Characterization and Prediction of Image Quality

Brian W. Keelan
Eastman Kodak Company
Rochester, New York

Abstract

In the broadest sense, the benefits of modeling are at least threefold: (1) cycle time compression through reduced prototyping and experimentation; (2) identification of non-obvious optimal solutions that might be missed by empirical testing over a restricted range; and (3) education and training of practitioners through virtual experimentation. At Eastman Kodak Company, predictive modeling of image quality has proved to be of great value in all three regards and has been regularly used in formulating business strategies, guiding design decisions, establishing product aims, budgeting system tolerances, and benchmarking. This paper will provide an overview of selected topics pertaining to image quality modeling, including: (1) definition of numerical scales of image quality tied to physical standards; (2) design of low-noise judging techniques calibrated to these scales and suitable for untrained observers; (3) development of a method for combining results from single-attribute experiments to predict multivariate quality; (4) generation of objective metrics correlating with perceptual attributes of conventional and digital imaging systems; (5) construction of Monte Carlo models that predict image quality distributions based on imaging system specifications; and (6) verification of model predictions through independent expert assessment of images from prototype handout and customer intercept studies.

Introduction

The nature and scope of imaging is undergoing dramatic change as we enter the digital imaging age. Portions of the formerly distinct photographic, electronic, software, television, computer, and printing industries are converging into a more generic imaging industry. The ways in which images are used are increasing in number and diversity, and the flexibility associated with digital imaging is leading to an increasingly complex field of opportunity. The rapid product introduction cycle time of the electronics industry sets the standard for the new imaging industry, leading to an urgent need to streamline strategic, design, and development processes. In this more horizontal industry, the ability to effectively exchange specifications and evaluations based on a common framework will become critical to the success of supplier-manufacturer and partnering relationships. Each of these industry trends is leading to an increasingly acute need for methods of quantifying, communicating, and predicting perceived image quality.

At Eastman Kodak Company, a consistent and integrated approach to image quality characterization and prediction has been developed and integrated into product development and strategic planning processes. Previously, information pertaining to the company's image quality research has usually been treated as proprietary in nature, but with the emergence of a more interactive imaging industry, disclosure of some aspects of the work has been deemed desirable. This paper will provide an overview of our approach to image quality and modeling by addressing the following topics: establishing image quality standards; performing psychophysical experiments calibrated thereto; designing objective metrics correlating to perceptual attributes; predicting the overall quality of samples affected by multiple attributes based on knowledge of the impact of each attribute in isolation; constructing software to enable powerful system image quality modeling of capability and performance; and verifying accuracy of predictions so generated. Most of the research in this field has been in support of the development and refinement of powerful software models that predict the performance of general imaging systems, in the hands of consumers, based on engineering specifications, component measurements, usage factors, etc. The following presentation, given by R. B. Wheeler,[1] will provide practical examples of the utility of such predictive models of image quality.

Establishing Image Quality Standards

Physical standards are an important part of image quality characterization. A primary standard, with a wide range of quality and a variety of scene content, can serve to anchor a numerical image quality rating scale. Primary standards can be initially rated by trained experts, using interval or ratio scales. These scales should be tested for uniformity of JND (just noticeable difference) values by forced-choice paired comparison experiments across the full range spanned by the physical standard, and adjustments made as necessary to achieve a constant JND increment (interval scale) or percentage (ratio scale). This simplifies subsequent interpretation of the scale by permitting assessment of the significance of differences in a trivial manner. The resulting values should be correlated with assessments by customers to ensure robust behavior. Samples exhibiting high variability in assessments by either trained experts or customers should be excluded. Samples falling especially far from a regression curve through the consumer vs trained expert assessment data may include attributes viewed as being of different importance to the two groups. While indicating a potential area for further research, such samples might bias assessments made against the standard, and so should be eliminated if possible.

Primary and secondary standards may be either univariate (consisting of images varying only in a single attribute) or multivariate (with quality affected by covarying attributes). Univariate standards are useful for evaluating both overall quality and the impact on quality of an individual attribute in the presence of other attributes. In addition, because their quality varies with a single attribute, generation of arbitrary quality positions through image simulation is facilitated. Secondary standards for specific purposes may be more restricted in quality range and/or scene content than primary standards, and may contain different attributes. The numerical quality ratings of images assessed against the standards should not vary systematically with the sensitivity of the observer making the assessment in most cases. Generally observers who are more sensitive will see larger quality differences between both the test samples and standards samples, which effects approximately cancel.

