Extending Image Quality Models

Peter G. Engeldrum
Imcotek
Winchester, MA 01890
pge@imcotek.com

Abstract

Image quality is known to be multivalued with some visual attributes or "nesses." One example of a "ness" is colorfulness. Published research has shown that the image quality versus colorfulness function reaches a maximum and increasing colorfulness beyond the optimum level actually degrades image quality. The present formulations of image quality models—e.g. Minkowski metrics and the Generalized Weighted Mean Hypothesis—implicitly assume that a monotonic relationship exists between image quality and the values of the independent "nesses." This paper proposes an extension to these popular image quality model formulations to represent the non-monotonic case. The new image quality model extension is compared to results of image quality versus colorfulness scaling of printed images.

Introduction

Image quality can be cast in terms of the component perceptual attributes, the "nesses". These component "nesses" provide the basis for the judgment of image quality the overall excellence of the image. Image quality models are just mathematical formalisms that provide a "ness" combination rule enabling the prediction of image quality from the known values of the "nesses."

However, the most successful forms of image quality models assume that the "nesses" are monotonic with image quality. If the "ness" is on a "goodness" scale, the value of the "ness" makes a positive contribution to image quality. An example of such a "ness" might be sharpness. On the other hand, a "ness" on measured by a "badness" scale causes a decrease in image quality as the value of the "ness" increases. Graininess is such a "ness."

In a series of experiments, Fedorovskaya, et. al.[(1)], de Ridder[(2)], and Fedorovskaya, et. al.[(3)] , showed that image quality was not a monotonic function of saturation, chroma and colorfulness. (They used these three different "ness" descriptors in each of their studies, for apparently the same percept, which will be called colorfulness in this paper.) Their reported image quality vs. colorfulness function reaches a maximum and then decreases as the colorfulness increases, and thus does not conform to the underlying assumptions of the most useful image quality models.

The purpose of this study was twofold. First, all the scaling experiments described in references (1-3) were performed with images displayed on a CRT. Given that there is no consensus on the effects of viewing mode on various image percepts, we wanted to confirm the general image quality vs. colorfulness function for reflection prints. The second purpose was to develop a model for these quality-colorfulness functions that would be sufficiently general for a variety of possible non-monotonic image quality/perceptual attribute relationships.

Extended Image Quality Model

Review

Many different forms of image quality models are used. However, to date, the most successful form of image quality model is the "power" model attributed to Bartleson[(4)] and of the following general form, equation (1):

$$IQ = \left(a1 * ness1^{p} + a2 * ness2^{p}\right)^{1/p} \qquad (1)$$

Here, p = the exponent or "power" of the model, *a1* and *a2* are coefficients and *ness*1, *ness*2 are the component attributes, the "nesses," of image quality, IQ.

There are several interpretations of equation (1), but the two most common are the so-called Minkowski distance metric, and the Generalized Weighted Mean Hypothesis[(5)] (GWMH). In the Minkowski distance interpretation, image quality is the distance with respect to the origin of the attribute coordinate system. The GWMH, on the otherhand, postulates that image quality is a generalized average of the component "nesses."

Another view is that equation is a combination rule, the rule used by observers to combine the values of the component "nesses" into image quality. For "dimensional" consistency, one can think of the weight factor "ness" product as an equivalent value of image quality. For example, suppose we ask observers to give image quality judgments of a set of images when only one "ness" is varying within the set. In this case, we can interpret the resulting image quality scale as the equivalent image quality value of the "ness" that varies. For many "nesses," there are at least monotonic, and often linear, relationships between the judged image quality scale and the "ness" scale. But this is not always the case, as demonstrated in the experiments described in references(1-3). See Engeldrum[(5)] for further details on some these image quality model interpretations.

Proposed Model

In the case where only one "ness" varies, equation (1) implies that image quality will be a linear function of that "ness." For colorfulness at least, the data reported in references (1-3) clearly show that this is not the case. This lack of montonicity may hold for some other "nesses," such as sharpness.

