Metrics for Evaluating Spectral Matches: A Quantitative Comparison

J A Stephen Viggiano, jasv@acolyte-color.com
Acolyte Color Research, http://www.acolyte-color.com, West Henrietta, NY, USA

ABSTRACT

Several of the spectral match metrics considered by Imai, *et al.,* are compared for a large number of non-metameric pairs of spectra in order to assess how accurately they track human perception as predicted by CIELAB (*i.e.,* the extent to which the metric will exhibit a proportional relationship with CIELAB total color difference), and how precisely they do so (*i.e.,* the relative compactness of the distribution of a metric for spectral pairs which differ by a given level of CIELAB total color difference). Both properties are important attributes of a spectral match metric. Of particular importance in optimization problems is the precision as the total color difference becomes small.

We found that among the metrics considered, only unweighted RMS and Viggiano's Spectral Comparison Index provided precision for both large and small color differences. The Viggiano Spectral Comparison Index had the closest correlation to human perception, and, for the non-metameric spectral pairs examined in this study, assumed values close to 2,6 times that of CIELAB ΔE*.

The paper includes a definition of non-metameric spectra, and describes a method for generating them, which include variations in Lightness, Hue, and Chroma.

Keywords: spectral match, metamerism, multichannel visible spectrum imaging, index of metamerism, Spectral Comparison Index

INTRODUCTION

In multichannel visible spectrum imaging, the goal is usually expressed as producing a reproduction whose spectrum matches that of the original, on a point-by-point basis, throughout the visible portion of the electromagnetic continuum. From a practical point of view, a certain degree of mismatch must be tolerated. We are thus left with the task of quantifying the seriousness of this mismatch.

In addition, optimization problems, such as parameter estimation in color device modeling, require an optimization criterion. This is often a quantification of the degree of mismatch between a set of spectra as measured and a corresponding set of spectra as predicted by a model using a certain set of parameter estimates. The goal of the optimization is to select the parameter values which minimize the degree of mismatch. The same metric may be used for both applications.

Several metrics were identified and compared by Imai, Rosen, and Berns in an excellent paper presented at CGIV 2002. [1] Unfortunately, only six near-matches (three metameric and three parameric) were numerically compared, leaving some uncertainty as to the magnitudes produced by each. Further, it is unclear the extent to which each metric produces uniform results for visually equivalent degrees of mismatch. This paper seeks to address these questions.

When assessing the extent to which a particular metric tracks human perception, it is tempting to use one of the established metrics, such as CIELAB total color difference (ΔE^*_{ab}), CIE94, or CIE DE2000. Unfortunately, such color difference formulae, which are based on integrated tristimulus values, discard any metameric component. This metameric component is the *raison d'être* for multichannel visible spectrum imaging, so it cannot be ignored.

Accordingly, comparisons based on tristimulus values computed under a single set of conditions (observer and illuminant) will be of use only for pairs of spectra which are non metameric. The CIE 1976 L*, a*, b* color difference formula, whose value, ΔE^*_{ab}, is approximately visually uniform, may be used to assess the visual difference between pairs of non-metameric spectra. We use the ΔE^*_{ab} (henceforth, simply ΔE*) between pairs of non metameric spectra in this paper.

While we recognize that the difference between spectra which will be compared *in praxis* will have both metameric and non-metameric components, we feel it is important to evaluate how closely the metrics track human perception wherever it may be meaningfully quantified. As we have argued, this perceptual difference may be meaningfully quantified by ΔE* for non-metameric pairs of spectra.

Metrics for Spectral [Mis-]Match

Imai, *et al.,* considered the following figures of merit for spectral match (or mis-match): [1]

- Unweighted Root Mean Square (RMS) Difference
- Hernández-Andrés, Romero Goodness of Fit Coefficient (GFC) [2]
- Special Index of Metamerism after Fairman's Parameric Decomposition [3]
- Viggiano's Spectral Comparison Index (SCI) [4, 5]
- Weighted RMS (weights chosen as reciprocals of spectral reflectances of standard)

- Weighted RMS (weights equal to diagonal of Matrix **R**)

When considering only non-metameric pairs, Fairman's Parameric Decomposition will tend to produce null results, as it examines only the difference between the metameric black spectra. Because in this study all pairs of spectra being compared had identical metameric blacks, we dropped this metric from our investigation.

