Progress Towards a New Test Method Based on CIELAB Colorimetry for Evaluating the Image Stability of Photographs

Mark McCormick-Goodhart and Henry Wilhelm
Wilhelm Imaging Research, Inc.
Grinnell, Iowa/USA

Abstract

Traditional methods of image permanence testing have tracked changes in image appearance (changes in density, color balance, and/or Dmin stain) with densitometric endpoint criteria. Densitometric filter sets were optimized for specific colorant systems, for example, the Status A filter set to measure chromogenic photographic materials like those processed in RA-4 chemistry. Modern digital printing systems employ a wide variety of pigment or dye-based colorants with markedly different spectral properties, and some use more than cyan, magenta, yellow, and black colorants to achieve enhanced color gamut and improved continuous tone properties. Densitometry is no longer wholly appropriate because equivalent neutral densities expressed in R, G, and B density units can lead to significant deviations from visually perceptible neutral gray tones in modern imaging systems. CIELAB colorimetry eliminates this technical problem, but designing a colorimetric method and an endpoint criteria set that correlates well with visually perceived changes in pictorial image quality is not trivial. This paper discusses psychophysical testing issues related to the development of a new test method based on CIELAB colorimetry for evaluating the image stability of photographs.

Introduction

From a historical perspective, Status A and M sets were tuned to the typical spectral characteristics of chromogenic dyes. Because broad parity existed within the industry in terms of the spectral properties of the various dye sets, in almost all cases the materials to be tested produced visually neutral gray step wedges when equal R, G, and B density values were measured. However, in the current digital era of highly diverse colorant sets, even within a single technology class such as inkjet materials, we can no longer rely on reasonable consistency of neutral gray patch fabrication based on specified densitometric aim points. Table I illustrates the significance of the problem with some modern inkjet systems. Visually neutral gray color patches were produced on the different media using ICC color profiling procedures and measured with a Gretag Spectroscan spectrophotometer to collect LAB, ΔE, and Status A density data. For comparative purposes, data for two chromogenic color print materials are also included in Table I. These data were measured on IT8.7/2 and Kodak Q60 color targets produced on the respective print papers.

The historical ANSI test method for evaluating light stability and dark storage stability only tracked losses of density from a single initial aim point of 1.00, although more starting aim points were allowed. ANSI/NAPM IT9.9-1996 is a test method not a specification.[1] Hence, published light-and-dark-storage stability test results have not to date had to conform to an industry-wide standard. Wilhelm Imaging Research, Inc (WIR) has for many years employed two aim points: 1.00 and 0.60 used in conjunction with a published endpoint criteria set.[2] The method and criteria set were reasonable because the traditional dye-based chromogenic systems lost density uniformly across their full tonal scale, resulting in a more or less parallel shift (light fade) or linear slope change (thermal aging) over the majority of the densitometric curve. Today, this type of behavior cannot be assumed. Catalytic fading, non-uniform printed dot dispersions, three, four, six, seven, eight (or more) ink colorant sets with different blending levels, and varied black component placement by GCR techniques means that the full tonal scale performance cannot be reliably inferred from measurements of just one or two initial density points. For these reasons, it is essential to evaluate the full image tonal scale behavior using a colorimetric method.[3]

Colorimetry solves the technical issue of identifying visually neutral colorant mixtures across the diverse range of modern materials. It is tempting to believe that the use of color difference models such as ΔE might serve as the analytical basis for a new image permanence test method. However, the current color difference models were intended to judge small incremental differences between two color patches presented side-by-side at a standardized subtended field of view (i.e., the 2° or 10° observer) and against a uniform surrounding field of neutral gray. Color difference models such as ΔE and its many variants oversimplify color and brightness perception in visually complex scenes (e.g., photographs). Color scientists today are actively developing new models that more accurately predict the appearance of color under more complex lighting and surround conditions.[4] However, the perception of pictorial image quality is even

Table I: Errors in Status A measurement of visually neutral grays.

