Predicting Multivariate Image Quality from Individual Perceptual Attributes

Brian W. Keelan
Eastman Kodak Company
Rochester, New York, USA

Abstract

A general multivariate formalism has been developed for predicting the overall quality of an image given the impact that each influential quality attribute would have in isolation. The quality change associated with each attribute is expressed in just noticeable differences (JNDs). These component changes are combined using a variable-power Minkowski metric, in which the value of the power reflects the degree to which dominant contributions suppress lesser effects. The formalism satisfactorily explains the overall quality of multivariate images from diverse experiments involving different psychometric tasks, viewing modalities, and attributes, including image structure, digital artifacts, and color and tone reproduction.

Introduction

A persistent issue in the field of image quality perception has been understanding how the overall quality of a multivariate sample is related to its individual attributes. Without a general rule for combining the effects of different contributors to image quality, psychometric experiments grow exponentially in size as the number of attributes considered increases, and predictive modeling becomes impractical.

Various attempts have been made to predict multivariate quality from a knowledge of the univariate (isolated) impact of contributing attributes. In 1964–65, Prosser, Allnatt, and Lewis found impairments, defined to be harmonically related to the complement of 1–5 scale quality ratings, to sum directly in experiments on monochrome television images;[1,2] however, the crude quantization of the scale and range effects limited the usefulness of these findings. In 1982, Bartleson modeled overall quality as a Minkowski sum (n^{th} root of the sum of n^{th} powers) of sharpness and the complement of graininess, each being expressed on a 1-9 interval scale,[3] but his mathematical treatment was not extensible to additional attributes without empirical adjustments. In 1992, de Ridder proposed the use of Minkowski metrics based on fractional quality loss compared to the maximum quality loss produced by each attribute, but the maximum values were specific to an experiment rather than general in nature and a somewhat arbitrary renormalization was required to fit the data.[4] None of these methods, nor others proposed in the external literature, have been successful in explaining the results of more than a single experiment for which they were optimized.[5]

In contrast, the multivariate formalism described in this paper, based on a variable power Minkowski metric, is successful in explaining results from four different experiments. This formalism is described in some detail in Chapter 11 of the author's forthcoming *Handbook of Image Quality: Characterization and Prediction,*[6] from which the figures in this paper are taken with the permission of the publisher, Marcel Dekker, Inc. The *Handbook of Image Quality* also provides many examples of the application of multivariate predictions to practical problems.

The Multivariate Formalism

The basic tenet of the multivariate formalism is that a universal relationship exists between a list of quality changes arising from a set of independent attributes, in isolation, and the overall quality change when all attributes are present in the same sample, provided that attribute and overall quality changes are expressed in strictly equivalent units. The choice of quality change units employed in the multivariate formalism is JNDs of overall quality. It is important to distinguish between JNDs of an attribute (a measure of detectability of differences in appearance) and JNDs of quality arising from that attribute (a measure of the significance of the appearance difference in terms of quality). A JND of quality arising from an attribute generally corresponds to a larger stimulus change than does one JND of that attribute, because it is usually possible to detect a change in appearance before the change is of significance in terms of its impact on quality. JNDs of individual attributes are not strictly equivalent units because equal detectability of different attributes does not imply equal impact on quality.

The JNDs used in this paper are designated as 50% JNDs, because they correspond to a stimulus difference that is distinguished 50% of the time. In a forced choice paired comparison, if stimuli differ by one 50% JND, a 75%:25% proportion results, because the 50% who see the difference all provide the "correct" answer, and half of the remainder guess correctly by chance. With minor reinterpretation, JNDs may also be used to quantify preferential attributes.[7]

Four mathematical requirements are placed upon the universal relationship between a list of JNDs of quality changes from individual attributes and the overall quality change in JNDs. These four requirements, listed below, are quoted directly from Ref. 6.