To model this sort of behavior, a two-function product formulation is proposed. The two functions are necessary to account for the asymmetry in the quality vs. colorfulness data from our experiment and the data in references (1-3). The first function, a symmetrical Gaussian-like function, is designed to capture the basic convex downward shape of the IQ-colorfulness data. Specifically we choose equation (2):

$$f1(x, x0, a, b) = \exp\left[-\frac{|x - x0|^{2/a}}{b}\right] \qquad (2)$$

Here x = the "ness," $x0$ = the peak of the IQ-ness curve, a = parameter the controls the rate that the function decays, and b is a width parameter. If $a = 1$ and $b = \sigma^2$, then $f1()$ is the familiar Gaussian function. To account for asymmetries that occur in typical data, an asymmetrical weight function—the well-known logistic function—is selected. With a simple change in the sign of a parameter, the logistic function can weight either "side" to increase the decay rate of the function. Equation (3) shows the functional form selected.

$$f2(x, x1, c) = \frac{1}{1 + e^{-c(x - x1)}} \qquad (3)$$

With two parameter values, both the location, $x1$, and the extent, c, of $f2()$ can be controlled. The sign of c determines if the slope of the function increases with increasing x, or with decreasing x.

The product of $f1()$ and $f2()$ and a scale factor, d, are taken as an overall "ness" weighting function. Equation (4), $fw(a,b,c,d,x0,x1)$, represents the complete empirical function for expressing non-monotonic, and non-linear, image quality vs. percept ("ness") relationships.

$$fw(a,b,c,d,x0,x1) = d * f1(x,x0,a,b) * f2(x,x1,c) \qquad (4)$$

The final parameter, d, is designed as a scale factor on the overall weight function. With six parameters to describe the weight function, there should be sufficient flexibility for fitting a wide range of real data.

For a multi-attribute image quality model along the lines of equation(1), equation(4) can be incorporated as shown in equation(5), but can be extended to as many "ness" components as needed. This formulation is based on the "equivalent image quality" idea of the component "nesses." Note that the aj weight and the scale factor, d, are redundant in this circumstance, and can be combined into one parameter.

$$IQ = \left(a1 * ness1^p + a2 * fw(a,b,c,d,x0,x1)^p\right)^{1/p} \qquad (5)$$

The author would like to be able to say that these functions represent some deep theoretical image quality basis, but he cannot. There is no particular theory governing the selection of these functions. They are purely empirical and seem to adequately describe the data.

Psychometric Scaling Experiments

Stimuli

Two images, one a still life, called "still", (Figure 1), and the second a scene of apartments, called "flats", (Figure 2) were used. The colorfulness of each of the scenes was manipulated in Adobe Photoshop® using the saturation tool. The nominal image was unadjusted, and the colorfulness of a series of 14 images was either increased or decreased by incrementing the Photoshop saturation slider by increments of 10 units. This yielded images with a range of colorfulness from almost a neutral black and white to extreme colorfulness that was clearly excessive.

Figure 1 "Still" image.

The resultant image files were compressed using JPEG and a Photoshop quality level of 90. Since the final images were sampled at 300 24-bit pixels per inch, the JPEG compression had negligible effect in the image quality.

EZprints, an Internet photofinisher, made photographic prints of the compressed image files. Overall, the prints were a close visual match to the CRT images, but no colorimetric measurements were made on either the prints or the CRT images to determine the exact nature of Photoshop's saturation control.

Figure 2 "Flat" image.

The photographic prints were surrounded by a one-inch gray cardboard matte board and backed with a white card stock. This arrangement forced the adapted, and assumedly adopted, white point to be in the images and not the paper border In addition, the backing increased the opacity of the sample image. Using a cardboard surround also made handling easier and kept the images clean and free of scratches and fingerprints.

On the bottom edge of the gray frame, a small downward–pointing triangle was fastened. This served as a reference point for the scaling process.

Observers
Nine observers, five males and four females, participated in this scaling study. Five observers scaled both sets of images. Of the remaining four two scaled the "still" image and two scaled the "flats" image. Seven observers scaled each image.

Four of the observers had considerable professional experience in judging the attributes of image quality. The remaining five were inexperienced in making quality judgments.