DEFINITION AND COMPUTATION OF NON-METAMERIC SPECTRA

While the definition of metameric spectra is straightforward (two spectra which differ in the visible portion of the spectrum, but produce identical tristimulus values for a given combination of observer and illuminant), it is less straightforward to define the conditions which apply to non-metameric (but different) spectra.

A mathematical definition of *metameric* spectra are spectra which have identical fundamental spectra, but different metameric black spectra. A logical definition, then, of *non-metameric* spectra would be two spectra which have identical metameric black spectra, but possibly different fundamental spectra. The fundamental of a spectrum is its projection onto observer/illuminant space, and may be computed as follows:

$$\beta_F = W^t \cdot (W \cdot W^t)^{-1} \cdot W \cdot \beta \tag{1}$$

where β is the column vector containing the spectral radiance ratios (*e.g.,* reflectances);

W is a 3-rowed matrix containing weights for tristimulus integration (the combined effect of observer and illuminant); and

β_F is the column vector containing the fundamental spectrum.

(The matrix product which pre-multiplies the vector β in Equation (1) is Cohen's Matrix **R**.) [6]

The metameric black is the residual which is not accounted for by the projection, and is computed as the residual:

$$\beta_B = \beta - \beta_F \tag{2}$$

where β_B is the column vector containing the metameric black spectrum. One may then define two spectra as non-metameric if they have the same metameric black spectrum. This is the definition of non-metameric spectra we shall use in this investigation.

The last two factors in Equation (1) are the tristimulus values of the spectrum, which may be placed in a three-element column vector **x**:

$$x = W \cdot \beta \tag{3}$$

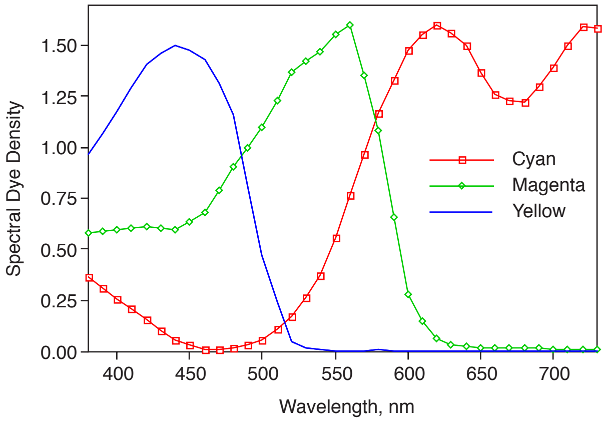

Figure 1: The dye density spectra, from which the Standard spectra were computed, are shown at their maximum concentrations.

This implies that, given an observer/illuminant combination, the fundamental spectrum depends solely upon the tristimulus values:

$$\beta_F = F \cdot x \tag{4}$$

where **F** is a 3-rowed matrix, and is computed as:

$$F = W^t \cdot (W \cdot W^t)^{-1} \tag{5}$$

A method of generating a Trial spectrum which is non-metameric to a given Standard spectrum, based on the definition given above, is:

> Given a Standard spectrum, and the tristimulus values of the Trial spectrum, compute the metameric black spectrum of the Standard (via Equations [1] and [2]) and add it to the fundamental spectrum of the Trial spectrum (via Equation [4]) to obtain the Trial spectrum.

EXPERIMENTAL

In schema, our experiment was:

- Generate 4096 reflectance spectra to serve as standards;
- Generate non-metameric companions (trial spectra) for each that differ by twelve different values of ΔE* (0,1, 0,5, 1, 2, . . ., 10) in the following ways:
    - Fundamental spectra differ in Lightness
    - Fundamental spectra differ in Chroma
    - Fundamental spectra differ in Hue
- Compare the original (Standard) spectra to each of the Trial spectra generated, using the several comparison metrics
- Determine which metric has the greatest precision and the closest correlation to the visual-based difference.

Evaluation Criteria

For the purpose of this investigation, we define the precision of a spectral match metric as its ability to produce a tight cluster of values for non-metameric standard/test pairs which differ by a certain, constant, amount in human perception. We define the accuracy as the extent to which a constant of proportionality characterizes the relationship between a metric's value and the corresponding perceptual difference for non-metameric pairs of spectra. Naturally, it is important that metrics be both precise and accurate.

