What Differences Do Observers See In Colour Image Reproduction Experiments?

Pei–Li Sun and Ján Morovic
Colour & Imaging Institute, University of Derby
Derby, United Kingdom

Abstract

A large number of psychophysical experiments have been conducted in which observers judged the quality of reproduction either in terms of accuracy or pleasantness. However, all that these experiments say is how well a set of approaches that was used for reproducing some originals performs. What would be more important is to understand in what way the various colour reproduction methods fail so as to attempt an improvement in those areas. The aim of this paper is therefore to try to understand what factors contribute to judgements made by observers in experiments where they judge the quality of colour reproduction. Having an understanding of these then provides a new kind of basis for developing better colour image reproduction solutions. To this end the present paper describes the experimental method, data analysis and initial results of a psychophysical experiment where observers were asked to identify what differences they saw between a range of reproductions and their corresponding originals.

Introduction

Trying to reproduce the appearance of an image on an imaging medium capable of reproducing only a smaller colour gamut than that of the original image is a challenge that has been around for a considerable amount of time.[1] Even though a large number of solutions to it have been proposed, each of them tends to work well only under some circumstances. For example, some gamut mapping solutions are particularly suited for making printed reproductions of originals present on transparencies whereas others work well when trying to match images using various printing technologies. Then there are gamut mapping algorithms (GMAs) that work well for certain images but not for others and consequently a fully automatic cross–media colour image reproduction system is still a utopia.

The Structure of Gamut Mapping Studies

Looking at the way in which GMAs have been developed shows that in the vast majority of cases researchers start with some idea of how to change an image's colours so as to make them fit a reproduction imaging medium's gamut. In many cases this initial idea is based either on experience from doing colour reproduction in a trial–and–error way or on assuming on *a priori* theoretical grounds – that the reproduction ought to have certain properties (e.g. that there should be a certain balance between lightness and chroma changes applied to the original). Some other studies first try to understand how observers would gamut–map images and they then attempt to model their behaviour. Understanding how observers would gamut–map images can either be done by looking at what colour reproduction professionals do with the tools used commercially[2,3] or by developing tools that allow images to be modified in terms of some of their appearance parameters and then getting naïve observers to make adjustments to reproductions so as to make them more accurate or pleasant.[4–6]

Evaluation of Colour Reproduction

What most studies then have in common is that they implement their idea of how gamut mapping should be done and test its success by psychophysical means. These psychophysical experiments are most often pair comparison, category judgement or ranking ones and in them a group of observers is asked to make a judgement about the extent to which a given reproduction exhibits a certain property. Most often the criterion for judgement is the reproduction's similarity to an original or its pleasantness in isolation.

As can be easily appreciated, the task of making such judgements about the overall accuracy or pleasantness of reproduction is not a simple one and the final judgement will depend on a number of individual factors that are in the end (even if unconsciously) given relative weights. Furthermore, the results of these psychophysical experiments merely give information about the performance of the various colour reproduction strategies tested but they do not suggest ways of improving them.

The aim of this paper is therefore to present results of work carried out to understand what it is that observers take into account when making judgements about the quality of colour reproduction. Having an understanding of this will then provide an alternative basis for developing gamut mapping algorithms which can attempt to improve performance in terms of the factors given most importance by observers.

Experimental Method

The key dilemma in designing psychophysical experiments is that between not making the task sufficiently clear to observers (which can result in a great deal of inter–observer variation) and by specifying it too narrowly (in which case the results are virtually implicit in the instructions). As in this case the ultimate aim was to understand what observers thought when making judgements about colour reproduction, an approach was taken that might err on the side of giving instructions that are too vague. While this makes subsequent analysis more difficult it has the potential to understand more deeply the factors affecting observers in this context.

