Evaluating Chromatic Adaptation Transform Performance

Sabine Süsstrunk[1,2] and Graham D. Finlayson[2]
(1) School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland;
(2) School of Computing Sciences, University of East Anglia, Norwich, United Kingdom

Abstract
The performance of many color science and imaging algorithms are evaluated based on their mean errors. However, if these errors are not normally distributed, statistical evaluations based on the mean are not appropriate performance metrics. We present a non-parametric method, called the Wilcoxon signed-rank test, which can be used to evaluate performance without making any underlying assumption of the error distribution. When applying the metric to the performance of chromatic adaptation transforms on corresponding color data, we can derive a new CAT that statistically significantly outperforms CAT02 at the 95% confidence level.

Introduction
Chromatic Adaptation Transforms (CATs) are used in color science and color imaging to model illumination change. Specifically, they provide a means to map *XYZ* under a reference source to *XYZ* under a target light such that the corresponding *XYZ* produce the same perceived color.

The color science and imaging community has mostly adopted the linear von Kries adaptation model to compute this illumination change.[2,3,8,10] This model states that the color responses of corresponding colors under two illuminants are simple scalings apart.[12] For example, if *RGB* and *R'G'B'* denote the color responses for an arbitrary surface viewed under two lights, then the von Kries model predicts that $R'=aR$, $G'=bG$, and $B'=cB$. In modern CATs, the scaling coefficients *a, b,* and *c* are the ratios of the color responses of the illuminants, i.e. $a=R_w/R'_w$, $b=G_w/G'_w$, and $c=B_w/B'_w$. However, the CATs differ in the color space in which this scaling is applied.

It is well known that the von Kries model operating in *XYZ* color space poorly describes corresponding color data (applying the scaling on *XYZ* tristimulus values is often referred to as the "wrong von Kries"). Thus, most modern CATs proposed in the literature are based on colorimetric color spaces,[6] i.e. color spaces that are derived as a linear transformation of *XYZ*:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \mathbf{M}^{-1} D \mathbf{M} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (1)$$

where $\mathbf{M}$ is a nonsingular (3x3) matrix linearly transforming *XYZ* values to *RGB* responses, $\mathbf{M}^{-1}$ its inverse and *D* the diagonal matrix containing the scaling coefficients.

The color space in which the scaling takes place, i.e. the linear transformation from *XYZ* to *RGB*, is usually derived based on some error minimization. Li et al.[8] iteratively optimized the coefficients of matrix $\mathbf{M}$ to produce minimum CIELAB color differences between predicted and observed results over a set of eight corresponding color data sets.[9] A modified version called CAT02 that excluded some of the successive haploscopic experimental data in the minimization was chosen for the CIECAM02 color appearance model.[10] Fairchild[2] used Munsell samples to calculate corresponding colors under illuminants A and D65 using the non-linear Bradford CAT[7] of CIECAM97s. He then developed a linear CAT by minimizing the CIELAB differences to the predictions of the Bradford CAT on this corresponding color data set. Finlayson and Süsstrunk[3] used spectral sharpening that minimizes *XYZ* leat-square errors to derive a linear CAT from Lam's corresponding color data set.[7]

Several studies evaluated the different linear CATs mentioned above to find if one outperforms the other.[1,8,10,11] In these studies, the performance criterion is based on the mean CIELAB prediction error. However, a single summary statistic does not always adequately summarize the underlying distribution. Having a lower mean does not necessarily imply that one algorithm is always better than the other.

In section 2, we discuss the underlying assumptions made when using a performance metric based on a mean error and propose a more appropriate statistical evaluation, namely the Wilcoxon signed-rank test, for populations that are not normally distributed. In section 3, we derive a new linear CAT that outperforms CAT02 at the 95% confidence level when tested on Lam's corresponding color data set. Section 4 concludes the article with a summary and some guidelines for evaluating color experiments.

Color Error Analysis
When evaluating chromatic adaptation transforms, we are interested in which transform best maps illumination change. A number of psychophysical experiments, collected by Luo and Rhodes,[9] provide us with *corresponding color* data. Corresponding colors are pairs of tristimulus values, based on one physical stimulus, which appear to be the same color when viewed under two different illuminants. A "good" CAT's prediction of the tristimulus values of a corresponding color under a test illuminant, obtained by mapping the tristimulus values under the reference illuminant to the test illuminant, is thus (close to) identical with the actual corresponding color obtained by the psychophysical experiment.