It is especially important that the precision be high when the match is close, as the optimization criterion is most critical in an optimization problem at this point. This means that the dispersion of metric values must be small when the perceptual difference between non-metameric pairs of spectra is small.

Note that the GFC criterion assumes a value of unity for a perfect match, while the others, including ΔE*, yield zero. We complement the GFC by subtracting it from unity. We refer to this modified metric as "CGFC," for "Complemented GFC."

Given a proportional relationship, if a metric value is divided by ΔE*, a constant value (the constant of proportionality) should result. Because the different metrics shall have different scalings, it is not appropriate to compare the standard deviations of their proportionality constants in order to assess precision and accuracy. However, the Coefficient of Variation (CV), which is the standard deviation divided by the mean, is dimensionless and shall remove the non-uniformity. We use this as our evaluation criterion.

In order to assess precision, the CVs at a single level of ΔE* are compared. In order to assess accuracy, the constant of proportionality should be invariant with respect to ΔE*, and the metric value should exhibit a proportional relationship with ΔE*. It is highly desirable that a metric for spectral match possess both precision and accuracy.

One is left with the question of how large a Coefficient of Variation must be to be considered "big." Because they assume only positive values, the spectral match metrics will probably exhibit skewed probability distributions. Such skewed distributions as Exponential and Gamma/Erlang/Chi-Square may be appropriate models, and offer clues. The Exponential distribution has a Coefficient of Variation of unity, while a Chi-Square distribution with 4 degrees of freedom (a Gamma/Erlang distribution with shape parameter 2) has a Coefficient of Variation of $^{\sqrt{2}}/_{2}$, or about 0,7. Coefficient of Variation values greater than these may be considered large.

Rejection of Physically Implausible Spectra

Because the metameric black component of the spectrum of a real surface color contains negative values (unless the spectrum is its own fundamental), there is no guarantee that, when added to a different fundamental spectrum, the resultant spectrum will be positive valued. Therefore, we reject all trial spectra thus generated which have one or more negative values.

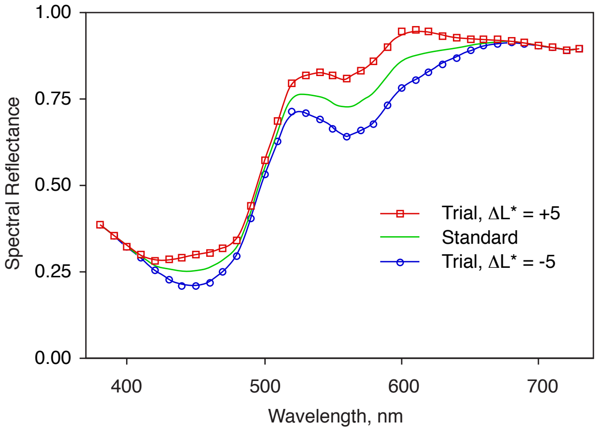

Figure 2: A Standard spectrum (middle curve) is flanked by two non-metameric Trial spectra which differ from the Standard by 3 units in L (upper and lower curves). All three spectra share a common Metameric Black spectrum, and, under the definition advanced in this paper, are regarded as "non-metameric."*

Although it is likewise possible to generate spectra with values greater than unity, we elected to retain these in the study. Such spectra can and are caused by fluorescence, self-luminosity, differences in lighting, and other causes.

Generation of Standard Spectra

We used the Beer-Bouger-Lambert law to generate a series of Standard Spectra. Three subtractive dyes were posited (Figure 1 shows the optical density spectra of the three dyes at maximum concentration). As a modest simplification, we assumed a perfect diffusing substrate, so the Standard spectra were generated according to:

$$\beta\ (\lambda) = exp\ [-c \cdot k_c(\lambda) - m \cdot k_m(\lambda) - y \cdot k_y(\lambda)] \qquad (6)$$

where c, m, and y are the concentrations of cyan, magenta, and yellow dyes; $k_c(\lambda)$, $k_m(\lambda)$, and $k_y(\lambda)$ are the effective spectral extinction coefficients of the three dyes at unit concentration; and β(λ) is the reflectance (radiance ratio) at wavelength λ. The effective extinction spectra were computed from the dye density spectra illustrated in Figure 1 by multiplying by the natural logarithm of 10. (We use the term "effective" to denote the fact that they account for two passes through the dye layer, rather than one.)

