Testing color difference evaluation methods for color digital images

Qingfen Tong (仝清芬), Haisong Xu (徐海松)*, and Rui Gong (宫睿)

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

Corresponding author: chsxu@zju.edu.cn

Received January 29, 2013; accepted April 12, 2013; posted online July 3, 2013

A psychophysical experiment is carried out to evaluate the color difference between pairs of color digital images and their modulated versions, which are displayed on a professional liquid crystal display (LCD) monitor, using the category judgment method. The modulated images for six original images are generated with variations in five attributes, namely, lightness, chroma, hue, resolution, and sharpness, considering both their chromatic and spatial characteristics. Several color difference evaluation methods, namely, CIELAB, CIEDE2 000, CAM02-UCS, S-CIELAB, and iCAM, are compared based on the experimental data. The results demonstrate that the performance of iCAM is the best for predicting image color difference.

OCIS codes: 330.0330, 330.1720, 330.6110.

doi: 10.3788/COL201311.073301.

There are various methods reported on the traditional color difference calculation such as CIELAB[1], CIEDE-2000[2], and CAM02-UCS[3]. These methods are all established based on uniform color patch samples. However, an image is formed by a large number of pixels with different colors, making it quite different from evaluating the uniform color samples; such samples are also difficult to measure directly with physical instruments. As such, a method must be developed to evaluate the color difference for color digital images.

In their research on image color difference evaluation, Song *et al.*[4−6], tested traditional color difference formulas in the image applications. Their results show that the mean perceived color difference is roughly $2.5\Delta E^*_{\text{ab}}$, but those for different test images are quite different depending on the image contents, indicating that it is necessary to consider the spatial characteristics of human vision system and the complexity of image contents for image color difference evaluation. Zhang *et al.*[7] proposed the S-CIELAB image color difference model using an image space filter according to the human visual contrast sensitivity function. Moreover, Johnson *et al.*[8] proposed a framework as a color image difference metric. The framework extends the idea of spatial extension in the S-CIELAB model by adding several pre-processing steps, including spatial filtering, adaptation, localization, as well as local and global contrast detection. The modular image difference metric is then incorporated into the image color appearance model (iCAM) to address image appearance, difference, and quality within a single model[9,10].

In this letter, the color differences among CIELAB, CIEDE2 000, CAM02-UCS, S-CIELAB, and iCAM for each image pair, which are based on the visual experiment for the standard test images and their modulated versions, are calculated and compared with the subjective evaluation results of the observers. The calculation and comparisons are performed to further investigate the image color difference evaluation methods.

Six test images were selected to cover the colors of the familiar objects, four of these were ISO SCID 300 images, including N1 (Woman), N2 (Street), N3 (Fruits), N6 (Flower), one from a CIE TC8-03 sRGB image named Ski, together with an additional image "Tree" for the colors of sky and plants (Fig. 1). These images were all clipped to ensure they had the same size of 15 × 20 (cm). To generate test images similar to the original images under a limited extent of difference, the six original images were manipulated in terms of five attributes, i.e., lightness, chroma, and hue angle corresponding to L^*, C, and H of CIELAB color space, respectively; resolution (R); and sharpness (S), according to the functions and parameters listed in Table 1. Five linear manipulations in the L^* and C channels were performed to investigate the effect of linear changes in lightness and chroma. Reductions and increases in lightness and chroma contrast were simulated using the sigmoid and inverse-sigmoid functions in $L^*$and C, respectively[11]. The variations of image resolution were manipulated using the bicubic resampling method, equivalently regarding 2 × 2, 3 × 3, 4 × 4 pixels as 1 × 1 pixels. One method based on a high frequency emphasis filter[11], was applied to increase sharpness. A total of 216 test images (6 images × 36 manipulations) were produced.

The visual experiment was conducted in a dark room.

