Perceptual image quality: Effects of tone characteristics

Peter B. Delahunt
Exponent Inc.
149 Commonwealth Drive
Menlo Park, California 94025

Xuemei Zhang
Agilent Technologies Laboratories
3500 Deer Creek Road, MS 26M-3
Palo Alto, California 94304

David H. Brainard
University of Pennsylvania
Department of Psychology
3401 Walnut Street, Suite 302C
Philadelphia, Pennsylvania 19104

Abstract. *Tone mapping refers to the conversion of luminance values recorded by a digital camera or other acquisition device, to the luminance levels available from an output device, such as a monitor or a printer. Tone mapping can improve the appearance of rendered images. Although there are a variety of algorithms available, there is little information about the image tone characteristics that produce pleasing images. We devised an experiment where preferences for images with different tone characteristics were measured. The results indicate that there is a systematic relation between image tone characteristics and perceptual image quality for images containing faces. For these images, a mean face luminance level of 46–49 CIELAB L* units and a luminance standard deviation (taken over the whole image) of 18 CIELAB L* units produced the best renderings. This information is relevant for the design of tone-mapping algorithms, particularly as many images taken by digital camera users include faces.* © 2005 SPIE and IS&T. [DOI: 10.1117/1.1900134]

1 Introduction

Consumers of digital cameras and related products desire high-quality images. Consumer preference for images, however, is not easy to predict. Even if it were technically feasible, creating a perfect reproduction of the light that arrived at the camera would not guarantee the most preferred rendering of the original scene. For example most professional portraiture employs a large degree of image enhancement, and the results are almost always preferred to a veridical rendering. This may occur because most consumers judge the attractiveness of an image without direct reference to the original scene, so that their judgments are based on memory, either of the specific scene or of generic scenes. There is evidence that memory for colored objects can be unreliable.[1–3]

Paper 03007 received Jan. 21, 2003; revised manuscript received Aug. 7, 2003; accepted for publication Aug. 17, 2004; published online May 12, 2005.
1017-9909/2005/$22.00 © 2005 SPIE and IS&T.

Digital images may be modified through the application of image processing algorithms, but what modifications make images look better is not well understood. One approach to this problem is to study directly the effect of image processing on image preference. We recently examined the perceptual performance of demosaicing algorithms in this manner.[4] Previous work has also studied the relation between image colorfulness and human observer quality/naturalness ratings.[5–7] Here we apply similar experimental methods to study the relation between image tone characteristics and perceptual image quality.

Tone mapping refers to the conversion of input luminance values, as captured by an acquisition device (e.g., a digital camera), to luminance values for display on an output device (e.g., a computer monitor). Luminance values in a natural image can range over about five orders of magnitude.[8] This compares to a much smaller range of about two orders of magnitude available with a computer monitor under typical viewing conditions. Even for the usual situation where the image acquisition device quantizes the number of luminance levels to match the number of levels available on the output device, tone mapping can still improve the appearance of an image. The relation between input and output luminance values produced by a tone-mapping algorithm is called a *tone-mapping curve*.

Tone mapping changes the *tone characteristics* of the image. By tone characteristics we mean the distribution of the luminance values of the image's pixels, without regard to how the pixels are arranged spatially. In general, tone characteristics can either be assessed globally (over the entire image), or locally (over some smaller region of interest). Within an image region (either global or local), tone characteristics are completely described by the *luminance histogram* of the region. This specifies the number of image

pixels within the region that have each possible output luminance value. In this paper, we will consider both global and local tone characteristics.

Previous work on tone mapping has focused on comparisons of the performance of different tone-mapping methods. Much of this work was conducted in the context of film-based photography, where practical considerations limited attention to global tone-mapping methods in which a single tone-mapping curve was applied to the entire image (see review by Nelson[9]). Bartleson and Breneman[10] suggested that a good tone-mapping curve established a 1:1 relation between relative perceived brightness values in the scene and the rendered image, where relative brightnesses were computed using a modified power function derived from research on brightness scaling.[11] Their curve corresponded closely to curves that received high ratings in a psychophysical study performed by Clark.[12] Further work by Hunt and co-workers[13,14] suggested that the Bartleson and Breneman principle[10] should be modified depending on the viewing conditions (in particular the surround of the image) and suggested that although a linear relation between scene and image relative brightnesses was appropriate for reflection prints, a power-law relation between relative brightnesses was more appropriate for transparencies. The widely used zone system for photographic tone mapping (reviewed in Reinhard *et al.*[15]) relies on perceptual judgments of how regions in the original scene appeared to the photographer.