Once defined, the numerical image quality rating scale should be invariant with time, although the physical standards associated with the scale may need to be updated occasionally to reflect current imaging technologies and applications. In contrast, the degree of satisfaction or acceptability associated with a particular value on the rating scale will change with time because customer expectations are influenced by advances in imaging technology and changes in intended application. Because most analyses (and customer purchases) involve comparison of a pertinent reference (control) system with a test system, changes in customer expectations may be reflected in the choice of the reference system. If desired, customer expectations can be monitored periodically by category sort experiments using adjectives such as good, fair, poor, etc., or classifications such as acceptable vs unacceptable. The selection of adjectives by consumers is influenced by the samples presented (a range effect), so a representative selection of images for the application of interest should be sought.

Calibrated Psychophysical Experiments

A quality ruler is a mechanism for making precise, rapid, visual assessments of quality that are calibrated against a standard scale. Quality rulers comprise a series of images of known quality that vary systematically, usually in a single perceptual attribute. The images are spaced by one to several JNDs and usually span a wide range of quality. The viewing conditions associated with the quality ruler must be constrained so that the quality calibration is not compromised. For example, if the ruler samples vary in image structure (e.g., sharpness or noisiness), the viewing distance must be fixed. Quality rulers may be organized in sets with each ruler depicting a different scene, which is individually calibrated for quality. The scenes should span a suitable range of subject matter and image characteristics. The attribute varying in a quality ruler ideally should be capable of strongly influencing quality; should not possess high observer or scene variability; and should be correlated strongly with routinely available objective measurements. One result of a quality ruler experiment is that the samples assessed constitute a set of derived (secondary) standards. These may, in turn, be assembled as new quality rulers.

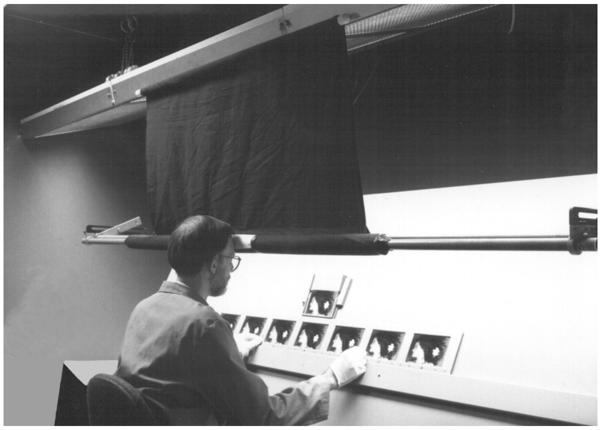

Figure 1. Hardcopy Quality Ruler

One implementation of a quality ruler that we have used extensively involves a sliding ruler with slots for reflection prints, allowing the test sample and a particular ruler print to be compared while adjacent to each other, in the same lighting and at the same viewing distance, without the observer moving (excluding using his or her hands to slide the ruler). Figure 1 shows a hardcopy quality ruler setup. The test image (above) is compared to the ruler, which slides in a Teflon track so that the image being compared to the test sample can be set directly below it. Viewing distance is controlled by a padded headrest. Lighting is at approximately 45-degree incidence and viewing is normal to the image. The black cloth reduces stray light, as does the dark lab coat worn by the observer.

Another implementation of a quality ruler that we use particularly for studies of color and tone reproduction involves two matched high-quality monitors. The observer indicates which of a pair of images (one test image and one ruler image) is higher in quality, and software selects a new ruler image to be displayed depending on the answer given, so as to converge on the ruler position of equivalency.