A psychophysical experiment was therefore set up in which observers were shown an original on a CRT display and a number of printed reproductions of it in a viewing booth (whereby they were always presented with a pair of these so as to make the conditions more similar to a pair–comparison experiment already carried out using the same stimuli).[7]

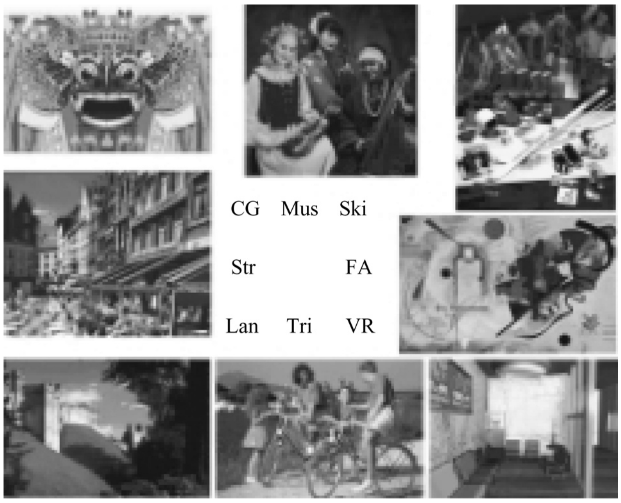

Figure 1. Test images used in experiment.

Eight test images were used as originals in this study (Figure 1) and each of them was reproduced using the following four gamut mapping algorithms: CARISMA,[8] GCUSP,[8] SKNEE[5] and WCLIP which is a weighted minimum ΔE gamut clipping algorithm. Fifteen colour–normal observers then participated in the experiment according to the following instructions which were read to them:

> *In this experiment, you will be shown an original image on a monitor and two printed reproductions of that same image. You will be asked several questions about them.*
> *The questions will be:*
> *1. What differences do you see between the original and the left (or right) reproduction? Each of the difference can be either for the whole image or just for some part of it.*
> *2. Now, I would like to ask you about the importance of the various kinds of differences you just listed. Please tell me which of them is the most important one?*
> *3. Please make a judgement about the other differences on a scale where 5 represents this most important difference and 0 represents differences of no importance.*

As can be seen, the task for the observers consisted in first identifying the differences they saw between the original and a reproduction and then making a judgement about how important these were in a way that made this judgement relative to a single reproduction. Here the identification of differences can lead to an understanding of their perceptibility (based on looking at the frequency of reporting particular differences) whereas the importance judgements can tell us about the acceptability of the various types of difference.

The reason for making the importance judgement in a relative way is that in a small–scale pilot experiment observers had difficulty with making absolute importance judgements. To make the task easier for the observers, it was therefore decided to make the importance judgements relative to a given image and observers, which presented no difficulties to the observers.

In the second part of the experiment, observers also performed a category judgement of how close the various reproductions were to the corresponding originals. In this experiment, observers were read the following instructions:

> *For each of the images you will be shown we would like you to tell us how accurate a reproduction of the originals you think it is. Please give your opinion on a scale of numbers from one to nine where one represent the most accurate image and nine represent the least accurate image you can think of. Use numbers between one and nine to represent equal intervals of accuracy so that the difference between any neighbouring categories should be the same.*

This category judgement data served two purposes–firstly, to see how closely observer judgements under the present experimental circumstances were related to the previous pair comparison experiment and secondly, it could be used for weighting the responses from the first part of the experiment so as to give differences that were present in images that were further from the original more weight.

Data Analysis

The data gathered in the above experiment is rich in information about various aspects of the cross–media colour image reproduction process and it will be analysed so as to shed light on a number of issues. The following is a discussion of some of the challenges that need to be met when attempting to analyse and summarise the results of this experiment.

Contribution From Individual Observers

In the first part of the experiment, the observers were given freedom to indicate the image differences for either the whole image or some part of it. One challenge in analysing and summarising the data comes from some observers mainly providing overall results and others providing results for many parts of an image. Hence if the data is summarised by simply adding up the results from each observer, those who listed a greater number of differences will contribute more to the conclusions. On the other hand, it might in fact be desirable to have a stronger contribution to the overall results from those who have reported more differences as it might be that the ones who listed fewer just did not perform the task as conscientiously as others.

