Comparison of Color Difference Perception on Soft-Display Versus Hardcopy

King F. Choi
Eastman Kodak Company, Rochester, New York

Introduction

As computers and color monitors are becoming cheaper and more ubiquitous, the need to understand color perception on soft-display and how it relates to color perception on hardcopy becomes crucial. While the demand for "what you see in soft-display is what you get in hardcopy" is mounting, many hurdles are yet to be overcome. The difficulties are a result of the differences in their color gamuts, viewing conditions, nature of lumination (emissive vs reflective), color reproduction method (additive vs subtractive), etc. Due to these intrinsic differences, the problem of softcopy/hardcopy matching in a general sense could not be adequately tackled without answering some of the more fundamental questions of color perception in the two different media of interest.

On the other hand, even if the general problem of softcopy/hardcopy matching may not be completely resolved in the near future, a partial solution will still be very valuable if it could be established that results of certain psychophysical experiments done in soft-display under controlled viewing conditions could be translated to similar experiments done in hardcopy under its normal viewing conditions. Two advantages are the time and cost savings in setting up the experiments. Psychophysical experiments involving images or color patches in hardcopy often take months to set up due to the difficulty of generating the desired samples. In comparison, the soft-display environment, once it is calibrated, is more stable and results in faster and cheaper setup for similar experiments.

Another advantage of conducting experiments in soft-display is that the subjects only have to use the keyboard and/or mouse instead of physically manipulating the test objects. This again results in noticeable savings in time. As a by-product, data entries are done by the subjects during the experiments thus eliminating the process of key-punching and related scribal errors. Furthermore, other interesting statistics such as keying sequence and timing data can be collected for later analysis if necessary.

With all the above mentioned incentives, an experiment was conducted to address one of these cross-media questions, namely, color difference perception in hardcopy vs soft-display. While similar experiments have been done,[1] this one concentrates on color difference with a delta E of around five to ten. More specifically, this color difference experiment is done on soft-display mimicking that done by Sayer and Skipper[3,4] on photographic reproductions to compare the results of the two methods.

Experiment Setup

The equipment of this experiment consists of a Macintosh Quadra 700 computer, a SuperMac Thunder/24 color display board, and a 20" SuperMac SuperMatch color monitor with a resolution of 1152x870 pixels. The luminance of the monitor's reference white was about 25 foot-lamberts, and the correlated color temperature was that of D5000 with the chromaticity coordinates x = 0.339 and y = 0.363. A software routine is used to convert from specified CIELAB values to monitor RGB code values. It has been verified that the average delta E between the measured and requested CIELAB values is less than one.

The experiment was conducted in a dark environment. The monitor was put on a desk resulting in eye-level viewing for the subjects. The viewing distance was maintained at about 16 inches by fixing the position of the subject's chair (no head-bar was used). The diameters of the circular color patches were 0.56 inch and subtended a viewing angle of 2 degrees.

Ten Eastman Kodak Company employees who were experienced in making critical color difference judgements were subjects in this study. Seven of them also participated in the previous hardcopy experiment by Sayer.[3] A trial session was given to each subject to get acquainted with the experiment as well as to adapt to the viewing condition. The subjects were first asked to establish a reference set of color differences to be used throughout the experiment. This reference set consists of a center neutral patch ($L^* \approx 65$), and six surrounding neutral patches ($L^* \approx 35, 40, 45, 50, 55, 60$) against a 20% gray background with labels in reference white color. The distance between the center neutral patch and each surrounding patch was 0.56 inch. The subjects were allowed to use any positive numbers that they felt comfortable with. To make data entry easier, the subjects were encouraged to use the number pad with the ENTER key to enter their responses, and they were allowed to modify their previous entries if necessary at any time. The reference set was displayed on the top left corner of the screen throughout the experiment.

After the establishment of the reference set, the subjects were shown a series of test patch-pairs to judge the

color difference using numbers consistent with those used for the reference set. The distance between the test patch-pair and the reference set was about 8 inches which subtended a 30° viewing angle to the subjects. Figure 1 shows the layout of the display with the reference set at the top left corner, the test patch-pair in the center, and a control box at the top right allowing subjects to examine previous patches.