In addition to being used to determine overall quality, quality rulers may be used to reliably assess the attribute represented in the ruler. For example, if a quality ruler is based on variations of noise, it is also an effective tool in assessing the noise level in images. This is often helpful in separating multivariate phenomena. When using a quality ruler to evaluate the single attribute varying in the ruler, the task involves selective appearance matching. This is in marked contrast to the case in which overall quality is assessed, where it is critical that observers avoid simply matching appearance, and instead try to evaluate each image based on its merits. Although in research applications we normally produce test samples that match the quality rulers in scene content through digital image simulation, surprisingly, the quality rulers have been found to work equally well in evaluating images with different scene content.

The RMS (root mean square) uncertainty in a single quality ruler assessment (one observer rating one sample) is estimated to be about 2.5 JNDs, compared to a theoretical minimum of one JND. Limited data indicates higher values for magnitude estimation (ca. 4 JNDs) and category sort (ca. 8 JNDs), as might be expected based on the number of reference samples provided (many in a quality ruler, usually one in magnitude estimation, and zero in category sort). From the 2.5 JND single-assessment RMS error value, estimates of precision of the mean may be made for pooled data from different numbers of observers and scenes. For example, if 6 assessments are pooled, the standard error of the mean would be approximately $2.5/\sqrt{6} = 1.0$ JNDs. Typical quality ruler assessment times are about two images per minute (compared to 4 per minute for category sort and an intermediate number for magnitude estimation), with about 85% of observers falling within 30% of the mean figure.

Design of Objective Metrics

One of the primary reasons for doing calibrated psychophysical testing is to facilitate generation and verification of objective metrics that strongly correlate with an attribute's impact on quality. When such a metric is successfully derived, it allows replacement of perceptual assessments by predictive analysis, which is usually faster, less expensive, and more robust because it is based on calibrated data from many observers and scenes. Objective metrics should be designed in a fashion that is perceptually relevant (as opposed to simply being a mathematical fit) to improve the likelihood that extrapolated behavior will be accurate. They should be verified for a range of practical and limiting cases so that failure modes can be identified and perhaps reduced through refinement of the metric. Where possible, objective metrics should be defined in a fashion that they can be computed from standardized measurements of a test target that can be propagated through the system. It is also very helpful when objective metrics can be readily calculated from engineering design parameters. In rare cases an engineering parameter itself may be a useful objective metric, but typically they are only one contributing factor among many, and so overall image quality may not even be monotonically related to them in some cases of practical interest. A classic example of a very incomplete descriptor of any perceptual attribute is device resolution.

We have been able to relate observed quality loss in JNDs to a number of objective metrics using a single, three-parameter equation. For simplicity, consider the case of an image artifact (although cases involving preference, such as color reproduction, can also be treated with this equation). It is assumed that below a certain threshold the defect is undetectable and does not affect quality; and that, well above threshold, the JND increment (the difference in objective metric values needed to produce a stimulus difference of one JND) approaches a constant. The latter condition is desirable because it makes the objective metric more easily interpretable, and improves the robustness of extrapolations (because they are based on straight lines). We often find that metrics expressed in terms of "linear" quantities must be logarithmically transformed to meet this criterion, as might be expected based on the Fechner-Weber Law. Thus, the JND increment is constant well above threshold but diverges at threshold, beyond which infinitely large reductions in objective metric yield no further change in quality loss because the artifact is already subthreshold. A simple hyperbolic transitional behavior between these two regimes is assumed:

$$\frac{dO}{dQ} = \Delta O_\infty + \frac{R_t}{O - O_t}$$

where Q is quality in JNDs, O is the objective metric, O_t is its value at threshold, ΔO_∞ is the asymptotic JND increment, and R_t is a curvature parameter that affects how quickly the transition is made between subthreshold and asymptotic behavior. Integrating dQ/dO from O_t to a given value O yields the total quality loss at O:

$$\Delta Q(O) = \frac{R_t}{\Delta O_\infty^2} \ln(1 + \frac{\Delta O_\infty (O - O_t)}{R_t}) - \frac{O - O_t}{\Delta O_\infty} \qquad (1)$$

with the sign convention that quality loss is negative. It may be shown that the curvature parameter R_t is the radius of curvature at threshold. An example of the application of this equation is given in the following section.