Deviations from actual and predicted values can be expressed with some error measure. As we are interested in color appearance, a perceptual measure seems the most appropriate. ΔE, which is calculated as the Euclidian distance in CIELAB, is indeed such a metric. As CIELAB is not perfectly perceptually uniform, ΔE94, ΔECMC, ΔE2000 were later derived that add different weights depending on hue, saturation, and/or lightness of the color samples to be evaluated.

These error measures can tell us how accurately a particular CAT maps a color to a different illuminant, and they allow us to easily compare the relative performance of different CATs on a single corresponding color pair. However, we are generally more interested in the performance over a large set of corresponding colors, as a CAT should predict many corresponding colors under many different illuminants. Often, a single summary statistic is chosen, such as the mean (or root mean square) ΔE, averaged over the data sets. If the mean error for one CAT is found to be lower than the mean error for the other CAT, then the conclusion is drawn that the first CAT is better than the second.

There are two potential problems with using the mean as a single summary statistic. First, the mean value is not an appropriate statistic when the errors are not normally distributed.[13] Figure 1 shows the histogram of CAT02[10] prediction errors (ΔE94) on Lam's corresponding color data.[7] Figure 2 plots the quantiles of this error distribution against quantiles of a standard normal distribution. It is clear from the histogram (Figure 1) that the errors are not normally distributed. If they were, then the plot of the quantiles would follow a straight line (Figure 2).

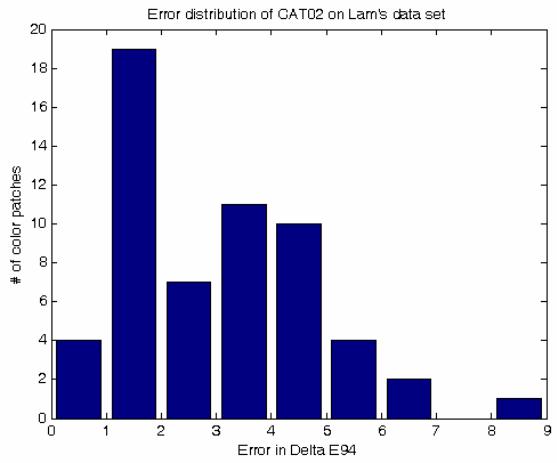

Figure 1. The distribution of CAT02 prediction errors (ΔE94) on Lam's corresponding color data.

If the error distribution is not normal, the error *median* is a better measure to reflect the central tendency of the samples, as it is not influenced by extreme values.[13]

Second, the fact that one CAT has a lower mean (or median) value than another is not sufficient information for drawing the conclusion that one CAT outperforms the other. An alternative is to use the whole error distribution. We can use the mean performance of the CATs to formulate a hypothesis and then test this hypothesis, as was done by Finlayson and Süsstrunk,[4] who employed a student t-test and found an infinite number of CATs that perform equally well for a given confidence interval. However, if the error distribution is not normal, we need to use a nonparametric (or distribution-free) method to test the hypothesis that a better median predicts a better CAT performance. A non-parametric alternative to the student t-test is the Wilcoxon signed-rank test, which makes no assumptions about the nature of the underlying error distributions, but takes into account the sign and rank of the error difference.

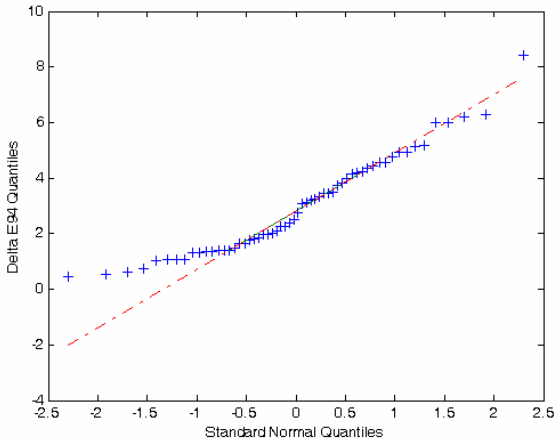

Figure 2. Quantiles of the error distribution plotted against the quantiles of a standard normal distribution.