So it seems that a compromise might be the solution, by using a function to initially equalise the data. For example, the scores given by an observer for a reproduction could be summed first (N_o) and then converted to a new, equalised scale via a logarithmic function (Equation 1):

$$I_e = \frac{N_e}{N_o} \cdot I_o \text{ where } N_e = 10 \cdot \log_{10}(N_o) \qquad (1)$$

where the *e* and *o* subscripts refer to equalised and original data respectively and *N* and *I* are the sum of scores and importance value respectively.

By means of this equalisation, the ratio between maximum and minimum contribution from different observers for each reproduction was reduced from 4.17 to 1.88 on average, and the Chi-square values of the data distributions in z-scale against a standard normal distribution were reduced from 77.25 to 32.68 where 37.65 is the threshold for $\alpha = 0.05$ (i.e. the 95% confidence level). As the equalised data was now more like a normal distribution, taking its "mean" will also be a better measure of the central tendency of observer opinions.

Figure 2 shows examples of two observers' responses, the left observer reported only one difference ("pale") with an importance of 5 (i.e. $I_o = 5$). For this observer the equalised score will be 7.0 (i.e. $I_e = 7.0$). The right observer, on the other hand, indicated 8 differences, each having an importance of 5 (i.e., $N_o = 40$). The equalized score for each of the items will be 2.0 (i.e. $I_e = 2.0$). The equalised total contribution N_e for the two observers therefore will be 7.0 and 16.0 respectively and this is a compromise between ignoring the fact that one observer saw more differences than the other and an attempt to give individual observers similar influence over the final result. As the maximum No in the data gathered in this experiment is 45, the difference of contribution by individual observers will not be very large after the above equalisation.

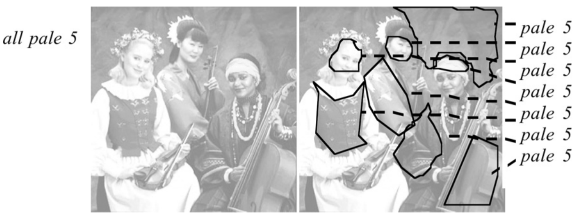

Figure 2. Example of two observers' responses.

Terminological Translation

As was to be expected, observers used various terms for describing the differences they saw between the original and reproduced images. On the one hand, there are some that have similar meaning (e.g., "blur" and "loss of details") and these can subsequently be grouped. On the other hand, some terms indicate a combination of a number of more basic differences (e.g., "pale" can be translated as the change of both "lightness" and "chroma") so that these can be separated for further data analysis.

In order to analyse image differences in general and also in terms of individual aspects, some of the term–categories need to be merged or separated. In the case of merging, one can use a general term to cover several relative terms for reducing the total number of categories. For example, "too light" and "too dark" can be merged as "lightness difference" (L diff.). In terms of separation, for instance, one can divide "colour difference" into "lightness" (L), "chroma" (C) and "hue" (H) differences.

However, a question then arises about how to transfer the importance scores into the separated or merged sub–categories. As there is no information from the observers about how to share the score into the sub–categories, they will be treated equally here. For example, to separate an importance score of 5 for a "colour difference" being reported by an observer, a score for each of the L, C and H differences would be (5/3).

Table 1. Three term-levels for data analysis.