The wavelength range we used was 380 to 730 nanometers, with a sample every 10 nanometers. The dye concentrations c, m, and y assumed sixteen different levels, between zero and unity.

A standard spectrum generated using this technique appears as the middle curve in Figure 2.

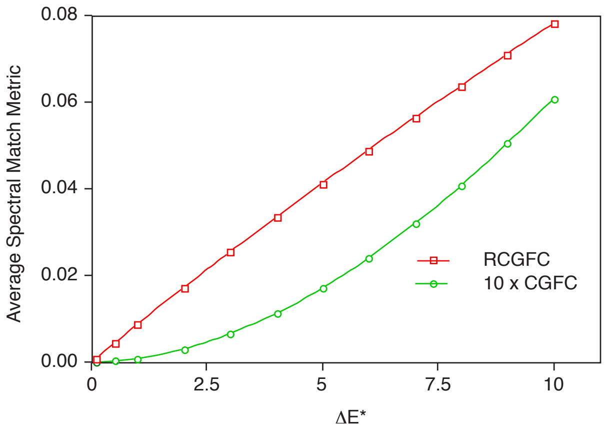

Figure 3. Average spectral match metrics as functions of ΔE for the Complemented Hernández-Andrés, Romero metric (CGFC, bottom curve) and its square root (RCGFC, upper curve). Because of a difference in scale, the latter is plotted at 10 times its actual value. This plot illustrates the non-linear manner in which the CGFC metric tracks ΔE*. The RCGFC enjoys a reasonably proportional relationship with ΔE*.*

Generation of Trial Spectra

The tristimulus values of the Standard spectra were determined, and Trial spectra were generated which differed from the standard in a number of ways: L* higher, L* lower, Chroma higher, Chroma lower, and Hue angle was increased and decreased. While this is by no means an exhaustive list of possible perturbations, we felt it was nevertheless sufficiently representative for this study. There were 12 levels by which each perturbation was performed, to produce ΔE*s of 0,1, 0,5, 1, 2, …, 10, for a potential of 72 Trial spectra from each Standard spectrum.

When the perturbation was in L*, the CIELAB coordinates of the trial spectrum were computed as:

$$\begin{aligned} L^*_t &= L^*_s \pm \Delta E^* \\ a^*_t &= a^*_s \\ b^*_t &= b^*_s \end{aligned} \tag{7}$$

where L^*_s, a^*_s, and b^*_s are the CIELAB coordinates of the standard, and L^*_t, a^*_t, and b^*_t are the CIELAB coordinates of a particular trial spectrum. The CIELAB coordinates of the trial were then converted back into tristimulus values, and Equation (4) was applied to obtain the fundamental spectrum of the Trial. Finally, the metameric black spectrum of the Standard was added to obtain the Trial spectrum.

To effect perturbations in Chroma, the CIELAB coordinates of the trial spectra were computed as:

$$\begin{aligned} L^*_t &= L^*_s \\ a^*_t &= a^*_s \cdot (C^*_s \pm \Delta E^*) / C^*_s \\ b^*_t &= b^*_s \cdot (C^*_s \pm \Delta E^*) / C^*_s \end{aligned} \tag{8}$$

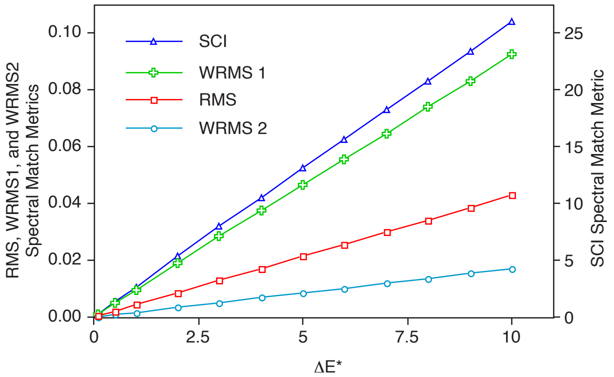

Figure 4. Average spectral match metrics as functions of ΔE.These metrics all exhibit excellent linear tracking of ΔE*. Because of its very different scale, the Viggiano Spectral Comparison Index uses the vertical axis to the right. The other metrics use the axis at the left. WRMS1: RMS Reflectance Difference weighted by reciprocal of spectral reflectance of Standard. WRMS2: RMS Reflectance Difference weighted by diagonal of Matrix R.*

where C^*_s is the Chroma of the standard. We discarded any perturbations for which ΔE* was greater than C^*_s (when this limit is exceeded, the color difference will have a component in Hue, and not just in Chroma). For the perturbations which were retained, the CIELAB coordinates were converted back into tristimulus values, thence into a fundamental spectrum, as was done for perturbation in Lightness.