Suppose we want to compare the performance of two CATs. We use each CAT to predict the corresponding colors under the test illuminant of a given data set. We calculate the error, using one of the error measures described above, between the actual and predicted corresponding colors. Let A and B be random variables representing the prediction error, and μ_A and μ_B their respective median. The Wilcoxon signed-rank test can be used to test the hypothesis that $\mu_A=\mu_B$, i.e. we hypothesize that both CATs have the same performance. We call this the null hypothesis H_0. To test this hypothesis, we consider the difference of the independent error pairs $(A_1\text{-}B_1)...(A_N\text{-}B_N)$ for N different corresponding color pairs. We rank the error pairs according to their absolute differences, and then assign a plus (+) or minus (-) sign to the ranks depending if $A_i>B_i$ or $A_i<B_i$. If H_0 is correct, then the sum of the ranks W will approximate zero. If W is much larger (or much smaller) than zero, the alternative hypothesis H_1, namely that $\mu_A>\mu_B$ or $\mu_A<\mu_B$ is true. We can test the null hypothesis H_0 against the alternative hypothesis H_1 at a given significance level α. We reject the null hypothesis and accept the alternate hypothesis if the probability of observing the error differences we obtained is less than or equal to α. For example, if $\alpha=0.05$ and the probability p we calculate is 0.04, then we can reject H_0 at the 0.05 significance level. That amounts to rejecting the null hypothesis 95% of the time.

CAT Experiment

We used a spherical sampling technique[4] to evaluate if we can find a chromatic adaptation transform that outperforms CAT02, using the Wilcoxon signed-rank test as performance metric. In the case of trichromatic (RGB and XYZ) imaging applications, the basis functions span a three-dimensional space. If the lengths of the vectors are normalized to unity, then different vector combinations can be illustrated with their end-points that lie on the surface of a sphere. Trying all possible combinations of three points distributed over the surface of the sphere allows us to find all possible solutions to a given problem. The advantage over other optimization techniques is that spherical sampling assures a global minimum is found, and that not only one, but a set of solutions can be retained if so desired.

We used Lam's corresponding data set and an error measure of ΔE94. While it is obvious that the choice of error measure could influence the results, two studies have found that for the corresponding color data sets considered, which ΔE error measure was chosen did not change the overall trends.[8,11]

Table 1 summarizes the mean values and the *p*-values found using the Wilcoxon signed-rank test as performance metric. The prediction errors of the best CAT found through spherical sampling was compared to CAT02[10] and the Sharp CAT.[3] As can be seen from the results, the best CAT (*W*-CAT) outperforms CAT02 at the 95% confidence level ($p<0.05$). However, the difference in median between *W*-CAT and the Sharp CAT are not statistically significant. Figure 3 shows the corresponding RGB color matching functions.

Table 1: Median ΔE94 values for Lam's data set, and probability *p*-values resulting from the Wilcoxon signed-rank test.

CAT	Median ΔE94	*p*-value
W-CAT	2.61	
CAT02	2.67	0.04
Sharp	2.69	0.60

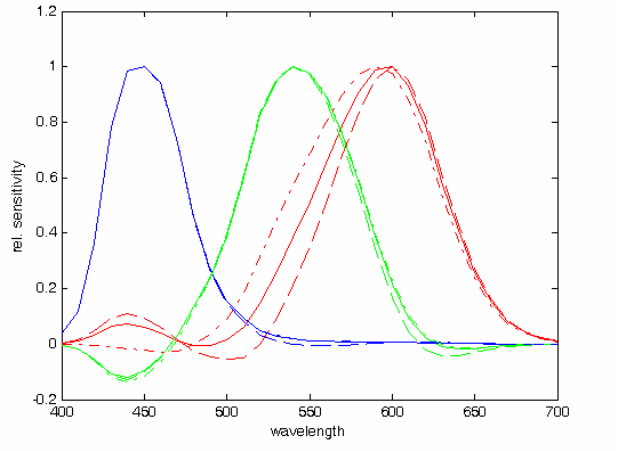

Figure 3. The RGB color matching functions of W-CAT (solid line), Sharp CAT (--), and CAT02 (-.-).

