

Image Distortion Maps[1]

Xuemei Zhang, Erick Setiawan, Brian Wandell
Image Systems Engineering Program
Jordan Hall, Bldg. 420
Stanford University, Stanford, CA 94305

Abstract

Subjects examined image pairs consisting of an original and a reproduction created using either JPEG compression or digital halftoning. Subjects marked locations on the reproduction that differed detectably from the original. We refer to the distribution of error marks by the subjects as *image distortion maps.*

The empirically obtained image distortion maps are compared to the visible difference calculated using three color difference metrics. These are color distortions predicted by the widely used mean square error (point-by-point MSE) computed in RGB values, the point-by-point CIELAB ΔE color difference formula (CIE, 1971), and S-CIELAB, a spatial extension of CIELAB that incorporates spatial filtering in an opponent colors representation prior to the CIELAB calculation (Zhang & Wandell, 1996).

For halftoned reproductions the RMS, CIELAB, and S-CIELAB error sizes correlated with the locations marked by subjects reasonably well, given the freedom to select a threshold level separately for each image. The S-CIELAB metric had the most consistent threshold levels across images; the RMS metric had the least consistent levels. For JPEG reproductions, all three metrics provided poor predictions of subjects' marked locations.

1. Introduction

One application of image fidelity models is to predict the reproduction quality at different locations within an image. To test the accuracy of such models, it is necessary to have a database of experimental measurements establishing where subjects perceive image reproduction errors. In this paper we report a set of measurements of perceived reproduction errors for a set of natural images and reproductions of these images created using (a) digital halftoning (void and cluster), and (b) image compression (JPEG-DCT).

After describing our experimental methods and results, we evaluate how well three different color difference metrics predict the data set. The three metrics are (a) a root-mean-squared (RMS) error metric applied to the red, green and blue (RGB) framebuffer entries, (b) the CIELAB color difference metric, and (c) the S-CIELAB color difference metric, which is a spatial extension of CIELAB. We evaluated these metrics in order to understand the conditions in which they can be usefully applied to predict the appearance of distortions in digital image reproductions.

The three metrics made predictions that described subjects' marked locations of halftone errors reasonably well, given the freedom to choose a separate threshold level for each image. The S-CIELAB metric had the most consistent threshold level across images; the RMS metric had the least consistent levels. All three metrics provided very poor predictions of subjects' responses to the JPEG-DCT reproductions.

2. Methods

Six 24-bit digital color images were used as originals. The images included faces, objects, buildings, and natural scenes. Reproductions were created for each image using two methods. One set of images was created using a void-and-cluster halftone method containing 32 levels. The dithered reproduction error includes dot noise that is particularly salient in some, but not other, regions of the image. A second set of images was created using JPEG-DCT with the standard quantization tables (Q factor of 50). Hence, a total of twelve reproductions were used in these experiments.

Each original-reproduction image pair was presented on a CRT screen controlled using a Macintosh computer. The images were viewed in a dark room with light from the computer screen only. Subjects viewed the display at a 12 inch distance. Calibrations were performed using a PhotoResearch PR-704 Spot SpectraScan spetral scanner and a PhotoResearch Auto Telephotometer, using software routines from the Psychophysics Toolbox on Macintosh by Brainard (1997). From the monitor calibration data, we computed the CIE XYZ representations of each image as shown on the CRT display. These XYZ values were used to compute the point-by-point CIELAB and S-CIELAB error values. Point-by-point RMS errors were computed from the framebuffer values and needed no calibration in-

[1]Supported by a donation from the Hewlett-Packard Corporation.

formation.

Subjects were undergraduate students at Stanford; they were paid for their participation. Subjects were tested for normal color vision using a set of isochromatic plates (Ishihara Plates). Twenty-four subjects were tested on the 12 image pairs.

Subjects identified regions in the image where the original and reproduction appeared to differ. They marked these regions using the mouse. Subjects could mark regions using a small (10 pixels in diameter, 0.4 deg), medium (30 pixels; 1.2 deg) or large (50 pixels, 2 deg) circular spot. The finite size of the marker limits the spatial resolution of the measured image distortion map and we account for this in the data analysis below. Each image pair was presented to the subjects once or twice in randomized order.

