Preferred Color Reproduction of Images with unknown Colorimetry

Scot R. Fernandez and Mark D. Fairchild
Munsell Color Science Lab, RIT
Rochester, New York

Abstract

Observer preferences in color reproduction have been the center of debate for many years. Through a series of psychophysical experiments we are trying to better understand what observers are evaluating while looking at an image. The first experiment was a survey of observers rating the quality of thirty-seven different image characteristics that relate to the color dimensions of images. The data collected has demonstrated that observer preferences remain relatively constant while judging color attributes between media and for various types of images. Furthermore, it is apparent that the rating of mid-tone image area content and the overall image characteristics are similar. The second experiment is an international rank ordering of image quality. The goal of this experiment is to build colorimetric tolerances on observer preference and determine if cultural biases exist. The preliminary results of the experiment indicate that on average a difference in culture does not demonstrates much difference in preference.

Introduction

With the recent prevalence of digital imaging, many of the constraints of traditional imaging systems have been lifted. Unfortunately, with the newfound flexibility of digital imaging, new complexities in quantifying color quality have been generated. Often minimizing the delta E is the standard goal in understanding the limits of color quality and color reproduction in images. The intent of minimizing a delta E metric or maximizing the colorimetric accuracy between an original image or scene and its reproduction through a cross-media reproduction system is known as a colorimetric reproduction objective. A colorimetric objective will produce a reasonable reproduction, but further work is required to understand why it doesn't always produce the best reproduction of an image. For example, previous research efforts support the idea that observers would prefer object colors to be reproduced with greater saturation in comparison to the original, and that certain memory colors such as grass, skin, and sky are remembered with slightly different hues and with greater purity.[1] Furthermore, it is known that an observer maintains the ability to rate the quality of an image with or without the original image present.[2] Without the original image present observers are rating the quality of an image in reference to some psychological concept of an idealized image.[3] So the goal of our color reproduction intent should sometimes be to match the psychological concept of an image, preferred image reproduction, rather than some arbitrary image said to be the original, a colorimetric image reproduction.[4]

Preferred image reproduction techniques should be viewed as an enhanced or customized version of a colorimetric objective. Thus, when evaluating preferred image reproduction, we need to move from a delta E metric to the degree of apparent match between a reproduced image and its internal memory reference, which has been labeled as naturalness. It is commonly understood that pictorial image quality has a positive correlation with naturalness, so an image of high quality is one that has a high degree of naturalness.[2,4,5]

Experimental

The goal of this research is through a series of psychophysical experiments, to better understand, observer preferences for color reproduction of pictorial images and provide guidance for the reproduction of images from an unknown colorimetric origin. The second experiment in this series is being run in conjunction with three other research facilities around the world in order to identify any cultural biases in the psychological nature of this research.

Experiment I - Survey of Image Characteristics

The goal of the first experiment was two fold, first to determine the importance of commonly used image characteristic terms, and secondly to see if preference scales can be applied to these terms. If this can be done, then individual image characteristics can be grouped together to reflect the manner in which observers evaluate an image. Another motivation behind understanding how observers group image characteristics is to better understand the adjustment controls needed to produce a preferred color reproduction of an image.

This psychophysical experiment was conducted four times, under various viewing conditions to determine if observer preferences were maintained. The first and fourth mode utilized hard copy images created on a Kodak 8670 PS thermal printer. For the first mode, the prints were viewed in a controlled daylight viewing room, set to model standard graphic arts viewing conditions (approximately 2000 Lux at 5000° K). Our industry sponsor conducted the fourth mode of the experiment, and their experimental setup utilized various daylight viewing conditions. The second and third mode of the experiment incorporated soft copy images. The second mode was conducted on an Apple 22" Cinema Display (LCD), viewed in a darkened room, and the third mode utilized a 19" SGI CRT display, also viewed in a darkened room. Each mode of the experiment asked the observer to rate the quality of thirty-seven different image characteristics for a series of eleven different images. The thumbnails below represent the image set for experiment I.

