Visual Evaluation at Scale of Threshold to Suprathreshold Color Difference

Haisong Xu,[1]* Hirohisa Yaguchi[2]

[1]State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou 310027, China

[2]Department of Information and Image Sciences, Chiba University, 1–33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan

Received 11 November 2003; revised 28 June 2004; accepted 24 September 2004

Abstract: Visual evaluation experiments of color discrimination threshold and suprathreshold color-difference comparison were carried out using CRT colors based on the psychophysical methods of interleaved staircase and constant stimuli, respectively. A large set of experimental data was generated ranged from threshold to large suprathreshold color difference at the five CIE color centers. The visual data were analyzed in detail for every observer at each visual scale to show the effect of color-difference magnitude on the observer precision. The chromaticity ellipses from this study were compared with four previous published data, of CRT colors by Cui and Luo, and of surface colors by RIT-DuPont, Cheung and Rigg, and Guan and Luo, to report the reproducibility of this kind of experiment using CRT colors and the variations between CRT and surface data, respectively. The present threshold data were also compared against the different suprathreshold data to show the effect of color-difference scales. The visual results were further used to test the three advance color-difference formulae, CMC, CIE94, and CIEDE2000, together with the basic CIELAB equation. In their original forms or with optimized K_L *values, the CIEDE2000 outperformed others, followed by CMC, and with the CIELAB and CIE94 the poorest for predicting the combined dataset of all color centers in the present study.* © 2005 Wiley Periodicals, Inc. Col Res Appl, 30, 198–208, 2005; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/col.20106

Key words: color difference; visual evaluation; color discrimination threshold; color-difference comparison; method of staircase; method of constant stimuli (MCS); color-difference formula; chromaticity ellipse; CRT color

*Correspondence to: Dr. Haisong Xu (e-mail: chsxu@zju.edu.cn)

INTRODUCTION

Various advanced color-difference formulae such as CMC,[1] CIE94,[2] and the latest CIEDE2000[3] have been proposed since the CIELAB[4] was recommended by the CIE in 1976. These formulae were originally developed by fitting some experimental data sets, such as RIT-DuPont,[5] Witt,[6] Leeds,[7] BFD,[8] and so on. However, they can only be used under one "reference" set of viewing conditions, that is, a pair of samples viewed under a high luminance of daylight simulator with hairline separation against a medium-gray background. To form a universal color-difference equation for industrial applications to consider all viewing conditions, the above formulae can be evaluated by fitting different visual data obtained under different viewing parameters such as background colors, separations, textures, media, and colour difference magnitudes. Though the surface color industries mainly deal with small color difference, large color difference is gaining importance nowadays in applications of color reproduction, industrial design, and color communication. The present study investigates a wide range from threshold to large suprathreshold color difference at the same time.

Two psychophysical experiments dealing with color discrimination threshold and suprathreshold color differences were conducted using the methods of interleaved staircase and constant stimuli. In this study, the CRT colors were used because of their efficiency, time saving, less labor needed, cheaper cost, and flexibility compared with object-color samples.[9–10]

EXPERIMENTAL PROCEDURES

All color stimuli used in this study were generated on a Sony Multiscan G500 CRT monitor under the control of a Cambridge Research Systems VSG 2/4 graphics board, with 15-bit luminance-calibrated look-up tables. The CRT dis-

TABLE I. The CIELAB values for each color center under the D65 and CIE1931 Standard Colorimetric Observer conditions.

Color center	L^*	a^*	b^*	C^*	h^o
Gray	61.65	0.11	0.04	0.12	20
Red	44.38	36.91	23.33	43.67	32
Yellow	86.65	−6.92	47.15	47.66	98
Green	56.09	−32.13	0.44	32.13	179
Blue	35.60	4.83	−30.18	30.56	279

play was first calibrated carefully using the probe equipped with the VSG system and was verified by the Minolta CS-1000 spectral radiometer. The display accuracy and stability of the CRT colors were confirmed in the colorimetric characterization procedure prior to the whole experiment.

Two experiments, color discrimination threshold and color-difference comparison, were included, both of which were carried out at the five CIE color centers[11,12] (a Gray, a Red, a Yellow, a Green, and a Blue in CIELAB color space). The CIELAB values of these color centers are listed in Table I. Figures 1(a) to 1(c) show the sample distributions in CIELAB a^*b^*, a^*L^*, and b^*L^* planes respectively. For each color center, the stimuli measured were evenly distributed along 12 directions every 30° in a^*b^* plane [Fig. 1(a)] and along 8 directions every 45° in a^*L^* and b^*L^* planes [Figs. 1(b) and 1(c) respectively]. Observations were performed in a darkened booth at a fixed viewing distance of 500 mm from the CRT to the eyes of observers. The experiment for each color center was separated into three sessions, one for each plane. Each session, started with a 3-min dark adaptation and a 1-min background adaptation, lasted less than 25 min. A session of more than 25 min would cause observer fatigue. A panel of eight observers with normal color vision took part in the experiments. All observers were university students and were naïve to the purposes of the experiments, and most of them had no experience for such observations.

The viewing conditions of the two visual experiments have been described in detail in the authors' earlier article.[13] The experimental design is described below.

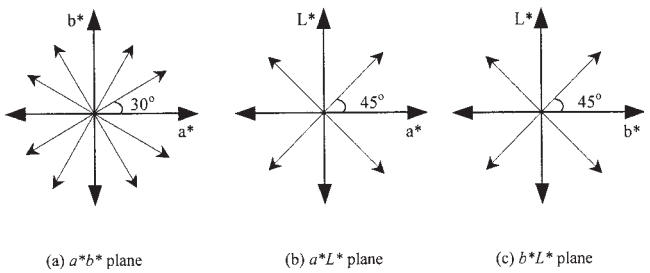

FIG. 1. Color stimuli distribution in CIELAB space. (a) a^*b^* plane; (b) a^*L^* plane; (c) b^*L^* plane.

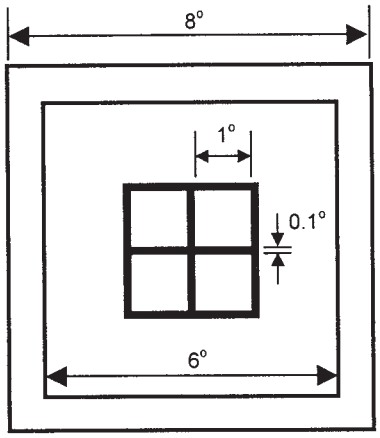

FIG. 2. The test stimulus arrangement used in color discrimination threshold experiment. A four-square array, with 0.1° separation and frame, was presented on a 6° background of center color, surrounded by a bright border of 8° visual angle, out of which was black screen. See text for details.

