† and Yu Wang†*
**Hewlett Packard Company, Barcelona, Spain*
†Colour & Imaging Institute, University of Derby, Derby, United Kingdom

Abstract

The choice of test images is an important part of the experimental evaluation of colour reproduction algorithms. As previous studies show low levels of correlation between individual image results, this study explores the impact of test image choice on the overall results of gamut mapping algorithm evaluation experiments. A set of fifteen test images is used in an experiment designed according to CIE TC8–03 Guidelines to evaluate the performance of six GMAs. The results first show the significant difference between how accurately individual images can be reproduced. Most importantly the results also show that five–image subsets can be selected so as to have each of the six GMAs in the top group for that set. Hence it is necessary to use greater numbers of test images in future experiments to obtain more robust and repeatable overall results.

Introduction

The justification of most colour science methods comes from their verification in psychophysical experiments where groups of observers are asked to perform tasks on the basis of visual stimuli they are presented with. As with all experiments the details of all their parameters play an important role and greatly affect the success of generalising experimental findings. The aim of the present paper is therefore to explore the role of a key parameter in experiments evaluating the performance of colour reproduction techniques. This parameter is the choice of test images for which the performance of such techniques is judged psychophysically.

The reason for focusing on this parameter is that numerous previous studies have shown that the performance of image colour reproduction methods is different for distinct original images. Looking at the results of evaluating the performance of gamut mapping algorithms in previous studies[1–8] shows that the coefficient of determination (R^2) between the results obtained for individual test images within a given study has a mean of 0.34 and a minimum even of 0.08. An implication of this weak correlation between the performances of a group of GMAs for different images is also that the overall performance of GMAs for a set of test images is potentially strongly dependent on the choice of that test image set. Furthermore looking at those 21 papers published at this Color Imaging Conference over the last ten years where GMAs are psychophysically evaluated shows that the median number of test images was five. There were even studies that used only one or two images and only three[9–11] of the above papers used the greatest number of test images – seven. As such the situation in previous GMA evaluation studies is such that the correlation between the results of individual images is very low and a only small number of images are used. It is then an important question to see whether the results of such studies in terms of how well individual algorithms perform for a set of images are determined more strongly by the GMAs themselves or whether they are more the result of what set of test images were used.

The primary aim of this study is therefore to explore how great an effect the choice of test images can have on the overall performances of GMAs judged on their basis. An understanding of this aspect of GMA evaluation experiments could have important implications on the choice of test images in a way that leads to robust and repeatable results. Furthermore if a strong effect is found then it could also contribute to explaining why there are such significant discrepancies between the findings of different gamut mapping studies.

The following parts of this paper will first provide the details of an experiment in which the effect of test image choice was studied, followed by a detailed exploration of the results.

Experimental Setup

This experiment in which the accuracy of six GMAs was judged for fifteen test images was carried out in accordance with the CIE TC 8–03 *Guidelines for the Evaluation of Gamut Mapping Algorithms*[12] (further referred to as *Guidelines*). One exception to this, however, is that the version of the ‘Ski’ image, specified by the *Guidelines* as obligatory, that was used here was a rendition of the colorimetric values given for the original transparency rather than the sRGB version which should be used for display to print workflows. The reason for this difference is that the sRGB version of the obligatory test image was not available at the time when this experiment was conducted. While this is a drawback in terms of comparing these results with other *Guidelines*–compliant studies, it does not diminish the findings from the point of view taken in this paper.

Test Images

The fifteen test images (Fig. 1) used in this experiment cover a range of image characteristics like having predominantly low or high chroma or being high or low key in tonal terms. Furthermore the images are of different types, including a business graphic, an art reproduction and a range of indoor and outdoor scenes featuring memory colours as well as objects whose colour might not be familiar to observers. Of these, the 'ski' and 'ysales' images are provided by the CIE TC 8–03 on Gamut Mapping (http://www.colour.org/tc8-03/).

bike *car* *cheese* *flower*

gold *ski* *plate* *ysales*

fa *fruit* *K3W* *mus*

lan *plane* *street*

Figure 1. Thumbnails of test images.