Level	Level 1 (raw)	Level 2 (relative)	Leval 3 (general)
Term	Colour difference	as left	L, C, H diff.
	L > (too light)	as left	L diff.
	L < (too dark)	as left	L diff.
	C > (higher chroma)	as left	C diff.
	C < (lower chroma)	as left	C diff.
	H (hue shift)	as left	H diff.
	Pale	C < and L >	C and L diff.
	Greyish	C < and L >	C and L diff.
	Faded	C < and L >	C and L diff.
	Not self-luminance	C < and L <	C and L diff.
	Contrast < (higher)	as left	Contrast diff.
	Contrast > (lower)	as left	Contrast diff.
	Less depth	Contrast <	Contrast diff.
	Detail > (too much)	as left	Detail diff.
	Detail < (loss)	as left	Detail diff.
	Blur	Detail <	Detail diff.
No.	16	10	5

Three term-grouping levels have been set and shown in Table 1. Level 1 contains 16 raw terms that were directly what observers reported and in Level 2 the number of items is reduced to 10 by using the method mentioned above. In this level, some items are the same in attribute (e.g., L) but different in direction (e.g., L< and L>). The number of items is reduced to 5 in Level 3 by merging those items having the same attribute (e.g., L) but different direction.

Conversion To Absolute Scale

The results produced by the above processing only give the relative importance of various image differences (i.e., their values are normalised within a given reproduction). As the worst difference in a good reproduction is less important than that in a poor reproduction, the absolute importance of image differences is of more interest when trying to improve colour reproduction. To convert the relative results into an absolute scale, the absolute judgements made in the category judgement (CJ) part of the experiment will be subtracted by one (as category one represents zero difference) and then used as weights.

A mean-category-value method[9] was then used for the data analysis of the CJ data and visual importance of image difference (ΔV_i) is defined as the mean CJ value minus one. A ΔV_i value of zero means no important difference between original and reproduction and larger values of it mean that the importance of difference is increasing. The results of ΔV_i values for the CJ experiment are shown in Figure 3.

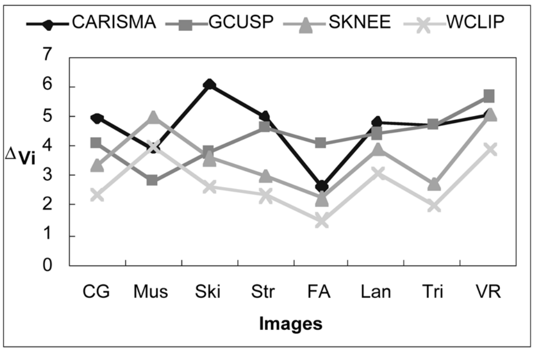

Figure 3. Mean ΔV_i of 4 GMAs for 8 images.

Looking at Figure 3 and the thumbnails of the test images shown in Figure 1, for example, suggests that images having large numbers of dark and saturated colours are more likely to result in higher ΔV_i.

Data Summary

As the ΔV_i values obtained from the category judgement experiment are the total importance of differences between an original and its reproduction, the ΔV_is of individual objects can be obtained by dividing the total .Vi in proportion to the importance judgements made for individual objects within the image (Equation 2).

$$\text{object's}\,\Delta V_i = \left(\frac{\text{total}\,I_e\ \text{of the object}}{\text{total}\,I_e\ \text{of the reproduction}}\right)\cdot \text{reproduction}\,\Delta V_i \quad (2)$$

Questions About Present Data Analysis

While choices in the above method of analysing the data were made with the intention of having results that best represent the overall judgements made by the group of observers who took part in the present experiment, it is very clear that each one of them is one of many alternative choices. As a result of this it is easily possible to object to each one of the approaches taken above and to say that, for example, the way of combining data from individual observers should have been done differently. Clearly it is not possible, or indeed desirable, to dispute this and instead the response here is to acknowledge that alternative ways of analysis are possible and that the main purpose here was to identify the above four issues that need to be addressed for the present kind of data and to propose a possible way of dealing with them.

An alternative objection that might be raised it is to suggest that the present experiment should only have been used as a pilot one for the purpose of identifying categories of differences and that a more rigidly structured experiment could then be performed that would yield data that could be analysed in an established way. While this is certainly the case, the present approach was preferred as it results in data that involves a smaller degree of uncontrolled deliberation than would be the case if observers were presented with categories that were not their own.

