

Maximum likelihood difference scaling of image quality in compression-degraded images

Christophe Charrier,¹ Laurence T. Maloney,² Hocine Cherifi,³ and Kenneth Knoblauch^{4,5,*}

¹Université de Caen Basse Normandie, LUSAC EA 2807, Groupe Vision & Analyse d'Images, 120 rue de l'Exode, 50000 Saint Lô, France

²Department of Psychology, Center for Neural Science, New York University, 6 Washington Place, 8th Floor, New York, New York 10003, USA

³Université de Bourgogne, 9 avenue Alain Savary, BP 47870, 21078 Dijon, France

⁴INSERM, U846, Stem Cell and Brain Research Institute, Département Neurosciences Intégratives, 18 avenue du Doyen Lépine, 69500 Bron, France

⁵Université de Lyon, UMR-S 864, Lyon 1, 69003, Lyon, France

*Corresponding author: knoblauch@lyon.inserm.fr

Received August 20, 2007; accepted August 20, 2007;
posted August 24, 2007 (Doc. ID 86698); published October 1, 2007

Lossy image compression techniques allow arbitrarily high compression rates but at the price of poor image quality. We applied maximum likelihood difference scaling to evaluate image quality of nine images, each compressed via vector quantization to ten different levels, within two different color spaces, RGB and CIE 1976 L*a*b*. In L*a*b* space, images could be compressed on average by 32% more than in RGB space, with little additional loss in quality. Further compression led to marked perceptual changes. Our approach permits a rapid, direct measurement of the consequences of image compression for human observers. © 2007 Optical Society of America

OCIS codes: 100.0110, 333.0330, 330.1690, 330.4060, 330.5020, 330.5510.

1. INTRODUCTION

Vector quantization (VQ) methods [1] allow arbitrary compression of digital images. They are widely employed in encoding video [2,3]. These methods share a common structure. The image is divided into pixel blocks, and these blocks are approximated by a smaller set of images drawn from a fixed code book. The number of bits needed to specify the original code block is replaced by the typically smaller number of bits needed to specify its corresponding code. The compression rate is the ratio of the file size of the uncompressed image over the size for the compressed image, denoted γ . The compression rate is varied by decreasing the size of the codebook but typically at the cost of increasing the discrepancy between a given block and the code that replaces it [1,4].

In this article, we describe a method, maximum likelihood difference scaling (MLDS), for estimating supra-threshold differences across a range of images. We apply it to the problem of assessing the perceptual effects of image compression with VQ methods. The MLDS method, described below, is based on simple, forced-choice judgments and requires remarkably few trials to obtain quantitative estimates of the effects of any degree of image compression.

In Fig. 1, we illustrate the effects of different rates, γ , of VQ compression on a sample image. There is an evident trade-off between compression rate and perceived image quality. An optimal compression method would maximize compression while minimizing subjective perceptual distortion.

MLDS [5] provides a method for quantifying super-

threshold perceptual differences between pairs of images in Fig. 1. On each trial the observer saw four images, such as depicted in Fig. 2, taken from the series in Fig. 1. In the example, the top left image is compressed by a factor of 6, and the image to its right by a factor of 15; the bottom left by a factor of 18 and the image to its right by a factor of 27. The right image in the upper pair is compressed 2.5 more than the left, while the right image in the lower pair is only 1.5 times more compressed than the corresponding left image. The observer is asked to compare the upper pair of images with the lower and to judge whether the perceptual change is greater in the upper pair or in the lower. The difference scale is based on 210 judgments of the kind illustrated in Fig. 2, with different choices of quadruples of images on each trial. In Section 2 we explain in detail how we fit the resulting data and derive difference scales, but first we describe how to interpret the difference scale based on the observer's judgments.

An example of a difference scale taken from our experiment is shown in Fig. 3. The horizontal scale marks degree of compression, and the vertical scale, labeled Difference Scale Value, is based on the observer's judgments. The curve is J-shaped, with a shallow plateau from 0% to 15% followed by a steep climb. The image compression algorithm has little effect on the difference scale values when $\gamma \leq 12$; the difference scale is flat or nearly so. Above that point, however, small changes in γ result in progressively larger increases in the scaled differences between images. So long as $\gamma \leq 12$, the benefits of image compression come with very little change in perceived image qual-

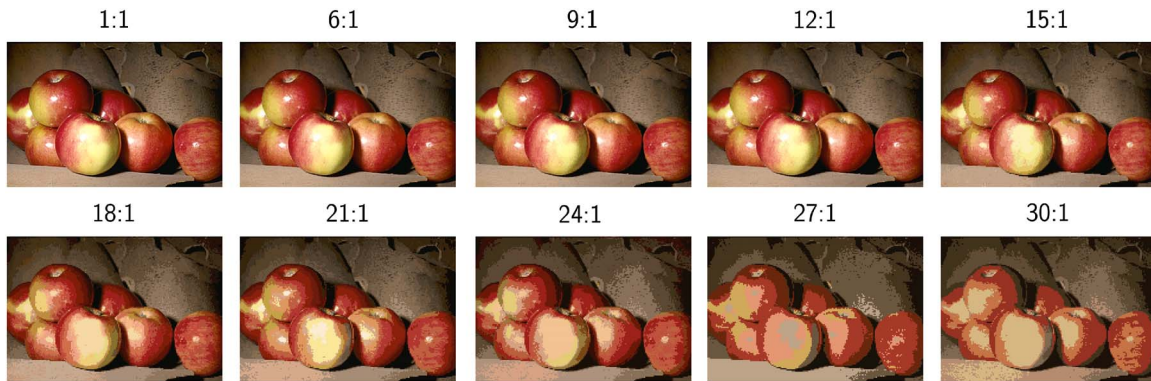


Fig. 1. (Color online) Effects of VQ compression. The original image (0% compression) is shown after VQ compression using a codebook based on the LBG algorithm applied to the image encoded in $L^*a^*b^*$ color space (see text). Larger compressions lead to evident decreases in image quality.

ity. Of course, the results of the difference scaling do not tell us whether the observer prefers the uncompressed image ($\gamma=1$) to the most compressed image ($\gamma=30$). It does tell us that up to a compression rate of 12, the observer sees little change in the images and that above this point, he sees marked change.