COLOR DISCRIMINATION THRESHOLD EXPERIMENT

Stimuli

Figure 2 shows the color discrimination threshold experiment. Each test stimulus is a square array of four $1^o \times 1^o$ squares with a small black dividing line of 0.1° visual angle. The squares, with a black frame of 0.1°, were presented on a 6° background set as the color of each color center. During the experiment, only one of the four squares was set as the test color and the remaining three squares to have the same color as the background. The visual subtended angle of the test stimulus at the center of CRT was about 2°, so that the CIE1931 Standard Colorimetric Observer (or 2° observer) was used in calculations. As illustrated in Fig. 2, the pattern was surrounded by a white border of 8°, with a luminance of 100 cd/m^2 and the chromaticity of D65. This border was displayed to define the white point for the test pattern and to have the CRT stimuli appear as simulated surface colors[14] or related colors rather than aperture colors. Outside of the white border was black screen.

Procedure

Each assessment had 2 sec, including two periods of 200 ms for showing background color and subsequently showing a black gap before and after the 1200-ms presentation of test stimulus. The responding time generally was less than 2 sec for all observers so that observer judgment was not influenced by the limited time for presenting test stimuli. The background color and black gaps between trials effectively prevented the possible cues from affecting the observer judgment caused by the color changing process and observer's adaptation to the stimulus color. During gaps all areas, including the background and the four-square array, were covered with black except for the surrounding border, which remained to hold the complete adaptation of the observer to the white point.

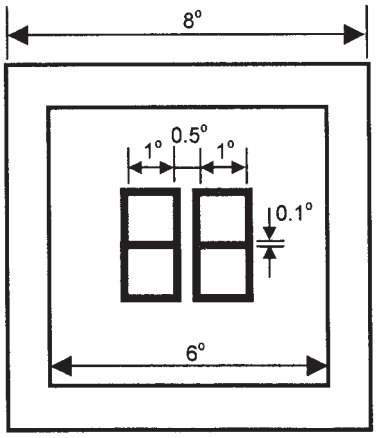

FIG. 3. The test stimulus paradigm used in color-difference comparison experiment. The construction was almost the same as that described in the legend to Fig. 2, except that two color pairs were presented at the center of the background, which color was set as Munsell N5 instead of the test center color. See text for details.

In each trial, the test color was presented on one of the four squares selected randomly by the software, with the other three squares remaining the same color as the background. The test color was determined according to the predicted color distance from the test color center by the psychophysical method of staircase. Each test color was assessed 3 times by each individual from a panel of eight observers as mentioned earlier. To avoid possible bias by observers from the presentation sequence of test stimuli, an interleaved staircase method was used. Each session involved a group of stimuli in four randomly selected directions with random presentation sequence. In the initial trial of the staircase, an obviously discriminable step was presented. The step size then decreased systematically until a criterion value, determined by pilot experiments to produce an efficient staircase, was reached to generate 10 repeats. The averages of the 10 reversals were calculated as thresholds.

The visual task of the observer was to judge the position of the square where a color different from the background color was perceived and then to press the corresponding key on the keyboard as his/her response; this stored the result and started the next trial. The step size changed systematically, but the levels of color discrimination threshold are different in individual directions for every observer, so the numbers of steps were not the same in different color directions for each observer; the mean steps were about 30. In total, 12,600 observations were carried out by each individual observer, that is, 28 directions $\times$ 3 assessments $\times$ 30 steps $\times$ 5 color centers.

COLOR-DIFFERENCE COMPARISON EXPERIMENT

Stimuli

Figure 3 illustrates the arrangement of the test stimulus pattern in color-difference comparison experiment. The stimulus arrangement was similar with that used in the previous staircase experiment for color discrimination threshold, except that at the center of the 6° background were two color pairs instead of the square array and that the background here was set to Munsell N5 neutral gray with the chromaticity of D65. The two color pairs, designated as reference and test pair respectively, consisted of two $1^\circ \times 1^\circ$ squares in upper and lower positions, with a black frame of 0.1° and a separation of 0.5° visual angle between them. The total visual angle of these two pairs was 2.7° (width) $\times$ 2.3° (height), so also met the demand for applying CIE1931 colorimetric system to calculations.

One color of the reference pair was selected as gray with the chromaticity of D65 and a lightness (CIE Y) of 30, the same as the CIE Gray, and the second color differed from it along $+L^*$ axis in CLELAB space. Thus the color differences of the reference pair was only the luminance variations ($+\Delta L^*$) or called gray scales along $+L^*$ axis of CLELAB space. For the test pairs, the color differences were the selected color distances from the test color centers according to a predetermined step size by a pilot experiment. Hence the test pairs were formed by the center colors and those stimuli evenly distributed around them in the CIELAB color space.

Procedure

In color-difference comparison experiment, every trial of observations began with a 200-ms gap and ended at receiving the response from each observer with no time constrain, so the duration of one trial was different for individual observers. During gaps only the reference and test pairs were replaced by black colors, whereas the surrounding border and background were remained there for each observer to maintain the complete adaptation to the white point and background throughout the entire experiment.

This experiment was designed based on the principle of the psychophysical method of constant stimuli (MCS). The test stimuli were chosen via a pilot experiment, so that the color differences of the test pairs ranged from "always judged to be greater than the reference" to "never judged to be greater than the reference," with the majority lying between these two extremes.[12] Each observer was asked to judge whether the color difference of test pair was greater or less than that of reference pair and then to enter the answer via a keyboard. The judgments were repeated sufficient times to assign each test pair a probability of being judged to have a color difference greater than the reference pair. An iterative algorithm called probit analysis,[5,15,16] a maximum-likelihood model that relates experimental response functions to occurrence probability estimates, was used to find the most precise estimate at the tolerance of 50% rejection probability. This corresponded to the color-difference value visually equivalent to the reference color difference.

The color difference of the reference pair was set as 4, 8, and 12 CIELAB ΔE units as the reference scales in the present study. Each test pair was assessed 20 times, carried out in two separate sessions with random orders of color stimulus presentation, by individuals on a panel of eight

TABLE II. Observer precision for individual observers and each visual scale from the staircase experiment for color discrimination threshold and the MCS experiment for suprathreshold color difference.

Experiment	Visual scale	DK	YK	YT	KF	IN	HT	NM	KS	Mean
All	All scales	18	25	25	35	19	22	18	34	24
Staircase	Threshold	22	38	28	31	26	32	27	37	30
MCS (Method of Constant Stimuli)	$\Delta V = 4$	15	25	17	41	12	14	11	12	18
	$\Delta V = 8$	11	21	11	31	14	17	11	6	15
	$\Delta V = 12$	12	17	16	31	16	17	17	10	17
	Mean	13	21	15	34	14	16	13	9	17
Mean	All scales	15	25	18	34	17	20	16	16	20

observers as used earlier. During the experiment, the stimulus arrangements of the left or right position of the reference and test pair and of the upper or lower position in the two color pairs were all randomly arranged in different trials to avoid any experimental bias. For each color center, each direction for each reference scale, nine test pairs, corresponding to nine color stimuli distributed in this direction, were compared with the reference pair. Following the statistical method of probit analysis, the equal color-difference contours for each of the reference visual scales, 4, 8, and 12 CIELAB ΔE units, were obtained, which are analyzed under RESULTS AND DISCUSSION.

