Testing the Softproofing Paradigm

Alexis Gatt, Stephen Westland and Raja Bala[+]
Centre for Colour Design Technology, University of Leeds, Leeds, United Kingdom
[+]Xerox, Rochester, New York, USA

Abstract

The purpose of this article was to assess the suitability of softproofs as a surrogate for the final print in judging colour-reproduction quality. A complex viewing apparatus was specially designed for this study to ensure that no cognitive cues were visible to observers and that the surround conditions for viewing softcopies and hardcopies were in very close agreement. Two experiments targeting judgements related to colour quality were carried out: one relating to colour accuracy, and one relating to colour preference. Each experiment was conducted using two workflows: one involving hardcopy stimuli, and the second involving softcopy simulations of those hardcopies. Overall, the general conclusion that can be drawn is that judgments made on the basis of softproofs are transferable to prints. While results based on pictorial images are very robust, the intrinsic characteristics of business graphics make them more prone to highlight the intrinsic differences and abilities of the reproduction devices of interest, and thus affect the judgements derived from such stimuli. Providing that the viewing conditions are very carefully equated and that a significant number of test images is used, softproofs are suitable as surrogates for the final print in judging the quality of colour reproduction.

Introduction

Softproofing is well known in the graphic arts industry as a means for reducing expensive and time-consuming iterations on the final printing process. However, limitations still exist in this area. Many customers are unwilling to make critical color decisions based on a softproof (i.e. a softcopy simulation of a hardcopy) of the final print. Reasons for this range from simple factors such as colour errors from inaccurate device characterisation to more complex factors such as inherent differences in the appearance of displayed vs. printed content, and the resulting expectations by the user. Modern colour-appearance models have already very significantly improved the quality of cross-media colour reproduction, but results[1] previously obtained fundamentally challenged a very common practice in both research and industrial settings, i.e. the use of a softproof as a surrogate for the final print in judging colour quality. MacDonald *et al.* assessed performance of gamut mapping algorithms (GMAs) by asking observers to rank the accuracy of a hardcopy reproduction in terms of similarity of appearance to the softcopy original. The feasibility of using softproof simulations of the hardcopies was also investigated by conducting the same experiment and simulating the colour appearance of the prints on a monitor. It was found that the relative performance of each algorithm in the two studied cases was very different: a finding that challenged a widely held belief that a colorimetric match is equivalent to a visual match under controlled viewing conditions.[2]

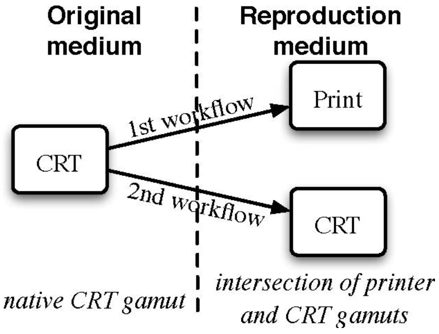

Figure 1. Overview of the two reproduction workflows used in both experiments used to generate hardcopy stimuli (1st) and softcopy simulations of them (2nd).

A series of experiments was therefore designed to address the fundamental question as to what effect the underlying media has on the appearance of colour stimuli. Two of these experiments will be introduced in this article. They aimed at determining whether judgments made on the basis of softproofs are transferable to prints. Each experiment had therefore to be duplicated in order to collect decisions based both on hardcopy stimuli (termed 1st reproduction workflow) and on softcopy simulations of hardcopy stimuli (2nd reproduction workflow). Since GMAs are by nature destined for cross-media colour reproduction applications, an assessment of their performance would provide an ideal way to compare the two studied reproduction workflows. However, it is important to understand that their intrinsic absolute performance is of no value in this specific case. Rather, the analysis should focus on the correlation of the results obtained by each workflow, instead of considering each case individually. Figure 1 summarizes the whole process.