In Subsection 1.A we describe MLDS. While we focus on VQ compression methods in this article, the methods and analyses we present are readily applicable to evaluating image quality for any compression scheme or method that leads to progressive distortion of images. In particular, it is well adapted to situations in which the range of compression rates is high and the loss of quality is severe. MLDS has been applied in other domains, including perception of surface gloss [6] and face discrimination [7].

We evaluate two applications of VQ that differ in choice of color space. The color spaces considered are RGB and CIE 1976 $L^*a^*b^*$ [8], pp. 166–169. A previous analysis of JPEG-compressed images found that images encoded and compressed in $L^*a^*b^*$ space exhibited less degradation of image quality than the same images encoded and compressed to the same degree in RGB space [9]. By use of MLDS we can explicitly quantify the effect of choice of color space in applications of VQ.



Fig. 2. (Color online) An example of a single difference scaling trial. On each trial, the observer sees two pairs of images and judges which pair (upper or lower) has the greater perceived difference. The upper pair corresponds to images compressed by factors of 6 (left) and 15 (right), the lower pair to images compressed by factors of 18 (left) and 27 (right).

A. Maximum Likelihood Difference Scaling

In this subsection we develop the model of the observer's judgments in the psychophysical task on which MLDS is based. In each experimental condition, the experimenter selected a particular base image I_1 and compression method. We describe the experimental conditions in detail in the Methods section below. The experimenter then selects N compression levels $\gamma_1 < \gamma_2 < \dots < \gamma_N$, where $\gamma_1=1$. He or she then prepares N images I_1, \dots, I_N , where I_j for $j > 1$ is I_1 compressed by a factor of γ_j (See Fig. 1).

On each trial the experimenter presented an observer with quadruples $(I_a, I_b; I_c, I_d)$ and asked him to judge which pair, I_a, I_b or I_c, I_d , exhibited the larger perceptual difference. It will prove convenient to replace $(I_a, I_b; I_c, I_d)$ with the simpler notation $(a, b; c, d)$. Over the course of the experiment, the observer saw many different quadruples, a subset of nonoverlapping quadruples. We used the set of all possible nonoverlapping quadruples $a < b < c < d$ for N stimuli, but this choice is not critical to the

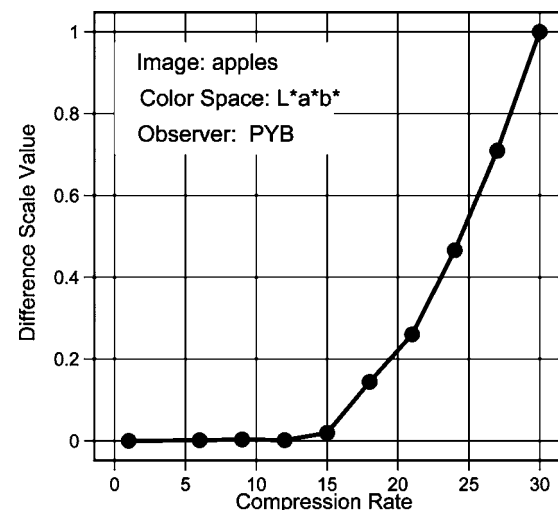


Fig. 3. Example of a difference scale. On the horizontal scale we plot degree of image compression γ , and on the vertical we plot difference scale values derived from a psychophysical procedure, MLDS [5]. The difference scale values are estimates based on the observer's judgments of superthreshold perceptual differences between the images portrayed in Fig. 1. See text. Compression rates up to a factor of 12–15 result in little perceived difference. Above a factor of 15, the difference scale values increase markedly with increased compression rate.

method [5]. By restricting the set of quadruples in this way, we avoided the possibility that two of the images presented to the subject would be identical.

In our experiments, the number of distinct compression levels was always $N=10$ and except for the first (uncompressed) level, equally spaced on the compression scale. The observer completed $P=210$ trials in each condition, permitting a judgment for each nonoverlapping quadruple once. On half of the forced-choice trials, chosen at random, the pairs were presented in the order $(a, b; c, d)$ and on the other half, $(c, d; a, b)$.

The experimenter estimated scale values $\psi_1, \psi_2, \dots, \psi_N$ corresponding to the stimuli, I_1, \dots, I_N , as follows. Given a quadruple, $(a, b; c, d)$, on a single trial, one might assume that the observer would consistently judge I_a, I_b to be farther apart than I_c, I_d precisely when

$$|\psi_b - \psi_a| > |\psi_d - \psi_c|; \quad (1)$$

that is, the difference scale values predict judgment of perceptual difference, and human judgments of these differences never vary from trial to trial.

