Color Correction Meets Blind Validation for Image Capture: Are We Teaching to the Test?

Don Williams; Image Science Associates and Peter D. Burns; Burns Digital Imaging

Abstract

It is common practice for digital image capture systems to use a small number of de-facto-standard test targets. Unfortunately, however, color (spectral-) characteristics of the colorants used may differ from those for the population of object/scenes to be captured. This can lead to poor color calibration of the system. A second limitation of current color-capture evaluation arises when the same set of color stimuli (color patches) are used to calibrate the color capture and to evaluate the residual color errors. When the same color-target is used, the reported color-encoding errors will usually be lower than those observed in normal image capture. This is because we are, in effect, 'teaching to the test', as when a student is prepared for test taking, rather than subject mastery. We can approach this under-reporting of color error as a measurement bias. We can treat color-correction (e.g. by a color-profile) as being a statistical model relating the detected image values and their intended ('correct') pixel values. Using a statistical approach we adopt a validation method aimed at determining the extent to which this model relationship between variables (the regression model) provides an acceptable description of the data. For our color-imaging case, the equivalent step would be to test the computed color-correction (ICC profile) with color patches that are independent of those used to build the profile. We demonstrate a candidate strategy for selecting these test colors, and an example of a validation set of colors chosen to be distinct from the calibration set in the popular ColorChecker SG.

Introduction

Color image capture normally includes a color-correction step that transforms detector signals into corresponding pixel values. For digital cameras and scanners, we usually base the color-correction operation on captured images of reference color charts. From a colorimetric description of the input reference color patches (e.g. CIELAB coordinates, $L^*a^*b^*$) and the corresponding (unprocessed) pixel values, we compute the color-correction parameters required for accurate color image encoding. This usually takes the form of either a custom or a popular output referred ICC profile. (e.g. sRGB, AdobeRGB, etc.) We can cast the building of an ICC color profile as a statistical modeling operation, where the model takes on the form specified by the profile elements, e.g., look-up tables, color matrix, etc.

It is common practice for digital image capture systems to use a small number of de-facto-standard test target. Unfortunately, however, color (spectral-) characteristics of the colorants used may differ from those for the population of object/scenes to be captured. This can lead to poor color calibration of the system. The selection of collection-specific test targets for improved color-capture has been addressed in the literature.[1-4]

In evaluating the goodness of any modeling there is normally a validation effort aimed at determining the extent to which the regression model provide an accurate description of the variables involved. A popular way to implement this regression is by way of an ICC color profile. In effect, this color profile acts as a color dictionary that translates triplets of captured RGB code values into equivalent colors as defined by the Profile Connection Space (PCS). The mathematical models to do so can be varied and sometimes complex. In the absence of a custom profile, often, standardized profiles are used.

A popular way to evaluate the quality of this color calibration is to simply compare the translated color of each patch in PCS to the measured reference color of the actual target. While this is an instinctive approach, it yields, by definition, an optimal residual color error for that model since the regression model is designed to minimize such errors. One is effectively 'teaching to the test' when evaluating digital capture color performance using the same colors for which the color-correction was performed.

We suggest using a validation approach where the color performance is tested with an independent and different set of color patches. Borrowing from medical clinical trials, these can be thought of as control (calibration) and treatment (validation) groups. While color calibration- or profiling *validation* is not often discussed in the literature, it can provide valuable information regarding the quality of image capture, and the likelihood of color artifacts during normal operation of the image capture system.

Validation Color Patch Selection Set

We recognize that the strategy for selecting a validation set of colors is open to infinite opinions. Being reasonable and without focusing on building a 'killer' validation target, we restrict our patch selection for validation using a set of criteria already included in the SG target. They are,

- The same number of total patches
- Identical set of gray patches (61)
- Same number of chromatic patches (79)
- Same number of patches within L*(10) slices
- Semi-Gloss surface
- Remained within the gamut of the existing CCSG

The differences between the two sets are:

- Different set of chromatic patches
- Select patches from the Natural Color System (NCS) index[5]

We selected the chromatic patches by inspecting the CIELAB a^*b^* plots of each of eight L^* slices (L^*=10 increments). We identified gaps between the existing SG coordinates and selected an appropriate color from the NCS library of colors, as measured for this study. Figure 1 shows two example L^* slices illustrating the SG calibration colors and the same number of NCS validation colors with that L^* slice.