3. Experimental Results

3.1. Data compilation

Two subjects' data were eliminated because one of them was not cooperative (drew little faces or wrote words on the images instead of marking places of visible error), and the other marked everywhere on every image. The error marks produced by the remaining subjects were pooled for each pair of images. From these pooled data we calculate a probability of a mark covering each pixel in the reproduction, which we call *image distortion maps*. Figure 1 shows the image distortion map for an original and its reproduction. The probability that a pixel is marked is represented by the gray level: Light regions correspond to frequently marked areas (high visible differences) and dark regions correspond to infrequently marked areas (low visible difference).

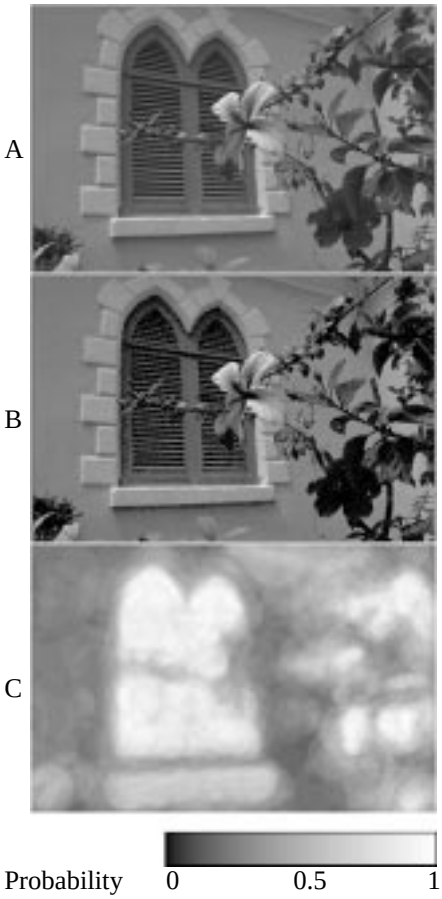

Figure 1: *The image distortion map for an original and its reproduction. (A) The original color image is shown in grayscale. (B) The halftoned reproduction using void-and-cluster matrix is shown in grayscale. (C) The image distortion map measured by pooling the data from all observers is shown. Light regions indicate areas with a high probability of being marked, dark regions indicate a low probability of being marked.*

3.2. Consistency of subjects' responses

Next, we evaluated how well different subjects agreed in their judgments concerning the locations of reproduction errors. We use the inter-subject consistency as a criterion against which to measure how well each of the color difference metrics predict the image distortion maps.

We estimate the variability of an image distortion map as follows. First, we assume the number of marks at each pixel is described by a binomial distribution. The maximum likelihood estimate of the binomial parameter of each pixel, p, is the value of the image distortion map. Using this estimate of the binomial probability, we compute the negative log likelihood ($\mathcal{N}$) of the marks made by each subject. The average $\mathcal{N}$ of all subjects data measures how well the image distortion map agrees with each subject's data. We use the average $\mathcal{N}$ to measure the reliability of the image distortion map, and thus to serve as a bound on how well we expect the models to perform. We do not expect a model to predict the data any better than the reliability of the image distortion map.

Table 1 lists the value of $\mathcal{N}$ averaged across subjects for each image. To make the numbers comparable across images, the likelihoods were normalized by image size and by the number of times the image was presented. Larger values in the table indicate larger inter-subject variability. The values in parentheses are percentage of the image area covered by subjects' marks, averaged over all subjects.

The $\mathcal{N}$ values for two of the JPEG-DCT compressed images are significantly smaller than the other images. This occurred because the reproductions appeared very similar to the original in these cases (as indicated by the percentage of mark coverage) and there was inter-subject agreement concerning the locations of the small number of vis-

ible reproduction errors.