There were 320 test patch-pairs grouped into eight color centers of red, green, blue, cyan, magenta, yellow, neutral, and skin color. Each patch-pair consists of a center patch and a test patch. For each color center, there are 20 test patches which are about 5 ΔE from the center patch and another 20 patches which are about 10 ΔE from the center patch.

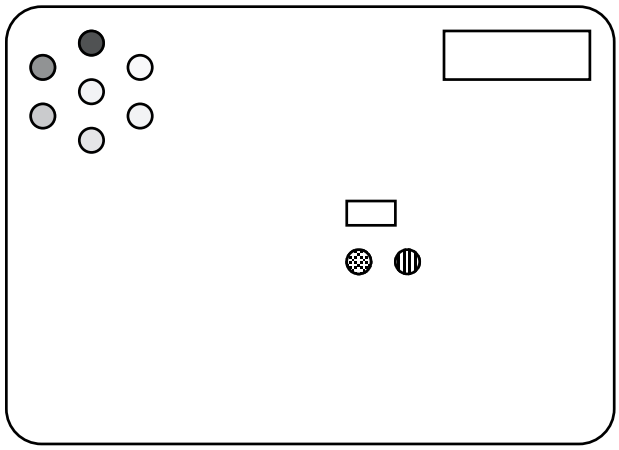

Figure 1. Screen layout for the experiment

The subjects were presented all 40 patch-pairs of one color center followed by those of another color center and so on. The order of the color centers were randomized for each subject, as were the presentation order of the patch-pairs within each color center. When entered, the response for each test patch-pair judgement was displayed in the rectangular box just above the test patch-pairs for easy viewing, immediate feedback, as well as to provide the reference white during the experiment. The response was also shown on the control box on the top right hand corner of the screen along with other buttons for backing up to the previous patch-pair, etc. The duration of the experiment was about one hour, and the subjects did not encounter any problem with the experiment.

Data Analysis

The data collected from the soft-display experiment were compared with the four datasets collected from the hardcopy experiment done about two years ago by Sayer.[3] Among the four datasets, two of them were collected from a duplicate experiment involving a group of ten Eastman Kodak Company employees who were experienced in making critical color difference judgements, and two of them were from a duplicate experiment involving a group of ten Virginia Polytechnic Institute and State University students who had little or no experience in making color difference judgements.

The data analysis can be broken down into three steps namely, (1) data editing, (2) data normalization, and (3) statistical test of significant differences.

Data Editing

The first step of the data analysis is the examination of the data collected in the soft-display experiment. It was discovered that one subject who used single-digit numbers throughout the experiment had four test patch-pairs judged substantially larger than ten. A review of the keying sequence collected showed that they were the results of intended modification which ended up in two digits instead of over-writing the first one. These four data were corrected before the actual data analysis. No other data were modified or deleted.

Data Normalization

Since each subject used his/her own scaling, color difference magnitudes used by the subjects need to be normalized in order to have meaningful comparison. One common method of normalization is the use of geometric mean as suggested by many researchers in Psychophysics.[2,5] However, to use the geometric mean in this experiment would mean largely ignoring the reference set data. In reviewing the color difference values used by subjects in judging the reference set, it was found that many used arithmetic or near-arithmetic series (e.g., 10, 20, 30, 40, 50, 60) while some others used geometric or near-geometric series (e.g., 5, 10, 20, 40, 80, 100). There appears to be certain arbitrariness in assigning those numbers. Furthermore, many subjects commented that during the judging of the test patch-pairs, they identified the two reference patch-pairs which appeared to bracket the color difference of the test patch-pair and then performed a mental interpolation between the two numbers. Since the subjects were strongly encouraged to look at the reference set in making color difference judgements, and the reference set is neutral with known delta E values between the center patch and the surrounding patches, it was decided to normalize the color difference responses to units of Neutral Equivalent Delta E (NEDE).

Figure 2 illustrates the translation from raw score to NEDE. First the measured delta E (n1, n2, ..., n6) of the reference set were plotted against the raw scores (r1, r2, ..., r6). Then, given a raw score **r**, the NEDE value **n** was found by linear interpolation between the bracketing reference points.