Judging the quality of a reproduction cannot be performed in absolute terms, but is always dependent on a specific application. Accuracy is a criterion that plays a major role in typical color reproduction framework, and formed the basis for the first experiment. Observers were asked to judge how precisely a hardcopy reproduction (or its simulation) approximates, in terms of colour appearance, a softcopy original. However, final outputs should not only provide a faithful rendering of the original scene, but should also be pleasing since they will most probably be used as stand-alone. The second experiment investigated the overall pleasantness of the reproductions produced by each GMAs that were tested. Observers were presented with two reproductions generated by two different GMAs *on the same medium* and asked to choose their preferred one in terms of colour appearance. From a more conceptual point of view, this preference experiment is fundamentally different from the previous one, although the set-up and procedures employed are very similar. It exclusively targeted "within-media" judgments, or more specifically how such decisions made on the basis of prints exclusively are transferable to softproofs only, whereas the previous experiment investigated how "cross-media" judgments transfer to "within-media" ones. The combination of both should provide a clear overview of the suitability of softproofs as a surrogate to hardcopies.

Experimental Set-Up

Viewing Apparatus

For this set of experiments, the viewing conditions need to be equated as accurately as possible in order to maximize the apparent similarity of the two stimuli presented to observers. Nothing except the media used to generate the colour stimuli themselves should differ in order to ensure the quality of the results. However, softcopies and hardcopies are generally observed under very different conditions. Being self-luminous, softcopy images can be viewed almost anywhere with standard office conditions being most typical. On the other hand, a viewing booth is generally recommended in order to assess hardcopies, since the viewing conditions can then be controlled more accurately. However, the presence of any cognitive cues that might help observers to discriminate the type of medium used to generate stimuli could seriously impair the pertinence of the results obtained in this study. The proposed solution consists in placing a screen between observers and stimuli that contains two apertures through which the stimuli will be displayed. The size of the apertures can be adjusted in order to hide anything that is not part of the stimuli themselves, such as the bezel of the screen, or the walls of a viewing booth. Both sides of the apparatus could be used to display either a hardcopy or a softcopy stimulus, so that the location of the two stimuli being compared could be changed at will. Figure 2 shows a schematic diagram of the experimental set-up that has been employed, which has the double advantage of hiding all cognitive cues and equating the surround conditions very accurately.

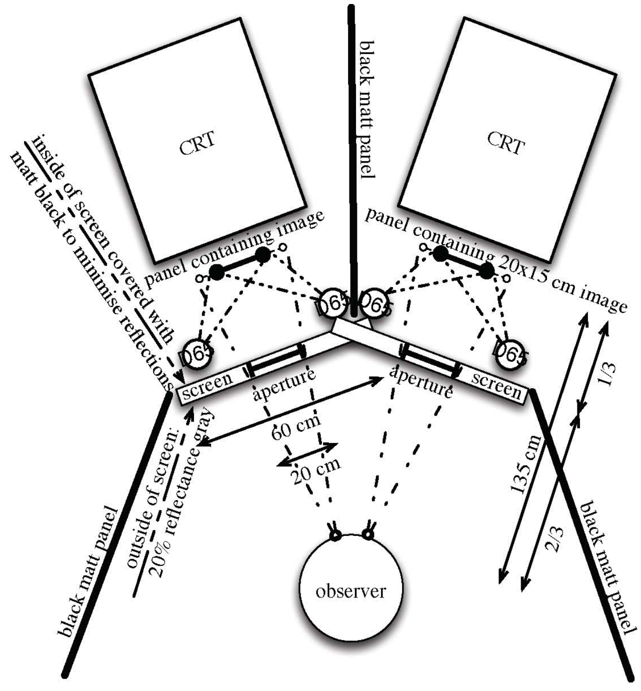

Figure 2. Schematic diagram of the viewing apparatus.