Finally, the perturbations in Hue were effected via:

$$\begin{aligned} L^*_t &= L^*_s \\ a^*_t &= C^*_s \cdot cos\ (h_s \pm \Delta h) \\ b^*_t &= C^*_s \cdot sin\ (h_s \pm \Delta h) \end{aligned} \tag{9}$$

where h_s is the Hue angle standard, and the Hue angle difference Δh was computed by re-arranging a variation (in which ΔH* = ΔE*, so the Chromas are equal) of the formula in Séve: [7]

$$\Delta h = acos\ [1 - \tfrac{1}{2}\ (\Delta E^* / C^*_s)^2] \tag{10}$$

Again, we discarded any perturbations for which ΔE* was greater than the Chroma of the standard.

Finally, as was noted earlier, any trial spectrum which exhibited a negative reflectance at one or more wavelengths was rejected as physically implausible.

Two Trial spectra appear in Figure 2. They flank the Standard spectrum, and each differs from it by 3 units in L*.

For a given level of ΔE*, there are then a potential of 4096 x 6 = 24 576 different Trial spectra. Some shall be rejected, so the actual number tested shall be lower than this, but a large number should nevertheless remain.

ΔE*	Number of Trials	ΔE*	Number of Trials
0,1	24 572	5	23 460
0,5	24 572	6	23 057
1	24 568	7	22 699
2	24 536	8	22 192
3	24 221	9	21 727
4	23 859	10	21 206
		Total	280 669

Table 1. The number of Trial spectra generated for each level of ΔE are shown.*

RESULTS

A total of 280 669 Trial spectra were generated, 24 572 (out of a possible 24 576) at each of the two smallest levels of ΔE* (only the Chroma and Hue perturbations for the plain substrate were rejected at these levels of ΔE*). A summary of the number of Trial spectra generated for each level of ΔE* appear in Table 1.

The complement of GFC did not directly linearly track ΔE*, but its square root did to a reasonable extent, as is illustrated in Figure 3. Accordingly, we have replaced the complemented GFC metric with its square root, which we refer to with the symbol RCGFC. All other metrics exhibited proportional tracking of ΔE*; see Figure 4.

Table 2 contains the Coefficients of Variation (CVs) for the metrics which remain. Lower CVs imply greater precision. The CVs are presented for each of the 12 levels of ΔE* considered in this study.

DISCUSSION

Only two metrics, unweighted Root Mean Square difference and the Spectral Comparison Index, exhibited acceptable levels of precision at all levels of ΔE*. Of the two, the Spectral Comparison Index exhibited 34 percent greater precision. The diagonal of Matrix **R**-weighted RMS difference was imprecise at all levels of ΔE* considered. The remaining metrics, RCGFC and reciprocal-weighted RMS, exhibited acceptable levels of precision at larger levels of ΔE*, but were unacceptably disperse for the smaller values of ΔE*.

As a key applications for spectral match metrics is that of optimization criterion, the precision of the metric becomes increasingly important as the match is refined to increasingly smaller levels of perceptual difference. (Although there may be little point in doing so, one could start an optimization using nearly any reasonable metric, and switch to one with better precision as the optimization progresses. Doing the reverse would tend to produce a less desirable solution.)

One could argue that accuracy, or the degree to which a spectral match metric serves as a surrogate for a perceptual or other measure, is important at the beginning of an optimization process, and its precision becomes more important as the process (hopefully) converges to its solution. Ideally, a metric would be possessed of both attributes. Only the unweighted RMS and Viggiano SCI metrics performed well in both areas.