Printer	Ink	Paper	Aim points (Lab and Status A)			Actual Measured values (ΔE and Status A red, green, blue, Visual)				
			L*,a*,b*	ΔE	O.D.	ΔE	Red	Green	Blue	Visual
Lexmark Z53	Lx	Kodak Ultima Glossy	51, 0, 0	0.0	0.71	1.8	0.62	0.66	0.69	0.69
Epson 890	Ep	Ep Prem Glossy Photo				0.7	0.69	0.74	0.80	0.71
HP 7150	HP	HP Prem Plus Photo G				0.8	0.68	0.72	0.77	0.72
Epson C82	Ep	Matte Heavyweight				1.0	0.61	0.69	0.73	0.70
Epson C82	Ep	Xerox Prem Brt white				1.6	0.62	0.70	0.73	0.70
Epson 2000P	Ep	Ep Prem Luster Photo				3.2	0.58	0.76	0.90	0.71
IT8.7/2 target	N.A.	Fujicolor Paper				0.8	0.73	0.74	0.74	0.72
Q-60 target	N.A.	Ektacolor Paper				2.6	0.73	0.74	0.70*	0.71
			L*,a*,b*	ΔE	O.D.	ΔE	Red	Green	Blue	Visual
Lexmark Z53	Lx	Kodak Ultima Glossy	36, 0, 0	0.0	1.05	1.9	0.94	0.97	1.02	1.03
Epson 890	Ep	Ep Prem Glossy Photo				0.6	1.02	1.08	1.17	1.06
HP 7150	HP	HP Prem Plus Photo G				0.8	0.97	1.03	1.12	1.03
Epson C82	Ep	Matte Heavyweight				1.9	0.94	1.01	1.02	1.03
Epson C82	Ep	Xerox Prem Brt white				1.2	0.96	1.02	1.04	1.02
Epson 2000P	Ep	Ep Prem Luster Photo				1.7	0.90	1.08	1.29	1.04
IT8.7/2 target	N.A.	Fujicolor Paper				0.9	1.06	1.06	1.07	1.04
Q-60 target	N.A.	Ektacolor Paper				2.7	1.08	1.08	1.04*	1.05

* patches appeared visually slightly blue (b* ~ -2.5) which accounted for larger ΔE and lower blue channel density value

more complicated. Tone and color reproduction quality in a photograph depends on the relationships between many tones and colors as well as memory colors (e.g., skin tones, neutrals and near neutrals, green grass, blue sky, etc.) in order to establish the contrast and spatial representation of the image. These relationships provide clues to realism or at the very least a plausible and appealing aesthetic. The human observer judges many scene attributes simultaneously including highlight and shadow detail retention, plausibility of color temperature (color balance), overall scene contrast, overall scene brightness, specific appearance of memory colors, etc.

One approach to a new colorimetric method for the evaluation of image permanence is to update the existing WIR criteria set by "translating" the endpoint criteria into colorimetric equivalents. This task cannot be accomplished precisely because densitometry and colorimetry units do not have commutable properties. Nevertheless, a fair approximation along with some updated improvements in the handling of pure colors and skin tone values can be accomplished. A second approach requires a longer term study that is presently ongoing. WIR is undertaking a new psychophysical scaling study designed to rank observed changes in image appearance and then derive a metric based on colorimetric data that correlates well with the psychophysical test results. The task is complex not only because it takes time to generate a representative group of pictorial images and successive iterations showing various stages of aging (either by simulation or actual accelerated aging tests), but also in terms of the design of the psychophysical scaling method. The scaled outcome depends highly on the way the observer is conditioned to the task. Observer conditioning is an especially important issue with regard to image permanence testing because the researcher designing the psychophysical test needs to be able to separate initial image quality attributes from final (aged) image quality attributes. Confounding the two components can lead to significant variability in the scaling results and cause a loss of precision in the determination of a quantifiable metric that correlates well with the psychophysical test results.