1. A list of only one quality difference must map to itself, i.e., if only one attribute affects quality, the overall quality difference must equal the quality difference arising from that attribute. This first identity requirement is a mere formality, but is stated for completeness.
2. Adding an element equal to zero to the quality difference list does not change the overall quality difference. This second identity requirement seems trivial, but it is easily violated inadvertently, e.g., by using an average of the elements of the list in the relationship.
3. When attribute quality differences are small in magnitude, they approximately sum. This additivity requirement is intuitively plausible; if each of three attributes in isolation degrades quality by one JND, it seems reasonable to expect that the overall quality loss corresponds to approximately three JNDs, because such subtle changes are unlikely to affect one another very much.
4. When one or more attribute quality differences are large in magnitude, modest changes in other attribute quality differences have little impact on overall quality. This criterion is called the suppression requirement because the presence of a serious degradation suppresses the impact of minor degradations on overall quality. Conversely, and more importantly, if an imaging system has several minor flaws and one major flaw, fixing the minor flaws will not yield much improvement because the major flaw largely determines the quality. In common parlance, the suppression requirement reflects the notion that the worst problem dominates.

There are many possible combination rules meeting these requirements. Among the simplest are those involving Minkowski metrics, which are generalized distance metrics, and in our application would take the form:

$$\Delta Q_m = -\left(\sum_i \left(-\Delta Q_i\right)^{n_m} \right)^{1/n_m} \qquad (1)$$

where ΔQ_i is the quality change arising from the i^{th} attribute, ΔQ_m is the overall quality change, and n_m is the power of the metric (which need not be an integer). All quality changes here are assumed to correspond to degradations and so to be negative, hence the negative signs in Eq. (1), which ensure that only positive numbers are raised to a power. When $n_m = 2$ the Minkowski metric is a root-mean-square (RMS) sum and so corresponds to a normal Euclidean distance.

Equation (1) meets the two identity requirements (#1 and #2). The suppression requirement (#4) is met when $n_m > 1$, but the additivity requirement (#3) is only met when $n_m = 1$. This result can be seen by considering the sum of N equal components, which is N raised to the $1/n_m$ power times as large as any individual component, instead of N times as large, as should be the case in the additive regime. Consequently, the additivity and suppression requirements cannot both be met simultaneously by a Minkowski metric. This consideration suggests the use of a variable power Minkowski metric. In particular, we use a power of the form

$$n_m = 1 + c_1 \cdot \tanh\left(\frac{(-\Delta Q)_{\max}}{c_2} \right) \qquad (2)$$

where $(-\Delta Q)_{\max}$ is the most severe component degradation. The constants c_1 and c_2 are determined by empirical optimization, as described subsequently. Because the hyperbolic tangent of a positive argument ranges from zero to one, the power n_m varies from unity to $1 + c_1$ continuously as the greatest attribute quality loss increases. Placing an asymptotic upper bound on the power helps to insure robust behavior at greater degradations.

Experimental

Data from four independent experiments for which it is possible to convert all assessments to JNDs of quality are considered in this analysis. The first of these experiments chronologically is that of C. James Bartleson[8] in 1982, in which reflection prints covarying in modulation transfer function (MTF) and film granularity were rated for sharpness, graininess, and overall quality on 1-9 scales using a categorical sort procedure. A subsequent internal Eastman Kodak Company study carried out by W. Mitchell Burke in 1983 provided sufficient information for these three rating scales to be converted to JNDs of overall quality. In addition, using similar methodology, Burke studied samples covarying in MTF, film granularity, and camera negative exposure. The latter primarily affected quality through tonal clipping (truncation) of shadow detail in underexposures, although there was also a second-order influence on image structure characteristics. These results constitute the second data set analyzed herein.