Conclusions

Many color algorithms are evaluated using the mean error as a statistically relevant performance metric. However, the underlying assumption that the error distribution is normal was shown to not always be true.[5] Thus, we believe that using the median as a singular quality indicator, and the Wilcoxon signed-rank test as a performance metric that also takes into account the underlying error distribution, is more applicable to many performance evaluations in color science and color imaging. Thus, the distribution of errors should first be analyzed before the statistical evaluation method is chosen.

We analyzed the error distribution of the predicted corresponding colors using CAT02, the chromatic adaptation transform chosen for CIECAM02, applied to Lam's corresponding color data set. We found that the errors do not follow a standard normal distribution. Using the Wilcoxon signed-rank test as performance metric and a spherical sampling technique, we derived a chromatic adaptation transform *W*-CAT that outperforms CAT02 at the 95% confidence level.

We are not claiming here that *W*-CAT outperforms CAT02 in all instances; this still needs to be evaluated. However, it is interesting to note that a performance metric more suited to the error distributions challenges the assumption that all modern CATs perform equally well.

When comparing W-CAT to Sharp CAT, we cannot find a statistically significant difference in performance between the two. Looking at the corresponding color matching functions in Figure 3, we notice that *W*-CAT is "sharper" in the red, i.e. more narrowband than CAT02. While not quite as sharp as the Sharp CAT, the peaks in the red are approximately at the same wavelength, while CAT02's peak is at shorter wavelength. Recall that the Sharp CAT is derived through *XYZ* error minimization of Lam's corresponding colors[3] and not through optimization of a perceptual ΔE error. This leads to the conclusion that at first approximation, sharpening is well suited to derive transforms that can predict corresponding colors.

References

1. A.J. Calabria and M.D. Fairchild. Herding CATS: A comparison of linear chromatic adaptation transforms for CIECAM97s. Proc. CIC9, pg. 174, 2001.
2. M.D. Fairchild. "A revision of CIECAM97s for practical applications." COL. Res. and Appl., 26(6), 418, 2001.
3. G.D. Finlayson and S. Süsstrunk. Performance of a chromatic adaptation transform based on spectral sharpening. Proc. CIC8, pg. 49, 2000.
4. G.D. Finlayson and S. Süsstrunk. Spherical sampling and color transformations. Proc. CIC9, pg. 321, 2001.
5. S. D. Hordley and G. D. Finlayson. Reevaluating colour constancy algorithms. Proc. ICPR17, vol. 1, pg. 76, 2004.
6. ISO 22028-1:2004. Photography and graphic technology - extended colour encodings for digital image storage, manipulation and interchange - Part 1: architecture and requirements.
7. K.M. Lam. Metamerism and Colour Constancy. PhD thesis, University of Bradford, 1985.

8. C. Li, M.R. Luo, B. Rigg, and R.W.G. Hunt, “CMC 2000 chromatic adaptation transform: CMCCAT2000,” COL. Res. and Appl., 27(1), 49, 2002.
9. M.R. Luo and P.A. Rhodes, “Corresponding colour datasets,” COL. Res. and Appl., 24(4), 295, 1999.
10. N. Moroney, M.D. Fairchild, R.W.G. Hunt, C. Li, M.R. Luo, and T. Newman. The CIECAM02 color appearance model. Proc. CIC10, pg. 23, 2002.
11. S. Süsstrunk, J. Holm, and G.D. Finlayson. Chromatic adaptation behavior of different RGB sensors. Proc. Electronic Imaging, vol. 4300, pg. 172, 2001.
12. J. von Kries. Theoretische Studien ueber die Umstimmung des Sehorgans. Festschrift der Albrecht-Ludwig Universitaet, pg. 145, 1902.
13. R.E. Walpole, R.H. Myers, and S.L. Myers. Probability and Statistics for Engineers and Scientists. Prentice Hall International, 6th edition, 1998.

Author Biography

Sabine Süsstrunk received her BS in scientific photography from ETHZ, Switzerland, her MS in Electronic Publishing from RIT, USA, and her PhD in Computing Science from UEA, UK. She is currently Professor for images and visual representation in the School of Computer and Communication Sciences at EPFL, Lausanne, where her work has focused on digital photography and color image processing. She is a member of IS&T, IEEE, and ACM.