	CV of Spectral Match Metric				
ΔE*	RMS	RCGFC	SCI	WRMS1	WRMS2
0,1	0,82	10,58	0,60	11,15	54,79
0,5	0,82	2,19	0,60	2,30	10,99
1	0,82	1,20	0,60	1,26	5,54
2	0,82	0,78	0,60	0,81	2,86
3	0,83	0,67	0,60	0,70	2,01
4	0,83	0,63	0,61	0,66	1,61
5	0,84	0,61	0,62	0,64	1,39
6	0,84	0,59	0,63	0,64	1,26
7	0,85	0,58	0,64	0,64	1,17
8	0,85	0,57	0,65	0,64	1,11
9	0,86	0,56	0,66	0,65	1,07
10	0,86	0,56	0,67	0,65	1,04

Table 2. The Coefficients of Variation (CVs) of the spectral match metrics are presented as functions of of ΔE. Lower Coefficients of Variation imply greater precision. WRMS1: RMS Reflectance Difference weighted by reciprocal of spectral reflectance of Standard. WRMS2: RMS Reflectance Difference weighted by diagonal of Matrix R.*

CONCLUSIONS

Of the five metrics originally considered, only the Hernández-Andrés GFC failed to exhibit reasonable linear tracking of ΔE* for non-metameric spectra. It was discovered, however, that the square root of its complement did. Unfortunately, this metric and the resulting derived metric are insensitive to shifts in magnitude (two spectra which differ by a multiplicative factor have the same scores as a perfect match). While this is not a problem for the intended purpose of this metric (indeed, it is a feature, as the issue of magnitude is considered separately when comparing spectral power distributions of illuminants), it is a significant problem in the applications generally considered for spectral match metrics.

Of the other metrics, the unweighted RMS difference and the SCI performed with reasonable precision for all levels of ΔE* considered. The precision of the SCI was approximately one-third greater than that of unweighted RMS. Therefore, when a general match criterion is desired, unweighted RMS difference is a reasonable choice among those considered. In situations when a specific match metric, which depends upon the color under consideration, can be used, the Spectral Comparison

Index is the choice among all metrics examined in this study.

For non-metameric spectral pairs, Viggiano's SCI assumed values which were in the neighborhood of 2,6 times that of CIELAB total color difference. Because the CIELAB total color difference for metameric spectra (or the metameric component of parametric spectra) is identically zero, a similar comparison for pairs of spectra with a metameric component would be meaningless.

LITERATURE CITED

1. Francisco H Imai, Mitchell R Rosen, and Roy S Berns, Comparative study of metrics for spectral match quality. *Proceedings of CGIV 2002: the First European Conference on Colour Graphics, Imaging, and Vision,* p 492-496.
2. J Hernández-Andrés and J Romero, Colorimetric and spectroradiometric characteristics of narrow-field-of-view clear skylight in Grenada, Spain. *Journal of the Optical Society of America A,* 2001. **18** : 412 - 420.
3. Hugh S Fairman, Metameric correction using parameric decomposition. *Color Research and Application,* 1997. **12** : 5 : 261-265.
4. J A Stephen Viggiano, The comparison of radiance ratio spectra: assessing a model's "goodness-of-fit." *Advanced Printing of Conference Summaries: SPSE's 43rd Annual Conference,* 1990, p 222-225.
5. J A Stephen Viggiano, A perception-referenced method for comparison of radiance ratio spectra and its application as an index of metamerism. *Proceedings of AIC Colour 2001: The 9th Congress of the International Colour Association,* p 701-704.
6. Josef B. Cohen and William E Kappauf, Metameric color stimuli, fundamental metamers, and Wyszecki's metameric blacks. *American Journal of Psychology,* 1982. **95** : 537 - 564.
7. Robert Séve, New formula for the computation of CIE 1976 Hue Difference. *Color Research and Application,* 1991. **16** : 3 : 217 - 218.

BIOGRAPHY

J A STEPHEN VIGGIANO is Principal and Founder of Acolyte Color Research, for which he provides consulting services, algorithm design, and color- and image quality evaluation services. Prior to closing its image science division, RIT Research Corporation had employed Steve for over 14 years, where he had risen to the rank of Principal Scientist. Steve has taught graduate and undergraduate courses in image reproduction theory, printing inks, paper, color, and research methods at RIT.

Steve holds an AB degree in Mathematics from Thomas Edison College, and Master's Degrees from RIT in Printing Technology and Mathematical Statistics. He is a member of CIE TC8-02 (Color Differences in Images, for which he authored the section on statistics) and TC8-03 (Gamut Mapping).