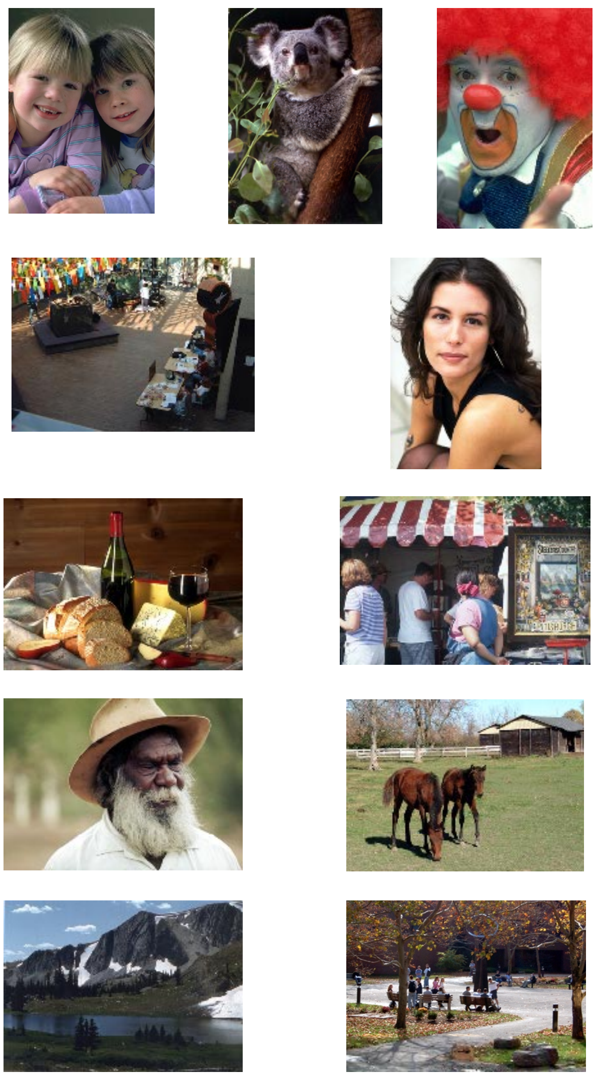

Figure 1. Image set for Experiment I – (From left to right, top to bottom 1. Kids, 2. Koala, 3. Clown, 4. Indoor Scene, 5. Model, 6. Food, 7. Art Fair, 8. Bearded Man, 9. Horses, 10. Mountains, 11. Campus

The thirty-seven image characteristics being evaluated were divided into six subgroups, each corresponding to a specific image area. The image characteristics and their breakdown into image area subgroups are presented below in Figure 2. The image areas evaluated were three groupings of Overall Image Characteristics, Shadow Image Area, Mid-tone Image Area, and Highlight Image Area.

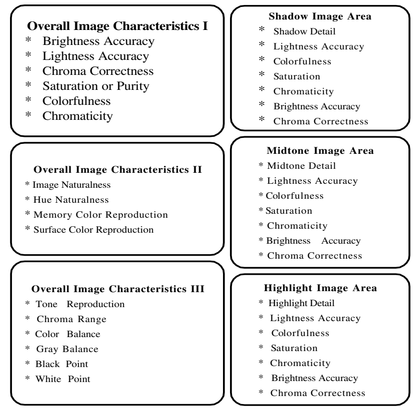

Figure 2. List of image characteristics and groupings.

The image characteristics were rated utilizing an ordinal scaling system, consisting of responses"1– 5" and "NA", 1 meaning poor, 5 meaning excellent, and "NA" meaning that the observer felt that the characteristics did not apply to the image or image area subgroup.

For modes I – III, conducted at RIT, a user interface written in IDL randomized the image order as well as the question subgroup order for each observer. The observers used the interface to record all of the ratings of each image characteristic in reference to the question subgroup. A pictorial example of the user interface for the experiment can be seen in Figure 3. The fourth mode of the experiment conducted at Xerox, utilized the same image characteristics and sub-groupings, so the only difference was that their question ordering was static.