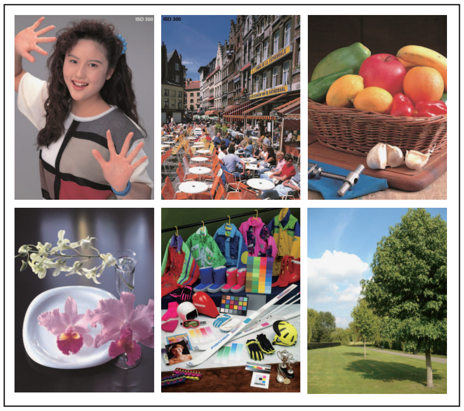

Fig. 1. Six original test images for the psychophysical evaluation.

Table 1. Image Manipulation Functions and their Corresponding Parameters

Attribute	Description	Parameter
L^*	Linear: $L^* = k \times L_0^*$ Sigmoid: $L^* = \dfrac{100}{[1/(1+M^E)] \times 1 + [M/0.01 \times L_0^*]^E}$ Inverse-sigmoid: $L^* = 100 \times M \times \left[\dfrac{1 - 0.01 \times [1/(1+M^E)] \times L_0^*}{0.01 \times [1/(1+M^E)] \times L_0^*}\right]^{-1/E}$	k=0.8, 0.9, 0.95, 1.05, 1.1 M=1.23, E=1.45, in sigmoid and inverse-sigmoid as SS and ISS, respectively, M=0.75, E=1.90 as SM and ISM, M=0.63, E=2.35 as SL and ISL, respectively.
C	Linear: $C = k \times C_0$ Sigmoid: $C = \dfrac{100}{[1/(1+M^E)] \times \{1 + [M/(C_0/C_{\max})]^E\}}$ Inverse-sigmoid: $C = 100 \times M \times \left[\dfrac{1 - 0.01 \times [1/(1+M^E)] \times (C_0/C_{\max})}{0.01 \times [1/(1+M^E)] \times (C_0/C_{\max})}\right]^{-1/E}$	
H	Offset: $h_{\text{out}} = h_0 \pm k$	k=2.5°, 5°, 10°
R	Bicubic resampling method	2×2, 3×3, 4×4
S	$L_{\text{s}}^* = \text{ifft2}[\text{fft2}(L_0^*) \times \text{filter}]$ $\text{filter} = 1 + 1.5 \times \left\{1 - \exp\left[\dfrac{-x^2}{2 \times (a \times p)^2}\right]\right\}$ x: spatial frequency a: resolution of the display	p=1/3, 1/5, 1/7, 1/9, 1/11

A 24–inch EIZO professional liquid crystal display (LCD) of ColorEdge CG241W, with a resolution of 1 920 × 1 200 pixels, was characterized using the Gain-Offset-Gamma model[12] under illuminant D65 and colorimetric accuracy of $0.92\Delta E_{\text{ab}}^*$. The image pairs were simultaneously displayed on the LCD with a neutral gray background; the lightness was equal to 22.5, where the resolution of a single image was 549 × 732 pixels. The viewing distance was set at 80 cm to obtain a horizontal view angle of 22.5° and a vertical view angle of 14.25°

In each experiment session, an image pair, including an original and one of its manipulated images, was presented to the observer in a random order. The position of the presented images on the left or right of the screen was also randomized to minimize the effect of the non-uniformity of display and the adaptation of the observers. The psychophysical method of category judgment with a 7-point grade was employed, and a panel of 10 observers with normal color vision was invited to assess the visual color difference sensation of the displayed image pairs according to the descriptions of category shown in Table 2. The 10 observers were all graduate students from Zhejiang University, of which 6 were males and 4 were females, with ages ranging from 20 to 33 years. In this experiment, each of the 10 observers evaluated all the 216 image pairs, and then 5 of them assessed half of all the image pairs again to estimate the observer repeatability. A total of 2 700 visual judgments comprising (10 observers×216 image pairs)+(5 observers×108 image pairs) were collected.