Viewing booths, while providing a very convenient and accurate way of illuminating reflection prints, employ a diffuse illumination which creates flare, which can significantly reduce the dynamic range achievable by a hardcopy. A directional light source is essential to preserve the full contrast and the saturation of prints, but it creates specular reflections which strongly affect the appearance at certain angles. The CIE "45/0" geometry of viewing illumination and viewing[3] using directional light sources was therefore adopted in order to overcome those issues. Since the presence of ambient lighting would generate viewing flare which could seriously affect the colour reproduction abilities of the self-luminous displays (SLDs), the experimental room did not contain any ambient light. Additional panels were also added to the screen in order to prevent stray light affecting both media and also to minimise cognitive cues given to observers. The most appropriate viewing technique in the studied case is simultaneous binocular viewing since the adaptation state of observers will be single and steady and because having both stimuli in the same field of view will maximize their discrimination ability. Both media were also set to be on the same radius with regard to the observer's point of view, and thus also having their axis perpendicular to the observer's viewing axis, since the SLDs may have an angular dependency. This is also required by the CIE standard "45/0" viewing geometry adopted.

Geometrical Set-up

The physical size of the built apparatus did not permit the viewing distance to conform to the usual recommendations (typically 80 cm) so it was extended to 135 cm. This increase was compensated by the large size of the stimuli used, 20 x 15 cm, which corresponds to a solid angle of 8.5 degrees vertically and 6.5 degrees horizontally. The whole visual field that can be seen through the aperture, i.e. the stimulus plus its border and its background, subtended an angle of 13 degrees horizontally and 10 degrees vertically. The hardcopy stimuli were mounted on a removable cardboard placed 1 cm in front of the monitor screen.

Colorimetrical Set-up

Hardcopies were illuminated by two daylight simulators (fluorescent tubes) whose chromaticities approximated those of the CIE Standard Illuminant D65. The screen and panels were painted in matt black in order to create totally dark surround conditions. The colour reproduction media were used to generate both the stimuli and also the background since it would be impossible to provide illumination to an external background without affecting the SLDs, nor to avoid cognitive cues created by the presence of edges. A small white border surrounding the stimulus itself was also included in order to steady the state of adaptation of observers. The chromaticities of the background for both media were set to match those of the paper's white border illuminated by the daylight simulators. The level of illumination for hardcopies was set to match exactly the luminance emitted by the SLDs, and the reflectance factor of the background was set to 0.20 that of the white border. Overall, the absolute colorimetric parameters for both media were equated as carefully as possible.

Colour Reproduction Framework Performance

Softcopies in these experiments were displayed on two *Lacie electronblue IV* 19" CRT monitors at a 100 dpi resolution. The performance of the generated GOG model[4] were typical of standard CRTs (average: 0.9 ΔE^*_{ab}, maximum: 1.9 ΔE^*_{ab}). Hardcopy reproductions were made using a *Xerox Phaser 7300* printer on *Xerox*'s *Glossy Coated Paper* substrate. A 5th order polynomial regression[4] was used to characterise it, and its colorimetric performance was higher than that of the CRT (average: 3.0 ΔE^*_{ab}, maximum: 8.8 ΔE^*_{ab}), but a fairly typical result for printers. The overall performance of the cross-media reproduction framework was assessed by reproducing a set of 59 colours contained in the intersection of both media's gamuts, and measuring each of them on both media under experimental conditions. The results (average: 3.1 ΔE^*_{ab}, maximum: 9.2 ΔE^*_{ab}) being not significantly worse than the accuracy of the printer, the quality of the built framework was therefore considered as adequate. Given the current reproduction abilities of printers, it would be illusory to expect to implement a reproduction framework where all errors would lie within the 3 $\Delta E*_{ab}$ limit that is usually considered as the just-noticeable-difference observable for pictorial images.[5]

Gamut Mapping Algorithms

The accuracy and pleasantness of three well established GMAs, GCUSP,[6,7] LLIN[8] and MINDE,[9] were evaluated in this set of experiments. All gamut mapping computations were performed in the CIECAM02 *Jab* colour-appearance space.[10] The colour gamut of both media being very different, the target gamut was restricted to the intersection of each medium gamut, in order to make the comparison between those media meaningful. It is important to remember that the GMAs are merely used as a tool for testing the softproofing paradigm. Their absolute performance is of no value in this specific study.