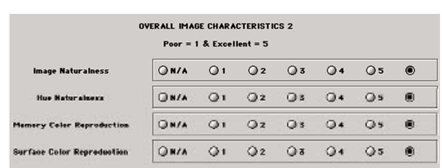

Figure 3. Example of user interface for Experiment I modes I-III.

For the first three modes of the experiment, the observer population consisted of students, faculty, and staff members, and for the fourth mode at the observer population consisted of employees of Xerox. Table 1 presents the breakdown for each observer group.

Table 1. Breakdown of Observer population for each mode of Experiment I

Mode	Hard Copy (RIT)	Soft Copy (LCD)	Soft Copy (CRT)	Hard Copy (Xerox)
Number of Males	20	18	9	11
Number of Females	10	12	2	2
Percentage of Experienced	63.0%	66.0%	100.0%	100.0%
Age Range	20 - 45	20 - 46	23 - 38	not recorded

Experiment II - International Image Characteristic Ranked Order

The first goal of this experiment is to build tolerances in colorimetric dimensions of observer preferences for color reproduction of scenic images, and the second goal is to research if any psychological biases of rating image quality can be linked to cultural differences.

This psychophysical experiment is a rank ordering of image quality being completed at four different research facilities around the world: Chiba University (Japan), University of Derby (UK), Xerox (USA), and RIT (USA). Except for three image substitutions the image set is identical to the first experiment. The images removed were Kids, Food, and Campus, and were replaced with Harmony, Church, and Dinner. Thumbnails of these images can be seen below in figure 4.

Figure 4. Image Substitutions for experiment II (from left to right, top to bottom) Dinner, Church, Harmony

The images were adjusted along eight different CIELAB dimensions. Four of these dimensions affected Color Balance (additive shifts of a* and b*), the other four manipulations were Lightness (a Gamma adjustment of L*), Contrast (a Sigmoid adjustment to L*, with an anchor at 50.0 L*), Chroma (multiplicative adjustment to C* at a constant Hue Angle), and Hue rotation (Hue Angle rotation at a constant Chroma value). The eight manipulations were applied to the eleven images to generate eighty-eight sheets of randomly ordered six-image sheets that varied around the nominal image. Each sheet demonstrated the effect of a single adjustment applied globally, and consisted of three steps above and below the original image. The increments were clearly perceivable, but not objectionably large, and the nominal image was generally not presented. The increments used are presented in Table 2.

Table 2. Adjustment ranges and increment values for Experiment II

	Starting Value	Ending Value	Increment
Gamma adjustment	0.55	1.30	0.15
Sigmoidal adjustment	0.55	1.55	0.20
Chroma adjustment	0.75	1.30	0.11
Hue Angle adjustment	-0.07	0.11	0.03
a* adjustment	-7.50	7.50	3.00
b* adjustment	-7.50	7.50	3.00
Direct adjustment	-7.50	7.50	3.00
	-7.50	7.50	3.00
Indirect adjustment	-7.50	7.50	3.00
	7.50	-7.50	-3.00

The sheets were printed on a Fujix Pictrography 3000, at a resolution of 300 dots per inch. This printing system was characterized using a 10x10x10 LUT, and a tetrahedral interpolation technique. The printer's forward characterization was utilized to pass the RGB Images into CIELAB space, were all of the manipulations where done and the inverse characterization was utilized to convert the CIELAB images back to RGB. This workflow of starting in the printer's gamut minimized gamut issues. A pictorial representation of a print sheet from the experiment is presented below in Figure 5.

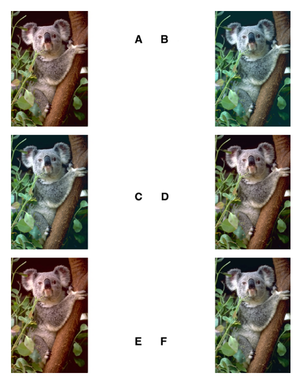

Figure 5. Sample of a sheet from Experiment II. This sheet represents an adjustment of color balance.

The observers of each sub-population were then asked to rank the sheet of images from best to worst using an online user interface. The sub-population statistics are presented below in Table 3.