In total, 75,600 observations were carried out by each observer, that is, 28 directions $\times$ 9 pairs $\times$ 20 observations $\times$ 5 color centers $\times$ 3 magnitudes.

RESULTS AND DISCUSSION

Performance Factor

In this study, a performance factor (PF), first devised by Luo and Rigg[17] and then modified to PF/3 by Guan and Luo[18] as given in Eq. (1), was adopted to ease the comparison between two sets of data.

$$\mathrm{PF/3} = 100[(\gamma - 1) + V_{\mathrm{AB}} + \mathrm{CV}/100]/3, \tag{1}$$

where CV and (Gamma) were proposed by Coates *et al.*,[19] and VAB derived by Schultz,[20] respectively, and their definitions are as follows.

$$\mathrm{CV} = \frac{\sqrt{\frac{1}{N}\sum(X_i - fY_i)^2}}{\bar{X}} \times 100, \tag{2}$$

and

$$f = \frac{\sum X_i Y_i}{\sum Y_i^2}, \tag{3}$$

$$\log_{10}(\gamma) = \sqrt{\frac{1}{N}\sum\left[\log_{10}\left(\frac{X_i}{Y_i}\right) - \overline{\log_{10}\left(\frac{X_i}{Y_i}\right)}\right]^2}, \tag{4}$$

$$V_{\mathrm{AB}} = \sqrt{\frac{1}{N}\sum\frac{(X_i - FY_i)^2}{X_i F Y_i}}, \tag{5}$$

and

$$F = \sqrt{\sum\frac{X_i}{Y_i}\Big/\sum\frac{Y_i}{X_i}}, \tag{6}$$

where N is the number of compared pairs, and X_i and Y_i are values of pair i. The PF/3 measure indicates a disagreement in percentage between two sets of data, such as the observer precision between the visual results of individual observers and the mean of the values for the panel of eight observers. A higher PF/3 value implies a worse agreement between datasets, and a PF/3 of 30 indicates a disagreement of about 30%.

Observer Precision

As mentioned above, the observer precision in PF/3 measure was calculated between each observer's and the mean visual results for each visual scale from the staircase experiment for color discrimination threshold and the MCS experiment for suprathreshold color difference, as listed in Table II. The mean value of precision for all observers was 24 PF/3 units ranging from 18 to 35 for the most and least precise observers, respectively. The mean error for all visual scales was 15 to 34 with a typical value of 20 PF/3 units. These results were somewhat better than those found by Cui and Luo[21,22] with CRT colors and by Guan and Luo[18,23] with surface colors, which all were based on the psychophysical experiments using gray-scale method. This agrees with the study on comparison of constant stimuli and gray-scale methods for color difference scaling by Montag and Wilber.[24]

The observer precisions for individual scales of suprathreshold color difference from MCS experiment were not very different (with a mean value of 17 PF/3 units ranging from 9 to 34 for all observers) and were obviously superior to those of the threshold (22 to 38 with a mean of 30 PF/3 units) from staircase experiment. This indicates that the suprathreshold judgments showed greater precision comparing to the color discrimination threshold and that different suprathreshold magnitudes hardly affect the observer precision. The different observer variations for threshold and suprathreshold color difference should be because of the different psychophysical methods used in the two experiments, which implies that the method of constant stimuli is more precise or repeatable than the staircase method. In any case, these overall precisions are thought to be typical for

TABLE III. Comparison of chromaticity ellipses at threshold with those for suprathreshold color difference.

Visual scale	Center	A	A/B	Θ	$\sqrt{\pi AB}$	Average $\sqrt{\pi AB}$	Scaling factor	PF/3
Threshold	Grey	2.53	2.98	115	2.60	2.97	1.00	
	Red	3.19	2.84	76	3.36			
	Yellow	2.80	2.04	92	3.47			
	Green	3.26	3.42	136	3.12			
	Blue	2.11	2.67	110	2.29			
$\Delta V = 4$	Grey	5.46	1.96	115	6.91	11.64	0.25	12
	Red	8.79	1.61	75	12.28			13
	Yellow	9.19	2.00	103	11.53			19
	Green	13.67	2.11	169	16.69			21
	Blue	10.29	2.85	120	10.81			21
	Mean							*17*
$\Delta V = 8$	Grey	9.73	2.10	114	11.90	18.29	0.16	36
	Red	14.68	1.69	83	20.02			22
	Yellow	16.50	2.25	98	19.50			9
	Green	17.24	1.68	153	23.60			18
	Blue	15.37	2.75	123	16.43			16
	Mean							*20*
$\Delta V = 12$	Grey	12.47	1.84	114	16.28	24.13	0.12	36
	Red	19.64	1.93	78	25.07			20
	Yellow	23.37	2.31	101	27.24			14
	Green	24.38	1.80	135	32.19			20
	Blue	18.30	2.66	120	19.89			21
	Mean							*22*

such visual experiments using the method of staircase and constant stimuli.[21]

Magnitude Effect of Color Difference on Chromaticity Ellipse

The experimental results from this study were also fitted as chromaticity ellipses. The parameters for each ellipse, in terms of semimajor axis (A), ratio of semiaxes (A/B), orientation angle (θ), and the square root of ellipse area, are given in Table III. The ellipses for all color-difference scales, together with those most inner ones of threshold, at the five color centers in a^*b^* plane are presented in Fig. 4. The qualitative analysis can be found in the author's earlier article.[13] A quantitative method developed by Strocka *et al*.[25] was also used to compare the predicted ΔE values between two ellipse equations. Five hundred color-difference pairs were randomly generated and their ΔE values were calculated, respectively, using the two ellipse equations compared. The sizes of all five ellipses for each suprathreshold magnitude were adjusted by a single scaling factor, 0.25, 0.16, and 0.12 for $\Delta V = 4$, 8, and 12 CIELAB ΔE units respectively, to have the same scale as the threshold ellipses. The results are also given in Table III in terms of the PF/*3* measure.

At Gray and Red centers, the prediction errors by ellipse equations for visual scale (ΔV) of 4 CIELAB ΔE units were better than those for ΔV of 8 or 12. For other centers, there was no clear tendency. The mean variations for all color centers were 17, 20, and 22 for ΔV of 4, 8, and 12, respectively, with a weak sequence from worse to better as the visual scale changing from small to large, though there was no great difference of disagreement. The overall predicting accuracy of ΔE is considered to be good comparing to the observer precision of 24 PF/3 units.