Media, Viewing Conditions And Gamuts

Originals in this experiment were rendered on an Apple 21" Studio Display and reproductions were made using a Canon BJC-6100 bubble-jet printer on plain paper. The display's white point chromaticity was set to match the paper's chromaticity when illuminated by a D65 simulator in a viewing booth. The level of illumination in the viewing booth was set so as to give the paper as similar a luminance to the display's white point as possible. Furthermore images had white borders and were viewed against mid–grey backgrounds on both media. The display and viewing booth were set up side–by–side in a dark room and viewed from approximately 75 cm. As such the viewing conditions match type (a) defined in the *Guidelines*. Under these conditions the CAM97s2[13] colour gamuts of the two media are as shown in Fig. 2.

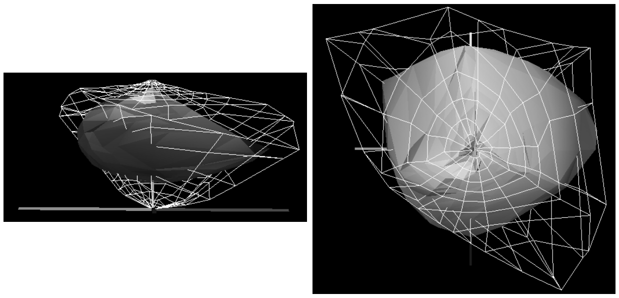


Figure 2. Gamuts of display (mesh) and print (solid).

Gamut Mapping Algorithms

The accuracy of six GMAs was evaluated in this experiment. The first pair were the hue–preserving minimum ΔE (HPMINDE) and SGCK algorithms specified in the *Guidelines* as obligatory. The second pair were spatial GMAs proposed by Bala *et al.*[14] where XSGM is the basic form of the algorithm and XIGI-SGM includes an initial lightness compression. The third pair of GMAs were the MSGM4 multi–resolution spatial GMA described elsewhere[15] and MSGM2, whish is its simplified version. Note that all of these GMAs were performed in the CAM97s2 *Jab* colour appearance space.[13]

Psychophysical Method

A category judgement technique[16] was used in the psychophysical evaluation of GMA accuracy. The accuracy of image reproduction was judged on an equi-interval accuracy scale with values from zero to six. Here zero represents the least accurate reproduction an observer can imagine and six represents the most accurate reproduction. Observers were asked to judge into which category each of the six reproductions of each of the fifteen test images belonged based on the accuracy with which it reproduced the corresponding original. A total of 15 colour–normal observers participated in the experiment.

Overall Results

Before looking at the results of the experiment it is first useful to evaluate the performance of observers who took part in it. Inter–observer repeatability was determined based on five observers repeating ten randomly–chosen judgements twice. The mean of these differences between the first and second judgements was 0.66 accuracy units. Taking the mean of all observer judgements and computing the mean of individual observer differences from it also established inter–observer agreement. This mean difference was 0.89 accuracy units, which means that on average observers repeated their judgements and also agreed with each other to within one category unit.

Overall GMA Accuracies

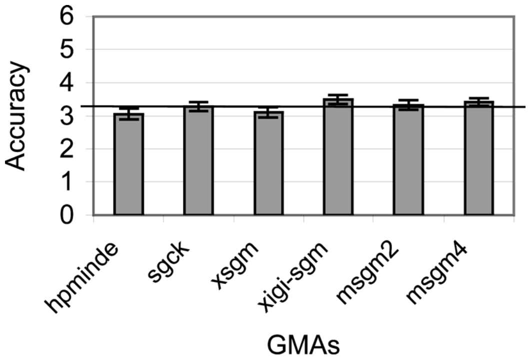


Figure 3. Mean category values for individual GMAs.

The first aspect of the results that will be considered here are the accuracy scores obtained by pooling together the category judgements made for all the test images as it will be this result, whose robustness will be explored later. From these results (Fig. 3) it can be seen that the differences between the six GMAs are not large when the whole test image set is considered. Nonetheless, the XIGI-SGM and MSGM4 algorithms are significantly more accurate than the other methods, while not being significantly different from each other. HPMINDE and XSGM, on the other hand, are significantly worse than the other methods.