Observer Conditioning

Psychophysical ranking of specimens is a well known industrial tool for deriving customer perception of quality. However, the outcome of such a study is dependent not only on the choice of specimens but also on how one frames

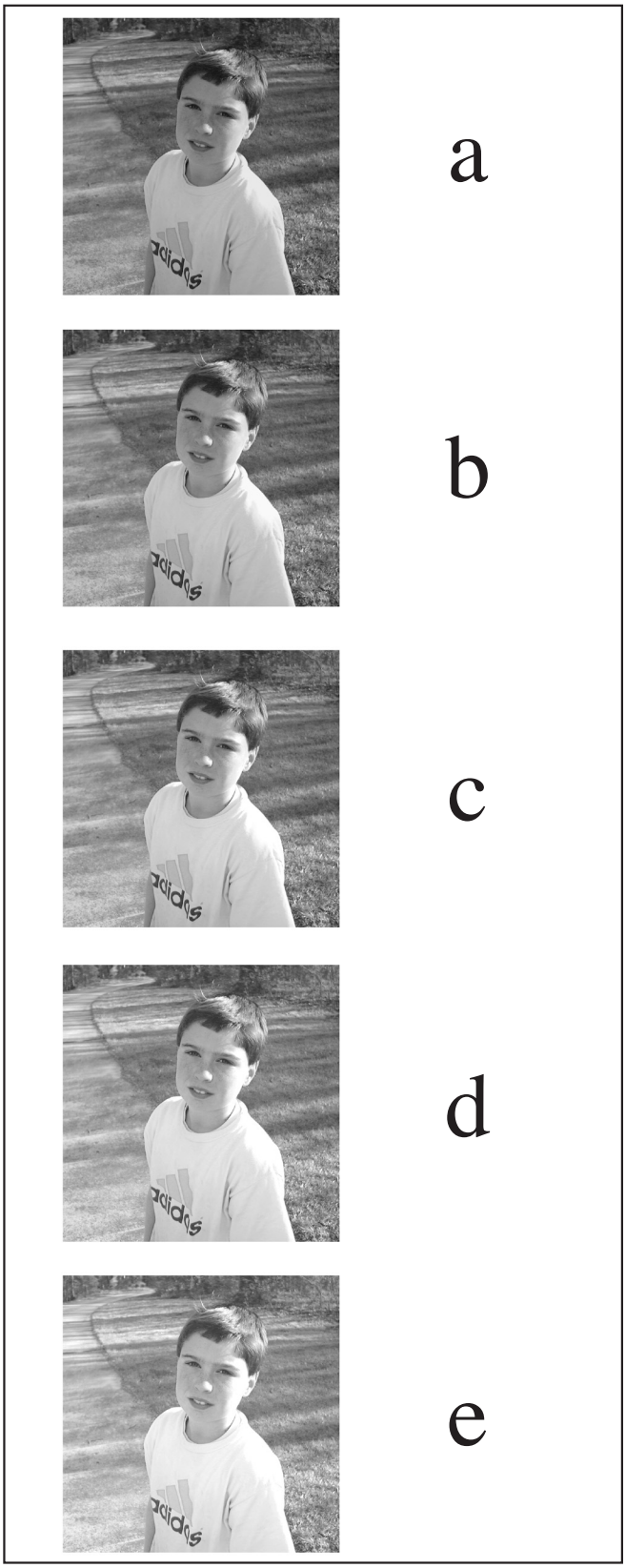

Figure 1. Image series increasing linearly in overall lightness by 5L units on the CIELAB scale. The reproduction in this paper will compress the values but noticeable differences should still be apparent.

the question that is being asked of the ranking observers. In other words, the way the observer is conditioned to the scaling task is critical to what the scaled results signify and to the variability in the ranking scores. Consider the two following ranking approaches.

Ranking Method #1 - The observer is asked to perform a paired comparison, judging a sample "aged" print side-by-side an "unaged" reference print. He or she is asked to assign a value from 1 to 5 to each sample aged print according to the following perceptual guidelines:

1) No noticeable difference
2) Just noticeable difference
3) Noticeable difference
4) Very noticeable difference
5) Extremely noticeable difference

Ranking Method #2 - The observer is shown only one print at a time, the total population of aged prints being presented in random order. The observer is asked to assign a quality value from 1 to 5 based on the following perceptual guidelines:

1) Excellent image quality
2) Good image quality
3) Satisfactory image quality
4) Poor image quality
5) Extremely poor, totally unacceptable image quality