Comparison with Other Studies

The visual data produced by the present CRT experiments were further compared with four previous published data-

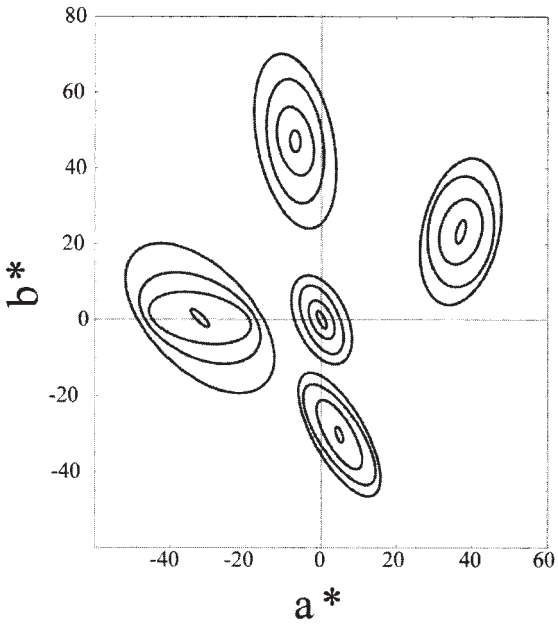

FIG. 4. Chromaticity ellipses for all visual scales at the five color centers plotted in a^*b^* plane. At each color center, from inner to outer are threshold, visual scales of 4, 8, and 12 CIELAB ΔE units, respectively.

TABLE IV. Comparison between visual results of this study and the previous published data of Cui and Luo, RIT-DuPont, Cheung and Rigg, and Guan and Luo.

Data set	Center	*A*	*A/B*	θ	$\sqrt{\pi AB}$	Average $\sqrt{\pi AB}$	Scaling factor	*PF/3*
Cui and Luo $\overline{\Delta E^*_{ab}} \approx 2.5$ (compared with the present study at $\Delta V = 4$)	Grey	1.12	1.69	109	1.53			10
	Red	2.99	2.01	35	3.74			12
	Yellow	2.28	1.81	85	3.01	2.95	3.95	26
	Green	2.83	2.17	172	3.41			9
	Blue	2.98	3.01	124	3.05			32
	Mean							*18*
RIT-DuPont $\overline{\Delta E^*_{ab}} \approx 1.0$ (compared with the present study at threshold)	Grey	1.33	1.53	102	1.90			24
	Red	2.39	1.71	37	3.25			22
	Yellow	2.13	1.71	78	2.88	2.91	1.02	15
	Green	2.44	1.86	169	3.17			18
	Blue	3.43	3.29	119	3.35			23
	Mean							*20*
Cheung and Rigg $\overline{\Delta E^*_{ab}} \approx 3$–4 (compared with the present study at $\Delta V = 4$)	Grey	0.97	1.66	90	1.34			7
	Red	2.72	2.18	40	3.26			23
	Yellow	3.24	2.07	101	3.99	3.17	3.68	25
	Green	2.87	2.27	176	3.37			33
	Blue	4.32	3.90	119	3.88			34
	Mean							*24*
Guan and Luo (GHM mode) $\overline{\Delta E^*_{ab}} \approx 13$ (Compared with the present study at $\Delta V = 12$)	Grey	1.49	1.43	127	2.20			8
	Red	1.85	1.32	59	2.86			10
	Yellow	1.80	1.29	112	2.80	2.75	8.79	23
	Green	2.02	1.28	157	3.16			11
	Blue	1.76	1.33	106	2.71			17
	Mean							*14*

sets of Cui and Luo[21,22] with CRT colors, and RIT-DuPont,[5] Cheung and Rigg,[26] and Guan and Luo[23] with surface colors. The average ellipse parameters from the above earlier studies are summarized in Table IV. The aims of these comparisons are to analyze the reproducibility of this kind of visual experiments using CRT colors and the variations between CRT and surface data.

Reproducibility of Experimental Data Using CRT Colors

According to the detailed experimental conditions, the visual results for $\Delta V = 4$ of this study were considered to be comparable with Cui and Luo's CRT data with an average color difference of about 2.5 CIELAB ΔE units. Table IV shows that, although the size of ellipses from the two studies were rather different, the orientations (θ) and shapes (A/B) were very near, considering the different viewing parameters used in the individual experiments, except for the Red center, at which a discrepancy was seen in the θ value.

The Strocka method was again used to compare the two sets of ellipse equations from the present and Cui and Luo's studies. The sizes of all five ellipses in Cui and Luo data set were adjusted by a single scaling factor of 3.95 to have the same scale as the present ellipses of $\Delta V = 4$. The visual comparison can be seen in Fig. 5, in which the adjusted

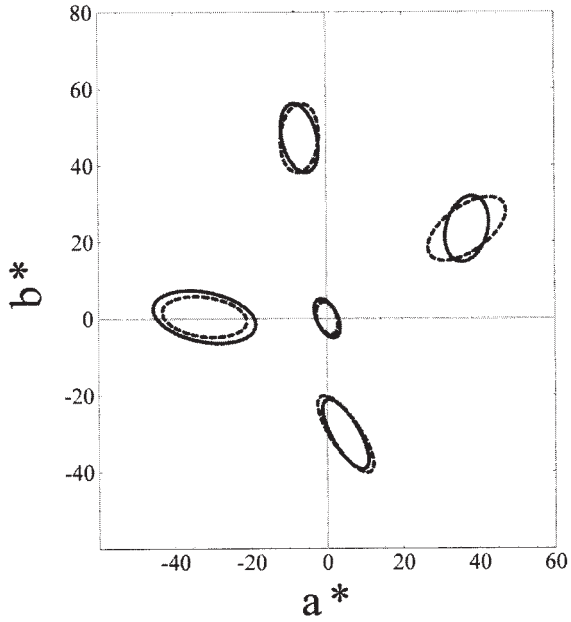

FIG. 5. The visual comparison between the adjusted ellipses of the Cui and Luo data set with an average color difference of about 2.5 CIELAB ΔE units (dashed line) and those of $\Delta V = 4$ from this study (solid line) in a^*b^* plane for all color centers.

ellipses of Cui and Luo data set, plotted in dashed line, are drawn with those of $\Delta V = 4$ from this study, in solid line, for all color centers. The quantitative comparison results are also given in Table IV in terms of PF/3 measure. The mean variation for the all five centers was 18, which shows the present data agrees well with those of Cui and Luo comparing to the observer precision of 24 and the present suprathreshold prediction accuracy of 17 PF/3 units for $\Delta V = 4$ against threshold. The above analysis indicates that the reproducibility of visual evaluation experiments using CRT colors is good.