The third and fourth experiments considered have been carried out recently using quality rulers, which directly yield JNDs of overall quality. The hardcopy quality ruler has been described previously.[9] In brief, it consists of a series of stimuli varying in only a single characteristic (in this case, MTF), and spaced apart by about three JNDs of overall quality. The images are mounted in order of quality, in a sliding fixture that allows any of the reference images to be brought into close physical proximity with a test sample depicting (in the present work) the same scene, but varying in different attributes. The observer slides the ruler back and forth until the point of equality between the ruler and test sample is identified, from which the JNDs of overall quality of the test sample directly results. The softcopy quality ruler, which has also been described previously,[10] is based on the same general principles, but the implementation differs because of the nature of the display. Reference images varying in a single characteristic (again, MTF in the present case) and spaced by approximately one JND of

quality are computed and stored on disk. A randomly chosen reference image and the test image are displayed on carefully matched monitors and the observer indicates which image is of higher quality. A binary search through the reference samples ensues, until the overall quality of the test sample has been bracketed to the desired precision.

In the third experiment, hardcopy samples covarying simultaneously in four attributes were assessed against a hardcopy quality ruler. Building on the progression already provided by the two-attribute MTF and film granularity, and three-attribute MTF, film granularity, and camera negative exposure work, this study involved variations in MTF, digital isotropic noise, tonal clipping, and streaking. Isotropic noise is a general term describing noise that does not vary significantly with spatial direction; film granularity is one example but electronic sensors often produce noise that is approximately isotropic as well. The tonal clipping variations emulated that resulting from misexposures in digital still cameras. Streaking is a digital artifact usually associated with output devices having linear arrays of marking elements that are not perfectly matched, causing stripes of differing density parallel to the direction of travel during writing. The noise power spectrum (NPS) of streaking so defined is a broadband feature lying along the spatial axis perpendicular to the streaks produced.

Three levels of each of the four attributes were chosen to produce approximately 4, 8, and 12 JNDs of quality loss in isolation. A total of 42 positions were digitally simulated in each of four scenes. Twelve univariate positions containing each level of each attribute were included, as was a null image, having no introduced degradation. The remaining 29 multivariate positions sampled from among the $3^4 = 81$ possible combinations of simultaneous non-zero degradations in all four attributes. Twenty-two observers assessed the stimuli, and for this analysis, responses were averaged over all observers and scenes. This yielded mean values in JNDs of the overall quality change for the multivariate and univariate positions; the latter measured the average impact of each of the contributing individual attributes in isolation.

In the fourth experiment, images covarying in color balance and tone reproduction (primarily midscale contrast) were assessed using the softcopy quality ruler to determine the applicability of the multivariate formalism to preferential attributes. Four color balance positions, involving different hue shift directions from the average preferred position, were chosen to have approximately 2, 4, 8, and 12 JNDs of quality loss. Three tone reproduction positions were selected that had approximately 2, 4, and 8 JNDs of quality loss. All univariate and multivariate combinations were simulated, including a null sample, so the experiment constituted a full factorial design with 20 positions. Two additional positions outside the factorial design were included for other reasons, leading to a total of 22 positions, which were simulated in each of 6 scenes. Twenty-three observers assessed the stimuli and the results were pooled over scene and observer, yielding mean JNDs of overall quality change for each of the bivariate and univariate positions.

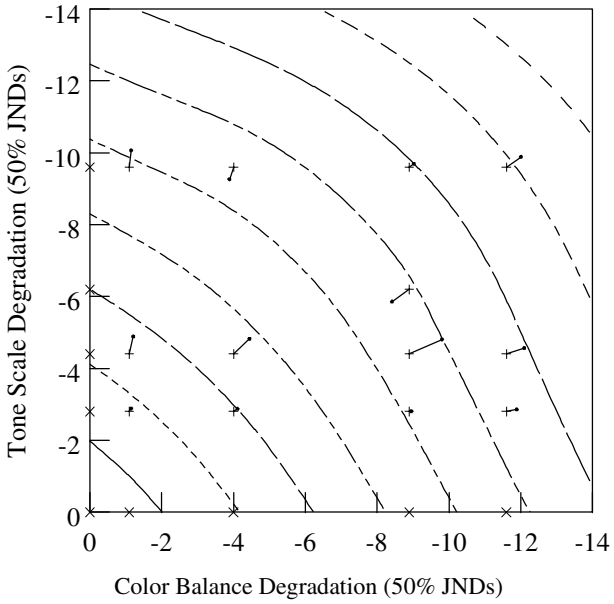

Figure 1. Overall quality contours for samples varying in color balance and tone scale (principally contrast).