	JPEG $\mathcal{N}$(%coverage)	Halftone $\mathcal{N}$(%coverage)
face	0.1984 (0.260)	0.2423 (0.653)
flowers	0.2666 (0.333)	0.2545 (0.513)
hats	0.2515 (0.314)	0.2015 (0.591)
house	0.0474 (0.029)	0.2093 (0.586)
rafting	0.1667 (0.170)	0.2197 (0.311)
wall	0.0471 (0.029)	0.1520 (0.314)

Table 1: *Consistency of image distortion map data. We use the average negative log likelihood, $\mathcal{N}$, of an individual subject's performance given the image distortion map to measure inter-subject variability. Larger values represent larger variability. The values in parentheses are percentage of coverage by subjects' marks for each image. See text for details.*

4. Predictions

4.1. Error models

The RMS error values were computed as point-by-point vector length of the RGB difference image between an original image and its reproduction. For example, the point-by-point RMS error at position (i, j) of the images is:

$$RMS_{ij} = \sqrt{(\Delta R_{ij})^2 + (\Delta G_{ij})^2 + (\Delta B_{ij})^2} \quad (1)$$

where ΔR, ΔG, and ΔB represent the difference in R, G, and B values between the original color image and the reproduction.

The point-by-point CIELAB errors were computed from XYZ values of an original image and its reproduction. We use the standard CIELAB color difference formula (CIE, 1971) to compute CIELAB errors. The result is an error map with one ΔE value per pixel.

The S-CIELAB errors were computed using the method described in Zhang & Wandell (1996) [2]. The result is also an error map with one ΔE value per pixel.

4.2. Simulations

The image distortion maps cannot be predicted directly from the model errors because of (a) the size of the markers, and (b) the variability in positioning the cursor. To predict image distortion maps from the model errors, we simulate the experiment using a two step procedure.

[2] Software for performing this computation is available on the internet at: http://white.stanford.edu/scielab/.

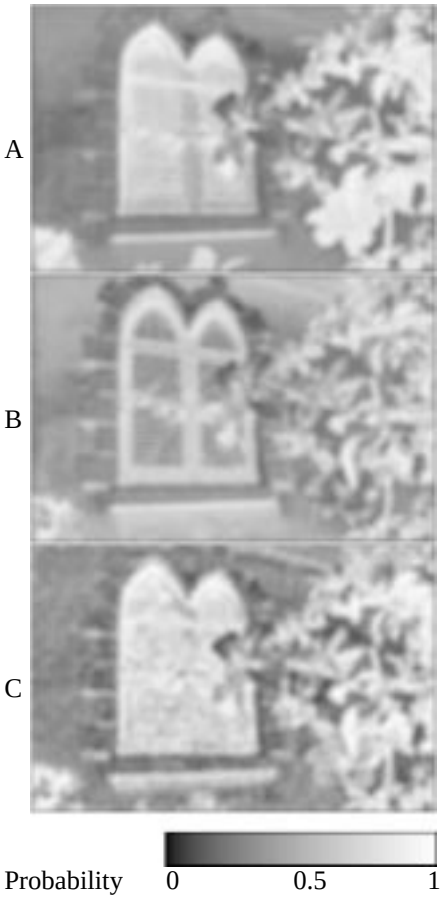

Figure 2: *The error maps computed using three different metrics. The three panels show error maps computed using the (A) RMSE, (B) CIELAB, and (C) S-CIELAB methods.*

1. We convert the error measure at each point to a probability of selecting that point, using the function

$$\hat{p} = 1 - exp(-(\frac{x}{t})^a) \quad (2)$$

The value x is the error measure computed from specific metrics, a is the acceleration parameter, and t is the threshold parameter. This function relates computed error measures to probability of each pixel being marked. The value t is the ΔE or RMS value at which the probability of marking a pixel is about 63%. The parameters a and t are estimated from the image distortion maps.

2. We convert the probability of marking each pixel to an image distortion map by convolving the pixel-marking probabilities with a kernel of the same size and shape as the smallest marker used by subjects in the experiments. The kernel values sum to one, thus preserving the mean. The convolution simu-

lates the effects of the marker size and the variability in marker placement.

Using this simulation, we derive a predicted image distortion map from the errors predicted by each of the metrics. In the next section, we compare the observed and predicted image distortion maps.

5. Evaluations

Images with halftone errors and JPEG errors are analyzed separately, because they represent different types of distortions. The halftone distortions generally are a random texture noise, while the JPEG distortions generally include blurring and blocking artifacts.

5.1. Halftone distortions

To fit the data with each model, we must specify the parameters a and t that relate the error value to the probability of marking a location. To begin the analysis, we chose a single acceleration parameter, a, for each model and used this parameter for all halftone image pairs. The threshold parameter, t, varied across images.