However, it is unlikely that human observers would be so reliable in judgment as to satisfy the criterion just given, particularly if the differences $|\psi_b - \psi_a|$ and $|\psi_d - \psi_c|$ were close. Maloney and Yang[5] proposed a model of difference judgment that allowed the observer to exhibit some stochastic variation in judgment. Let $L_{ab} = |\psi_b - \psi_a|$ denote the unsigned length of the interval (I_a, I_b) . The proposed decision model is an equal-variance Gaussian signal detection model [10], where the signal is the difference in the lengths of the intervals,

$$\delta(a, b; c, d) = L_{cd} - L_{ab} = |\psi_d - \psi_c| - |\psi_b - \psi_a|. \quad (2)$$

We assume that if δ is positive, the observer chooses the second interval as larger, and when it is not positive, the first. When the magnitude of δ is small relative to the Gaussian standard deviation, σ , we expect the observer, presented with the same stimuli, to give different, apparently inconsistent judgments. To summarize, the decision variable employed by the observer is assumed to be

$$\Delta(a, b; c, d) = \delta(a, b; c, d) + \epsilon = L_{cd} - L_{ab} + \epsilon, \quad (3)$$

where ϵ is a Gaussian random variable with mean zero and standard deviation $\sigma > 0$; given the quadruple $(a, b; c, d)$, the observer selects the pair I_c, I_d precisely when

$$\Delta(a, b; c, d) > 0. \quad (4)$$

In each experimental condition the observer completes P trials, each based on a quadruple $\mathbf{q}^k = (a^k, b^k; c^k, d^k)$, $k = 1, P$. The observer's response is coded as $R^k = 0$ (the difference of the first pair is judged larger) or $R^k = 1$ (second pair judged larger). We fitted the parameters $\Psi = (\psi_1, \psi_2, \dots, \psi_N)$ and σ by maximizing the likelihood,

$$L(\Psi, \sigma) = \prod_{k=1}^P \Phi\left(\frac{\delta(\mathbf{q}^k)}{\sigma}\right)^{1-R^k} \left(1 - \Phi\left(\frac{\delta(\mathbf{q}^k)}{\sigma}\right)\right)^{R^k}, \quad (5)$$

where $\Phi(x)$ denotes the cumulative distribution function of the Gaussian with mean 0 and variance 1 and $\delta(\mathbf{q}^k) = \Delta(a^k, b^k; c^k, d^k)$ was defined in Eq. (3). The details of the

fitting procedure are described by Maloney and Yang [5].

At first glance, it would appear that the stochastic difference scaling model just presented has $N+1$ free parameters: ψ_1, \dots, ψ_N together with the standard deviation of the error term, σ . However, any linear transformation of the ψ_1, \dots, ψ_N together with a corresponding scaling results in a set of parameters that predicts exactly the same performance as the original parameters. Without any loss of generality, we can set $\psi_1 = 0$ and $\psi_N = 1$, leaving us with the $N-1$ free parameters, $\psi_2, \dots, \psi_{N-1}$ and σ . When scale values are normalized in this way, we describe them as standard scales. We fitted parameter values by direct numerical optimization as described below.

2. METHODS

A. Observers

Two observers participated in the experiment. One was a coauthor, and the other was unaware of the purpose of the experiment. The data from each observer were not analyzed until all experimental conditions had been completed. Both had normal color vision (as assessed using Ishihara Plates) and normal acuity (20/20 on a Snellen Chart).

B. Apparatus

Color images were displayed on a SUN CRT display driven by a GC14/SX graphics card with a spatial resolution of 1152 by 900 pixels and a color depth of 24 bits. A gamma correction table was used to display color images with precise control of the luminance, color, and contrast. The gamma correction table was determined using a photometer as described previously [11]. On each trial the observer saw a 2×2 array of images (similar to those in Fig. 2) displayed in a $15 \text{ cm} \times 15 \text{ cm}$ area centered on the display screen. The viewing distance was 50 cm, and the display subtended 17.1 deg of visual angle.

C. Conditions

The nine images used are shown in uncompressed form in Fig. 4. These images include a variety of subject matter and differ in distribution of spatial and chromatic detail.

D. Compression Methods

We used two compression methods that differed only in the choice of color encoding (color space). The color spaces were RGB and $L^*a^*b^*$. The RGB color space is simply the gamma-corrected intensities (gun excitations) of the CRT monitor used. The CIE 1976 $L^*a^*b^*$ color space was an attempt to develop a uniform color space by nonlinear transformations of the CIE 1931 xyY color space. For a fixed luminance L^* , MacAdam ellipses transformed to this space are approximately circles of equal radius, but discrepancies remain [8], pp. 166–169. For our purposes, $L^*a^*b^*$ represents a color representation method that was constructed to capture important aspects of human color discrimination.

Each set of nine images was VQ-based compressed to ten different levels ranging from $\gamma=1$ (uncompressed) to $\gamma=30$ (compression by a factor of 30) and within each of the color spaces.