An Example of an Objective Metric

In this section, the artifact of misregistration is used to provide an example of objective metric design. Misregistration is a spatial shift between color records of an image, as can be seen in extreme form in Sunday comics, where the colors do not match up with the outlines drawn by the cartoonist. First, a trial objective metric is proposed. In this case each color record has an associated weight (the sum of the weights being normalized to unity), and treating these weights as if they were masses, a visual "center of gravity" of the color records, is determined from their relative displacements.

$$x_c = \sum_i w_i x_i$$

where x is an arbitrary coordinate, w is a weight, and i a color record index. A similar equation applies for the orthogonal y-coordinate. Next, a "moment of inertia" about that center is computed.

$$d_i^2 = (x_i - x_c)^2 + (y_i - y_c)^2$$

$$O^2 = \sum_i w_i d_i^2$$

The square root of this quantity, which is essentially a visual RMS error term, is converted to angular subtense at the eye, to approximately generalize the metric for viewing distances different from that studied.

Figure 2 shows quality loss against this objective metric of visual RMS misregistration subtense in arc-seconds, from the results of a quality ruler experiment. The regression, based on Eq. 1, was computed using samples with different degrees of green record shift (circles). It is usually not too hard to find an objective metric that varies monotonically with quality loss for a magnitude series; the challenge is in predicting what will happen when variations of a fundamentally different nature are made. For example, what quality loss would be expected if different color records were shifted, or multiple records were shifted in different relative directions? Such data are also shown in this figure, and all lie within or close to the 95% regression confidence limits, and within one JND of the prediction, indicating good performance of the objective metric.

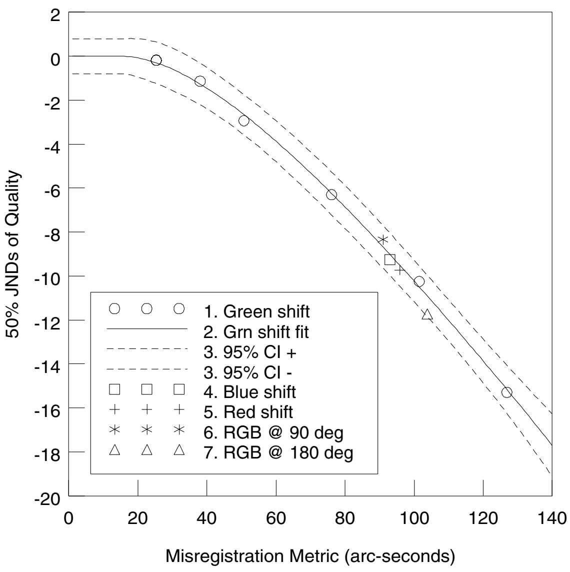

Figure 2. Quality Loss vs Misregistration Metric

Regressions for different subsets of observer sensitivity and scene susceptibility are useful in Monte Carlo calculations of system performance, as discussed later, as well as being helpful in making application-specific predictions. Figure 3 compares results for three different observer/scene subsets (full/full, more sensitive 50%/most susceptible 25%, and less sensitive 50%/least susceptible 25%). This level of variability is somewhat lower than we have observed for most image artifacts.

The primary use of objective metrics in our work has been to facilitate predictive system modeling of quality distributions that would be produced by various systems. However, objective metrics can also be useful in product literature, patent applications, etc. When benchmarking one component in a system, it is often helpful to fix the characteristics of the remainder of the system into one of a few representative positions, so that the resulting objective metric measures the component's performance in typical systems. These standardized metric values can be tracked over time to measure technological advance in a perceptually relevant fashion.