Figure 2 illustrates a comparison image set between the SG calibration target and the validator target we will call SGX. The similarities and differences are consistent with the descriptions cited above. All of the patches, except for a few at L^*>80, a^*<20, and b^* >90 (illuminant D50, 2 degrees) fit within the Adobe RGB color space.

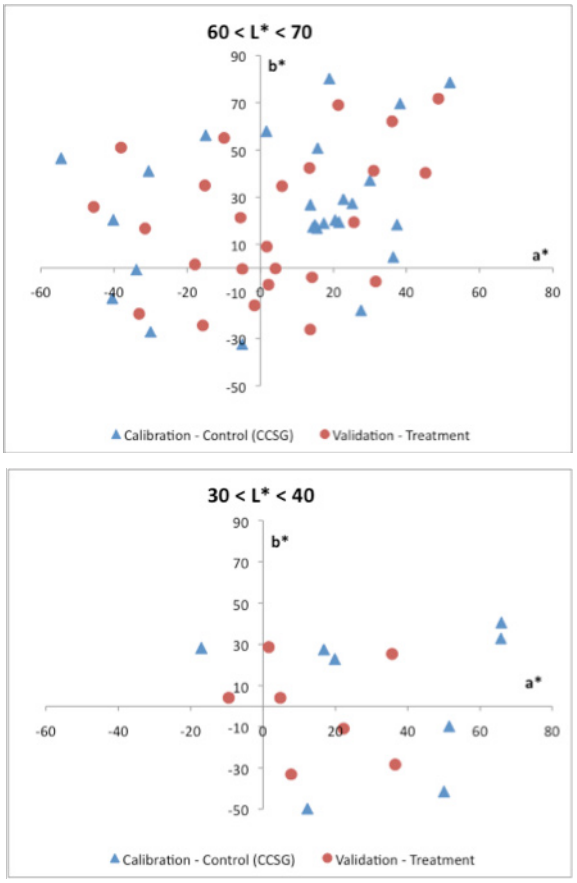

Figure 1: Comparison of calibration and validation CIELAB coordinates for two selected L slices*

Experimental

Custom colorimetric reference files were created for both targets. A raw digital image for both the Calibration and Validation targets was acquired (SG and SGX, respectively) from an Epson 10000XL scanner. Both were processed with a gamma 2.2 using all of the center 15 gray-patch values.

Rough Profiler, built on the open-source Argyll color-management system,[6] was used to create three different ICC profiles for each image using different methods (models). The models were labeled as,

1. Lab cLut, medium quality (Lcm)
2. Shaper + Matrix, medium quality (SMm)
3. Lab cLut, high quality (Lch)

At this point six different color profiles were now available. Each of the above three models for the two target images. The above profiles were then embedded into each of the two candidate image files and evaluated for color encoding accuracy.

Figure 2: Comparison images of the SG (top) and SGX (bottom) targets

Examples of the notation we will use to describe the image and profiling pairings are;

- **$SG_{SG\text{-}LCM}$** – SG image with a profile created using the SG target using a cLut medium quality profiling model.
- **$SGX_{SGX\text{-}LCM}$** - SGX image with a profile created using the SGX target using a cLut medium quality profiling model.
- **$SGX_{SG\text{-}LCM}$** – SGX image with a profile created using the SG target using a cLut medium quality profiling model.

The first two pairings above would be normal pairings of each target evaluated against a profile created by that target. It is the third pairing that is of interest where the validator target (SGX) is assessed for color encoding accuracy using a profile generated via the SG target. Results for several combinations of image target and ICC profile are presented in the Results section that follows.