Metrics:	RMS	CIELAB	S-CIELAB
face	0.2779	0.2696	0.2738
flowers	0.2832	0.2918	0.2811
hats	0.2454	0.2441	0.2411
house	0.2401	0.2419	0.2387
rafting	0.2584	0.2514	0.2483
wall	0.2122	0.2505	0.2183

Table 2: *Quality of fitting model outputs to data: Negative log likelihood errors for halftone images. The likelihood errors are normalized for image size so that the numbers are comparable across images.*

For each image, the overall model error in predicting the image distortion map for each image pair is listed in Table 2. These values can be compared directly with the inter-subject variability, $\mathcal{N}$, in Table 1. The table shows that when the threshold value is free to vary across images, each metric can predict the image distortion map about as well as the precision of the data.

Figure 2 is a more detailed examination of the predicted image distortion maps from the three metrics. The predicted distortions maps agree with the data at a coarse scale. On a finer scale, there are regions where each metric makes incorrect predictions. Figure 3 shows the regions where each metric deviates most from the data in the "Flowers" image. As expected, CIELAB fails mainly in the high frequency regions of the image (such as the blinds on the window), due to lack of spatial sensitivity mechanisms in the metric. Over relatively constant regions, both CIELAB and S-CIELAB did better than RMS (such as the large flower in front).

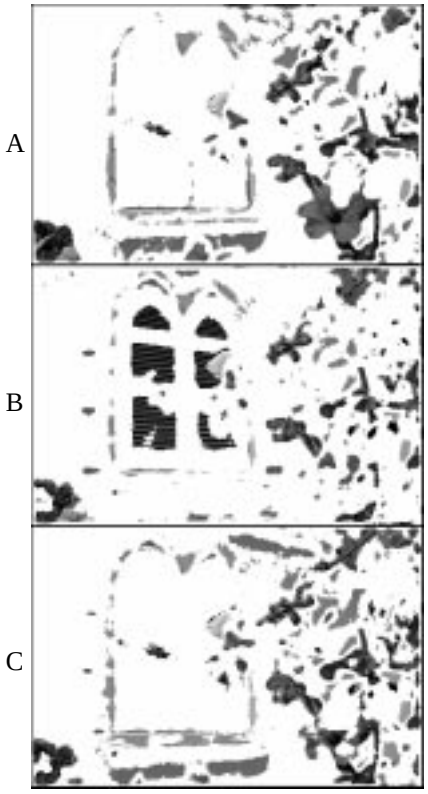

Figure 3: *Regions of largest deviation between model and data. Each panel indicates those image regions within the highest quintile of $\mathcal{N}$ for the RMS (A), CIELAB (B) and S-CIELAB (C) models. Regions in the lower four quintiles are shown as white.*

A useful image metric should make consistent predictions across different images, so that interpretation of the ΔE or RMS values are meaningful in practice. In fitting the models to the data, we have allowed different threshold parameters t in the psychometric functions for each image. The psychometric functions that relate model error measures with detection probabilities for all halftone images are plotted in Figure 4.

Figure 4 shows that S-CIELAB had the most consistent threshold predictions across images. The ΔE values corresponding to 63% detection (t values)are between 2 and 5 ΔE units, consistent with the traditional interpretation of ΔE values. The CIELAB values ranged from 4 to 20 ΔE units at 63% detection. Thus, a ΔE value of 10 was predicted to be nearly always visible in one image, but visible less than half the time in another image. The RMS threshold values were the most variable across images, ranging an order of magnitude, from 0.036 to 0.36 (the largest pos-