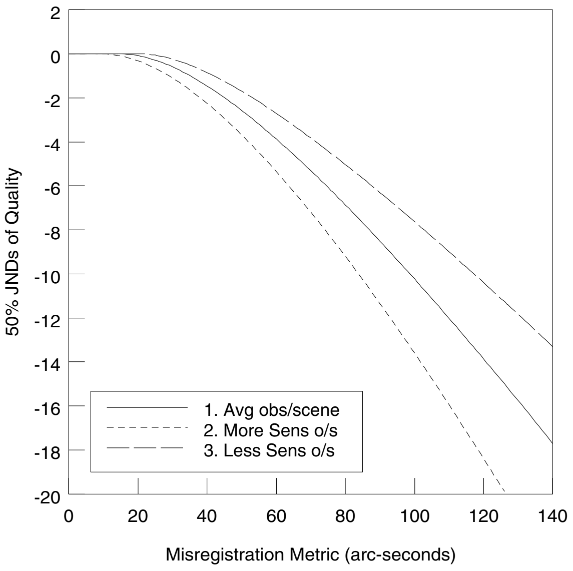

Figure 3. Observer and Scene Sensitivity Variations

Prediction of Multivariate Image Quality

It is not, in general, feasible to perform factorial experiments that fully map out the dependence of quality on different aspects of a pictorial imaging system. An approach referred to as the multivariate formalism allows this situation to be simplified, and has proven to be very reliable. This approach greatly facilitates the construction of predictive models of image quality. A set of seemingly distinct perceptual attributes (e.g., sharpness, noisiness, etc.) that span the quality variations of interest are identified. The dependence of overall quality on each attribute, varied one at a time, is determined. The results are expressed in terms of JNDs of quality change. Where feasible, the quality changes are related to correlating objective metrics. Once the effect of each attribute alone is known, a variable exponent Minkowski metric is applied to predict the impact on overall quality of all the attributes in combination.

$$\Delta Q = (\sum_i \Delta Q_i^{\varepsilon})^{1/\varepsilon}$$

The functional form of ε is chosen such that ε decreases and approaches unity as the ΔQ_i tend to zero. Consequently, this equation has the property that for small quality changes, the changes in JNDs are approximately additive, but when large quality changes are involved, the larger individual contributors dominate. This leads to the result that, if one problem is serious, fixing a minor problem will not significantly improve overall quality.

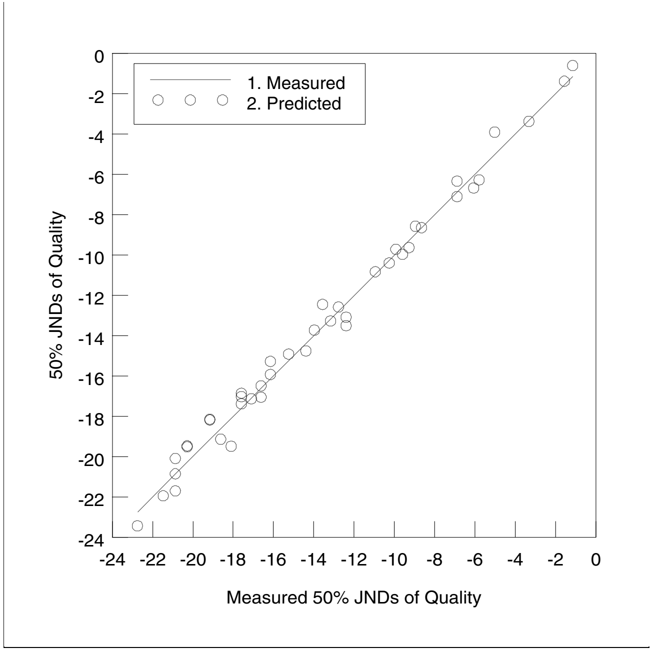

Figure 4. Predicted vs Measured Multivariate Image Quality

In Figure 4, data from an experiment widely covarying sharpness and noise are fit using a variable exponent Minkowski relationship involving only two degrees of freedom. The solid line is the line of equality; points above the line represent overpredictions of quality loss, those below underpredictions. This test spans the range from subthreshold noise and excellent sharpness to approximately 20 JNDs of quality loss, corresponding to a "poor" image. This relationship, with the same values of the two fit parameters and no new degrees of freedom, has been successfully used to predict the results of other experiments involving different attributes, and appears to be quite general in nature. The expression of the effect of each contributing attribute in terms of the universal units of JNDs of overall quality, thereby permitting a single law of combination to be written for most attributes, is a key step in this multivariate formalism.

To apply this multivariate formalism, it is not necessary to construct a scale of the attribute itself (e.g., a scale of sharpness, as opposed to quality because of sharpness). Occasionally, such information is useful, e.g., in advertising claims, or in psychophysical research regarding the perceptual nature of an artifact. In the cases of sharpness and noise, where such scales have been determined, the stimulus change required to produce a JND in a single attribute averages about half as large as that needed to produce a JND of overall quality.