Variations between CRT and Surface Data

The RIT-DuPont data, with an average color difference of about 1.0 CIELAB ΔE units using surface colors, were compared with the threshold data of this study. Generally, as listed in Table IV, the ellipse shapes (A/B) of the present study were more elongated than those of RIT-DuPont except for Blue center, at which the case was opposite. The orientations (θ) were similar for the two sets of ellipses. The sizes of the RIT-DuPont ellipses were almost the same as those of threshold from this study, which resulted in a scaling factor of 1.02, the nearest to 1.0 among the four compared data sets. The graphical presentations of the two sets of ellipses in Figs. 6(a) to 6(e) at the five color centers, respectively, give the same story. The evaluation with Strocka method resulted in a mean PF/3 value of 20, which is only a little worse than that between the CRT data of Cui and Luo and the present $\Delta V = 4$. Thus the threshold data of this study based on CRT display agree well with the RIT-DuPont data using surface colors, consistent with the qualitative analysis.

The average color difference of Cheung and Rigg data set was about 3–4 CIELAB ΔE units, so it was compared with the present visual results for $\Delta V = 4$. The Cheung and Rigg ellipse at Blue center was rather elongated comparing to the present one, which resulted in a relatively large prediction error of 34 PF/3 units as estimated using Strocka method at a large scaling factor of 3.68. But the mean value of 24 for all color centers implied a satisfactory agreement between the CRT and surface data. Figure 7 illustrates the comparison between Cheung and Rigg ellipses with about 3–4 CIELAB ΔE units and the present ones of $\Delta V = 4$ for all color centers.

The study of Guan and Luo for large color difference, with a mean of about 13 CIELAB ΔE units, included several viewing parameters, in which the data of GHM mode (Grey background, Hairline separation, and medium luminance) were considered to be comparable with the present ellipses for ΔV of 12 CIELAB ΔE units, although the chromaticity values of the color centers were not the same as the present ones. The very large scaling factor of 8.79 reflects on the rather different ellipse areas between the two datasets compared. However, the least mean PF/3 value of 14, among the four earlier studies, shows a very good consistency between the Guan and Luo data in GHM mode and the present $\Delta V = 12$ data, as can be seen in Fig. 8.

In conclusion, the variations between CRT and surface data were 14–24 PF/3 units, which cannot be thought to be serious as compared with that between CRT data of 18 and the total observer precision of 24. This indicates that the CRT colors can be used conveniently in visual experiments with no obvious disadvantage.

Testing Color-Difference Formulae

Three advanced color-difference formulae, CMC, CIE94, and CIEDE2000, together with the basic CIELAB equation were tested with respect to their performance in predicting different magnitudes of suprathreshold color difference from the MCS experiment at the five CIE color centers. The comparisons between color differences (ΔE) predicted by individual formula and the corresponding visual scales (ΔV) of 4, 8, and 12 CIELAB ΔE units were carried out by two methods: the first was to evaluate each formula using its original form with $K_L = K_C = K_H = 1$, and the second using an optimized K_L value with $K_C = K_H = 1$ to give the best fit to the visual scales. The test results, in terms of PF/3 measure, for each color center and individual formula and for individual visual scale ΔV (only in formulae's original form) and the combined data set of all suprathreshold color differences, respectively, are summarized in Table V with the best performance in each case printed in bold font for ease of comparison.

For the performance with the original forms of all color-difference formulae, the detailed analysis has been partly described in the authors' previous article.[13] The ellipses predicted by the three advanced color-difference equations, CMC, CIE94, and CIEDE2000, at all color centers are presented in Figs. 9, 10, and 11, respectively, for comparison with those of the basic CIELAB formula in Fig. 4. The CIEDE2000 outperformed all other formulae by a large margin at Grey, Yellow, and Blue center. It is worth pointing out that all formulae performed almost worst at Blue center among individual centers, but CIEDE2000, due to the rotation item involved in its equation, still showed excellent performance, even better than at Red and Green centers. At Red center, CIEDE2000 was only slightly poorer than the best CIE94. At the Green center, the performance differences among all formulae were small, but the CIEDE2000 was just better than the poorest CIE94. The poorest performance of CIE94, even worse than CIELAB at most centers except for the Red center, implies partly the influence of viewing parameters in this study, which differed from its original reference condition. For the combined data set of all color centers, the CIEDE2000 again performed best, and CMC was only somewhat poorer than the former. This confirms the effective improvement of this latest equation of CIEDE2000 and, meanwhile, the good prediction ability of CMC for large color differences.

The color-difference predicting performances of different formula at each ΔV of 4, 8, and 12 CIELAB ΔE units are consistent with those for the combined data set of all visual scales. For different visual scale, each formula performed similarly in general, but the PF/3 results were not the same for different formula and at different color center. At Gray center, the performances of all formulae were better for