After the raw scores had been converted to the NEDE scale, they were plotted according to the eight color centers and five subject groups consisting of the soft-display group (CRT), the experienced hardcopy groups (EK1 and EK2), and the inexperienced hardcopy groups (VT1 and VT2). Since EK1 and EK2 contain data from the same subjects, they should track each other very well, and they did. A similar comment can be made about VT1 and VT2. Moreover, it is observed that the data for the CRT match those of EK1 and EK2 better than VT1

and VT2. A tentative qualitative conclusion can be drawn that the CRT results match with those of EK1 and EK2. A further statistical test shows that this is indeed the case.

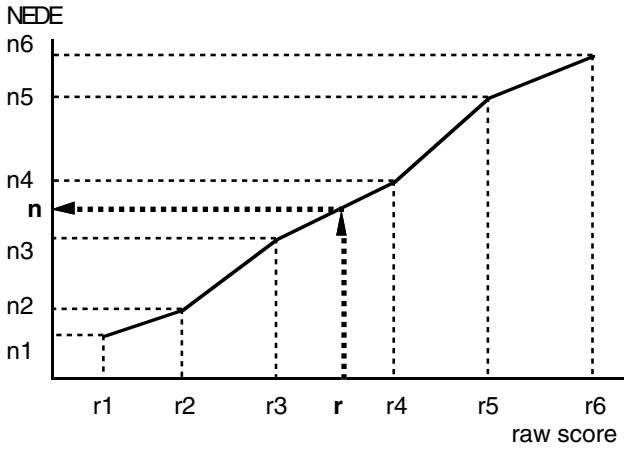

Figure 2. Raw score to NEDE translation

Statistical Test of Significant Differences

To test the hypothesis that there is no significant difference between the results of any two datasets, one might first use a standard two-tail t-test to find the p-value for any patch-pair, and then combine the p-values of the 320 patch-pairs into a single overall p-value. If the overall p-value is less than 0.05, then reject the hypothesis, otherwise, accept it. However, the process of combining 320 p-values implies the calculation of the joint probability of 320 t-distributions each with 18 degrees of freedom. It is not obvious how that can be done effectively.

As a result, an alternative method of analysis was used. First a standard two-tail t-test with p-value = 0.5 is done for each patch-pair. If the p-value is greater than 0.5, then it is labeled IN; otherwise OUT. The result is 320 IN/OUT, one for each patch-pair. Together, they form a binomial distribution with p = 0.5, q = 0.5, and n = 320. Since n is large, the binomial distribution approaches normal distribution with mean $\mu = np = 160$ and variance $\sigma^2 = npq = 80$, or standard deviation $\sigma = 8.94$. Therefore, if the total number of OUT is greater than $\mu + 1.645\sigma = 174.7$ (i.e., a p-value of 0.05), the hypothesis that there is no significant difference between the results of any two datasets will be rejected; otherwise, it will be accepted.

As it turns out, this analysis technique has a high power of discrimination in the sense that some datasets that were previously concluded to have no significant difference using another method of analysis are found to be significantly different using this new technique. This will be discussed in the following section.

Results

A program written in C on the Macintosh computer was used to analyze the data. The analysis was done for every dataset-pair. The results of the analysis are summarized in Table 1.

Table 1. Results of Pair-Wise Comparison

dataset-pair	# of OUT	sigma	p-value	sig. diff
CRT-EK1	161	0.112	0.455	no
CRT-EK2	160	0.000	0.500	no
CRT-VT1	238	8.721	0.000	yes
CRT-VT2	216	6.261	0.000	yes
EK1-EK2	90	-7.826	1.000	no
EK1-VT1	220	6.708	0.000	yes
EK1-VT2	182	2.460	0.007	yes
EK2-VT1	250	10.062	0.000	yes
EK2-VT2	229	7.714	0.000	yes
VT1-VT2	92	-7.603	1.000	no

It is expected that there should be no significant difference between EK1 and EK2; and between VT1 and VT2 since they are duplicate datasets from the same subjects. The result shows that this is indeed the case—both of them have p-values equal to 1.0. On the other hand, there is a significant difference between any EK dataset and the VT datasets. This conclusion is different from that of Sayer's,[3] and is due to different methods of analysis. The implication is that the analysis technique as employed in this experiment has a very high power of discrimination.