Table 3. Breakdown of Observer population for each sub-population of Experiment II

Ethnic Background	Chinese	European	American	Asian	American
Testing Location	Derby	Derby	RIT	RIT	XEROX
Number of Female Observers	2	2	6	2	1
Number of Male Observers	8	8	12	5	2
Age Range of the Observers	23 - 43	22 - 39	17 - 39	28 - 31	29 - 44

Discussion

Experiment I - Survey of Image Characteristics

The analysis of this experiment was done in several stages. The first stage of the analysis utilized a multivariate hierarchical clustering technique, with a complete linkage objective, in order to group the image characteristics based on similarity of observer response. Similarity in this case is a distance measurement from one list of image characteristic response to another. This analysis was done on both the Hardcopy (RIT), and Softcopy (LCD) experiments, and for each mode the responses from all the images and observers were pooled. The cluster analysis was then repeated five times, producing an analysis at eighteen clusters, ten clusters, seven clusters, five clusters, and three clusters. Repeating the cluster analysis for different numbers of clusters made it possible to understand the balance between many clusters with high similarity versus few clusters with lower similarity between image characteristic terms, and also gave an idea of when characteristics group together. Two of the main trends that were evident from both sub-populations of the experiment were that the image characteristics term for shadow image area content grouped together early, or at a high level of similarity, and that the mid-tone characteristics grouped themselves with the overall image characteristics.

The second stage of the analysis consisted of rank ordering of the image characteristics terms based on the average magnitude of the responses. The rank order was calculated twice for each mode, first including the “NA” response valued as a 0, second and excluding “NA” response. This analysis proved to be quite confusing and difficult to interpret.

In order to simplify the analysis the third stage of the analysis only looked at the percentage of “NA” responses for each image characteristic. The first pass of this stage pooled all the observations for all four modes and all images to generate an overall idea of what was deemed important and what wasn’t. The results of this plot are presented in Figure 6.

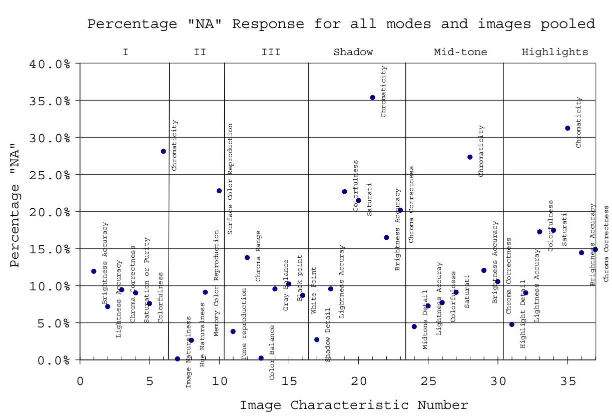

Figure 6. Percentage “NA” response for Experiment I, all responses pooled.

The second pass of the third stage of analysis tried to determine if any dependencies existed for viewing modes or images, and incorporated all observers. This evaluation of the data shared that no real image dependency, or viewing mode dependency existed for this data set. This was determined because the individual plots of each image and mode appeared very similar to average response. Next, we wanted to get an idea of how consistent our observers were. So another pass of this analysis was done on subset of the population using the same ten observers from the Hardcopy (RIT), LCD, and CRT versions of the experiment, evaluating for both image dependency and viewing mode dependencies. The results of these evaluations are below in Figures 7 and 8.

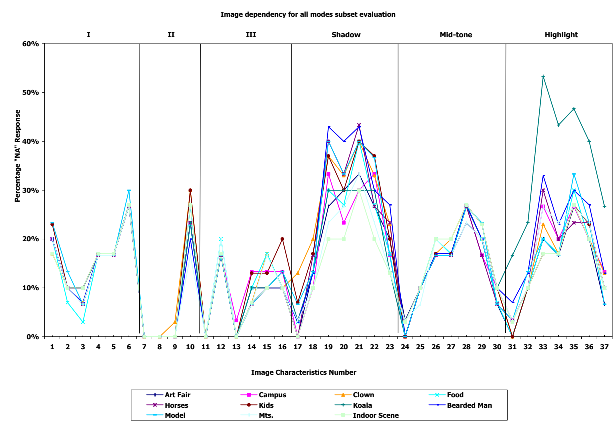

Figure 7. Evaluation of Image dependency of Percentage “NA” Response for sub-group of population.