Table 2. Descriptions of the Judgment Category for the Visual Experiment

Level of Color Difference	Grade
Imperceptive	1
Just Perceptible	2
Perceptible but Completely Acceptable	3
Reluctantly Acceptable	4
Just Unacceptable	5
Completely Unacceptable	6
Extremely Unacceptable	7

In this letter, the observer variations were computed using the coefficient of variation, CV, as defined as

$$\text{CV} = \frac{100}{\overline{y}}\left[\sum(x_{\text{i}} - y_{\text{i}})^2/n\right]^{1/2}, \tag{1}$$

where n is the number of assessed images. In addition, for intra-observer accuracy, x_{i} and y_{i} are the first- and second-judgment data respectively, while for inter-observer accuracy, x_{i} is the individual observer data, and y_{i} is the average data over all the 10 observers. The mean value of y_{i} data set is $\overline{y}$. A CV value of 30 means a 30% disagreement between the two sets of data.

The average CV of inter-observer accuracy is 33.1 (range from 27.6 to 42.7), while that of intra-observer accuracy is 36.4 (range from 21.7 to 48.6). With comparison to the published data[11,13], this observer variation is acceptable, and thus, the experimental data are valid and credible.

All the categorical data were transformed into the equal-interval scale values. First, the number of observers contained in each grade of the category was calculated to obtain the frequency matrix. After computing the cumulative frequency matrix and cumulative probability matrix, the Z-score matrix was computed through the inverse of the standard normal cumulative distribution. Finally, the Z-score matrix was transformed to interval scale values[11] equivalent to the visual judgment of the observer as regards the color difference for the 216 image pairs. Here, a bigger scale value reflects higher visual sensitivity for the color difference between the tested image pair. In Fig. 2, the ΔV of the vertical axis represents the average scale values of all the test images for different manipulation attributes. The error bars show the standard deviation of 95% confidence interval, thereby indicating the influence of the image contents.

For the linear manipulation of lightness and chroma

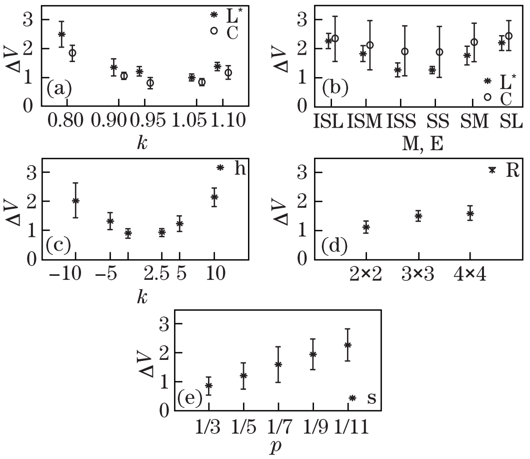

Fig. 2. Scale values of visual judgments for different manipulation methods. (a) Linear manipulation of lightness and chroma; (b) sigmoid and inverse-sigmoid manipulation of lightness and chroma; (c) offset manipulation of hue; (d) manipulation of resolution; (e) manipulation of sharpness.

with the same parameter k, the scale values for lightness is significantly higher than that of chroma (F=1.91 with F_c=1.86 based on the results of F-test analysis) (Fig. 2(a)). This finding indicates that the human vision system is more sensitive to lightness difference, similar to those of previous studies[4,6]. In addition, the effect of image contents on the visual judgments for linear manipulation of lightness and chroma is related to the parameter k. This means that the influence of image contents on the visual judgments is related to the amount of change in lightness and chroma. In Fig. 2(b), SS, SM, and SL represent the increase of lightness contrast and chroma contrast, respectively, from small to large extent using the sigmoid function. Meanwhile, ISS, ISM, and ISL produce the reduction of lightness contrast and chroma contrast using the inverse-sigmoid function. Based on the results, for the contrast manipulation of chroma and lightness with the same parameters, the scale values of lightness contrast manipulation are significantly lower than those of chroma (with F value of 0.39 and F_c value of 1.76 according to the F-test results). Moreover, the scale value rises with the increase of both chroma and lightness contrast variation. Once again, the influence of image contents is quite large for chroma contrast variations, and both the scale values and the effect of image contents for hue attribute increase with the rising of the hue offset (Fig. 2(c)).