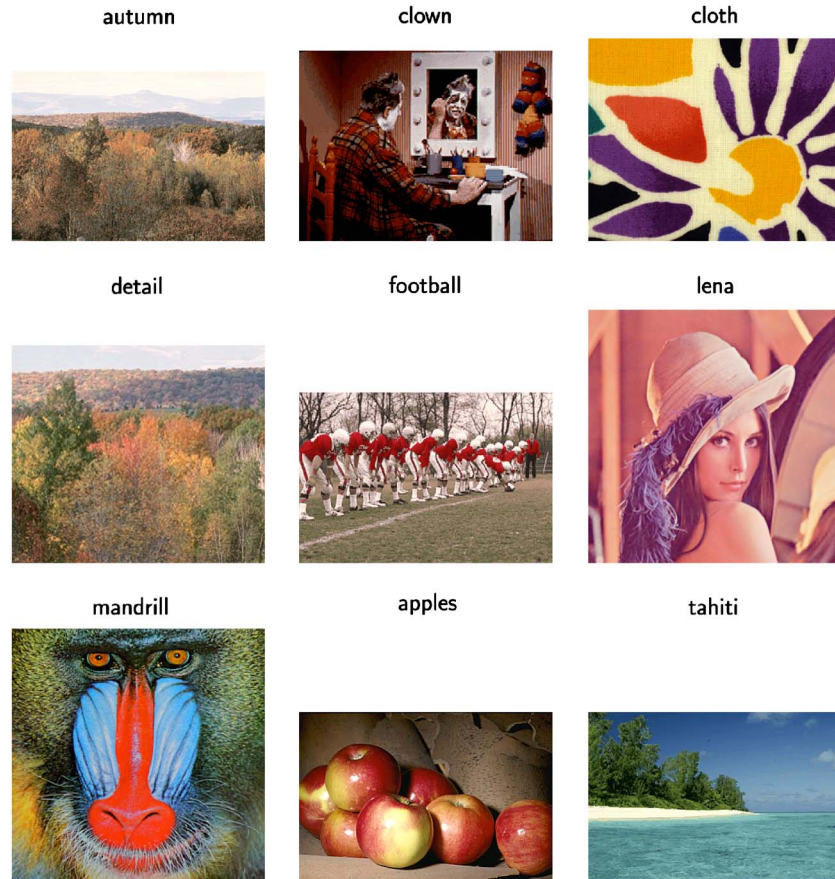


Fig. 4. (Color online) Images. The nine images used in the experiments are shown, with mnemonic labels. For each image and each color space, RGB and $L^*a^*b^*$, we estimated a difference scale based on each observer's judgments.

E. Procedure

On each trial the observer saw four images corresponding to a quadruple $(a, b; c, d)$ arranged in a 2×2 array as in Fig. 2. The pair a, b appeared on the top or bottom with equal probability. For each trial, the observer reported his response to the question, "Which pair of images (top or bottom) shows the greater perceptual difference?" The observer completed 210 trials in each of the $18 = 2 \times 9$ conditions of the experiment. Neither observer reported any difficulty in carrying out the task. Each session of 210 trials required about 15 minutes to complete.

3. ANALYSIS AND RESULTS

We fitted the MLDS model to each observer's data for each image in each color space condition, using a numerical optimization method to maximize likelihood as defined in Eq. (5), employing multiple starting points to minimize the possibility of encountering local minima. All computations were carried out in the statistical language R [12] using the optim function. In Appendix A we show that optimization of Eq. (5) can be recast as a generalized linear model (GLM) fit with an inverse Gaussian (probit) linking function [13]. We have integrated the functions necessary to perform these fits using either approach in an R package (MLDS) available from the Comprehensive R Archive Network (CRAN, accessible from <http://www.r-project.org/>) or from the corresponding author.

We first plot the difference scales for each image and both observers (Fig. 5). The error bars shown (± 1 SD) were estimated by a Bootstrap procedure [14] as described by Maloney and Yang [5], p. 577. In brief, given an observer's fitted probability of response for any experimental condition, we repeatedly simulated the observer's performance on the trials in that condition and fitted the simulated data just as we fitted the observer's original data. The error bars correspond to the standard deviations at each compression level of 10,000 repetitions of this process for an observer in each experimental condition. Each plot contains the difference scale for RGB [dashed (red online)] and for $L^*a^*b^*$ (solid, black). The difference scales for images compressed in the $L^*a^*b^*$ space are J-shaped, consisting of a roughly linear plateau followed by a region of roughly linear increase ("the cliff"). To estimate where plateau and cliff meet, we fitted a four-parameter model consisting of two lines with different slopes which we refer to as a J-function. The J-function model has slope a_1 from 0 to a breakpoint B ("the plateau"), slope a_2 from B to the end of the scale (the cliff) and intercept parameter a_3 :

$$\begin{aligned} \psi(\gamma) &= a_1\gamma + a_3, & \gamma \leq B, \\ &= a_2(\gamma - B) + a_1B + a_3, & \gamma > B. \end{aligned} \quad (6)$$

The four fitted parameters of the J-functions are a_1, a_2, a_3, B . B is an estimate of the end of the plateau, the