Next, there is significant difference between the CRT dataset and any VT dataset. This would be expected since the difference between experienced and inexperienced subjects alone can lead to significant difference as in EK vs VT dataset comparisons.

Finally, the p-values between CRT and any EK dataset is greater than 0.05 resulting in the acceptance of the hypothesis that there is no significant difference between experiments performed by experienced subjects using soft-display and hardcopy. This result agrees with the qualitative examination of the plots.

Since the actual test patch-pairs used in the soft-display experiment were slightly different from those in the reflection print experiment, the question of how these differences affect the result must be addressed. This was done by plotting the difference between the delta E values against the t-score of each test patch-pair for each dataset-pair. Figure 3 shows such a plot between the CRT group and the EK1 group. In this figure, only those patch-pairs that were OUT were plotted. Datapoints above the x-axis correspond to CRT patch-pairs having larger delta E than the respective EK1 patch-pairs. Datapoints on the right of the y-axis correspond to CRT patch-pairs being judged as having larger color differences than the respective EK1 patch-pairs. Here we notice that there are more points in the first and third quadrants signifying that there is correlation between the size of delta E to the perceived color difference. This implies that some of the datapoints were classified as OUT due to actual delta E differences between the color patches of the soft-display and the hardcopy experiments rather than medium difference. If the delta E differences between the color patches in the soft-display and the hardcopy experiments were minimized, then there will be fewer datapoints that are OUT resulting in a higher p-value.

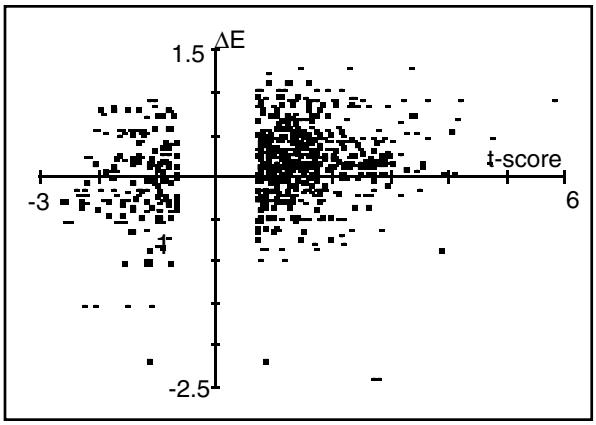

Figure 3. t-score vs ΔE between CRT and EK1

Conclusion and Discussion

From the result of the analysis, we conclude that there is no significant difference between color difference perception conducted in soft-display and hardcopy among similar subjects at L* ≈ 50 for eight color centers with moderate color difference of five to ten delta E. This agreed very well with the qualitative examination of the data.

To complete a "balanced" test, one should also conduct the soft-display experiment on inexperienced subjects. This was not done due to limitation of resources. Nevertheless, the fact that the soft-display and reflection print datasets from the experienced subjects passed such a stringent method of analysis gives us much confidence about generalizing the result to other subjects and perhaps to other L* levels and color centers as well.

References

1. Roy S. Berns, "Color Tolerance Feasibility Study Comparing CRT-Generated Stimuli with an Acrylic-Lacquer Coating," *Color Res. Appl.* **16**, 232-242 (1991).
2. George A. Gescheider, "Psychophysics: Method, Theory, and Application," 2nd edition, Lawrence Erlbaum Associates, 1985.
3. James R. Sayer, "The Perception of Moderate and Large Color Differences in Photographic Prints: An Evaluation of Five Color-Difference Equations," M. S. Theses, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1991.
4. James R. Sayer and Julie H. Skipper, "The Perception of Large Color-Differences in Photographic Prints: An Evaluation of Five Color-Difference Equations," Society for Information Displays, 1991.
5. S. S. Stevens, "Issues in Psychophysical Measurement," *Psychological Review*, **78**, 426-450 (1971).