In summary, the present approach tries to extract an overall picture from data that more closely relates to the views of observers but for which there are a number of alternative ways of analysis.

Overall Results

The ΔV_i values of objects in the test images used here could be summarised in many ways. To see the performance of the four GMAs in general, Level 3 terms will be used for summarising global and local ΔV_is. The relative visual differences of lightness (L), chroma (C), hue (H), contrast and detail for the four GMAs together with the overall results are shown as pie charts in Figure 4 and a bar char in it also shows the mean ΔV_is of the four GMAs to provide an idea of importance in absolute terms.

In the top four pie chars, white areas represent the "global differences" and the grey ones represent "local differences". As can been seen, CARISMA's problems were mainly in terms of local lightness and hue differences, GCUSP suffered from loss of global and local chroma and both SKNEE and WCLIP show similar features having relatively larger problems with loss of detail. Both GCUSP and SKNEE intend to be hue-preserving GMAs and they should therefore have no "H difference." However, the fact that they do can be explained as appearance models are not hue constant in their colour spaces, as there are characterisation errors of the media involved and as observers might not have a clear idea of the components of a large colour difference.

Referring to the overall (i.e. combined global and local) results, it is clear that existing GMAs have problems mainly in terms of how they treat lightness and chroma. By combining lightness, chroma and hue as "colour difference", the importance of "colour difference", "contrast" and "detail" is shown to be 84%, 12% and 4% respectively. This in turn suggests that GMAs should be improved first by focusing on colour difference, then on preserving contrast and finally on maintaining detail. These findings also agree with our previous study into the role of 3D colour histograms in colour reproduction[7] as well as with a study on the importance of colour in image quality.[10]

At this stage it is also important to note that it might be impossible to reduce some of the differences identified here by observers due to the inherent limitations of the medium difference between the original and reproduction media used here.

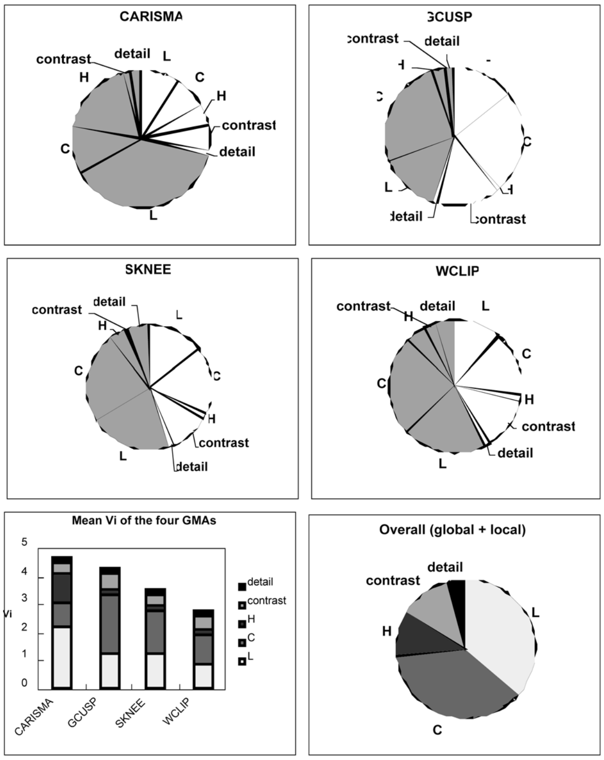

Figure 4. Individual and overall visual importance of differences for 4 GMAs.

Properties of Local Differences

As has been shown in Figure 4, more than 50% of the important errors were perceived in parts of images rather than for entire images. This results in a need for understanding the properties of the problematic areas in originals so as to model them for improving GMAs. The following therefore is an investigation of some of these properties.