Results

Table 1 lists the median and maximum ΔE_{2000} values for the important target-ICC profile combinations. For any particular statistical model, three sets of metrics are cited. Using the Lab cLut-medium quality model as an example, the SG-SG table numbers indicate lower median and maximum ΔE_{2000} values when the same IIC profile is used for the target from which it was derived. The same applies to the SGX-SGX combination. These two can act as baseline metrics for assessing the relative magnitude of the difference for the SGX-SG combination. This was the important validation combination for which this experiment was performed. The SGX-SG values were calculated using the SGX

validation target but with an embedded ICC color profile calculated using the SG target.

Using the median and maximum ΔE_{2000} values alone as indicators there does not appear to be a very large difference between the two. While there is a slight increase in encoding error by using the mismatched target-ICC profile validation combinations (i.e. SG-SGX combinations) it is not as large the authors expected. This applies to all three models. Indeed, the validation set for the shaper-matrix combination actually had a lower overall ΔE_{2000} compared to the target for which it was designed.

Table 1 – Summary ΔE_{2000} for image-profile combinations

ΔE2000	Target type		Target profile source	Model
	SG	SGX		
median	2.18 (2.13)*	2.30 (2.34)*	SG	Lcm
max	4.33 (4.34)*	4.84 (4.77)*		
median		2.06	SGX	
max		3.29		
median	3.14	2.87	SG	SMm
max	10.64	7.55		
median		2.85	SGX	
max		7.60		
median	2.16	2.30	SG	Lch
max	3.60	4.99		
median		2.03	SGX	
max		3.16		

To better normalize the *ΔE2000* data we eliminated the common neutral values from the calculations. These data are shown in the parenthetical values of Table 1 for the *Lcm* model alone. This is also shown graphically in the pseudo-color illustrations of Figs. 3 and 4. Again, there is not a very large difference.

Figure 3: ΔE2000 map of SGSG-LCM image-profile combination with neutrals toggled off

Figure 4: ΔE2000 map of SGSG-LCM image-profile combination with neutrals toggled off

It is worthwhile though to evaluate the box-whisker plots associated with the neutral normalized data sets for the SG-SG and SG-SGX combinations. These are shown enlarged in Fig. 5 below. (see Appendix for explanation of how to interpret these plots).

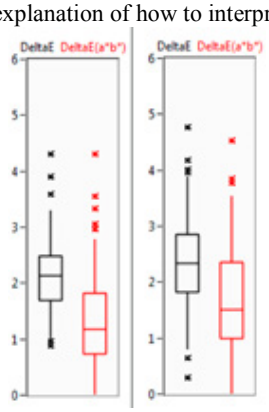

Figure 5: Box-whisker plots of ΔE_{2000} for SG-SG (left) and SG-SGX (right) target-ICC profile combinations

Evaluating these in context of the pseudo-color illustrations shows that there is a higher variance of ΔE_{2000} errors with the validation set than for the calibration set. So, while the overall accuracy (i.e. median) of the data for the SG-SGX is lower, albeit small, the variance of error is higher (i.e. lower precision) for this set. In combination the additive difference of lower accuracy and precision suggests a recommendation for performing independent calibration-validation tests, as outlined in this experiment.

This also suggests the need for better summary and specificity measures for evaluating color encoding error for digital capture. We have chosen to use box-whisker plot because they can describe central tendency, distribution, and outlier data quickly in a simple graphic. As shown in Fig. 3 and 4, it is also possible to communicate better specificity on color errors by including toggled masks that exclude color patches that are not important to the use at hand.

Conclusions

Methods and practices used to design and evaluate color image capture rely on several statistical concepts. We derive color correction based on sets of *sample* colors drawn from the *ensemble* of likely objects. Our choice of method is the selection of a *statistical model*. Color calibration is the fitting of the sample data to derive *estimates* of the parameters. System evaluation of color capture is reported in terms of *statistics* of (residual) error *distributions*. Consistent with this statistical view, we suggest that the evaluation of image color capture should employ methods which would be common practice for other statistical modelling efforts, that of independent model-validation.

For color image capture, independent, or blind, verification involves using a validation color data set that is independent of the (model) set used to compute the color correction parameters. While these concepts have been discussed privately and presented before, there appears to be little discussion of common sense strategies for identifying independent sets of colors for validation experiments. To demonstrate and investigate the approach we chose a commonly used color test chart, the ColorChecker SG, for our color-correction (model building) step. For the validation color samples, we selected from the set of NSC samples, commercially available for custom targets. ICC color profiles were computed and used for color management of a desk top scanner.