In film photography, it is not practical to automatically adjust the tone-mapping curve between images at separate locations within an image, since the shape of these curves is governed by physical characteristics of the emulsions and film-development process. With the advent of digital imaging, a wider range of tone-mapping algorithms become of practical interest. On the other hand, in many digital cameras image quantization precedes the application of a tone-mapping algorithm, a feature that increases the challenges for successful tone mapping. Thus there has been renewed interest in developing tone-mapping algorithms (see, e.g., Refs. 8 and 16–18). Evaluation of these methods has again emphasized comparing the output of competing algorithms. A recent study by Drago *et al.*,[19] for example, applied seven tone-mapping techniques to four digital images and their performance was rank ordered based on observer preferences.

Algorithms that apply a fixed tone-mapping curve to any image have the feature that the tone characteristics of the images produced by the algorithm can vary widely, since these characteristics depend strongly on the input. Digital imaging presents the opportunity to develop algorithms using a different principle. Rather than defining the relationship between input and output luminances, one can specify target output tone characteristics and apply an image-dependent transformation that yields a good approximation of these characteristics. One early digital tone-mapping algorithm, histogram equalization, is based on this idea: the algorithm maps the luminance values in the input image to produce a desired luminance histogram in the output image. Although it seems unlikely that the optimal output histogram is completely independent of image content, the principle of specifying target output image tone characteristics has been incorporated into recent tone-mapping algorithms intended to improve upon histogram equalization. In these algorithms, the output histogram varies with an analysis of image content.[8,16]

The work we present here is intended to further explore the idea that effective tone mapping can be achieved through specification of desired output image tone characteristics. Rather than focusing on the development and evaluation of tone-mapping algorithms, we chose to address the underlying issue of whether we could identify output tone characteristics that produce perceptually attractive images, and whether such characteristics depend on image content. To this end, we report the results of two image preference studies and analyze how image preference is related to image tone characteristics.

The work presented here employs images captured with standard digital cameras and is directed at improving the quality of images produced from such cameras. We do not explicitly consider the case where the dynamic range of the capture and display devices varies greatly (see Refs. 8, 17, 18, and 20).

As most amateur digital photographs include people, our studies employ an image set that consisted mainly of images of people. We also wanted to include images of people from different ethnic backgrounds, since many earlier tone-mapping studies used images of Caucasians only (e.g. Refs. 12, 21, and 22).

2 Experiment 1

2.1 *Overview*

Experiment 1 was exploratory, with the goal of identifying systematic relationships between tone variables and image quality. We applied four different tone-mapping methods to each of 25 experimental images and measured the perceptual quality of the different renderings of each image. These algorithms produced output images with a range of tone characteristics. Image preference was measured using a pairwise comparison procedure. On each trial, observers indicated which of two presented images was the most attractive.

The pairwise comparison procedure is intuitive for observers and yields reliable data.[4] Note, however, that observers only make judgments about different renderings of the same input image. Thus some analysis is required to aggregate a data set large enough to explore the question of how an image's tone characteristics relate to its perceptual quality. To this end, the preference choice data were analyzed using a regression procedure[23] to yield metric differences in image quality between image pairs. The procedure yields difference ratings that are commensurate across input images. We then asked whether differences in specific image *tone variables* were predictive of the difference ratings. Here the term tone variable refers to a summary measure, such as mean luminance, that may be computed from the output luminance histogram.

We used 25 digitally acquired images and rendered each on a CRT computer monitor using four different tone-mapping methods. The four methods produced results that were perceptually different for most of the images, thus providing variation in image tone characteristics whose effect we could study.

2.2 *Methods: Image Acquisition*

Twenty-five images were used in Experiment 1. Twenty-one were captured in Santa Barbara, California and four were taken in Palo Alto, California. All of the images were captured under daylight, at different times of the day, throughout May 1999. The illuminant was measured immediately following the acquisition of each image by placing a white reflectance standard in the scene and measuring the reflected light using a Photo Research PR-650 spectraradiometer. Of the 25 images, 17 were portraits of people, 5 were landscapes, and 3 were of objects.