Mean Category Values of Image Reproductions

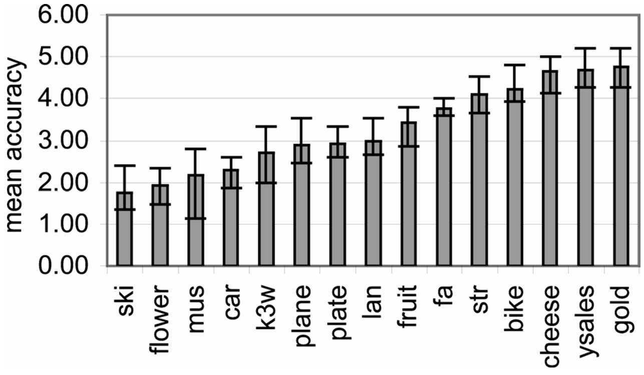


Figure 4. Mean category values for individual test images.

Before exploring the robustness of the above data, another aspect of these results is worth exploring. As a category judgement technique was used and as the categories in terms of which reproduction accuracy was judged were on an absolute scale, a comparison of how accurately each of the fifteen originals was reproduced can be made. For each test image Fig. 4 therefore shows the mean of all category values judged for its reproductions. This represents the mean accuracy achieved by all six algorithms for each of the test images in turn and allows for a comparison of how difficult these images are to reproduce. If all originals were reproduced equally well by the set of six GMAs used here, the mean category values for all test images would be the same. Fig. 4 however, shows that images vary greatly in terms of how well they can be reproduced. Having this kind of information is another advantage of the category judgement method, which, unlike the pair-comparison approach provides results on a single scale for all reproductions in the experiment. Images like ‘ski’ and ‘flower’ are reproduced with low accuracy by all the algorithms tested here while all reproductions of images like ‘cheese’, ‘ysales’ and ‘gold’ are very accurate.

The error bars in Fig. 3 further provide information about the range of GMA accuracies for each reproduction of a given original and the accuracy scores of the least and most accurate GMAs determine them for each of the original images. This shows how much the accuracies of GMAs vary for each of the originals and it can be seen, for example, that all the GMAs used here are much more similar to each other for the ‘fa’ image than for the ‘mus’ image. Furthermore it is clear from this figure that the differences between various GMAs are significantly smaller than differences between all reproductions of different originals. For example, the difference between the least and most accurate reproductions for the ‘mus’ image is smaller than the degree by which the least accurate reproduction of the ‘gold’ image is more accurate than the most accurate reproduction of the ‘ski’ image. In other words, image differences are significantly greater than GMA differences and this too points to the importance of how test image sets are chosen.

Given these results it is possible to look for image characteristics that could explain the differences in GMA performance. For example, images with many dark and saturated colours are more likely to be inaccurately reproduced.[17] On the other hand, images with many light colours are very likely to be reproduced well and the remaining images are ranked between these two extremes.

Correlation Between Image Results

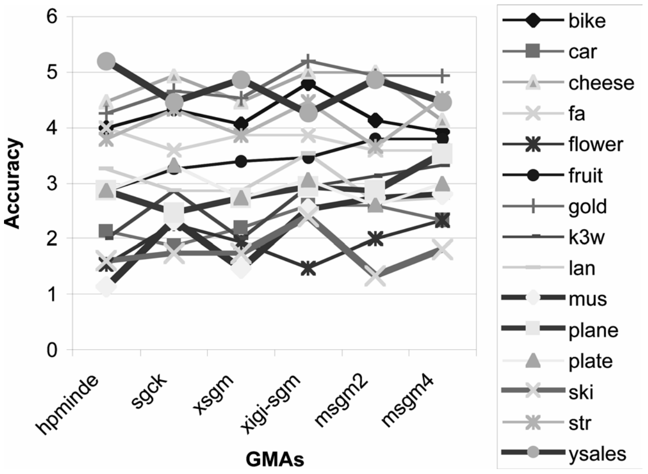


Figure 5. Accuracy scores of all GMAs for all images.

As the starting point of this study was the realisation that the results for individual images in previous studies are not well correlated, this section will look at the relationships between the results for the fifteen test images used here (Fig. 5). Considering all the 105 image pairs that can be formed for this test image set, the mean R^2 is 0.23, which is even lower than was found previously. Furthermore, 12 of the pairs even have R^2s of 0.00 and one pair ('lan' and 'flower') has a strong negative correlation of -0.89. On the other hand one of the pairs – 'mus' and 'k3w' – shows a strong relationship (R^2=0.97), which is not surprising as both these images show a group three women photographed in studios.