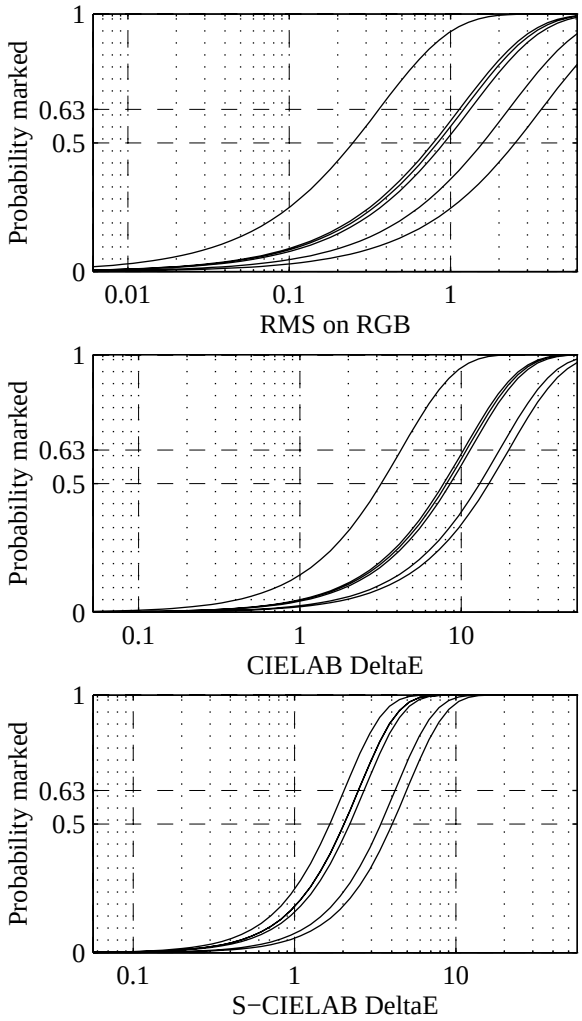

Figure 4: *Psychometric functions that relate ΔE or RMS values to detection probabilities, fitted to image pairs with halftone distortions. In each plot, one curve represents one image pair.*

sible RMS error is 1.732).

Figure 5 is a graphical evaluation of the cost (increased negative log likelihood error $\mathcal{N}$) of fixing the threshold level, t, across all images. A model with consistent thresholds across images pays no cost, and a model thresholds that vary greatly across images pays a large cost. As shown in the plot, when the threshold is fixed, the RMS model error is increased more than the S-CIELAB error. Hence, the RMS threshold is less consistent than S-CIELAB across images.

5.2. JPEG distortions

None of the models made satisfactory predictions of the marked errors in the JPEG-DCT reproductions. Figure 6 shows the probability a location is marked as a function

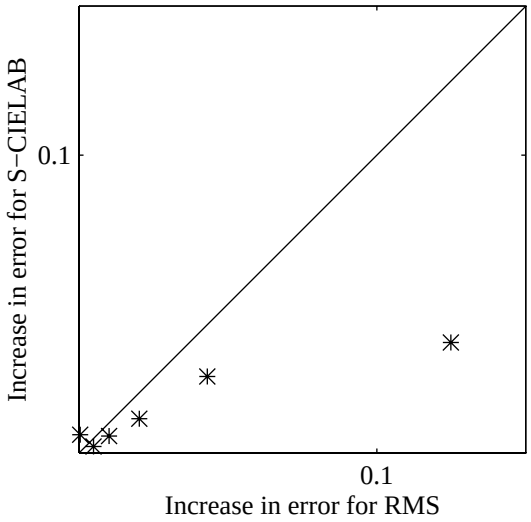

Figure 5: *The increased error caused by fixing the threshold parameter for the halftone images. Each point represents the increase in $\mathcal{N}$ for a single image as measured by the RMS model (horizontal axis) and S-CIELAB model (vertical axis). Fixing the threshold increases the RMS error more than the S-CIELAB error.*

of RMS, CIELAB, and S-CIELAB error measures. The subjects did not mark high predicted error regions much more frequently than lower predicted error regions. Hence, we have performed no further analyses on the models.

Given the nature of JPEG artifacts and the characteristics of the three color difference metrics, it is not surprising that none of them performed well in predicting visibility of JPEG-DCT distortions. The JPEG-DCT artifacts arise from (1) the coarse quantization of high frequency components, and (2) the block processing structure of the algorithm. In the case of quantization, the errors are typically correlated with lines or edges in the images, and therefore hidden by the effect of orientation selective masking and contrast masking (Legge & Foley, 1980; Losada & Mullen, 1994; Limb, 1979; Chaddha & Meng, 1993). The image distortions metrics evaluated here do not include effects of contrast masking or orientation selective masking. Hence, these metrics should not be expected to make accurate predictions about visibility of JPEG artifacts.