Scaling
Two psychometric scaling studies were conducted using the graphical rating scale (GRS) method[6]. Two images were scaled by each observer for two attributes, colorfulness and image quality. The first attribute-image combination that was scaled was randomized among the observers. Some observers scaled colorfulness-"still" image first, and then switched both attribute and image. Thus, the scaled attribute and image alternated during the course of the four scalings performed by each observer.

No anchors or references were used, and definitions of image quality and colorfulness were not provided to the observers. A random three-digit number was placed on the back of the print for identification.

Scaling was performed in two different locations over a period of four months. In both cases the light source was coolwhite fluorescent with an illuminance on the "ruler" of approximately 770 lux.

Instructions
The instructions given to the observers described the graphical rating scale task. In each of the two scaling studies, image quality and colorfulness, the set of instructions was the same. The following illustrates the exact instruction set for each "ness" scaled with the exception that the words "image quality" were substituted for "colorfulness" where appropriate.

Thank you for participating in our study.
We want to get your opinion of the colorfulness of several images. To do this, we ask that you view some samples and rate them according to your opinion of overall colorfulness.
In front of you is a ruler, with values from 0 to 100, that you will use to determine your ratings of colorfulness. A higher number on the scale indicates more colorfulness.
Place each sample above the ruler according to the amount of overall colorfulness. Arrange the samples so the higher colorfulness samples are on the right and the lower colorfulness samples are on the left. Please make sure the ***distance between the samples is proportional to the difference in colorfulness****. Use the downward pointing triangle on the sample as the sample position reference. If two or more samples have the same colorfulness, place one sample above the other. Feel free to adjust the samples until you think the distances between the samples represent the differences in colorfulness. You need not place the samples at the "tic" marks, please use the whole scale.*
Before you start, please look through all the samples. There is no time limit, and there are no right or wrong answers. We are seeking your opinion.
Are there any questions?

When the observer's indicated that they finished the scaling task, they were instructed to hand to the test conductor each stimulus along with the "ness" rating (number) assigned. Data were tabulated in a matrix where the rows are the observers and the columns the print samples (stimuli). Matrix entries are the ratings assigned by each observer to each of the prints.

Scale Generation
Computing the scale values for image quality and colorfulness is straightforward using the GRS method[6]. Since no anchors or image references were used, the first step is to account for the different "ruler" usage by each observer. This is accomplished by subtracting the observer's mean rating from each of the stimuli rating and dividing by the observer rating standard deviation[6]. Normalizing the ratings in this way puts each observer's rating data on a scale that has zero mean and unit standard deviation. The mean value adjusts for the average position used by the observer and the standard deviation adjusts for the range of the ruler used. Processing the data in this way eliminates the variation among the observers, and tends to reduce the variation in the computed interval scale.

Scale values for both image quality and colorfulness were determined by computing the column average of the normalized data matrix.

Results

Psychometric Scaling
It seemed prudent to evaluate the agreement among the observers for each of the attributes and images. Ordinarily one could perform an appropriate Analysis of Variance on this interval data to see if the observers were different. However, our normalization procedure eliminated any difference between observers, so we chose instead to test the agreement of *ranks* among the observers. For each of the four scaling studies (two attributes times two images), we computed the sample *ranks* of each observer and used Kendall's Coefficient of Concordance[7] to see if there was any disagreement. At a $\alpha = 0.05$ we found no difference in the *ranks* for the observers.

Figure 3 shows the image quality scale value versus the colorfulness–scale value for the "still" image. Figure 4 shows the same for the "flats" image. (The solid lines are the model fits explained below.) Note that for these two images, the image quality reaches a maximum and then decreases, the exact behavior found in references (1-3).

Figures 3 and 4 also illustrate, for these images, that the quality vs. colorfulness relationship is neither symmetric with respect to the maximum image quality, nor the same. It is this asymmetry and unequal behavior, also shown in

previous research[1,2,3] which suggested the functional form of the model in equation(5).