Location and Content

The summations of local ΔV_i for L, C and H of the four GMAs are shown as error images in Figure 5. In the images, the brighter the region, the higher the local ΔV_i. Referring to the error images and their corresponding originals in Figure 1, it seems location within an image is not a factor influencing the attention of observers. Instead large uniform backgrounds, especially sky, are more likely to be a region where observers identify important visual differences.

Surprisingly areas of flesh tones did not show important differences in the present study and there are at least two possible reasons: first, skin colours did not occupy large areas in the test images used here and second, the colours were not very chromatic and thus changed only slightly as a result of the gamut mapping process.

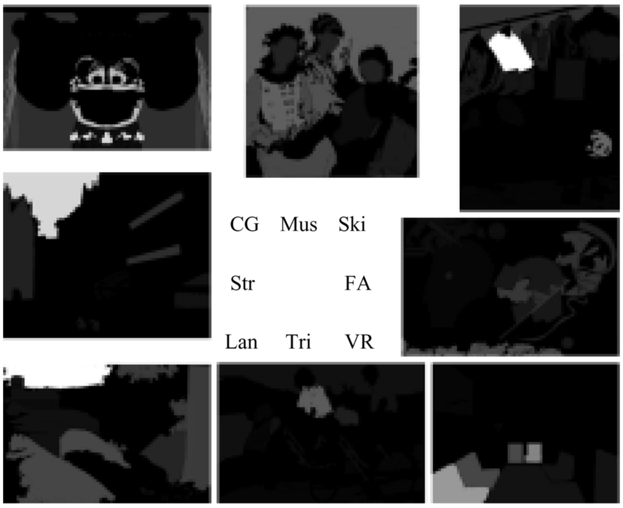

Figure 5. Visual importance for local regions.

Colour Distribution

Since these GMAs mainly change image colours in high chroma and dark regions, the colours of local error areas were, as expected, precisely from those colour regions. Large errors can also be perceived for light green in particular (refer to the green jacket in the Ski image and the green chair in the VR image). The reason for this phenomenon is that the medium gamuts differ greatly in this part of colour space and no GMA can overcome this inherent difference.

Spatial Frequency

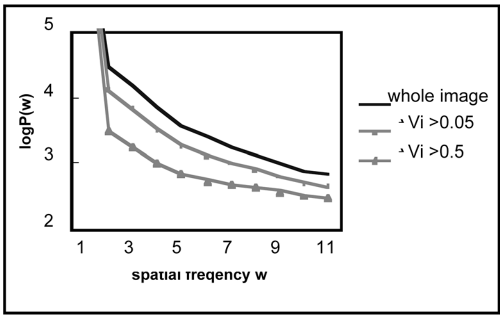

Figure 6. logP(w) spectrum for global or local regions.

Another property to take into account are the spatial characteristics of the test images and it is therefore of value to determine whether the regions of local differences have different spatial characteristics from the whole image. To make this kind of comparison, the LogP$_{Jab}$[7] metric was used, whereby it expresses the overall energy in an image's power spectrum of the spatial frequency domain based on 16 x 16 pixel blocks.

Furthermore, local error regions of two levels of importance were considered in turn whereby these had ΔV_i values larger than 0.05 and 0.5 respectively. The former level of importance was the case for about 34% of image area and the latter for about 8%. Results for these two levels of importance for local differences had logP$_{Jab}$ values of 3.46 and 3.31 for the $\Delta V_i > 0.05$ and $\Delta V_i > 0.5$

cases respectively, compared with a $\log P_{Jab}$ value of 3.69 for the entire image.

The logP(w) spectra, which show how much energy there is at different spatial frequencies (w), are shown in Figure 6 for the two types of local region as well as for the entire image. There it can be seen that power in higher frequencies is reduced for image regions that have more important differences. This shows that important local regions have lower spatial frequencies and this phenomenon should be taken into account in both GMAs and image difference formulæ.