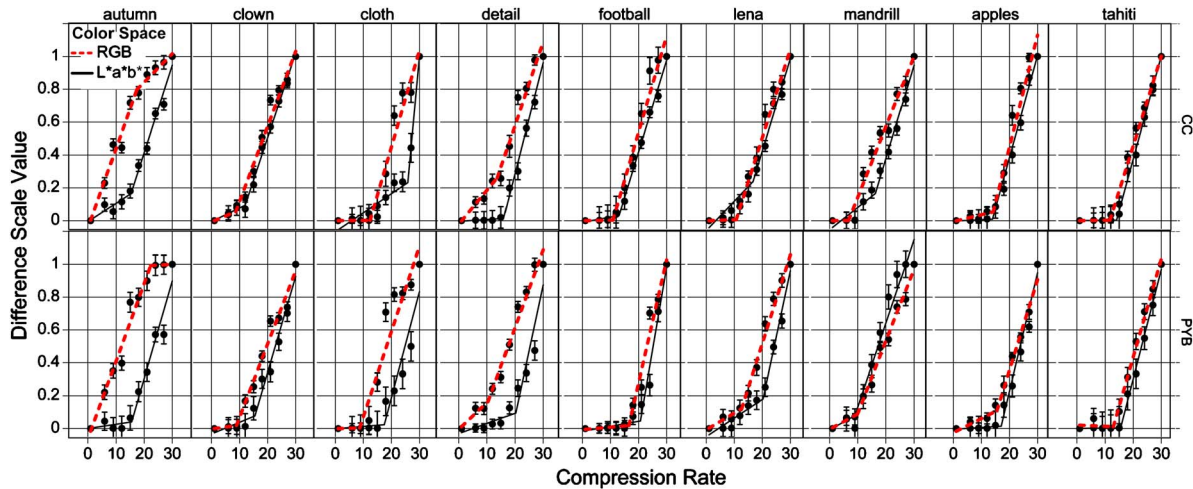


Fig. 5. (Color online) Difference scales for each image and observer. The difference scales for each image and observer are shown. The image labels correspond to those of Fig. 4. Observers’ initials are indicated in the right-hand labels for each row of panels. In each plot we show the scale corresponding to the VQ compression based on RGB color space [dashed, red (online)] and separately the scale values based on $L^*a^*b^*$ color space (solid, black) The confidence intervals shown (± 1 SD) were estimated by a Bootstrap procedure [14] as described by Maloney and Yang [5]. The fitted lines are J-functions, defined in the text.

point where any further compression leads to a change in slope of the difference scale (“the cliff”). These values are tabulated for each observer separately in Table 1. The fitted J-functions for the $L^*a^*b^*$ data are shown as solid lines in Fig. 6. For the RGB data, we repeated the analysis, and the fitted J-functions are plotted in Fig. 6 as dashed lines and the fitted parameters displayed in Table 2. Except for image “autumn” (Fig. 4) in RGB space, the values of a_1 are consistently smaller than the corresponding values for a_2 , justifying the labels “plateau” and “cliff.” In the case of “RGB/autumn,” the trend is reversed with an initial steep segment followed by a shallow segment. In effect, the initial plateau is absent and the breakpoint should be at 0, not at the high-value fit to the data that indicates a different phenomenon, which is the leveling off of quality loss at high compression rates for

this condition. We discuss the results for this image in more detail below.

We report the ratio a_2/a_1 as a measure of the change between plateau and cliff. The value $\beta = a_1B + a_3$, the height of the estimated difference scale at the break point, is reported as well. The values of the ratio a_2/a_1 , except in the two cases noted above, are consistently large, indicating that the cliff region is indeed steeper than the plateau for all images and observers. The value of β is an index of how much of the difference scale (0–1) is taken up by the plateau. Overall the J-functions for RGB space were above those for $L^*a^*b^*$ space, indicating that the perceptual differences in RGB were consistently greater in the early part of the compression range than those for $L^*a^*b^*$. We tested the hypothesis that $a_1=0$ (that the plateau is flat) for both color spaces across all observ-

Table 1. Estimated Parameters for J-Functions Fitted to Perceptual Scales for Images in $L^*a^*b^*$ Color Space

Observer	Image	a_1	a_2	a_3	B	a_2/a_1	β
CC	autumn	0.011	0.053	-0.328	15.3	4.7	0.165
PYB	autumn	0.003	0.059	-0.440	15.5	22.8	0.041
CC	clown	0.008	0.051	-0.264	12.0	6.2	0.094
PYB	clown	0.007	0.058	-0.430	15.5	8.2	0.077
CC	cloth	0.012	0.185	-2.313	25.9	16.0	0.232
PYB	cloth	0.001	0.065	-0.553	17.5	45.2	0.024
CC	detail	0.001	0.067	-0.532	15.9	61.2	0.014
PYB	detail	0.006	0.080	-0.777	20.2	12.9	0.095
CC	football	0.003	0.056	-0.353	13.1	17.5	0.032
PYB	football	0.003	0.100	-1.021	20.7	33.0	0.046
CC	lena	0.018	0.056	-0.381	16.7	3.1	0.238
PYB	lena	0.011	0.080	-0.745	20.3	7.1	0.180
CC	mandrill	0.014	0.057	-0.409	16.3	4.2	0.166
PYB	mandrill	0.010	0.051	-0.194	9.0	5.3	0.079
CC	apples	0.001	0.062	-0.429	13.9	77.2	0.009
PYB	apples	0.001	0.072	-0.609	17.0	65.8	0.014
CC	tahiti	0.000	0.063	-0.439	14.0	443.3	0.002
PYB	tahiti	0.000	0.067	-0.513	15.5	414.0	0.003

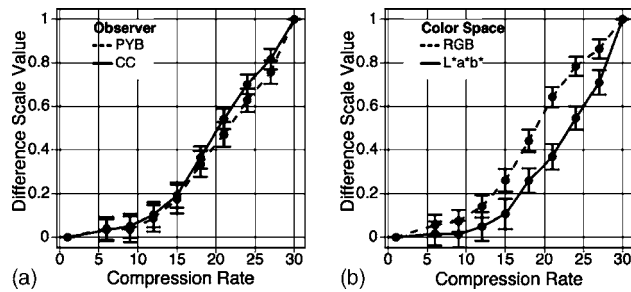


Fig. 6. Summary results. (a) Difference scales in Fig. 5 averaged across all images and both color spaces, separately for each of the two observers. Average observer results are in good agreement. (b) Difference scales in Fig. 5 averaged across all images and each observer separately for each color space.

ers and images by comparing mean squared errors and could not reject it [$F(40, 216) = 1.336$; $p = 0.100$].

We compared the breakpoint values in Fig. 5 across the two color spaces pairing the difference scale values for each image and observer. The difference was statistically significant ($t_{17} = 5.0721$; $p < 0.0001$, two-tailed), and the sign of the difference indicated the breakpoints to be at higher compression rates for L*a*b* space. Elimination of the image “autumn” from the above analyses did not affect the conclusions drawn from any of the tests reported above.