The 21 Santa Barbara images were taken with a Kodak DCS-200 camera and the 4 Palo Alto images with a Kodak DCS-420 camera. Both cameras have a resolution of 1524 $\times$1012 with RGB sensors arranged in a Bayer mosaic.[24] The DCS-200 captures the input light intensity using 8-bit linear quantization, whereas the DCS-420 captures with 12-bit precision. The 12-bit values captured by the DCS-420 are converted to 8-bit values on-camera via a nonlinear transformation. The relative RGB spectral sensitivities and response properties of both cameras were characterized as described elsewhere.[25] This characterization left one free parameter describing the overall sensitivity of the camera undetermined, as this parameter varies with acquisition exposure duration and f-stop. The images were cropped to a maximum size of 575 (w) by 800 (h) pixels to ensure that two renderings of each image could be displayed simultaneously on the computer screen used in our experiment.

2.3 *Image Processing*

2.3.1 *Dark level subtraction*

For the DCS-200, a dark level was subtracted from the raw quantized pixel values before further processing. The dark level was estimated from an image acquired with the lens cap on and computing the spatial average of the resulting image. The average for the red, green, and blue sensors were all 13.5 on the camera's 8-bit (0–255) output scale and this is the value that was subtracted. To estimate the dynamic range of the images, we compared the minimum and maximum pixel values for the green sensor. These typically occupied the entire allowable output range (approximately 13–255 before dark subtraction). Given that some pixels had values near zero after dark subtraction, it is not possible to express the dynamic range of these images as a meaningful ratio.

For the DCS-420, it was possible to linearize the output values using a look-up table provided as part of each raw image file. This was done prior to further processing. After linearization, the estimated dark level for the DCS-420 was close to zero and no explicit dark level subtraction was performed. The dynamic range of these images could be estimated by taking the ratio of the maximum to minimum linearized output value for the green sensor. These ratios varied from 17 to 140 across the DCS-420 images used in this experiment.

2.3.2 *Demosaicing*

Because the two cameras employed a mosaiced design, with each raw pixel corresponding to only one of the three sensor types, it was necessary to apply a demosiacing algorithm to convert the raw mosaiced image to a full color RGB image. We used a Bayesian demosaicing algorithm developed by Brainard and colleague[26–28] and summarized in a recent paper[4] where we evaluated the perceptual quality of demosaicing algorithms. (The performance of the Bayesian algorithm is controlled by a number of parameters. For the application here, the correlation between nearest-neighbor pixels was assumed to be 0.90, whereas the correlation between the responses of different sensor classes at the same image location was estimated from a bilinear interpolation of the mosaiced image. Finally, the algorithm assumed that there was additive normally distributed pixel noise with a standard deviation for each sensor class equal to 4% of the spatial average of responses for that class. The estimates at each location were obtained by applying the algorithm to a 5$\times$5 image region surrounding that pixel.) The demosaicing results for our images were in general quite good, with very few noticeable artifacts.

2.3.3 *Color balancing*

We by-passed the on-board color balancing of the cameras and used our measurements of the scene illuminants to color balance the images. Given the camera's RGB sensor relative spectral sensitivities and the measured illuminant, we were able to estimate the relative surface spectral reflectance of the object at each scene location. This was done using a Bayesian estimation procedure that will be described in a future report. Briefly, we constructed a normally distributed multivariate prior distribution for object surface reflectances by analyzing the Vrehl *et al.*[29] data set of measured surface reflectance functions. The analysis followed closely the method introduced by Brainard and Freeman[30] in their work on computational color constancy. Given the prior, estimating reflectances from the sensor responses is a straightforward application of Bayes rule. Using the estimated surface reflectance functions, we could then synthesize an image that consisted of the CIE XYZ tristimulus coordinates that would have been obtained had the surface been viewed under standard CIE daylight D65, up to an overall scale factor. This scale factor varied from image to image depending on the scene illuminant, acquisition exposure, and acquisition f-stop. Uncertainty about the scale factor is equivalent to uncertainty about the overall intensity of the scene illuminant and is thus handled transparently by the tone-mapping algorithms that we applied to render the images, which are designed to apply to images captured over a wide range of overall scene luminances. Note that image L* properties reported in this paper refer to L* values for the experimental images displayed on the experimental monitor, not to L* properties of regions of the original scene.