Overall though the weak correlation between GMA performances for the individual test images indicates that their choice will have a strong effect on the overall GMA performance judged on their basis.

GMA–Tailored Five–Image Subsets

As fifteen test images were used in this experiment and as this number is far greater than the five test images usual for this kind of evaluation, the present data lends itself to another kind of analysis. We can see what would happen in terms of overall GMA performance results for different five-image subsets of the present set of fifteen test images. This will tell us how image-set dependent the results are.

The five-image subsets chosen for this analysis were such combinations, which would be considered to be good test image sets in experiments that would use only five test images. In other words, each of the following sets was chosen with the aim of being as representative as is possible with five images. Further, image sets were chosen so as to try and favour each of the GMAs in one of the sets and the parts of Fig. 6 are labelled according to which GMA they are tailored to.

In Fig. 6a, the most highly ranked four images for MSGM4 are 'k3w', 'mus', 'plane', and 'street'. The 'ski' image is also included so as to comply with the TC8-03 *Guidelines*. For these test images, the MSGM4 algorithm gets the highest ranking and outperforms all the other algorithms at the 95 percent confidence level, except for XIGI-SGM. However, if we were to include the 'flower' image instead of the 'ski' image, MSGM4 would significantly outperform even XIGI-SGM.

In Fig. 6b, the best four images for MSGM2 are 'cheese', 'fruit', 'k3w' and 'mus'. 'Ski', which favours the XIGI-SGM and XSGM algorithms, is also chosen again for the same reasons as in the previous case. Based on this set of five images, the MSGM2 algorithm performs much better than it did in the whole set of fifteen images. However, if we look at the 95 percent confidence intervals, MSGM2's performance is statistically indistinguishable from those of MSGM4 and XIGI-SGM.

In Fig. 6c, the five images favouring XIGI-SGM are 'bike', 'gold', 'landscape', 'mus' and 'ski'. Noticeably there are two bright images among these five, which got higher category judgements from the observers. Consequently XIGI-SGM obtained the highest accuracy in this set of images. However, it is not high enough to outperform the other algorithms at the 95 percent confidence level.

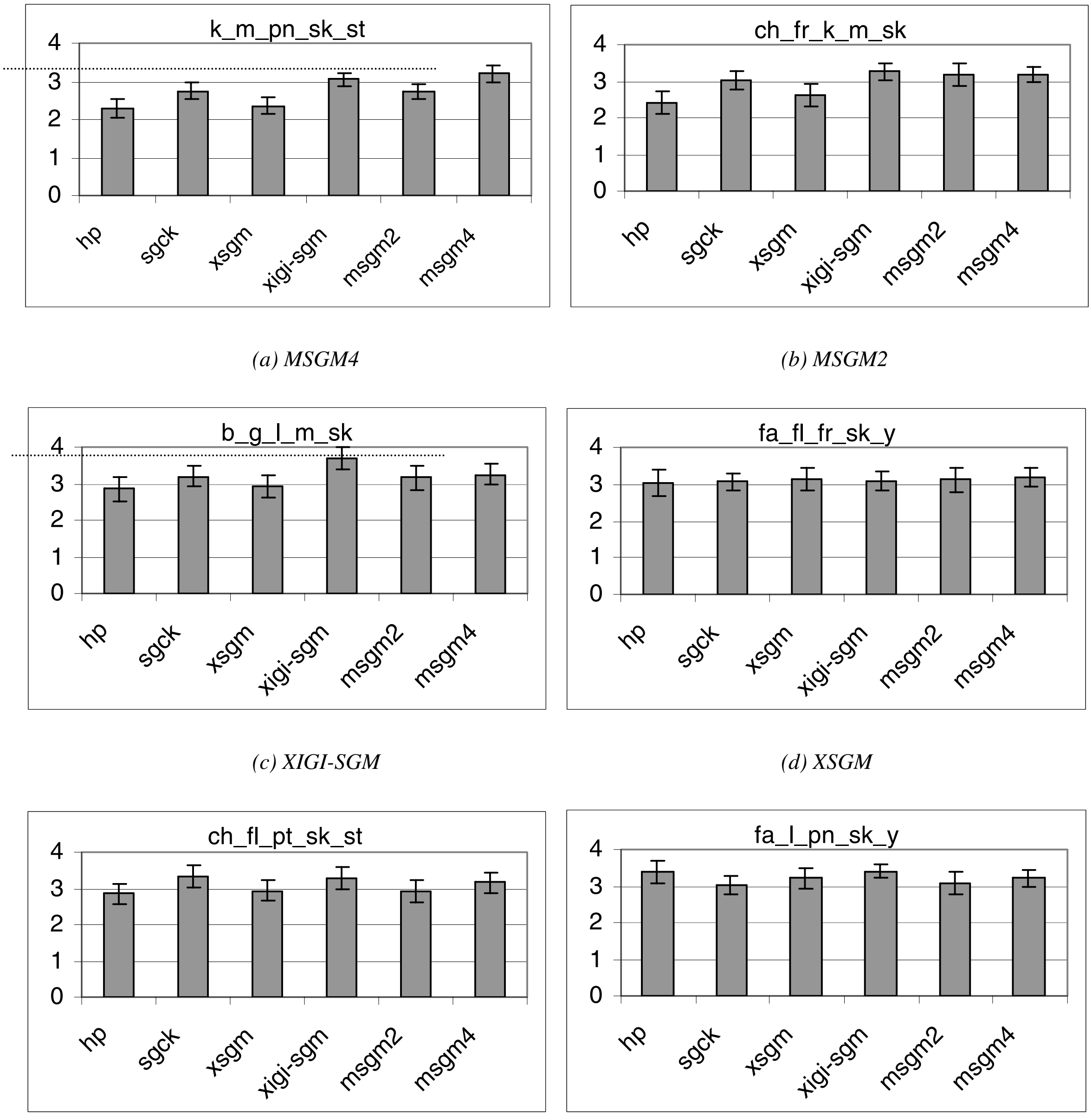
In Fig. 6d, the five images favouring XSGM are 'fa', 'flower', 'fruit', 'ski' and 'ysales'. However, even with these images, XSGM cannot outdo the other GMAs in overall performance.