Test Images

Pursuant to Morovic and Wang,[11] twelve images covering a wide range of image content and colorimetric characteristics were selected. The two main classes of stimuli, i.e. pictorial images and business graphics, were almost equally represented, as this set of experiments aims at encompassing the widest range of applications possible. The sampling was performed so as to incorporate as many image types as those identified by the experimental guideline published by the CIE TC 8-03.[12] For instance, the *cou* image incorporates many low key and heavy cast components destined to stress the abilities of both media to reproduce very dark colours.

Psychophysical Methods

A category-judgement technique was used in the first experiment to evaluate the GMAs accuracy. Observers were asked to judge the colorimetric precision of the reproduction, for both the hardcopy and its simulation, on a scale from -3 to +3. Some preliminary sessions were performed with some observers in order to determine two stimuli that were systematically categorized respectively among the best and worst. Each observer was subsequently shown those two stimuli before each session in order to calibrate their answers. A pair-comparison method was used in the second experiment comparing the pleasantness of the reproduction generated by each GMA. In both cases, observers' answers were converted into an interval scale according to Torgerson's law of categorical judgement[13] and Thurstone's law of comparative judgement.[14] 16 and 14 colour-normal observers participated in the first and second experiments respectively. For both experiments, three sessions were carried out in which all stimuli or stimuli pairs were shown in a random order in order to test repeatability. Observer repeatability was estimated in terms of the coefficient of variation for the accuracy experiment and the percentage of wrong decisions[15] for the preference experiment.

Results and Discussion

Overall Results

The overall scale values obtained by each GMA for the first and second experiment are plotted in Figure 3 and 4 respectively. Since the scale values obtained according to the two psychophysical techniques employed are not on an absolute scale, it is not reasonable to compare directly the absolute scale values from reproduction workflow 1 (light gray bars) with those from workflow 2 (dark gray bars). However, it is the correlation of the relative performances of the GMAs *between* the two workflows *within* each experiment that should be assessed. From these results, it can be seen clearly that the same overall conclusion can be made from both workflows. That is, in each experiment, the relative difference between each GMA's scale values is approximately the same regardless of the target medium used, as confirmed by the very high coefficient of determination R^2 between the scale values obtained for both workflows in each experiment (Tab. 1 col. 1). The preference experiment clearly differentiates between the performance of each of the GMAs. Furthermore, for each experiment, the rank order of the algorithms is almost the same for both workflows. The only minor discrepancy comes from the accuracy experiment results, where the GCUSP and LLIN GMAs are not significantly different for the CRT-CRT case, whereas they are just different for the second workflow. Nonetheless, this difference has the merit to highlight the only noticeable discrepancy between the two workflows, i.e. that the differences between judgements made on the basis of a hardcopy simulation (Figure 3) tend to be less spread than those obtained with real hardcopy stimuli (Figure 4). The relatively poor precision of printers, in terms of colour reproduction, may explain the higher dispersion present in the hardcopy workflow. Nevertheless, the general conclusion that can be unambiguously drawn from these results is that, providing that the viewing conditions are very carefully equated and that a significant number of test images is used, softproofs are suitable as surrogates for the final print in judging the quality of colour reproduction.

Another major goal of this study consisted in investigating whether the relative performance of the GMAs for both workflows was influenced by the type of images used.

Figure 5-A and 6-A shows the resulting scale values for business graphics for the accuracy and the preference experiment respectively, while those for pictorial images are illustrated in Figure 5-B and 6-B.

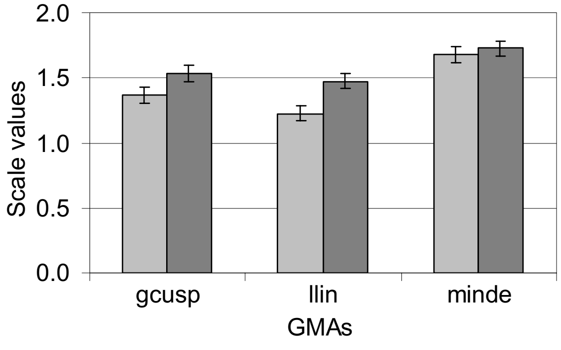

Figure 3. Accuracy experiment: average accuracy scale values of individual GMAs for CRT-print (light gray) and CRT-CRT (dark gray) workflows.