For resolution and sharpness manipulations, shown in Figs. 2(d) and (e), respectively, the variation of the scale values shows a linear tendency with the decrease of resolution and the increase of sharpness, of which the correlation coefficients are 0.875 and 0.999, respectively. Herewith, some relationship exists between the perceived image color difference and the change of resolution and sharpness, thus implying that the spatial attributes of image need to be considered when evaluating the image color difference. Moreover, with increase of image sharpness, the influence of image contents on the visual judgment increases at first, before reaching an approximately constant value. However, the effect of image contents on the visual judgment of different image resolutions is not significant, which may be due to the fact that the changes of resolution do not cause obvious variations in lightness and chromatic components.

The image color difference metric predicted by an ideal formula should be consistent with the subjective evaluation results of the observer. In this letter, the color differences were calculated between the original images and their modulated images pixel by pixel using CIELAB, CIEDE2 000, CAM02-UCS, S-CIELAB, and iCAM, respectively. The correlations between the calculated color differences and the corresponding visual evaluation values were analyzed by standardized residual sum of squares (STRESS)[14]. The average STRESS values of different manipulation attributes for all the test image pairs, together with their standard deviation values, are listed in Table 3. As can be seen, the smaller STRESS values represent better correlation, and larger standard deviation demonstrates greater impact of the image contents on the correlation. The average STRESS values of all the test image pairs for the five image color difference evaluation methods according to different attributes are depicted with the standard deviations (error bars) in Fig. 3.

As can be seen from Table 3 and Fig. 3, iCAM performs best in terms of the overall prediction accuracy for all the manipulation attributes; it is followed by S-CIELAB, CIELAB and CIEDE2 000, with CAM02-UCS being the poorest. The performance of iCAM is also the most outstanding in each manipulation component, except for the resolution manipulation. This is mainly because the pre-processing procedures before calculating the color difference pixel by pixel in iCAM are not applicable to this situation. The pre-processing procedures of iCAM used in this letter include spatial filtering and spatial localization, as introduced by Johnson *et al.*[8]. However, only the pre-processing procedures enhance the

Table 3. Average Values of STRESS and their Standard Deviations

	CIELAB		CIEDE2000		CAM02-UCS		S-CIELAB		iCAM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
L^*	27.3	5.1	31.1	5.5	28.9	5.8	24.9	6.4	17.0	2.9
C	26.5	9.8	33.2	7.5	30.3	13.0	30.9	7.7	16.7	5.0
H	20.8	8.2	20.9	7.9	42.5	2.3	20.8	7.9	17.4	9.1
R	14.8	4.8	14.6	4.6	10.9	4.6	21.1	6.7	33.1	14.3
S	17.1	4.1	16.5	4.2	15.3	9.7	21.6	4.0	13.8	5.1
Overall	32.9	6.1	34.2	5.1	54.5	7.8	30.4	6.8	21.1	4.8

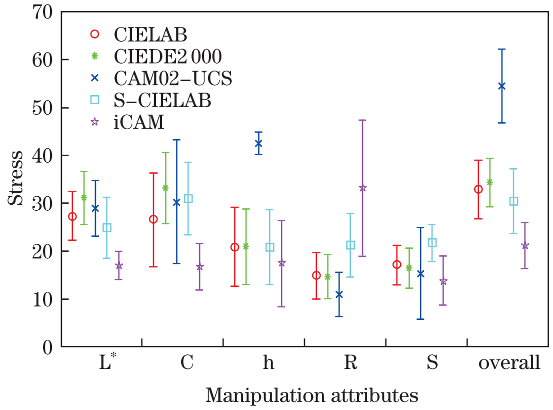

Fig. 3. Performance comparison of image color difference evaluation methods.

prediction performance of iCAM for other manipulation attributes. The result predicted by S-CIELAB is slightly worse than that by iCAM, indicating that only one spatial filtering in S-CIELAB may not be enough for image color difference evaluation. Therefore, iCAM is the most promising method for image color difference evaluation. However, further studies are expected to improve the applicability of the pre-processing procedures in iCAM.