Using either ranking method, the test scores of the sample population are then to be correlated with a derived numerical metric. Essentially, the metric is reverse engineered from mathematical components representing brightness, contrast, color balance, skin tone reproduction, etc. In the first ranking method, the observer is not asked to rate the initial quality of the reference print, and it is available at all times to be compared to the aged sample. Initial image quality is largely irrelevant. The observer is being asked to determine noticeable differences, not to pass judgement on how well he or she likes the print. In the second method, initial image quality is clearly a component of the outcome. If an "unaged" print is determined to have excellent initial image quality then within the limits of observer variability, incrementally aged samples may reach all other rank levels, from "excellent" to "extremely poor". On the other hand, if observers don't like a particular scene and/or it is printed to less than optimum initial image quality, none of the samples including the "unaged" print may achieve an "excellent" rating. Initial image quality is confounded with final image quality, so an additional statistical treatment of the data is required to separate these variables. Depending on how the population of "unaged" prints is produced there may or may not be enough data to determine what viewer tolerances are for excellent reproduction of each selected image. Consider Figures 1a-1e. The same scene is rendered at increasing

levels of scene brightness. From the darkest to the lightest rendition shown in Figure 1, shadows are becoming more open, mid-tones are lightening while maintaining contrast, and highlights are lightening and beginning to have contrast compression as more values get pushed to the limit of paper white. Note the gradually increasing loss of highlight detail particularly on the sunlit side of the child's face. This series was produced by converting a gray scale image to LAB colorspace, and raising the "L" channel curve uniformly by 5L unit increments. This incremental difference is noticeable on a computer monitor display of the images. Note that in the reproduction shown in this paper, the 5L increment will be reduced due to tonal gamut compression in the printing process, but a "noticeable" or "just noticeable" difference between each printed sample should still be apparent.

If the samples illustrated in Figure 1 are ranked by Method #1, then the sample that depicts the "unaged" condition does not have to be a "perfect print" in the mind of the observer, although it cannot be so extreme that further observations of change become difficult (e.g. so lacking in image contrast that further fading is hard to notice). It may or may not meet all of the observer's preferred choices for optimum rendition of the scene. For example, some observers might prefer Figure 1c while others may prefer 1b or 1d. However, if 1c is chosen as the default reference print, aging effects in the direction of 1b or 1d will begin to appear in the rank score. In comparison, if ranking Method #2 is used in the study and 1a-1e samples are shown randomly mixed with many other scenes, it is possible that more than one brightness level will still be judged as an excellent print. Viewers do indeed accept variations especially in brightness and contrast. However, this complicates the final image quality assessment as it relates to allowable change. If 1b and 1c are both rated "excellent" by most observers, for example, while 1d is judged to be "good" then 1b has more headroom to fade in a light fastness study than 1c before triggering a rank score of "good" because it begins the aging test as a darker print. This situation leads to significant variability and perhaps even a systematic error in the study unless there are enough sample statistics about initial image quality to subtract these differences from the final image quality scores.

Scene Selection and Probability Factors

A further complication to the development of an image permanence metric is the need to select a group of scenes that represent not only different photographic applications (flash photography, landscape, portrait/wedding, etc.) but also to sample the correct frequency of occurrence of tone reproduction quality indicators. The indicators include highlight, mid-tone, and shadow contrast, skin tone color reproduction, overall color balance, pure color hue and saturation accuracy, and overall scene brightness. These image appearance attributes are the basis for assessment of tone reproduction quality. For example, some scenes are entirely plausible at different average brightness levels because human observers are entirely accustomed to witnessing natural scenes rendered at different levels of brightness (again, see Figure 1) and color temperature adaptation. Memory colors also provide critical clues but are highly observer dependent. For example, Ferrari red is a memory color if a car in the scene happens to be a red Ferrari and the observer is a car enthusiast, but for other observers, the significance of the red automobile in the print is not very important. The pure red color could change dramatically and as long as other major indicators such as overall color balance in the scene did not become objectionable, ranking Method #2 would not necessarily elicit a large change in rank score while Method #1 might produce an "extremely noticeable" observer response. Again, consider a print system that ages with considerable loss of highlight detail while still maintaining color balance and contrast. In every scene that has no critical highlights, loss of highlight detail is not available as an indicator of change. In an extreme case where no pictorial prints in the study include important highlights, a metric that correlates well with the ranked results would require no weighted factor for loss of highlight detail. The computed metric would be flawed and the derived test standard would favor digital print systems that actually suffer significantly with regard to highlight loss. Conversely, if a highlight factor is given equal weight to other indicators because every print in the study has important highlights, then the computed image permanence metric overweights the contribution of highlight detail to image quality in the "average" pictorial image. Thus, there is a probability factor for each of the individual variables that are included in the final metric which is needed to weight the contribution of each variable to an average scene. Image permanence experts have long emphasized that individual prints may differ widely in actual longevity due to large differences in environmental conditions. However, there is also a significant image dependence whereby a specific scene printed on a specific printing system will fare better than average with all other factors being equal and another scene will fare worse than average simply due to image content and the nature of the print system's failure mode. The probability factors for each contributing image attribute must be incorporated in some way into the quantified image permanence criteria set in order to fairly rank different printing system failure modes.