6. Conclusions

Subjects identified visible reproduction errors in a collection of halftone and JPEG-DCT reproductions. The responses were summarized as image distortion maps. Using these maps, we evaluated three image distortion metrics: RMS, CIELAB, and S-CIELAB. The metrics all performed reasonably well in predicting relative size of vis-

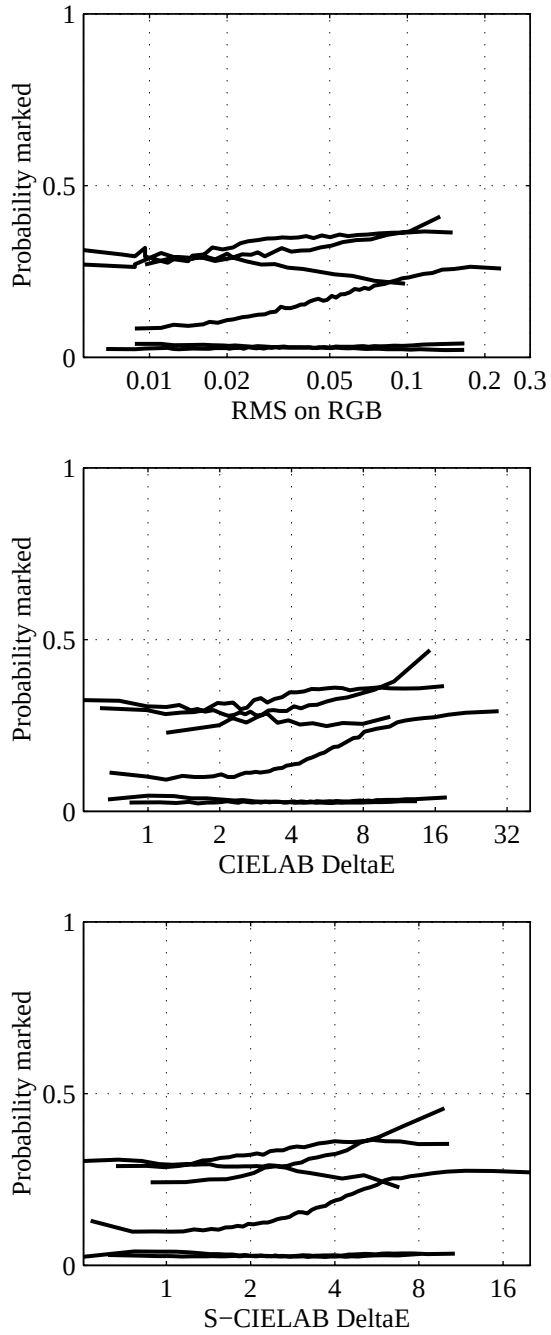

Figure 6: *Probability of marking a location as a function of predicted error for JPEG-DCT reproductions. The panels show the relationship for the RMS, CIELAB and S-CIELAB metrics. The different lines within each panel represent data for different images.*

ible halftone noise in individual images. The S-CIELAB metric made the most consistent predictions across images, and RMS was the least consistent. The CIELAB metric was designed to be used on large uniform targets only, therefore it made most mistakes at high spatial frequency regions of images. The S-CIELAB metric did much better at the high frequency regions of images due to its addition of spatial-color sensitivity mechanisms. The RMS metric was calculated on RGB frame buffer values, which is not a perceptually meaningful color space. It also does not incorporate spatial sensitivity in the calculation. Therefore, it is surprising that RMS did not fail completely in predicting halftone errors. This remains to be understood.

All three metrics failed to predict the image distortion maps measured with JPEG-DCT reproductions. We suspect this is due to the lack of contrast masking and orientation selective masking in these metrics.

7. References

1. D. H. Brainard, *Spatial Vision* **10**, 433-436 (1997).
2. N. Chaddha and T. H. Meng, *SPIE Proc. on Visual Communications and Image Processing* **Nov. 8-11** (1993).
3. International Commission on Illumination (CIE), *Suppl. No.2 to CIE Pub. No.15 (E.-1.3.1)* **TC-1.3** (1971).
4. G. Legge and J. Foley, *J. Opt. Soc. Am.* **70**, 1458-1471 (1980).
5. J. O. Limb, *IEEE Trans. on Systems, Man, and Cybernetics* **SMC-9(12)**, 778-793 (1979).
6. M. Losada and K. Mullen, *Vis. Res.* **34(3)**, 331-341 (1994).
7. X. Zhang and B. A. Wandell, *SID Symposium Technical Digest* **27**, 731-734 (1996).