A. Color Space Comparisons

In Fig. 6 we plot the data of Fig. 5 averaged across observers [Fig. 6(a)] and across color spaces [Fig. 6(b)]. There is evident agreement between the two observers in both Fig. 5 and Fig. 6(a), and, as just noted, the fitted J-functions for both observers had no plateau for the same image, autumn, and color space in Fig. 6. In Fig. 6(b), we see that compression in L*a*b* space results in a longer and shallower plateau.

4. SUMMARY AND DISCUSSION

We applied a psychophysical method, maximum likelihood difference scaling (MLDS) [5] to evaluate the percep-

tual changes in images with increasing compression rates, γ , ranging from 1 (no compression) to 30. On each trial observers saw two pairs of images (as in Fig. 2). We denoted the two pairs as a quadruple $(a, b; c, d)$ with pairs (a, b) and (c, d) . The observer was instructed to select the pair that had the greater perceived difference. Over the course of the experiment, the observer often saw pairs that were obviously different. That is, image a was evidently different from image b , and image c was evidently different from image d . However, the observer’s task was not to discriminate the two images in each pair but to order the perceived magnitude of superthreshold perceptual differences. The task underlying MLDS then is not discrimination of images but rather direct comparison of superthreshold differences. MLDS differs from image quality methods based on just noticeable differences (e.g., [9, 15]) in the judgment required of the observer.

We fitted the data by maximum likelihood methods to derive a difference scale. The use of maximum likelihood methods is desirable since it permits the experimenter to test nested hypotheses formulated in terms of simple parametric models. In Appendix A we show that MLDS can be treated as an example of the GLM [13], and this connection opens up possible analyses using standard GLM packages.

The fitted scale values summarize the relative magnitudes of superthreshold perceptual differences. The comparison of superthreshold differences is well suited to measuring the perceptual differences introduced by VQ compression. We found that the fitted difference scales for both color spaces that we considered had a characteristic shape for both observers and nearly all images that we referred to as a J-function. It consisted of a shallow plateau followed by a steep climb (“cliff”), and the extent of plateau region measured the degree of compression that could be tolerated with little or no perceptual difference. We could not reject the hypothesis that the plateaus had zero slope.

The images compressed in L*a*b* could be compressed roughly 12- to 15-fold with little or no perceived change.

Table 2. Estimated Parameters for J-Functions Fitted to Perceptual Scales for Images in RGB Color Space

Observer	Image	a_1	a_2	a_3	B	a_2/a_1	β
CC	autumn	0.049	0.017	0.225	17.4	0.4	0.798
PYB	autumn	0.047	0.001	0.453	22.3	0.0	0.992
CC	clown	0.007	0.046	-0.173	8.6	6.4	0.055
PYB	clown	0.002	0.045	-0.190	8.9	18.9	0.019
CC	cloth	0.000	0.059	-0.363	12.4	416.2	0.002
PYB	cloth	0.000	0.052	-0.224	8.7	109.5	0.004
CC	detail	0.021	0.051	-0.230	13.7	2.5	0.262
PYB	detail	0.016	0.047	-0.166	10.5	2.9	0.165
CC	football	0.000	0.059	-0.317	10.9	182.6	0.003
PYB	football	0.002	0.075	-0.614	16.5	49.8	0.018
CC	lena	0.001	0.054	-0.285	10.7	88.9	0.006
PYB	lena	0.010	0.054	-0.284	12.6	5.3	0.121
CC	mandrill	0.000	0.043	-0.144	6.7	171.2	0.001
PYB	mandrill	0.009	0.044	-0.176	9.9	4.7	0.089
CC	apples	0.004	0.068	-0.457	14.2	16.5	0.053
PYB	apples	0.009	0.055	-0.385	15.5	6.4	0.109
CC	tahiti	0.000	0.055	-0.318	11.7	224.1	0.003
PYB	tahiti	-0.001	0.060	-0.368	12.9	-79.4	0.012

In contrast, the images compressed in RGB color space had smaller plateaus, and for one image, “autumn” (Fig. 4), no detectable plateau. In effect, for this one image and color space, any substantial degree of compression led to evident perceptual change. The image “autumn” is a natural scene in which luminance and chromatic differences occur at a fine scale. It is possible that RGB encoding (which does not separate luminance from chromatic information) is particularly susceptible to compression-related image quality losses for this type of image. We note that the image labeled “detail” is a magnified subregion of autumn, and its scale shows a similar trend. The fitted equation does generate two slopes, but the first segment of the RGB curves is only slightly less steep than the second segment.

The color space $L^*a^*b^*$ is an example of an approximately uniform color space, designed so that equally discriminable lights are represented by points that are roughly equally far apart in Euclidean distance [8], pp. 166–169. The LBG algorithm that we used in selecting a code book for VQ compression uses Euclidean distance in color space to evaluate the match between pixel blocks and their encodings. It is therefore not completely surprising that an encoding based on a color space designed to represent more faithfully chromatic differences would exhibit less perceptual distortion. We emphasize that the MLDS results presented here clearly capture and quantify the differences in image quality that are due to choice of color space for every image and for both observers.

There is no *a priori* reason to believe that a metric, such as $L^*a^*b^*$, based on small color differences between simple stimuli would generalize to a situation concerning superthreshold color differences in complex images. However, the fact that it corresponds to a more nearly orthogonal coding of spectral information in an image is likely to be of importance in permitting an optimal compression rate [16].

The MLDS method proposed by Maloney and Yang [5] makes use of direct comparison of perceptual differences of image pairs. Observers can readily make such direct comparisons, and, by avoiding the use of image quality rating scales, the MLDS method avoids known problems associated with how humans use rating scales [17–21]. Shepard, in particular, argued that absolute judgments are highly variable while ratio or difference judgments are less so. It is plausible that image quality ratings, based on absolute judgments, would be less reliable than difference judgments (or difference of difference judgments).