To check the accuracy of the color balancing process, an image of a Macbeth color checker was taken using the Kodak DCS-420 digital camera. Raw RGB values (before demosaicing) were extracted for each of the 24 color checker patches. The Bayes color correction was used to estimate the XYZ values of the patches under CIE D65 illumination. These estimates were compared with target values computed from measured spectral reflectances of the color checker patches and the known spectral relative spectral power distribution of CIE daylight D65. Here the free overall scale factor was determined so that the two middle gray color checker patches (patch Nos. 21 and 22) matched

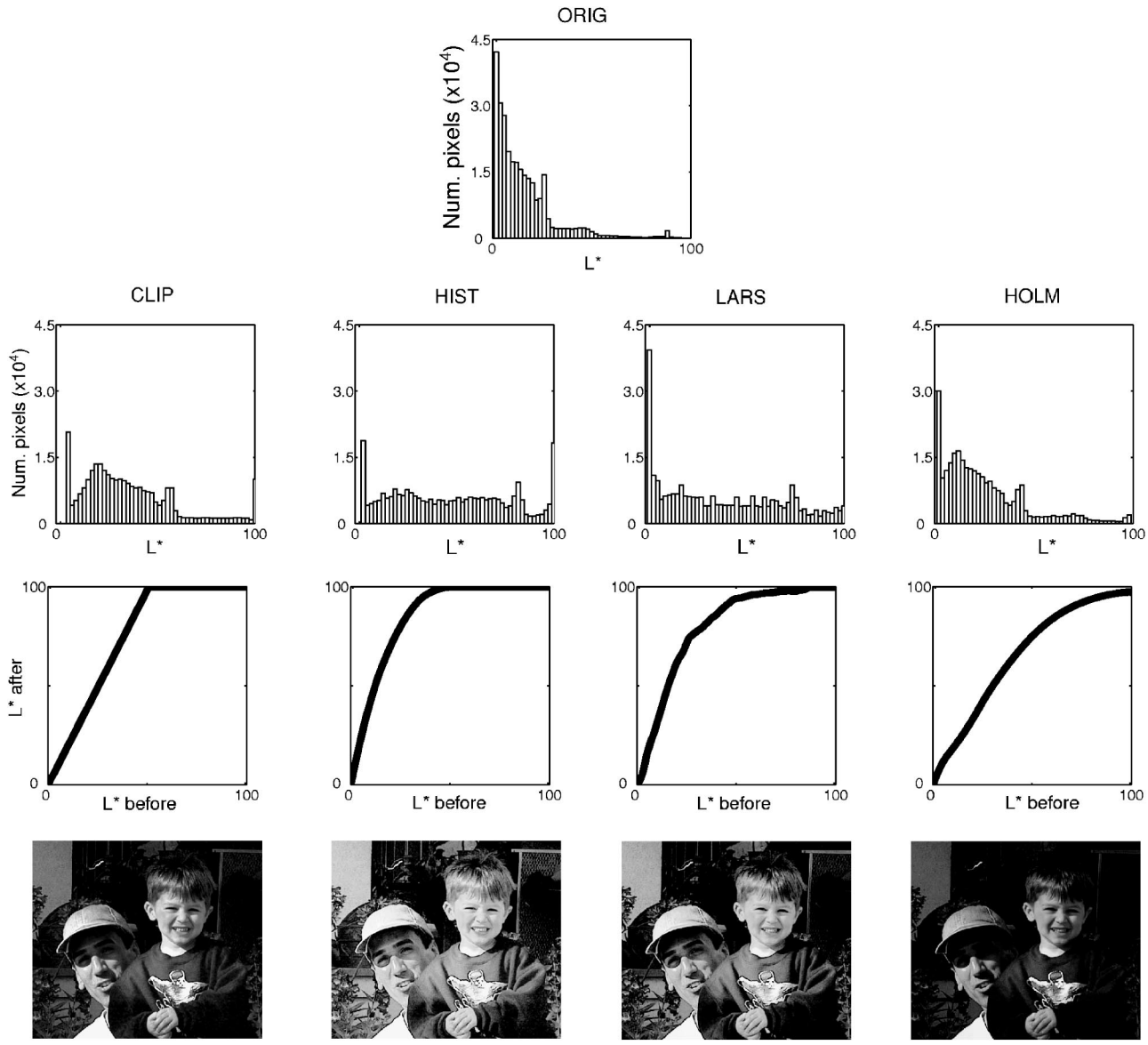

Fig. 1 Tone mapping. The top panel in the figure shows the global L* luminance histogram of the original image. The four panels in the first full row show the histograms after application of the four tone-mapping algorithms. The four panels in the middle row show the tone-mapping curves used by the four algorithms for the image shown. The bottom panels show the output images for each of the four algorithms.