For the remaining Figs. 6e and 6f, the corresponding algorithms accuracies are on the top group (i.e. not significantly different from the most accurate GMA) and they perform much better than they did in the whole set of 15 test images. However, neither of them can outperform all of the other GMAs.

From this analysis it can be seen that GMA performance greatly depends on the test images used in a psychophysical experiment. The effect is so strong that the choice of test images rather than GMAs that has a greater effect on overall GMA performance results. Hence for a given colour reproduction system with a fixed pair of media it is the choice of test images rather than the characteristics of GMAs that determine GMA performance. Test images can even be chosen to make any of a set of GMAs have a score that puts it into the top set of algorithms. To avoid such dramatic image-dependence, a larger set of test images is needed for GMA performance evaluation and further work needs to be carried out to see what number of test images leads to robust results that also transfer to results obtained using other test image sets.

Conclusions

The aim of this paper was to highlight the importance of how sets of test images are chosen when colour reproduction algorithms are evaluated. To this end a survey of existing work was carried out and the findings showed very weak correlation between GMA evaluation results for individual test images. Further it was found that relatively small numbers of test images are typical of these experiments. A set of GMAs was therefore evaluated on the basis of CIE TC8–03 *Guidelines for the Evaluation of Gamut Mapping Algorithms*. A key aspect of the experiment was that a relatively large set of test images (fifteen) was used. As the category judgement method was used, the results first showed that the ranges of GMA accuracies vary greatly for different originals whereby the least accurate reproductions of some images can be much more accurate than the most accurate reproductions of others. Looking at the correlation between image pairs in the chosen test set again showed very low values for the R^2 coefficient of determination. Finally the implications of this were illustrated by compiling sets of five test images taken from the total set of fifteen in such a way that each five–image subset favoured one of the six GMAs. This showed clearly that five images could be chosen so as to make the overall performance of each of the GMAs in turn be in the top group of statistically indistinguishable methods. These findings show that more than the typical five test images are needed for obtaining robust GMA evaluation results and that future work could be carried out to establish what minimum number of test images is desirable.



(a) MSGM4

(b) MSGM2

(c) XIGI-SGM

(d) XSGM

(e) SGCK

(f) HPMINDE

Figure 6. Mean GMA category values for selected five–image sets (error bars show standard error at 95 percent confidence level).