It is possible to estimate difference scales given numerical ratings of the magnitude of each pair in a quadruple [22]. However, to do so, the experimenter must develop a model of how the rater/observer generates a numerical rating. In the form proposed by Maloney and Yang, the observer’s task is simply a series of forced-choice judgments that can be modeled as Bernoulli random variables. Maloney and Yang [5] also showed that difference scaling is remarkably robust to failures of distributional assumptions.

One last point is that the MLDS scales estimated here were based on only 210 forced-choice judgments per scale. Each difference scale was estimated from data collected

in about 15 minutes. The ease of collection of significant amounts of data is another advantage of MLDS.

APPENDIX A: MAXIMUM LIKELIHOOD DIFFERENCE SCALING AS A GENERALIZED LINEAR MODEL

The maximum likelihood estimation in Eq. (5) can be rewritten as a GLM [13]. Since GLM packages are widely available, the reader may find it useful to fit difference scaling models in this way.

We first describe the GLM and then show in detail how to recast difference scaling as a GLM. As in the main text, there are N images I_1, \dots, I_N that are obtained by compressing a base image I_1 by factors $\gamma_1 < \gamma_2 < \dots < \gamma_N$, respectively, with $\gamma_1 = 1$. The observer is asked to look at pairs of images, I_a, I_b and I_c, I_d and select the pair that exhibits the greater perceptual difference. Each judgment corresponds to a quadruple $(a, b; c, d)$, and, over the course of the experiment, the observer judges a subset of all possible quadruples, possibly with repetitions. The difference scaling model assumes that the observer’s responses are based on differences between subjective scale values $\psi_1 \leq \psi_2 \leq \dots \leq \psi_N$: the observer judges the quadruple $(a, b; c, d)$ by forming the difference

$$\delta = |\psi_d - \psi_c| - |\psi_b - \psi_a|,$$

$$\Delta = \delta + \epsilon, \quad (\text{A1})$$

where ϵ is Gaussian with mean 0. In the main text we set the standard deviation of ϵ to be σ , another free parameter describing the subject. Here we will set the standard deviation to be 1, as this choice will prove convenient in formulating difference scaling as a GLM.

In the main text we noted that we could add a constant to all of the $\psi_1 \leq \psi_2 \leq \dots \leq \psi_N$ without changing the observer’s judgments, and so we can set $\psi_1 = 0$ without loss of generality. We also noted that we could scale all of the ψ values by a common positive constant as long as we also scaled σ . As a consequence we could set $\psi_N = 1$. Since we have fixed σ to be 1 here, we will not do this; ψ_N will be estimated from the data, and its maximum likelihood estimate will prove to be equal to that of σ^{-1} .

We order the entries in the quadruples such that we can omit the absolute value signs in Eq. (2), and it becomes

$$\delta = \psi_d - \psi_c - \psi_b + \psi_a,$$

$$\Delta = \delta + \epsilon, \quad (\text{A2})$$

The observer bases his or her judgment on $\Delta = \delta + \epsilon$, where ϵ is Gaussian with mean 0 and standard deviation 1. The observer therefore selects the second pair (c, d) with probability $\Phi(\delta)$.

Let $\tilde{\Psi} = (\psi_2, \dots, \psi_N)^T$, the column vector of the ψ values with $\psi_1 = 0$ omitted. We define a design matrix \mathbf{M} with $N - 1$ columns and one row for every quadruple $(a, b; c, d)$. If none of the values a, b, c, d is 1, we set the a, b, c, d th entry in the row to $+1, -1, -1, +1$, respectively [these are the coefficients of the corresponding entries in Eq. (A2)]. All of the remaining entries are set to 0. If any of a, b, c, d is 1,

we ignore it. For example, we show a few quadruples and the corresponding rows of the design matrix when $N=10$. Note that the design matrix has only nine columns and that the first column corresponds to ψ_2 , not $\psi_1=0$:

$$\begin{pmatrix} 1 & 3; & 5 & 7 \\ 7 & 9; & 4 & 5 \\ 1 & 6; & 7 & 8 \\ 3 & 4; & 9 & 10 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

In many statistical packages, such as R, it may be easier to create a design matrix with N columns, the first one corresponding to ψ_1 , and then to strip this column from the design matrix.

Then we can write in matrix form

$$\boldsymbol{\delta} = \mathbf{M}\tilde{\boldsymbol{\Psi}}, \quad (\text{A3})$$

where $\boldsymbol{\delta}$ is a column vector of differences δ , one for each trial. As in the main text, let $\Phi(x)$ denote the cumulative distribution function of a Gaussian random variable with mean 0 and standard deviation 1. Then we can write

$$\mathbf{p} = \Phi(\boldsymbol{\delta}) \quad (\text{A4})$$

as a column vector of probabilities of selecting the second interval on each trial. If we denote the responses of the observer as a column vector \mathbf{R} of 0's and 1's where 1 denotes second interval chosen, then

$$P[R = 1] = \Phi(\mathbf{M}\tilde{\boldsymbol{\Psi}}), \quad (\text{A5})$$

or, in terms of the expected values of the binary responses,

$$E[R] = \Phi(\mathbf{M}\tilde{\boldsymbol{\Psi}}), \quad (\text{A6})$$

which we rewrite as

$$\Phi^{-1}(E[R]) = \boldsymbol{\delta} = \mathbf{M}\tilde{\boldsymbol{\Psi}}. \quad (\text{A7})$$

Equation (A7) is in the form of a GLM [13]. In the present case, the responses of the observer can be modeled as Bernoulli random variables. The expected value of the response, δ , is related to the linear predictors through a nonlinear function, $\eta(\cdot)$, that is the inverse cumulative distribution function of the Gaussian. Equation (A7) is a form of probit analysis, a special case of the GLM. The GLM estimates are maximum likelihood estimates.