Bartleson[8] graphically depicted his two-dimensional results by plotting iso-quality contours against the two contributing attribute levels. A similar plot is presented in Fig. 1, based on the softcopy data; "X" symbols mark the univariate positions, and plus symbols identify the bivariate positions. The Statistical Analysis Software (SAS) "Gcontour" procedure[11] was used to generate the best fit contours shown in the figure; these contours are not based on the multivariate formalism equations, but rather are simply a graphical means of examining the data. The contours correspond to quality losses of 0, –2, –4, … –18 JNDs. To indicate the goodness of fit, dots are located on the contours of the response surface at the locations of the judged value of the test level and are connected with a radial line segment to the respective univariate levels of the tone and color balance degradations contained in the test level. For example, the nominal (-8, -8) test level, with actual univariate assessments at (-8.9, -9.6), was judged at –14.0 JNDs (solid dot) but would be predicted to be at –13.8 JNDs (plus symbol).

Figure 1 demonstrates the principles of additivity (item #3 in the previous section) and suppression (item #4). At lesser degradations (lower left corner), the contours are nearly straight lines at 45° angles, as expected if the attributes were additive. At greater degradations (farther up and/or right), the contours become more sharply curved, reflecting the suppression by the dominant attribute.

Results

The constants c_1 and c_2 in Eq. (2) will be empirically determined by optimizing the agreement of predictions of the multivariate formalism with the experimental data, with particular emphasis on fitting the results of the two more

recent and rigorously calibrated experiments. In Fig. 2, measured data from all four experiments, shown as individual symbols, are compared against the predictions of Eqs. (1) and (2), which plot as a 45° line. The Bartleson two-attribute data are plotted as circles; the Burke three-attribute data as plus symbols; the four-attribute hardcopy quality ruler data as asterisks; and the two-attribute softcopy quality ruler data as triangles. The data, which span a rather wide range of ≈22 JNDs, are all very well fit by the optimized multivariate formalism, with the exception of the very lowest quality results from the Bartleson experiment, which exhibit stronger suppression than is observed in the other experiments. It is likely that this apparent suppression is a result of range effects (saturation) near the ends of the categorical rating scale used in that experiment.

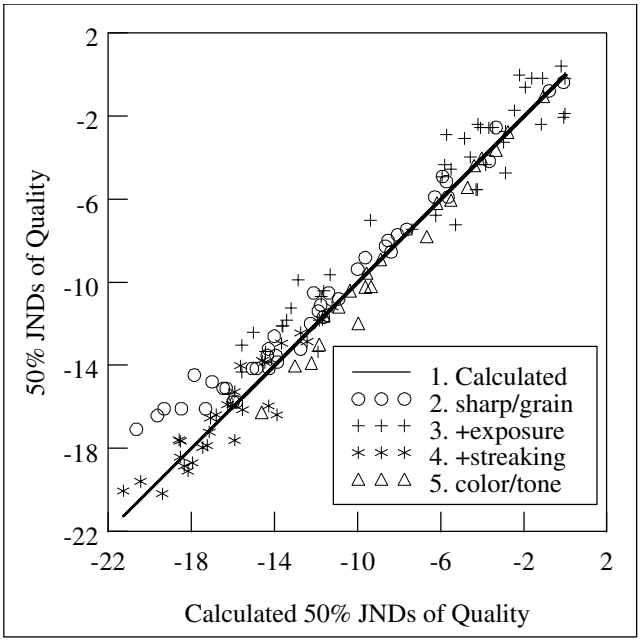

Figure 2. Calculated versus measured overall quality of multivariate samples from four independent studies.