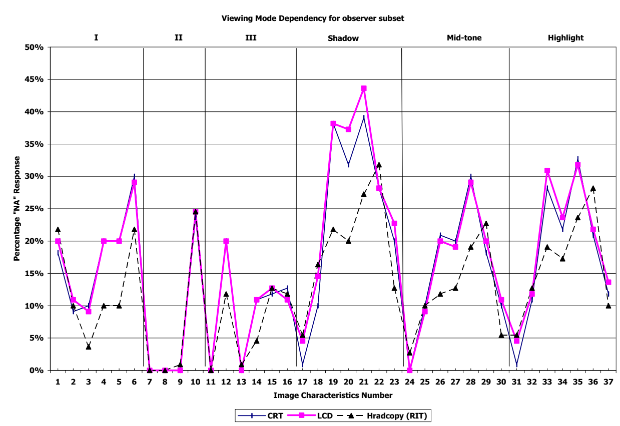

Figure 8. Evaluation of mode dependency of Percentage “NA” Response for sub-group of population.

From these plots it easy to see that no real image dependency or viewing mode dependency existed which reaffirms the findings of the evaluation on the similar analysis of the entire population. However, looking at the two plots reveals that individual observers are very consistent in their response particularly when evaluating overall image characteristics and mid-tone image characteristics. This coincides with the general trends of the cluster analysis. There is less consistency in the shadow and highlight regions, perhaps because these regions seem to be more viewing mode dependent then the other regions of interest.

Finally, the idea of rank ordering the average responses for image characteristics was redone but in a more restricted sense in that any terms with more than ten percent "NA" response from the subset population were thrown out of the evaluation. Twelve image characteristics remained after cutting at a ninety percent importance level, and Table 4 presents which image characteristics remained and the their rank ordering based on average magnitude of preference response. Table 5 presents in a percentage how similar each ranking are to each other.

Table 4. Rank ordering of important image characteristics from observer subset.

	CRT	LCD	Hardcopy (RIT)
Midtone Lightness Accuracy	2	1	1
Hue Naturalness	4	4	2
Midtone Detail	1	2	3
Image Naturalness	3	3	4
Midtone Chroma Correctness	5	5	5
Overall Chroma Correctness	6	6	6
Memory Color Reproduction	10	9	7
Gray Balance	7	7	8
Color Balance	9	8	9
Tone Reproduction	11	10	10
Highlight Detail	8	11	11
Shadow Detail	12	12	12

Table 5. Percent similarity of rank ordering of image characteristics deemed important by the observer subset.

	CRT	LCD	Hardcopy (RIT)
CRT	100%	89%	81%
LCD	89%	100%	89%
Hardcopy (RIT)	81%	89%	100%

This final step of the evaluation reiterates the idea that observers place greatest importance on overall and mid-tone image characteristics. It also reiterates that observers are fairly consistent from one viewing mode to another.

Experiment II - International Rank Ordering of Image Quality

The analysis of this data is still very much in the preliminary stages. Complete sub-populations of Americans in the US (AM-X, AM-R), Chinese in Europe (CH-D), European in Europe (EU-D), and Asians in the US (AS-R) have been obtained. So far the data on average supports the idea that little difference between ethnicity exists, however results from observer to observer within a population are quite variable. These data are also very image dependent. Presented below in Figures 9 and 10, are the population average response for all images that had their Gamma manipulated, and Chroma manipulated. For the evaluation the most preferred image was given a value of 6, and the least preferred image was given a value of 1. The nominal image, if present for these ramps, would fall between steps 3 and 4.