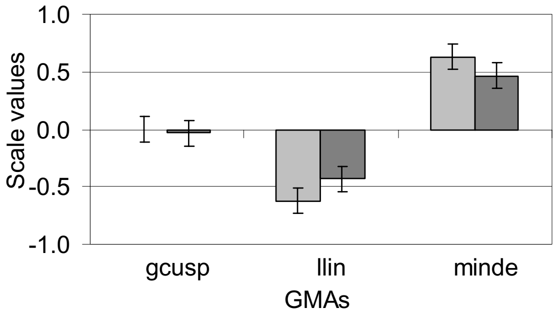

Figure 4. Preference experiment: average pleasantness scale values of individual GMAs for CRT-print (light gray) and CRT-CRT (dark gray) workflows.

Image Types vs. Individual Images Results

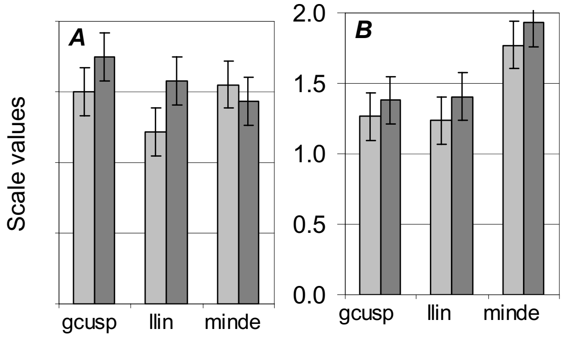

Figure 5. Accuracy experiment: average accuracy scale values of individual GMAs for business graphics (A) and pictorial images (B) (same colour code as Fig. 3)

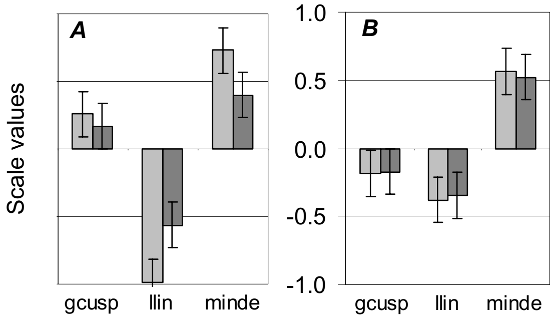

Figure 6. Preference experiment: average pleasantness scale values of individual GMAs for business graphics (A) and pictorial images (B) (same colour code as Fig. 3)

Table 1. Correlation between 1^{st} and 2^{nd} workflows for both experiments for different stimuli set.

	Overall	Pictorial Images	Business graphics	Business / outliers
Accuracy	0.998	0.996	0.492	0.984
Preference	0.996	0.989	0.980	0.987

From these results, the high correlation between the two workflows for the pictorial images is easily observable (Tab. 1 col. 2), which strongly confirm the conclusions previously drawn. However, business graphics images do not show this tendency. Their results for the second experiment clearly illustrate the origin of the higher dispersion of scale values previously observed in the print workflow. Their results for the first experiment not only confirm this fact, but also suggest that the overall GMAs ranking differs depending on the workflow used, as the rank of LLIN algorithm is significantly different between the two workflows for the accuracy experiment. This is also confirmed by the weak correlation observed (Tab. 1 col. 3). If that was the case, judgements based on softcopy simulations of a print would not be in agreement with those obtained with real hardcopies.

The analysis was further pursued by computing the scale values obtained by each GMAs for every test image, rather than image types. By comparing the ranking order and also the coefficient of determination of the GMAs between the two workflows for every individual image, it was found that a few did not concur with the overall results. For the accuracy experiment, two business graphics images, *pollution* and *air*, and one pictorial image, *cou*, generated markedly different results between the two workflows. Similarly, for the preference experiment, only a single image, *pollution*, was found to behave in a different way. Since most of them belong to the business graphics type, they might be responsible for the odd behaviour exhibited by this type of images. Indeed, removing those outliers from the business graphics dataset brought its results back in agreement with the overall conclusion, as Figure 7 illustrates. The correlation between the two workflows also increased dramatically (Tab. 1 col. 3).