Selection of the constants c_1 and c_2 in Eq. (2) to yield the result of Fig. 2 was achieved as follows. A nonlinear regression was run to optimize the fit to the two recent experiments and the two fit parameters were noted to be strongly negatively correlated, as might be expected based on a Taylor series expansion of the hyperbolic tangent. As long as $c_1 \geq 2$, predictions of the experimental data were fairly independent of the individual values of c_1 and c_2, depending only on their product. Therefore, the minimum value of $c_1 = 2$ was chosen to minimize the range of values assumed by the variable Minkowski power, to improve robustness. With this choice, only one fit parameter, c_2, remained. The regression was run again with c_1 fixed, yielding $c_2 = 16.9$. Consequently, Eq. (2) becomes;

$$n_m = 1 + 2 \cdot \tanh\left(\frac{(-\Delta Q)_{\max}}{16.9} \right) \quad (3)$$

The range of possible values of this Minkowski power is from one to three. Therefore, the quality changes arising from the individual attributes add in a fashion varying continuously from summing linearly for small changes, to adding as the cube root of a sum of cubes at large changes. When the maximum degradation equals $-19.9 \cdot \tanh^{-1}(1/2) \approx -9.3$ 50% JNDs, the degradations add as an RMS sum.

The success of the multivariate formalism in explaining the results of these four experiments is quite remarkable, particularly given the use of essentially only a single fit parameter. Both the diversity of types and numbers of attributes varied, and the variety of psychometric and display methods employed, support the general validity of the multivariate formalism.

Discussion

The implications of the multivariate formalism are most easily understood by considering several simple examples. Table 1 shows multivariate sums based on Eqs. (1) and (3) for several contrasting cases. The first row demonstrates the additivity requirement because the three small degradations nearly arithmetically sum (they do not exactly sum because suppression does not vanish until zero JNDs of degradation are approached). The second row shows that as degradations increase, the shortfall relative to additivity grows because suppression increases. Still, when the individual degradations are comparable in magnitude, the suppression is modest, as reflected by the multivariate sum (–6.8 JNDs) being significantly lower than the individual contributors (–3 JNDs each). In contrast, the following row shows that if one attribute accounts for most of the degradation, suppression causes it to dominate the multivariate sum disproportionately (compare –11.0 JNDs from the individual attribute to the total of –11.3 JNDs). Finally, the last row shows that redistributing the same arithmetic sum of degradation evenly among attributes leads to a superior quality position (–10.0 JNDs vs –11.3 JNDs). This point will be developed further in a subsequent example.

Table 1 Examples of Multivariate Sums

Attribute			Multivariate	Property
#1	#2	#3	Sum	Demonstrated
–1	–1	–1	–2.7	additivity (approximate)
–3	–3	–3	–6.8	symmetric suppression
–2	–2	–11	–11.3	asymmetric suppression
–5	–5	–5	–10.0	better balance

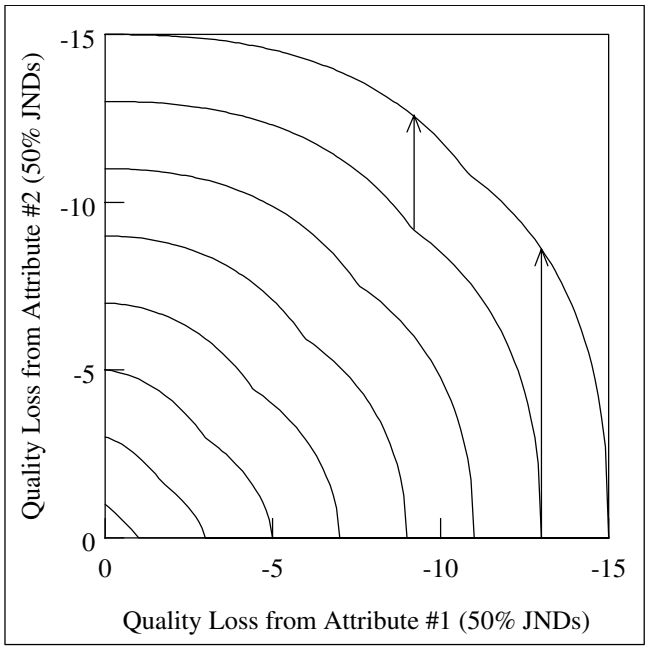

Figure 3. Predicted iso-quality contours for two attributes based on the multivariate formalism.