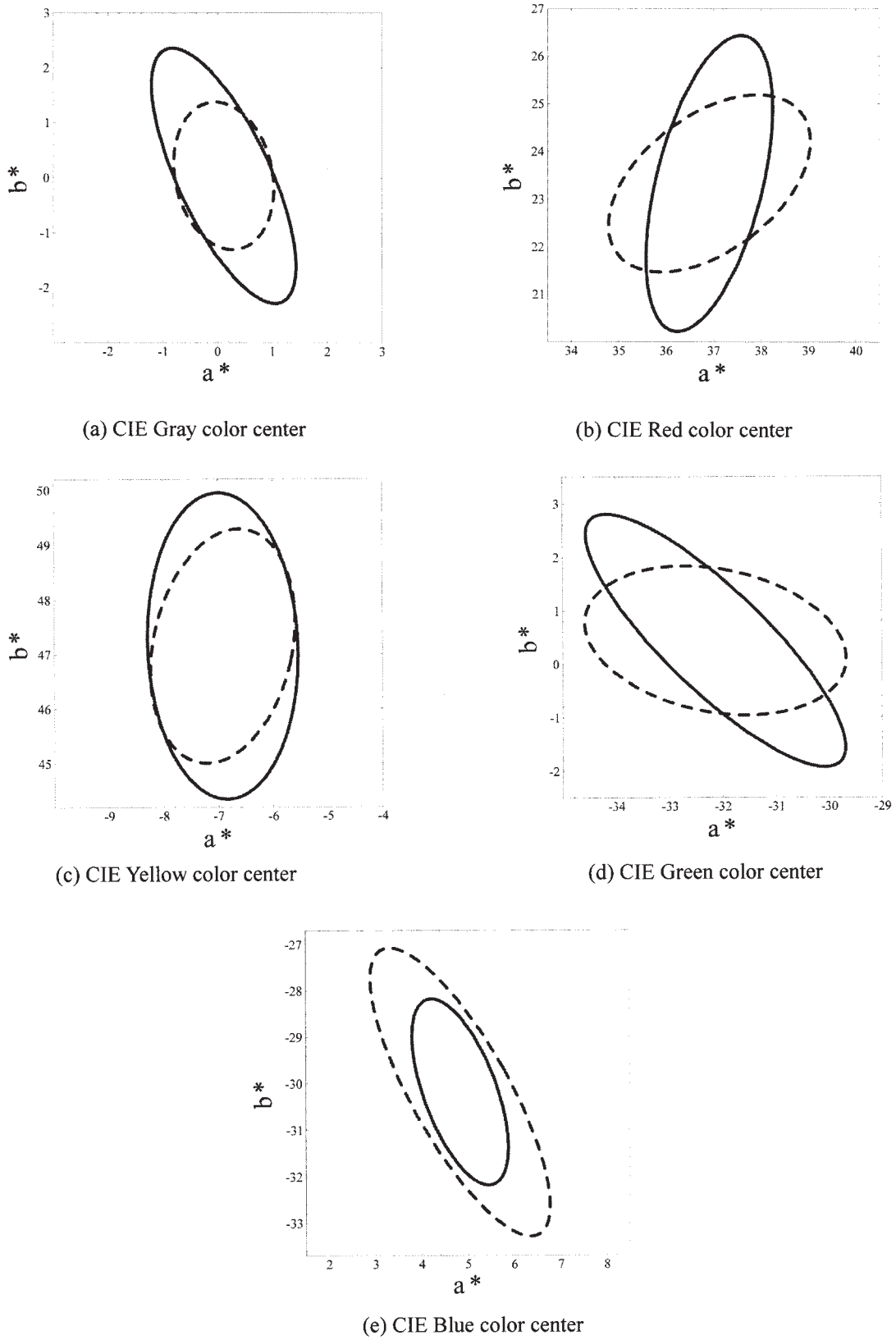

FIG. 6. Variations between the ellipses of threshold data using CRT stimuli in this study (solid line) and of RIT-DuPont data with an average color difference of about 1.0 CIELAB ΔE units using surface colors (dashed line) at the CIE color centers of (a) Gray, (b) Red, (c) Yellow, (d) Green, and (e) Blue.

small color differences than large ones, that is, their PF/3 values became larger with the visual scale changing from 4, passing by 8, and the on to 12. The CIELAB performed better at smaller color difference at Gray and Yellow centers, whereas at Blue center the order of its performance was in reverse, that is, the larger the visual scale, the smaller the

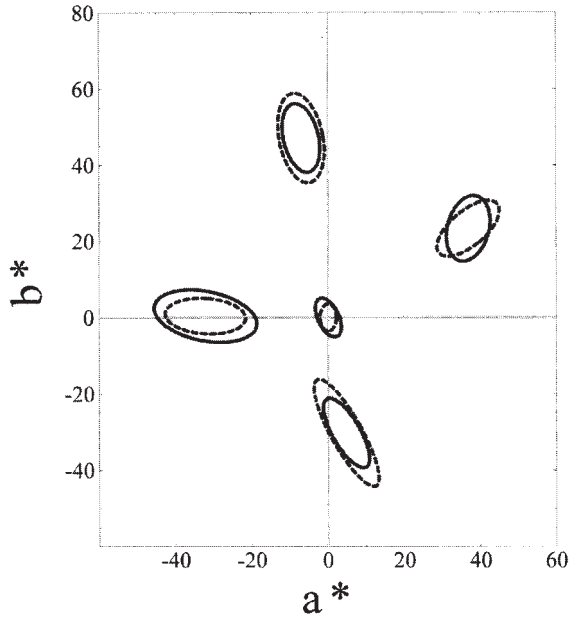

FIG. 7. Comparison between Cheung and Rigg ellipses with an average color difference of about 3–4 CIELAB ΔE units (dashed line) and the present ones of $\Delta V = 4$ (solid line) for all color centers.

PF/3 value. At the Red center, all formulae performed best for ΔV of 4 and worst for ΔV of 8, with the ΔV of 12 between them. The orders of PF/3 values for CMC and CIE94 formulae were always the same with each other at any color center, and those for CIEDE2000 were very different at all centers. The above analysis indicates that the effects of color-difference magnitude on the prediction performance of different formula are different due to their various modal configurations.

Using the individual optimized K_L values for each color center and each formula, the performances of all formulae except CIELAB were improved to some different extents. This reflects on the poor structure of CIELAB system and the correlation between the values of parametric factors and the practical viewing condition. The optimized K_L values of each formula were different for individual color centers. In general, for all equations the K_L values were greatest at Green center, followed by Gray and the two very similar Yellow and Blue centers. At the Red center, the K_L values were the least and all less than one as in their original forms. This indicates that these formulae predicted color differences with rather different performances in different color regions, which further implies the uniformity of corresponding colorimetric space is poor. The limited improvement of the overall performance for all equations, especially for CMC (32.8 to 32.3 PF/3 units), with optimized K_L shows that they were not so sensitive to K_L parameters (at least for the present data set). However, for the combined data set from all centers, the performance sequence form best to poorest was hardly changed, that is, CIEDE2000, CMC, CIE94, and CIELAB, in which only CIELAB fell from third to the last position and exchanged with CIE94 compared with those in their original forms.

The CIEDE2000 color-difference formula is recommended for use with the CIE 10^o color-matching functions and for the color differences under 5 CIELAB ΔE units. Whereas the performance evaluation here shows that this equation outperformed all other formulae in most cases even under these changed conditions of CIE 2^o color-matching functions and the large range of color differences from threshold to 12 CIELAB ΔE units as the maximum in this study. However, it is worth pointing out that these tests in this articles cannot be assumed simply to fully validate this formula for use in large color difference computations, because there are some mathematical discontinuities in the CIEDE2000 computation that may be significant for large color differences.[27]

Except for CIEDE2000 formula at Gray, Yellow, and Blue centers and CMC at Gray, Yellow, and Green centers, most PF/3 values of all equations at each center were rather large in comparison with the total observer precision of 24. This indicates that these CIELAB-based advanced color-difference formulae are far from the final goal, in the authors' opinion, and just the timely but important achievements leading to the ideal aim for industrial color-difference evaluation.