Overall, there was a 75% agreement in terms of ranking order between the two workflows for the accuracy experiment (60% for the business graphics stimuli, and 86% for pictorial images) and 92% for the preference experiment (80% for the business graphics stimuli, and 100% for pictorial images). Those results confirm once again that judgements made on the basis of pictorial images are very robust, and do not seem to suffer from any loss when simulations are used instead of hardcopies. On the other hand, business graphics images are more prone to discrepancies, although the majority of stimuli seem to concur overall.

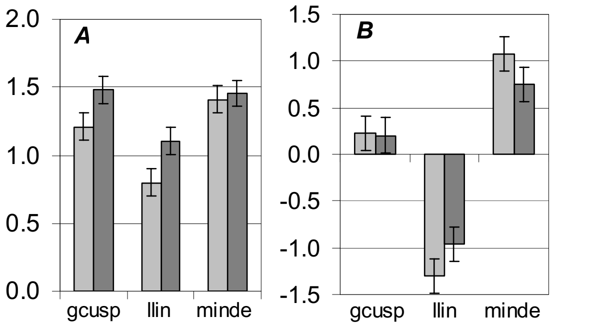

Figure 7. Business graphics images: average accuracy (A) and preference (B) scale values of individual GMAs for both experiments (same colour code as Fig. 3)

Several observers whose results concur with the observed discrepancies between the two workflows were asked to describe more precisely the reasons of their choices regarding the incriminated images. *cou*, an African sunset, and the only pictorial image exhibiting a different behaviour, really put the colour reproduction abilities of both media under serious stress, as it contains very dark areas alongside highly saturated sky colours. *pollution,* the only image to exhibit problems in both experiments, mainly consists of a single and saturated yellowish-red cast, which both media did not manage to reproduce in the same way, the printer giving emphasis to the yellow part while the CRT reproduced red better. The problem encountered in the a*ir* image resulted from the large and smooth red-to-green gradient it contains, which the lack of uniformity of the printer did not allow to render properly. A better printer may certainly have helped to improve the overall agreement of the results.

Despite the observed discrepancies, the overall conclusion still holds that judgements based on softproofs are transferable to print. However, business graphics images have a natural inclination to diverge, as their characteristics make them more prone to highlight the intrinsic differences and abilities of the reproduction devices of interest, and thus affect the judgements derived from such stimuli. Great care must therefore be taken when reporting performance based on this type of stimuli. A consequent stimuli set is not superfluous in this case.

Observer Repeatability

Table 2. Intra and Inter-Observer Repeatability

	Exp 1 (CV)		Exp 2 (%WD)	
	Print	CRT	Print	CRT
Intra	33.2	25.2	33	39
Inter	28.8	30.3	39	44

Table 2 indicates intra-observer and inter-observer repeatability in the experiments. For both measures of repeatability, a value of zero would indicate a perfect agreement. The relatively high magnitude of those metric may seem odd at first, but the built apparatus allowed having more than one reproduction pathway, i.e. each media could be displayed at will on the right or the left hand side. The display side was thus alternated for one out of the three sessions. Although it was attempted to equalise each configuration as accurately as possible, certain discrepancies remained between the two sides, which probably slightly contribute to increase the apparent dissimilarities between the two media.

The intra- and inter-repeatability for both workflows are otherwise remarkably similar. The values for the workflow involving prints tend to be slightly lower than for the CRT one, but this is to be expected as a lower colorimetric precision entail a higher discriminability between differences. However, the abilities of softproofs to replicate judgements made on the basis of hardcopies is successfully verified in terms of repeatability, both intra- and inter-observer.