Sometimes, the presence of one attribute significantly affects the perception of another attribute. For example, the presence of noise can mask (reduce the visibility of) artifacts that involve extended patterns, by visually breaking up the regularity of the defects. In such cases, samples containing varying degrees of both the affected and affecting attributes may be assessed to build an interaction model that permits prediction of attribute JNDs in the presence of other attributes.

Constructing System Modeling Software

Key aspects of imaging system models are methods for propagation of important quantities such as tone and color, MTF, and NPS from the original scene to the final viewed reproduction. Propagation of tone and color is based on sensitometry or digital transforms, with interactions between color channels (including chemical interimage effects) handled via equivalent matrixing or three-dimensional look-up tables. Propagation of MTF is usually via a linear systems approximation. Despite the fact that the assumptions underlying this theory are routinely violated to an easily measured degree by individual components of a system, the predictions made for full imaging systems are remarkably good. In some instances, the utility of this approach can be extended by empirical measurement or analysis guidelines. Extension of the standard approach to describe color systems, with channels interactions described by matrixing, is feasible.

Propagation of NPS is usually via the Doerner Equation,[2] extended to color systems via matrixing. Noise propagation depends on MTF and tone scale propagation as well, and so provides a very rigorous test of system models. Figure 5 compares print grain index[3] (a logarithmic visual RMS granularity metric) measured by reflection microdensitometry with predicted values for negative film printed optically onto color paper. The data spans the range from near-threshold noise to that encountered in enlargements from high-speed film. The predictions are based on first principles analytical modeling using component data such as film granularity, sensitometry, and dye set spectra; paper spectral sensitivity, dye set, MTF and granularity; printing lens MTF; and printing and viewing illuminants. The agreement between predicted and measured values is excellent, with the largest errors being on the order of one 50% JND of quality.

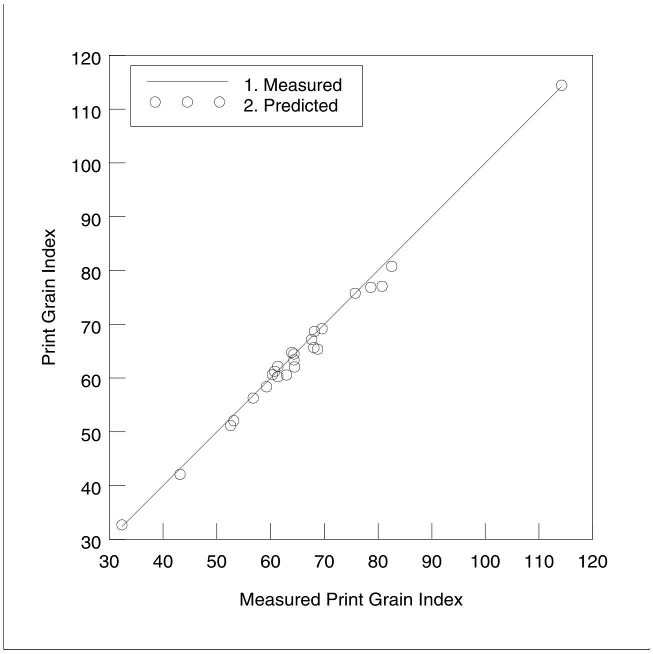

Figure 5. Predicted vs Measured Print Grain Index

To predict final image perception, colorimetric quantities, MTF and NPS are combined with visual models of varying complexity to yield objective metrics characterizing an array of image quality attributes. The prediction of system performance requires the merging of capability models with Monte Carlo techniques and availability of relevant variability data. Examples of the practical application of capability and performance modeling will be discussed in the following presentation.[1]

We have incorporated our image quality modeling capabilities into a single, unified software package. A graphical interface allows the user to construct a block diagram of their system. Each component icon has associated data entry screens, which are modified in real-time in accordance with the other components in the system and their specified characteristics. The user entries are extensively audited for validity so that nearly all invocations of the computational engine result in successful calculations. There is detailed on-line help with both information on use of the software and literature citations. The software is linked directly to a database of measurements that have been carried out in accordance with well-documented protocols. In addition, numerous built-in calculations permit estimates of measured quantities from a small number of engineering design parameters. For example, the MTF of a monitor having a good compromise between sharpness and raster line visibility can be readily estimated from the monitor pixel pitch. Such parametric calculations are extremely valuable early in product design for setting specifications, and also are helpful when carrying out more general analyses, such as for strategic planning.