The Gamma Adjustment plot in Figure 9, supports the idea that the different ethnicities appear to behave similar on average, and that the image set had reasonable tone reproduction. Figure 10, the plot of Chroma Manipulations demonstrates that all sub-populations on average wanted to see the image set with more chroma. Furthermore it is also interesting to note the order of peek preference for the chroma attribute of an image. According to these results the farther east that the observer is from the more chroma they prefer.

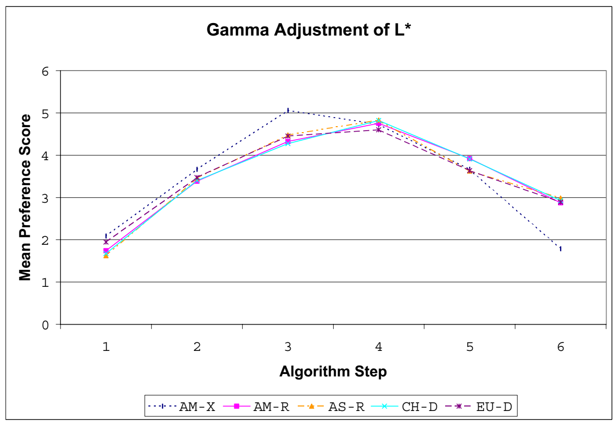

Figure 9. Average response of each sub-population for Gamma manipulation

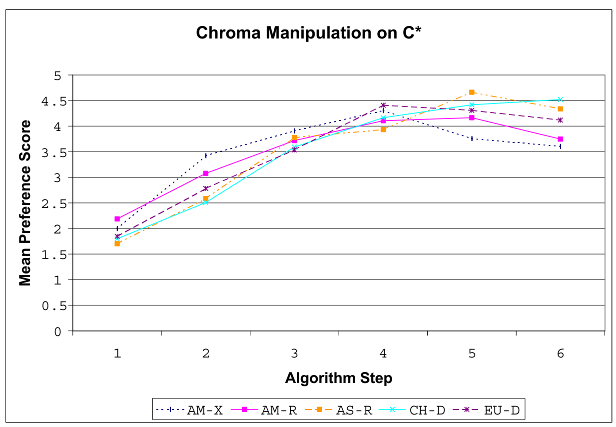

Figure 10. Average response of each sub-population for Chroma manipulation

Conclusion

Observer preference incorporated into current image reproduction techniques should be viewed as an enhanced or customized version of colorimetric reproduction objective. The idea of a need for a customized reproduction objective is the underlying theme of this research. The results demonstrated that that regardless of image content, or viewing mode a group of observers are consistent on reporting what image characteristics they deem are important. When evaluated on an individual level the amount of preference weighted to each characteristic is very different. The same concept seems to be apparent in the multi-cultural experiment also, in that on average the sub-populations are similar, however individual observers are very different. This leads one to believe that a colorimetric goal of image reproduction is a good first step, but now we need to try and build custom adjustments that will allow the colorimetric functions to incorporate different levels of observer preference.

Acknowledgement

The authors wish to thank Dr. K. Braun of Xerox Corp. for providing images and helpful comments.

References

1. RWG Hunt, IT Pitt, LM Winter. J. of Photographic Sci. 22, 144 – 149(1974).
2. RWG Hunt, The Reproduction of Color in Photography, Printing & Television. Fountain Press: Tolworth, UK, 1987.
3. T.J.W.M. Janssen and F.J.J Blommaert. J. Img. Sci. Technol. 44(2), 93-104(2000).
4. SN Yendrikhovskij, FJJ Blommaert, H de Ridder. Col Res Appl., 24, 52-67(1999).
5. H de Ridder, JIST, 40(6), 487 – 493(1996).

Biography

Scot R. Fernandez received his B.S. degree in Imaging and Photographic Technology from the Rochester Institute of Technology in 1999. This work on image quality and preferred color reproduction is part of his thesis requirement for a M.S. degree in Color Science at Rochester Institute of Technology. In September of 2001, he began working towards his Ph.D in Imaging Science at Rochester Institute of Technology. He is a member of the IS&T and ISCC.