Conclusions

The purpose of this article consisted in assessing the suitability of softproofs as a surrogate for the final print in judging colour quality. A complex viewing apparatus which has the double advantage of hiding all cognitive cues and equating the surround conditions very accurately was specially designed to meet the stringent requirements of this study. Two experiments targeting judgements related to colour quality were carried out twice, first for collecting decisions involving actual hardcopy stimuli, and second with softcopy simulations of those hardcopies. Overall, the general conclusion that can be drawn is that judgments made on the basis of softproofs are transferable to prints. However, some nuances need to be added to this statement depending on the type of stimuli used. While results based on pictorial images are very robust, the intrinsic characteristics of business graphics make them more prone to highlight the intrinsic differences and abilities of the reproduction devices of interest, and thus affect the judgements derived from such stimuli. This new set of results seems to contradict those previously obtained by MacDonald *et al.*[1]. However, this disparity may be explained by some fundamental differences in the experimental set-up, such as the colorimetric characteristics of the physical apparatus employed to illuminate hardcopies (viewing booth *vs.* specially designed environment) or the degree of accuracy achieved by the characterisation process. Future work will attempt to determine which parameters of the strict setup can be relaxed and how their relaxation could be compensated if the results are not consistent. The rest of this series of experiments aimed at determining whether observers can distinguish the media used if every cognitive cue is hidden and both stimuli and viewing conditions are colorimetrically equated. Results will soon be submitted.

Acknowledgment

We wish to thank first of all the Xerox Foundation for their financial support for this work. Many thanks go to Dr. Ján Morovi for his valuable help in designing the experiment. Last but not least, we would also like to thank all the observers at the Colour and Imaging Institute for their tremendous contribution.

References

1. MacDonald L. W., Morovic J. and Xiao K., A Topographic Gamut Mapping Algorithm, *Colour Imaging Science: Exploiting Digital Media*, MacDonald L. W. and Luo M. R. (eds) , John Wiley & Sons, 291-317 (2002).
2. Alessi P. J., An Update on Colour Appearance Model Evaluation for Hardcopy/Softcopy Image Comparison, *CIE Expert Symposium '96 on Colour Standards for Image Technology*, 183-186 (1996).
3. Hunt R. W. G., *Measuring Colour*, 3rd ed., Fountain Press (1998).
4. Balasubramanian R., Device Characterization, *Digital Color Imaging Handbook*, Sharma G. (ed), CRC Press (2003).
5. Stokes M., Fairchild M. D., and Berns R. S., Colorimetrically Quantified Visual Tolerances for Pictorial Images, *ISCC/TAGA Comparison of Color Images Presented in Different Media*, vol. 2, 757–777.
6. Morovic J, To Develop a Universal Gamut Mapping Algorithm, *Ph.D. Thesis*, University of Derby (1998).
7. Braun G. J. and Fairchild M. D., Image Lightness Rescaling Using Sigmoidal Contrast Enhancement Functions, *Journal of Electronic Imaging*, 8/4, 380-393 (1999).
8. Johnson A. J., Perceptual Requirements of Digital Picture Processing, *IARAIGAI symposium* (1979).
9. Morovic J. and Sun P. L., Non-Iterative Minimum ΔE Gamut Clipping, *Proc. of IS&T/SID 9th Color Imaging Conference*, 251-256 (2001).
10. CIE TC 8-01, A Colour Appearance Model for Colour Management Systems: CIECAM02, *Technical Report* (2003).
11. Morovic J. and Wang Y., Influence of Test Image Choice on Experimental Results, *Proc. of IS&T/SID 7th Color Imaging Conference*, 143-148 (2003).
12. CIE TC 8-03, Guidelines for the Evaluation of Gamut Mapping Algorithms, *Technical Report* (2004).
13. Torgerson W. S., A Law of Categorical Judgment, *Consumer Behaviour*, Clark L. H. (ed.), New-York University Press, 92–93 (1954).
14. Thurstone L. L., A Law of Comparative Judgment, *Psychological Review*, 34:273–286 (1927).
15. McLaren K., Colour Passing – Visual or Instrumental?, *Journal of Society of Dyers and Colourists*, 86, 389-393 (1970).