Discussion

In Method #1 where observers are asked to judge increasing levels of noticeable difference in a strict side-by-side comparison, the psychophysical test results are bound to have the minimum amount of variability that can

be achieved in a subjective ranking test. Furthermore, the scaled results more closely address the needs of artists, archivists, and museum curators than would the results of Method #2. For artistic intent and matters of historical integrity one needs to recognize when a print is deviating from its original state as it ages irrespective of its perceived initial image quality. In contrast, manufacturers of the digital technology used to make the print and perhaps some consumers may want to know how long a print will last before it is judged to be unacceptable. This question is a convolution of initial image quality with final image quality over time. If a print has less than optimum image quality to begin with but is judged acceptable, then two possibilities exist. It may be close to failure on a specific image attribute (e.g., color balance) and the aging process may push it further in the undesirable direction. Or the aging process may push the print in the desirable direction first before crossing over towards an undesirable state. For example, a portrait printed with Caucasian skin tones that are somewhat too yellow but acceptable to the print observer may be printed on a system that exhibits yellow losses in skin tones more quickly than losses of yellow in neutrals and near neutrals. Overall color balance is preserved longer than optimized skin tone reproduction in this system. Since the specific print sample begins life with too much yellow in the skin tones, it has an excess level of colorant that can fade before reaching optimum image quality. The resultant time to reach "failure" is longer than would have occurred had the print been printed "optimally" before aging. Without a large statistical analysis, it is unclear whether prints that benefit will cancel out prints that are at a disadvantage due to initial in image quality. To further complicate matters, many scenes may be rendered initially with very noticeable differences and yet be judged to have excellent print quality. Art historians, for example, are well aware that Ansel Adams made prints in the 1950's that differ greatly from his printing style in the 1970's. His later prints took on higher contrast, bolder, more epic qualities. If "Aspens, Northern New Mexico" printed earlier in Adam's career is compared to the same image printed toward the end of his career the differences are startling. Yet both are beautiful prints in their own way. Thus, it would appear that a parametric study designed to answer the question of image acceptability over time must incorporate enough random samples of varying initial print appearance of each scene to cancel out the distortions in the data caused by the initial image quality variances.

Lastly, the two different ranking methods differ greatly in application dependence. Method #1 is largely independent of the intended purpose or application of the photographs whereas Method #2 is highly dependent on the intended use of the prints. A group of expert observers assigned the task of assessing image quality for fine art or museum quality prints will score changes in image appearance more critically than consumers accustomed to low cost amateur photofinishing prints. Professional photographers in the portraiture/wedding photography business will similarly hold image quality to higher standards than consumer level photofinishing. The rank scores of Method #2 are thus highly application dependent. Hence, the choice of observers and selected images used in the study strongly influences the outcome. In contrast, all humans with normal vision have been training since birth to notice changes in hue, chroma, and lightness, and have an inherent ability to observe side-by-side differences in photographs even though they may find it difficult to appraise image quality. Provided that lighting and print size is uniformly controlled in the viewing area, fine art printmakers and professional photographers will not do significantly better than amateurs in ranking prints on a "no noticeable" to "extremely noticeable" scale when comparing prints in a side-by-side situation. A little training of the novice observer so that he or she is instructed to look for changes in color balance, contrast, highlight detail, etc., and ignore other aspects such as image gloss or print curl will serve the study well, but ranking Method #1 is essentially application independent. The observer will notice a change without being required to make further judgements on quality requirements for the intended use of the prints.