Acknowledgement

We would like to thank Tadashi Yoshida and Todd Newman (Canon Development Americas) for supporting this work. Many thanks go to Dr. Pei-Li Sun (Shih Hsin University) for his valuable help in conducting the experiment and to Dr. Raja Bala (Xerox) for advice on correctly implementing his GMAs. We would also like to thank all the observers at the Colour & Imaging Institute (UK) for their contribution to the psychophysical evaluation.

References

1. MacDonald L. W. and Morovic J. (1995) Assessing the Effects of Gamut Compression in the Reproduction of Fine Art Paintings, *Proc. 3rd IS&T/SID Color Imaging Conference*, 194-200.

2. Wei R. Y. C., Shyu M. J. and Sun P. L. (1997) A New Gamut Mapping Approach Involving Lightness, Chroma and Hue Adjustment, *TAGA Proceedings*, 685-702.
3. Braun G. J. (1999) *A Paradigm for Color Gamut Mapping of Pictorial Images*, PhD. Thesis, Rochester Institute of Technology, Rochester, NY.
4. Katoh N. and Ito M. (1999) Applying Non-linear Compression to the Three-dimensional Gamut Mapping, *Proc. 7th IS&T/SID Color Imaging Conference*, 155-159.
5. Pirrotta E., Newman T. and Lavendel L. (2000) A "Universal" Gamut Mapping Algorithm? *Proc. Colour Image Science 2000 Conference*, University of Derby, UK, 313-320.
6. MacDonald L. W., Morovic J. and Xiao K. (2002) A Topographic Gamut Mapping Algorithm, *Colour Imaging Science: Exploiting Digital Media*, MacDonald L. W. and Luo M. R. (eds.), John Wiley & Sons, Ltd., 291-317.
7. Motomura H. (2000) Gamut Mapping Using Color-Categorical Weighted Method, *Proc. 8th IS&T/SID Color Imaging Conference*, 318-323.
8. Chen H. S., Omamiuda M. and Kotera H. (2001) Gamma-Compression Gamut Mapping Method Based on the Concept of Image-to-Device, *J. Imaging Science and Technology*, **45**(2): 141-151.
9. E. D. Montag and M. D. Fairchild, Gamut Mapping: Evaluation of Chroma Clipping Techniques for Three Destination Gamuts, *Proc. of IS&T/SID 6th Color Imaging Conference,* 57–61 (1998).
10. P. G. Herzog and H. Büring, Optimizing Gamut Mapping: Lightness and Hue Adjustments, *Proc. of IS&T/SID 7th Color Imaging Conference,* 160–166.
11. G. J. Braun and M. D. Fairchild, General-Purpose Gamut-Mapping Algorithms: Evaluation of Contrast-Preserving Rescaling Functions for Color Gamut Mapping, *Proc. of IS&T/SID 7th Color Imaging Conference,* 167–172.
12. CIE TC8–03, Guidelines for the Evaluation of Gamut Mapping Algorithms, *Technical Report, Draft No. 18M*, (2002).
13. C. J. Li, M. R. Luo and R. W. G. Hunt, The CAM97s2 Model, *Proc. of IS&T/SID 7th Color Imaging Conference*, 262-263 (1999).
14. R. Bala, R. DeQueiroz, R. Eschbach and W. Wu, Gamut mapping to preserve spatial luminance variations, *J. Imaging Science and Technology*, **45**(5): 436-443 (2001).
15. J. Morovic and Y. Wang, A Multi–Resolution, Full–Colour Spatial *Gamut* Mapping Algorithm, *IS&T/SID 11th Color Imaging Conference, submitted for publication* (2003).
16. W. S. *Torgerson*, A Law of Categorical Judgment, *Consumer Behaviour*, L. H. Clark (ed.), New York University Press, New York, 92–93 (1954).
17. P. L. Sun, *Influence of Image Characteristics on Colour Gamut Mapping*, Ph.D. Thesis, University of Derby, UK (2002).

Biography

Ján Morovic received a Ph.D. in Colour Science at the Colour & Imaging Institute (CII) of the University of Derby (UK) in 1998, where the title of his thesis was *To Develop a Universal Gamut Mapping Algorithm*. He then worked as a lecturer in digital colour reproduction at the CII and is currently an image quality engineer at Hewlett Packard in Barcelona, Spain. He is also the chairman of the CIE Technical Committee 8-03 on Gamut Mapping.