Thus we may use GLM to estimate the maximum likelihood estimates $\tilde{\boldsymbol{\Psi}} = (\hat{\psi}_2, \dots, \hat{\psi}_N)$, and, together with $\psi_1 = 0$, we have maximum likelihood estimates of the scale values. These form a difference scale where $\sigma=1$ by assumption, and $\hat{\psi}_N$ is not normalized to 1. As noted above, $\hat{\psi}_N = \hat{\sigma}^{-1}$, and as a last step we can normalize the scale by dividing $(\psi_1, \hat{\psi}_2, \dots, \hat{\psi}_N)$ by $\hat{\psi}_N$ and setting $\hat{\sigma} = 1/\hat{\psi}_N$. The justification for these last steps is the invariance of maximum likelihood estimation under reparameterization [23], pp. 284–286.

Since the estimates in the main text were also maximum likelihood estimates of these same parameters, the two sets of estimates should agree within numerical error. We compared solutions using direct optimization [optim() in R] and GLM fits [glm() function in R] and found good

agreement. Of course, the GLM package is simply optimizing likelihood by numerical methods, and where the two methods disagree, one or the other (or both!) of the methods must have found a local maximum that is not the global maximum.

In the statistical and computing environment R, there is a choice between five built-in link functions for the binomial family, including the logit, probit, and cauchit (based on the Cauchy distribution). As of R version 2.4.0, it has become simple for the user to define additional links. In many circumstances, the choice of link is not critical, since over the rising part of these functions, they are quite similar. The difference scaling procedure, however, generates many responses at the tails, i.e., easily discriminable differences, and may be more sensitive to the choice of link.

ACKNOWLEDGMENTS

This research was funded in part by grant EY08266 from the National Institute of Health (to L. T. Maloney). The experiments were performed at the Université Jean Monnet in the Laboratoire d'Informatique Graphique et d'Ingénierie de la Vision. We made extensive use of the ggplot package in R [24] and thank Hadley Wickham for his helpful responses to all of our questions. In addition, we thank Vincent Lozano for aid in image conversion. Part of the work described here was presented at the European Conference on Visual Perception, 1998 [25].

REFERENCES

1. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression* (Kluwer Academic, 1991).
2. I. E. G. Richardson, *H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia* (Wiley, UK, 2003).
3. M. Flierl and B. Girod, *Video Coding with Superimposed Motion-Compensated Signals: Applications to H.264 and Beyond* (Kluwer Academic, 2004).
4. R. M. Gray, "Vector quantization," *IEEE ASSP Mag.* **1**, 4–29 (1984).
5. L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *J. Vision* **3**, 573–585 (2003). <http://www.journalofvision.org/3/8/5>.
6. G. Obein, K. Knoblauch, and F. Viénot, "Difference scaling of gloss: non-linearity, binocularity, and constancy," *J. Vision* **4**, 711–720 (2004). <http://journalofvision.org/4/9/4>.
7. G. Rhodes, L. T. Maloney, J. Turner, and L. Ewing, "Adaptive face coding and discrimination around the average face," *Vision Res.* **47**, 974–989 (2007).
8. G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulas*, 2nd ed. (Wiley, 1982).
9. N. Mulroney and M. D. Fairchild, "Color space selection for JPEG image compression," in *Proceedings of 1st IS&T/SID Color Imaging Conference* (Society for Imaging Science and Technology, 1993), pp. 157–159.
10. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1974).
11. C. Charrier, K. Knoblauch, and H. Cherif, "Perceptual distortion analysis of color image based coding," *Proc. SPIE* **3005**, 134–143 (1997).
12. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2007). ISBN 3-900051-07-0. <http://www.R-project.org>.

13. P. McCullagh and J. A. Nelder, *Generalized Linear Models* (Chapman & Hall, 1989).
14. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).
15. A. B. Watson and L. Kreslake, "Measurement of visual impairment scales for digital video," *Proc. SPIE* **4299**, 79–89 (2001).
16. G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," *Proc. R. Soc. London, Ser. B* **220**, 89–113 (1983).
17. D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement* (Academic, 1971), pp. 140–141.
18. D. H. Krantz, "A theory of context effects based on cross-context matching," *J. Math. Psychol.* **5**, 1–48 (1968).
19. D. H. Krantz, "A theory of magnitude estimation and cross-modality matching," *J. Math. Psychol.* **9**, 168–199 (1972).
20. R. N. Shepard, "On the status of 'direct' psychophysical measurement," in *Minnesota Studies in the Philosophy of Science*, Vol. 9, C. W. Savage, ed. (University of Minnesota Press, 1978), pp. 441–490.
21. R. N. Shepard, "Psychological relations and psychophysical scales: on the status of 'direct' psychophysical measurement," *J. Math. Psychol.* **24**, 21–57 (1981).
22. M. C. Boschman, "DifScal: a tool for analyzing difference ratings on an ordinal category scale," *Behav. Res. Methods Instrum. Comput.* **33**, 10–20 (2001).
23. A. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics*, 3rd ed. (McGraw-Hill, 1974).
24. H. Wickham, *ggplot: An Implementation of the Grammar of Graphics in R* (2006). R package version 0.4.0. <http://had.co.nz/ggplot>.
25. K. Knoblauch, C. Charrier, H. Cherifi, J. N. Yang, and L. T. Maloney, "Difference scaling of image quality in compression-degraded images," *Perception* **27**, S174 (1998).