CONCLUSIONS

A new visual data set at the scale of threshold to large suprathreshold color difference at the five CIE color centers in CIELAB space was generated using CRT colors based on the psychophysical method of interleaved staircase and constant stimuli, respectively. Detailed analysis shows that the observer precision at threshold was inferior to that of suprathreshold color difference, which implies the difficulty in color discrimination judgment and that the psychophysical

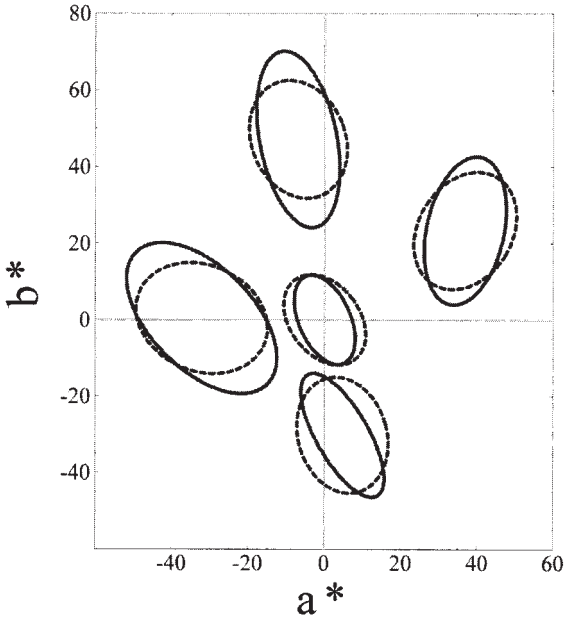

FIG. 8. The ellipse comparison between the Guan and Luo data in GHM mode with a mean color difference of about 13 CIELAB ΔE units (dashed line) and the present $\Delta V = 12$ data (solid line) for all color centers.

TABLE V. Performance of color-difference formulae in terms of PF/3 measure for individual color centers.

Color center	ΔV	CIELAB	CMC	CIE94	CIEDE2000
Original form with $K_L = K_C = K_H = 1$					
Grey	4	24.9	15.8	25.6	**14.3**
	8	26.0	16.1	26.7	**15.1**
	12	27.0	**17.1**	28.0	17.6
	All	*29.7*	*23.4*	*30.7*	***21.4***
Red	4	30.9	27.4	**25.0**	26.1
	8	39.7	36.1	**31.7**	33.7
	12	32.2	30.6	**29.9**	30.0
	All	*39.2*	*35.9*	***34.2***	*34.2*
Yellow	4	25.4	19.5	36.1	**17.8**
	8	26.0	18.1	34.7	**15.3**
	12	27.8	19.5	**34.8**	**16.3**
	All	*30.0*	*22.4*	*38.1*	***19.6***
Green	4	22.1	**16.2**	19.9	18.1
	8	**18.4**	**16.0**	21.8	20.5
	12	**21.5**	25.3	31.6	30.0
	All	*27.7*	***26.6***	*31.8*	*29.5*
Blue	4	27.7	30.6	33.5	**17.4**
	8	27.2	29.9	32.1	**17.0**
	12	25.1	30.5	32.6	**19.3**
	All	*33.5*	*36.2*	*38.8*	***24.8***
All centers	All	38.0	32.8	39.4	**32.6**
Optimized K_L with $K_C = K_H = 1$					
Grey	K_L	*1.18*	*1.38*	*1.17*	*1.36*
	PF/3	26.2	24.3	27.0	**18.5**
Red	K_L	*0.68*	*0.82*	*0.63*	*0.76*
	PF/3	**29.9**	34.3	42.7	35.4
Yellow	K_L	*1.38*	*0.99*	*1.05*	*0.88*
	PF/3	33.9	**22.6**	36.0	**22.8**
Green	K_L	*1.61*	*1.43*	*1.20*	*1.50*
	PF/3	37.7	**26.5**	**28.8**	27.6
Blue	K_L	*0.87*	*1.21*	*0.81*	*0.88*
	PF/3	33.7	31.7	46.2	**28.3**
All centers	K_L	*1.23*	*1.22*	*1.05*	*1.21*
	PF/3	40.0	32.3	38.4	**31.5**

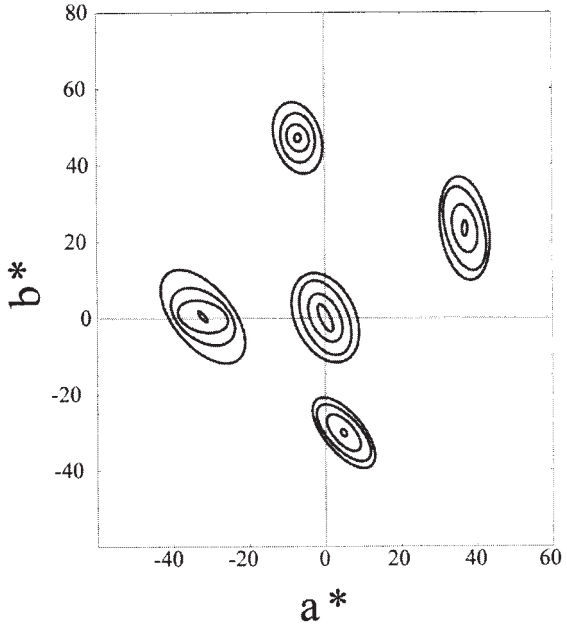

FIG. 9. Chromaticity ellipses predicted by CMC color-difference equation with its original form at all the five CIE color centers plotted in a^*b^* plane.

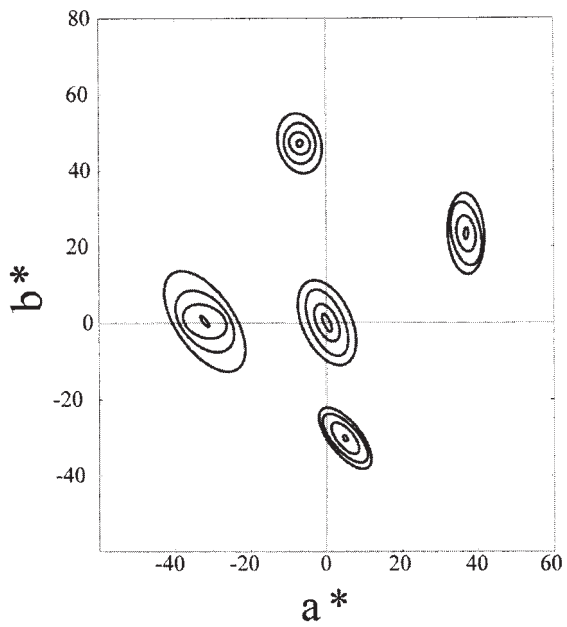

FIG. 10. Chromaticity ellipses predicted by CIE94 color-difference equation with its original form for all color centers plotted in a^*b^* plane.