Meanwhile, the performance of CAM02-UCS for different manipulation attributes is very unstable (Fig. 3). The ability of CAM02-UCS to predict resolution manipulation is the best due to the superiority of color appearance model, which considers the effect of background, surroundings, and so on. The prediction of CAM02-UCS for hue manipulation is rather poor. This can be attributed to the fact that the offset of hue is modulated in CIELAB color space, which has no effect on lightness and chroma attributes when calculating color difference, although the variation of hue in CAM02-UCS system can lead to chroma change[3]. Therefore, the hue difference between the pairs of corresponding pixels in an image pair calculated using CIELAB and CAM02-UCS may be quite different, which is also true for the overall hue difference of the image pair. Contrary to CAM02-UCS, the performances of CIELAB and CIEDE2 000 are relatively stable for different manipulation attributes, thus implying that the simple calculation results through the traditional color difference formula for different attributes is consistent. However, the overall performances of CIELAB, CIEDE2 000, and CAM02-UCS are inferior to those of S-CIELAB and iCAM, indicating that the traditional color difference formulas based on uniform color patch samples are not very suitable for image color difference evaluation.

An expected image color difference evaluation method should consider several aspects, such as the influence of background and surrounding as well as the contrast sensitivity of human vision system. Hence, it is important to conduct more in-depth research about the pre-processing steps (including spatial filtering, adaptation, and contrast detection) in iCAM to achieve an ideal image color difference evaluation method.

In conclusion, a visual experiment is carried out through the psychophysical method of category judgment, in order to test the performance of CIELAB, CIEDE2 000, CAM02-UCS, S-CIELAB, and iCAM for image color difference evaluation. The test images are manipulated with different parameters in five image attributes (i.e., lightness, chroma, hue, resolution, and sharpness) including both chromatic and spatial alterations. Based on the visual judgments for the image pairs of different manipulations, the correlations between the predictions by the five image color difference models and the subjective estimations of the observers are analyzed by STRESS and CV. The results demonstrate that iCAM outperforms the others, though it must be further studied to improve its applicability to the evaluation of image color difference.

References

1. Colorimetry, CIE Publication 15.2, Vienna: CIE Central Bureau (1986).
2. M. R. Luo, G. Cui, and B. Rigg, Color Res. Appl. **26,** 340 (2001).
3. N. Moroney, in *Proceedings of the 10th IS&T/SID Color Imaging Conference* 23 (2002).
4. T. Song and M. R. Luo, in *Proceedings of the 8th IS&T/SID Color Imaging Conference* 44 (2000).
5. L. W. MacDonald and M. R. Luo, *Color Image Science: Exploiting Digital Media* (John Wiley and Sons Limited, Chichester, 2002).
6. J. E. Gibson, M. D. Fairchild, and S. L. Wright, in *Proceedings of the 8th IS&T/SID Color Imaging Conference* 295 (2000).
7. X. Zhang and B. A. Wandell, J. SID **5,** 61 (1997).
8. G. M. Johnson and M. D. Fairchild, in *Proceedings of the 9th IS&T/SID Color Imaging Conference* 108 (2001).
9. M. D. Fairchild and G. M. Johnson, J. Electron. Imaging **13,** 126 (2004).
10. G. M. Johnson and M. D. Fairchild, Proc. SPIE **5007,** 51 (2003).
11. S. Y. Choi, "Modeling Color and Image Appearance under Flat Panel Display Viewing Conditions", PhD. Thesis (University of Leeds, 2008).
12. R. S. Berns, R. J. Motta, and M. E. Gorzynski, Color Res. Appl. **18,** 299 (1993).
13. Z. Wang and J. Y. Hardeberg, J. Electron. Imaging **21,** 2 (2012).
14. P. A. García, R. Huertas, M. Melgosa, and G. Cui, JOSA A **24,** 1823 (2007).