Verification of Modeling

The Advanced Photo System design and development were strongly influenced by system modeling, which plays an especially important role when working with partners, or when creating a system spanning many suppliers, both of which were applicable in this case. This system presented a particularly rigorous test of the validity of the image quality modeling because nearly every aspect of the system differed from that of its predecessors. The films, format size, cameras, magnetically encoded information, printer lenses, printing magnifications, and print sizes all departed substantially from existing photographic systems. Several perceptual effects needed to be accurately modeled to predict the perceived quality of panoramic prints because of their high aspect ratio. Panoramic prints are made at about 2.5× higher magnification than standard 4R prints from 35-mm format, leading to lower sharpness and higher noise. However, these factors are partially offset by the longer viewing distances and higher intrinsic quality associated with larger images.

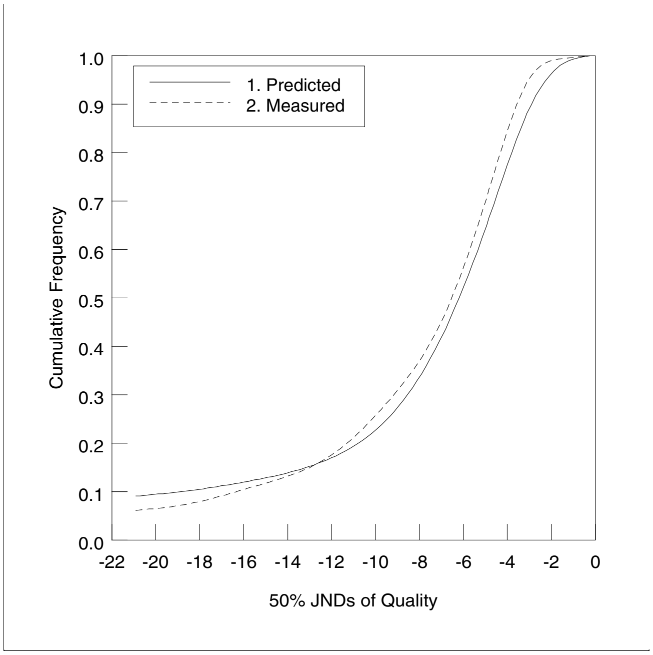

Figure 6. Predicted vs Measured Image Quality Distribution

Once the system was developed, large customer intercept surveys were performed, allowing verification of modeling predictions. Figure 6 compares the quality distribution of 8500 consumer images, evaluated against physical standards by trained experts, with the distribution predicted by our modeling at the time of product design. The x-axis is quality in JNDs, with higher quality to the right, and the y-axis cumulative frequency, i.e., the fraction of images having quality less than or equal to x. The modeling predictions are in terms of the same standard scale of quality as represented by the physical standards, so there are no adjustments whatsoever to the data shown in this figure. The agreement is good to one JND and/or 5% cumulative probability over the entire quality distribution, an extremely satisfying result. This superb agreement provides a dramatic confirmation of the success of the approach described herein.

Conclusion

A consistent and integrated approach to the characterization and modeling of image quality has yielded computer models capable of quantitative predictions of imaging system performance. This capability will be of even greater value, as digital imaging, with its inherent flexibility, becomes more prevalent.

References

1. R. B. Wheeler, “Use of System Image Quality Models to Improve Product Design”, this volume.
2. E. C. Doerner, “Wiener-Spectrum Analysis of Photographic Granularity”, J. Opt. Soc. Am. **52**(6), 669 (1962).
3. Kodak Publication E-58 (http://www.kodak.com/global/en/professional/support/techPubs/e58/e58.shtml); PIMA IT2.37-1999, WD#2, “Print Grain Index – Assessment of Print Graininess from Color Negative Films”.