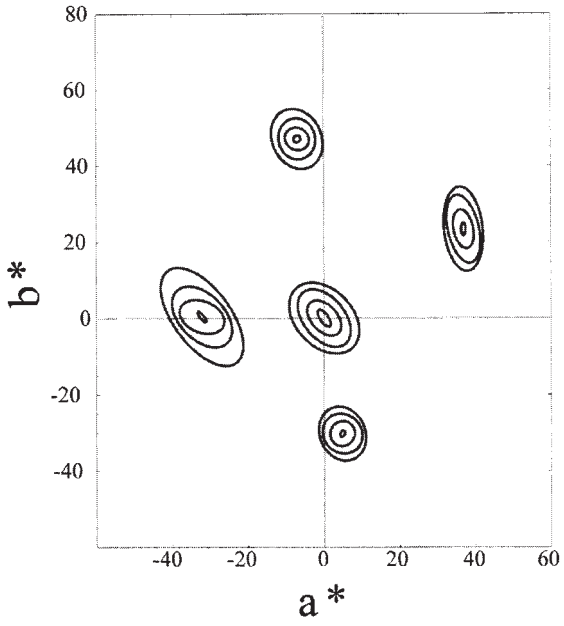

FIG. 11. Chromaticity ellipses predicted by CIEDE2000 color-difference equation with its original form for all color centers plotted in a^*b^* plane.

method of constant stimuli is more accurate or repeatable than the staircase method.

The experimental data in this study were compared with other four previous studies: CRT data by Cui and Luo, surface data by RIT-DuPont, Cheung and Rigg, and Guan and Luo. The results indicate that the reproducibility of average 18% for such kind of visual experiments using CRT colors is rather good and that the CRT data agree well with the surface data, with the mean variations of 14–24%. This suggests that CRT display is a convenient and reliable tool indeed in color vision study.

Three CIELAB-based color-difference formulae were tested using the new data set produced in this study. The CIEDE2000 and CMC equations outperformed CIE94 and CIELAB for the combined dataset of all color centers, no matter in their original forms or with K_L optimized. The former two formulae performed very similarly, but CIEDE2000 was the best among all formulae, particularly at Blue and Gray centers. In addition, the PF/3 values of all formulae at different visual scales indicate that the effects of color-difference magnitude on the prediction performance of individual formula are rather different because of their various modal configurations. However, the different optimized K_L values at each center and the relatively low prediction accuracy of all equations suggest the further progress in industrial color-difference study.

ACKNOWLEDGMENTS

The authors acknowledge the Scientific Research Foundation for the Returned Overseas Chinese Scholars of State Education Ministry and Zhejiang Province of P. R. China for supporting the project. The authors thank Professor M. Ronnier Luo for providing technical guidance and all the observers for their patience and time. Finally, the authors are also full of gratitude to the anonymous referees for their detailed and thoughtful comments, which helped to improve the article. Contract grant sponsor: SRF for ROCS, SEM and ZJP, P. R. China.

1. Clarke FJJ, McDonald R, Rigg B. Modification to JPC79 colour-difference formula. J Soc Dyers Col 1984;100:128–132.
2. CIE. Technical Report: Industrial Colour-Difference Evaluation. CIE Publication No. 116. Vienna: Central Bureau of the CIE; 1995.
3. Luo MR, Cui G, Rigg B. The development of the CIE 2000 colour-difference formula: CIEDE2000. Color Res Appl 2001;26:340–350.
4. CIE. Colorimetry, 2nd ed. CIE Publication No. 15.2. Vienna: Central Bureau of the CIE; 1986.
5. Berns RS, Alman DH, Reniff L, Snyder GD, Balonon-Rosen MR. Visual determination of suprathreshold color-difference tolerances using probit analysis. Color Res Appl 1991;16:297–316.
6. Witt K. Geometric relations between scales of small colour differences. Color Res Appl 1999;24:78–92.
7. Kim DH, Nobbs JH. New weighting functions for the weighted CIELAB colour difference formula. Proc Color 97 Kyoto 1997;1:446–449.
8. Luo MR, Rigg B. BFD (*l*:*c*) colour-difference formula. Part I—development of the formula. J Soc Dyers Col 1987;103:86–94.
9. Berns RS. Color tolerance feasibility study comparing CRT-generated stimuli with acrylic-lacquer coating. Color Res Appl 1991;16:232–242.
10. Montag ED, Berns RS. Visual determination of hue suprathreshold color-difference tolerances using CRT-generated stimuli. Color Res Appl 1999;24:164–176.
11. Witt K. CIE guidelines for coordinated future work on industrial colour-difference evaluation. Color Res Appl 1995;20:399–403.
12. Robertson AR. CIE guidelines for coordinated research on colour-difference evaluation. Color Res Appl 1978;3:149–151.
13. Xu H, Yaguchi H, Shioiri S. Correlation between visual and colorimetric scales ranging from threshold to large color difference. Color Res Appl 2002;27:349–359.
14. Indow T, Robertson AR, von Grunau M, Fielder GH. Discrimination ellipsoids of aperture and simulated surface colors by matching and paired comparison. Color Res Appl 1992;17:6–23.
15. Alman DH, Berns RS, Snyder GD, Larsen WA. Performance testing of color-difference metrics using a color tolerance dataset. Color Res Appl 1989;14:139–151.
16. Indow T, Morrison ML. Construction of discrimination ellipsoids for surface colors by the method of constant stimuli. Color Res Appl 1991;16:42–56.
17. Luo MR, Rigg B. BFD (*l*:*c*) colour-difference formula. Part II—performance of the formula. J Soc Dyers Col 1987;103:126–132.
18. Guan SS, Luo MR. Investigation of parametric effects using small colour differences. Color Res Appl 1999;24:331–343.
19. Coates E, Fong KY, Rigg B. Uniform lightness scales. J Soc Dyers Col 1981;97:179–183.
20. Schultz W. The usefulness of color difference formulae for fixing color tolerances, color metrics. Soesterberg: AIC/Holland; 1972. p 245–265.
21. Cui G, Luo MR, Rigg B, Li W. Colour-difference evaluation using CRT colours. Part I: data gathering and testing colour difference formulae. Color Res Appl 2001;26:394–402.
22. Cui G, Luo MR, Rigg B, Li W. Colour-difference evaluation using CRT colours. Part II: parametric effects. Color Res Appl 2001;26:403–412.
23. Guan SS, Luo MR. Investigation of parametric effects using large colour differences. Color Res Appl 1999;24:356–368.
24. Montag ED, Wilber DC. A comparison of constant stimuli and grey-scale methods of color difference scaling. Color Res Appl 2003;28:36–44.
25. Strocka D, Brockes A, Paffhausen W. Influence of experimental parameters on the evaluation of color-difference ellipsoids. Color Res Appl 1983;8:169–175.
26. Cheung M, Rigg B. Colour-difference ellipsoids for five CIE colour centers. Color Res Appl 1986;11:185–195.
27. Sharma G, Wu W, Dalal N. The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. Color Res Appl 2005; 30:21–30.