Figure 3 shows iso-quality contours like those of Fig. 1 and demonstrates the difference in suppression between balanced (individual degradations approximately equal) and imbalanced cases. The contours show overall quality losses of −1, −3, ..., −15 50% JNDs. As in Fig. 1, the nearly linear contours at lesser degradation (lower left corner) correspond to near additivity, whereas the strongly curved contours at greater degradations are caused by suppression. Although the contours look approximately circular, corresponding to a Minkowski power of two (which indeed is in the middle of the range of one to three allowed by Eq. (3)), particularly as more attributes vary, no fixed power Minkowski metric is capable of simultaneously fitting the data from the four experiments well. The small cusps in the contours along the diagonal of the figure, which will also be seen in the following figure, are of no perceptual or practical significance; they reflect the use of only the maximum degradation in Eq. (3), and could be smoothed out by including the second worst degradation. However, this modification does not improve the fit to experimental data and so the additional complexity is deemed unjustified.

Starting on the x-axis at the −13 JND contour, where the quality change from the first attribute is −13 JNDs, and that of the second is zero JNDs, and then increasing the severity of the degradation from the second attribute as shown by the long arrow, requires ≈ −8.6 JNDs of change in the second attribute to shift the overall quality by just −2 JNDs, to the −15 JND contour. In contrast, when starting from the position on the −13 JND contour where the contributions of the two attributes are equal (along the diagonal), only ≈ −3.4 JNDs of shift in the second attribute (short arrow) is required to change the overall quality by the same amount, i.e., −2 JNDs. When the attribute effects are approximately balanced, changes in either attribute will significantly affect overall quality, so no single attribute limits quality.

Frequently, image quality attributes may be affected in opposing ways by a process or a variation in system design parameters. For example, performing a digital spatial filtering operation can increase sharpness but at the expense of noise amplification. If an image had poor MTF but very low noise, such a sharpening operation might improve overall quality by better balancing the attributes. Figure 4 shows multivariate formalism predictions for the case in which quality losses arising from the two attributes are constrained to arithmetically sum to a constant amount, so that if one improves by a certain number of JNDs, the other becomes worse by the same amount. The x-axis shows the ratio of the first attribute to the sum of the attributes; this fraction varies from zero to one and is equal to one-half when the attributes are perfectly balanced. Each curve depicts the relationship for a different direct sum of attributes, having values of −1, −3, ..., −15 JNDs. At lesser degradations, the balance between the attributes has little effect on quality because the effects are nearly additive. In contrast, at greater degradations, the balance significantly influences overall quality, with the best quality occurring when the magnitudes of the two attributes are approximately equal. As indicated earlier, in well-designed imaging systems, no single attribute consistently dominates overall quality, but rather a balance is maintained.