Object Size

The size (area) of objects with local errors has been determined for the two importance levels used in the previous analysis - $\Delta V_i > 0.05$ (important regions) and $\Delta V_i > 0.5$ (very important regions). The size of each object was reported in terms of percentage of image area and the results are shown in Table 2. As can be seen, the variation of object size were very large, with a median of 4% of image area. The fact that the minimum is relatively small also suggests that for an object to have important differences, it does not necessary have to be large.

Table 2. Size Error Regions.

Size & Shape	Perc. of image area	
ΔV_i	> 0.05	> 0.5
Min.	0.4%	0.9%
Median	4.1%	4.2%
Max.	24.6%	20.3%

Conclusions

A new approach to understanding the cross–media colour image reproduction process has been described in the present paper, including a look at past attempts of giving colour reproduction a psychophysical basis. Then an experiment was described that resulted in information about what differences observers see between originals and reproductions and finally a number of challenges were discussed that need to be met when analysing the results of this kind of experiment.

Our initial experimental results showed that colour differences were more important than contrast and detail differences in cross-media reproduction, and that more than 50% of differences were present only in local regions, rather than being the case for entire reproductions. It was also shown that regions of important differences also have lower spatial frequencies than the entire test images used here and that the size and location of an object within an image do not have a strong impact on whether it's difference will be judged to be important.

In terms of future work, the statistics of observer judgements discussed in the present paper will be compared with colorimetric comparisons between originals and their reproductions. Having that further level of understanding will also significantly contribute to determining the implications of the present results.

Overall, the results of the present experiment will be useful not only for deriving better colour reproduction systems but also for deriving new image difference metrics which involve image analysis.

Acknowledgements

The authors would like to thank Tadashi Yoshida and Todd Newman (Canon R&D Center Americas, Inc.) for supporting this research and all the experimental observers for their patience.

References

1. J. Morovic and M. R. Luo, The Fundamentals of Gamut Mapping: A Survey, *Journal of Imaging Science and Technology*, **45**/3:283–290, (2001).
2. CARISMA, *Colour Appearance Research for Interactive System Management and Application – CARISMA, Work Package 2 – Device Characterisation, Report WP2–19 Colour Gamut Compression,* (1992).
3. P. Green and M. R. Luo, Developing the CARISMA Gamut Mapping Model, *Proc. of Colour Image Science 2000 Conference*, Derby, 244–256, (2000).
4. F. Ebner and M. D. Fairchild, Gamut Mapping from Below: Finding the Minimum Perceptual Distances for Colors Outside the Gamut Volume, *Color Res. Appl.*, **22**:402–413, (1997).
5. G. J. Braun, *A Paradigm for Color Gamut Mapping of Pictorial Images*, PhD. Thesis, Rochester Institute of Technology, Rochester, NY, (1999).
6. B. H. Kang, M. S. Cho, J. Morovic and M. R. Luo, Gamut Compression Analysis Based on Observer Experimental Data, in *Proc. 7th IS&T/SID Color Imaging Conf.*, IS&T, Springfield, VA, 295–300, (1999).
7. P. L. Sun and J. M. Morovic, 3D Histograms in Colour Image Reproduction, *Proc. of SPIE*, **4663**/9, (2002).
8. J. Morovic and M. R. Luo, Developing Algorithms for Universal Color Gamut Mapping, *Color Engineering: Vision and Technology*, L. W. MacDonald and M. R. Luo (eds.), John Wiley & Sons, UK, 253–283, (1999).
9. C. J. Bartleson, *Optical Radiation Measurement. Vol. 5 – Visual Measurements*, C. J. Bartleson and F. Grum (eds.), Academic Press Inc., 475, (1984).
10. E. D. Montag and H. Kasahara, Multidimensional Analysis Reveals Importance of Color for Image Quality, in *Proc. 9th IS&T/SID Color Imaging Conf.*, 17-21, (2001).