The results indicated that when independent color patches are used, the reported color errors are often greater than those for a single model and verification set. This is as we would have expected, although the differences were not large. We can interpret the magnitude of this difference as a measure of the smoothness of the relationship needed for accurate color-correction. A desk top scanner, where illumination is controlled, will usually shows stable and 'well-behaved' color characteristics. In the case of a camera or printer, larger differences might be expected.

Another interpretation of the magnitude of the differences between the single- and independent-set validations could be the selection of the number of color-patches to be needed for good color management.

The ideas and demonstrations in this paper are intended to suggest direction for further study of color-encoding performance reporting. For product design, image processing optimization, and standard evaluation, the independent validation of color-correction appears of practical value and straightforward to implement.

Acknowledgement

We have benefited from discussions with Prof. Roy Berns, Rochester Institute of Technology.

References

[1] D. Williams and P. D. Burns, Capturing the Color of Black and White, Proc. IS&T Archiving Conf., 96-100, IS&T, 2010

[2] G. Trumpy, Digital Reproduction of Small Gamut Object: A Profiling Procedure based on Custom Colour Targets, Proc. CGIV Conf., 143-147, IS&T, 2010

[3] D. Williams and P. D. Burns, Targeting for Important Color Content: Near Neutrals and Pastels, Proc. IS&T Archiving Conf., 190-194, 2012

[4] R. S. Berns, Artist Paint Target (APT): a tool for verifying camera performance, Technical Report, Studio for Scientific Imaging and Archiving of Cultural Heritage, Rochester Inst. Tech., 2014.

[5] Natural Colour System, index of all NCS 1950 original colours, http://www.ncscolour.com/en/design-architecture/web-shop/ncs-index/

[6] Argyll Color Management System, http://www.argyllcms.com/

Author Biography

Don Williams is founder of Image Science Associates, a digital imaging consulting and software group. Their work focuses on quantitative performance metrics for digital capture imaging devices, and imaging fidelity issues for the cultural heritage community. He has taught short courses for many years, contributes to several imaging standards activities, and is a member of the Advisory Board for the interagency US Federal Agencies Digitization Guidelines Initiative, FADGI.

Peter Burns is a consultant supporting digital imaging system and service development, and related intellectual property efforts. Previously he worked for Carestream Health, Eastman Kodak and Xerox Corp. He is a frequent conference speaker, and teaches courses on these subjects.

Appendix: interpreting box-whisker plots

The examples and wording here are largely taken from the National Instruments Labview manual on box-whisker plots. Only minor changes have been made. The box-whisker plots in this paper are taken from Labview software graphics.

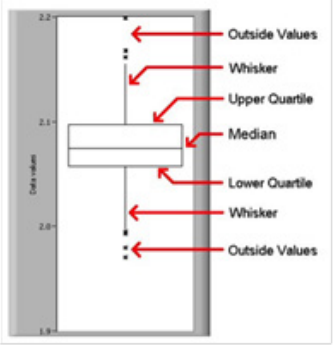

Figure 6: Example box-whisker plot

The large, divided rectangle in the middle forms the box around which supplemental statistical features are derived. The upper and lower quartiles of the data set determine the size and location of this box. The line that divides the box horizontally through the middle represents the median of the data set. The top edge of the box indicates the value corresponding to the upper quartile of the data. The upper quartile is the median of the upper 50% of the data values, or the values greater than the global median. The bottom edge of the box shows the value corresponding to the lower quartile of the data. The lower quartile is the median of the lower 50% of the data values, or the values less than the global median.

Vertical lines called whiskers extend from the middle of the top and bottom edges of the box. The whiskers are 1.5 times the inner quartile spread in length measured from the median. The inner quartile spread is the difference between the upper and lower quartiles of the data. The whiskers provide an arbitrary cutoff point to identify outlier values. Data points falling outside the whiskers but less than three times the length of the inner quartile spread are identified with small xs. Points beyond the whiskers are identified with large Xs.