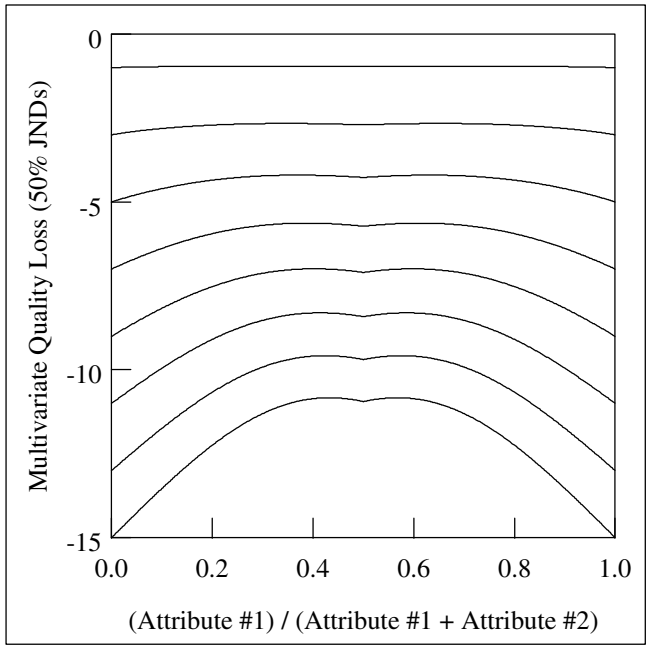

Figure 4. Predicted overall quality loss for two attributes that arithmetically sum to a constant amount.

Conclusion

The multivariate formalism, represented by Eqs. (1) and (3), relates total image quality to the impact of the contributing attributes in isolation, when all quantities are expressed in terms of 50% JNDs of overall quality. It successfully explains the results of four independent experiments involving : (1) two, three, and four simultaneously varying attributes; (2) both artifactual and preferential attributes, including sharpness, noise, digital artifacts, and color and tone reproduction; and (3) images displayed in both hardcopy and softcopy modes.

Acknowledgments

Karin Töpfer, Scott O'Dell, and Robert Cookingham provided the data from the softcopy experiment on color balance and tone scale (contrast). Julie Skipper and Donna DeMay designed and generated the four-attribute hardcopy study samples, and Donna Hofstra administered the psychometric experiment in which they were evaluated. The unpublished data from the study conducted by W. Mitchell Burke was used with the permission of James Weaver of Eastman Kodak Company.

References

1. R. D. Prosser, J. W. Allnatt, and N. W. Lewis, "Quality Grading of Impaired Television Images", *Proc. IEEE **111(3)***, pp. 187–188 (1964).
2. N. W. Lewis and R. D. Prosser, "Subjective Quality of Television Pictures with Multiple Impairments", *Electr. Letters **1(7)***, pp. 491–502 (1965).
3. C. J. Bartleson, "The Combined Influence of Sharpness and Graininess on the Quality of Color Prints", *J. Photogr. Sci. **30***, pp. 33–38 (1982).
4. H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments", *Human Vision, Visual Processing, and Digital Display III*, SPIE Vol. 1666, Society of Photo-Optical Instrumentation Engineers, Bellingham, Washington, pp. 16–26 (1992).
5. P. G. Engeldrum, "Image Quality Modeling: Where Are We?", *Proc. IS&T's PICS 1999 Conference*, Savannah, Georgia, Society for Imaging Science and Technology, Springfield, Virginia, pp. 251-255 (1999).
6. B. W. Keelan, *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, Ch. 11 (to be published in Spring 2002).
7. K. Töpfer , B. W. Keelan, S. F. O'Dell, and R. E. Cookingham, "Preference in Image Quality Modeling", this volume.
8. C. J. Bartleson, "The Combined Influence of Sharpness and Graininess on the Quality of Color Prints", *J. Photogr. Sci. **30***, pp. 33–38 (1982).
9. B. W. Keelan, "Characterization and Prediction of Image Quality", *Proc. IS&T's 2000 PICS Conf.*, pp. 197–203 (2000).
10. K. Töpfer and R. E. Cookingham, "The Quantitative Aspects of Color Rendering for Memory Colors", *Proc. IS&Ts 2000 PICS Conf.*, pp. 94–98 (2000).
11. SAS Institute, *SAS/GRAPH© Software*, Ver. 6, 1st ed., Vol. 2, SAS Institute, Inc., Cary, North Carolina, Ch. 24. (1990).

Biography

Brian Keelan obtained a Ph.D. in Chemistry from the California Institute of Technology in 1986. Since then, he has worked in the research laboratories of Eastman Kodak Company, where his efforts have focused on predictive computer modeling, image quality metrics, and psychometrics. He is the author of the *Handbook of Image Quality: Characterization and Prediction*, to be published by Marcel Dekker in Spring, 2002.