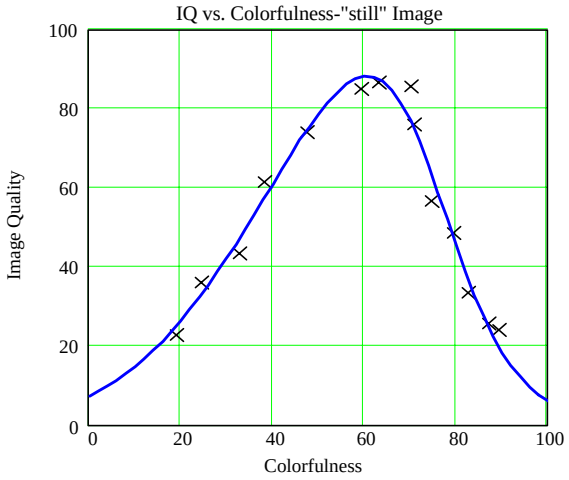

Figure 3 "Still" image results. Crosses are data and solid line is IQ model fit. RMS = 4.18.

Model

The image quality versus colorfulness data for both images was fitted via least-squares, using the Mathcad 2001Pro application. Although this is a convenient and practical approach, it is not strictly correct because both the image quality and colorfulness data have error. Least-squares theory assumes that the independent variable—colorfulness in this case—is without error and attempts to minimize the error in the independent, IQ data. This is clearly not the case, so the model parameter estimates may not be "optimum," in some sense.

Table 1 Extended Image Quality Model Parameters

Image	x0	x1	a	b	c	d	RMS
Still	136.7	74.75	0.573	947669	-0.124	153.2	4.18
Flat	56.21	95.24	0.804	10801	-0.107	75.42	3.99

Figures 3 and 4 show that the model fits to the scaled data for the two images. Table I contains the model parameters and the RMS deviation about the model for the two images. The RMS deviations are about 4.0 for both images. On a scale range of 100, this represents quite a small error considering the limited number of observers.

Conclusions

An extension to a popular image quality model has been proposed and tested with scaled colorfulness and

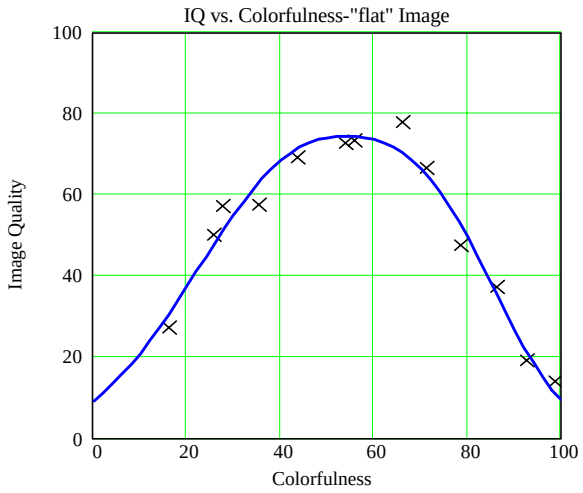

Figure 4 "Flat" image results. Crosses are data and solid line is model fit. RMS = 3.99.

image quality data. The fit of the model to the data is very good, with an RMS error, for a single attribute image quality scaling, of about 4.0. The general applicability of this new model extension awaits further experimental data.

References

1) Fedorovskaya, E. A., F. J. J. Blommaert, and H. de Ridder, *Perceptual quality of color images of natural scenes transformed into CIELUV color space*, IS&T & SID's Color Imaging Conference Proceedings, pg 37 (1993).
2) de Ridder, H. *Naturalness and Image Quality: Saturation and lightness variation in color images*, Jour. Imag. Sci. & Tech. 40:487(1996).
3) Fedorovskaya, E. A., H. de Ridder, and F. J. J. Blommaert, *Chroma variations and perceived quality of color images of natural scenes*, Color Res. & Appl. 22:96(1997)
4) Bartleson, C. J., *The combined influence of sharpness and graininess on the quality of colour prints*, Jour. Photog. Science 30:33(1982)
5) Engeldrum, P. G., *A framework for image quality models*, Jour. Imag. Sci. & Tech. 39:312(1995)
6) Engeldrum, P. G., **Psychometric Scaling: A Toolkit for Imaging System Development**, Imcotek Press, Winchester, MA 2000, ISBN 0-9678706-0-7
7) Kendall, M., & J. D. Gibbons, **Rank Correlation Methods - Fifth Edition**, Oxford University Press, NY, NY 1990.