Conclusion

In a psychophysical study where observers are conditioned to judge "before" and "after" print quality as described by ranking Method #2, the scored results are a measure of acceptability for the intended application. The results cannot be applied to other product applications. Significant variability and perhaps a systematic error can occur if the study does not provide enough random samples of initial scene reproduction to assess observer tolerances within rank, i.e., what constitutes excellent quality versus good quality, etc. In a worst case scenario, initial print reproduction attributes can be skewed to favor specific print system characteristics. In contrast, conditioning observers to judge noticeable levels of change ranging from "no noticeable difference" to "extremely noticeable difference" as described in ranking Method #1, produces results that are effectively application independent. A smaller statistical sample size can be used to produce results with lower variability in the test data. It is much easier to achieve consistent agreement between observers on that which is "noticeable" rather than that which is objectionable since limits of acceptability vary by intended purpose whereas the ability to notice change does not require further subjective

judgements regarding intended use. Clearly, both ranking methods have merit, and there would probably be overlapping of the scales if both types of observer conditioning were undertaken in a larger study. However, the extent of the overlap cannot be estimated without both ranking methods being performed on the same sample population, and the number of prints in the study must be greater to accommodate the greater variability in Method #2. Method #1 is more appropriate to museum and fine art requirements where historical or artistic intent must be preserved over the life of the print, whereas Method #2 is useful for specific applications such as amateur photography at the consumer level where the manufacturer may want to determine product acceptance limits.[5]

References

1. *ANSI/NAPM IT9.9-1996 — American National Standard for Imaging Materials — Stability of Color Photographic Images— Methods for Measuring*, American National Standards Institute, Inc., New York, New York, 1996.

2. Henry Wilhelm, "Yellowish Stain Formation in Inkjet Prints and Traditional Silver-Halide Color Photographs," *IS&T's NIP19 International Conference on Digital Printing Technologies, Final Program and Proceedings, pp. 444–449,* New Orleans, Louisiana, Sept. 28–Oct. 3, 2003.

3. Mark McCormick-Goodhart and Henry Wilhelm, "A New Test Method Based on CIELAB Colorimetry for Evaluating the Permanence of Pictorial Images," *available at Wilhelm Imaging Research, Inc. website (www.wilhelm-research.com)* as PDF file <WIR_CIELAB_TEST_2003_07_25.pdf>, June 16, 2003.

4. Mark D. Fairchild, "Color Appearance Models," *Addison Wesley Longman, Inc.,* ISBN 0-201-6346-3, MA, 1998.

5. David Oldfield, Gary Pino, Rise Segur, John Paul Twist, and Scott O'Dell, "VOC Based End-Point Criteria for Lightfastness of Hardcopy Prints," *IS&T's NIP19: International Conference on Digital Printing Technologies, Final Program and Proceedings, p. 396*, New Orleans, Louisiana, Sept. 28–Oct. 3, 2003.

Additional Reference

Hiroshi Ishizuka, Yoshio Seoka, and Yoshihiko Shibahara, "End Point Criteria for Evaluating Image Stability of Digital Prints," *IS&T's NIP19: International Conference on Digital Printing Technologies, Final Program and Proceedings, pp. 411–414,* New Orleans, Louisiana, Sept. 28–Oct. 3, 2003.

Paper by Mark McCormick-Goodhart and Henry Wilhelm (Wilhelm Imaging Research, Inc.) entitled: "Progress Towards a New Test Method Based on CIELAB Colorimetry for Evaluating the Image Stability of Photographs" appeared on pages 25–30 in:

Final Program and Advance

Printing of Paper Summaries

IS&T's 13th International Symposium on Photofinishing Technology

ISBN: 0-89208-249-6

February 8 and 9, 2004
The Riviera Hotel
Las Vegas, Nevada U.S.A.

Published by:

IS&T: The Society for Imaging Science and Technology
7003 Kilworth Lane
Springfield, Virginia 22151 U.S.A.
Phone: 703-642-9090; Fax: 703-642-9094
www.imaging.org