in average luminance between the color balanced and target values. The average CIELAB ΔE 94 difference between the estimated values and directly determined target values (average taken over the 24 patches) was 3.6 units, indicating that the algorithm worked well.

2.3.4 *Tone mapping*

Four tone-mapping algorithms (*Clipping*, *Histogram Equalization*, *Larson's Method*, and *Holm's Method*) were applied to the color balanced XYZ images. These are described below. Each method transformed the luminance of each image pixel while holding the chromaticity of each pixel constant. The relation between a particular measurement of input and output luminance is referred to as the algorithm's *tone-mapping curve*. In general, the tone-mapping curve produced by an algorithm depends on image content. Each of the algorithms used was global, in the sense that the same tone-mapping curve was applied to every pixel in the image.

It should be emphasized that our main goal was to use a variety of tone-mapping algorithms that would produce different tone-mapping characteristics. The performance of each algorithm was not of primary concern. All four methods led to acceptable (as judged by the authors) renderings for all of the images. In Fig. 1 we show an example of the histograms, tone-mapping curves, and output images produced by the four algorithms.

(1) *Clipping*: For the clipping method, the tone-mapping curve relating image luminance to display luminance was a straight line through the origin. Image lumi-

nances that were mapped to display luminances greater than the maximum available on the output device were clipped to the maximum. The slope of the tone-mapping curve was determined so that maximum display luminance was equal to five times the mean luminance of the tone-mapped image. This clipping method provides a simple baseline that works reasonably well.

(2) *Histogram equalization*: A widely used method that re-assigns luminance values to achieve a particular target luminance histogram (e.g., uniform or Gaussian) in the tone-mapped image.[31] This method efficiently uses the dynamic range of the display device, but can generate images that have exaggerated contrast and thus a harsh appearance. In our implementation, the target histogram was a Gaussian centered at the middle of the output range.

(3) *Larson method*: A more sophisticated version of histogram equalization. The idea is to limit the magnitude of luminance mapping, so that luminance differences within the image that were not visible before tone mapping are not made visible by it. Images tone mapped with the Larson method generally have a more natural appearance than when using the traditional histogram equalization method.

(4) *Holm's method* (Ref. 16): Part of a color reproduction pipeline created at Hewlett-Packard Labs for use in digital cameras. We used only the tone-mapping segment of the pipeline for consistency with the other methods. In Holm's method, the input image is first classified as one of several different types (e.g., high key or low key) using a set of image statistics. A tone-mapping curve is then generated according to the image type and image statistics, and this curve is applied to the whole image. This method incorporates preference guidelines that came from the inventor's extensive experience in photographic imaging.

2.3.5 *Rendering for display*

The images were presented on a CRT monitor. Conversion between the tone-mapped XYZ values and monitor settings was achieved using the general model of monitor performance and calibration procedures described by Brainard.[32] The calibration was performed using the PR-650 spectraradiometer. Spectral measurements were made at 4 nm increments between 380 and 780 nm but interpolated with a cubic spline to the CIE recommended wavelength sampling of 5 nm increments between 380 and 780 nm. CIE XYZ coordinates were computed with respect to the CIE 1931 color matching functions.

2.3.6 *Room and display setup*

The experimental room was set up according to the International Organization for Standardization Recommendations for Viewing Conditions for Graphic Technology and Photography.[33] The walls were made of a medium gray material and the table on which the monitor was placed was covered with black cloth. The room was lit by two fluorescent ceiling lights (3500 K) controlled by a dimmer switch set at a dim level. The illumination measured at the observer position was 41 lux. The experiment was controlled by MATLAB software based on the Psychophysics Toolbox.[34,35] The images were displayed on a Hewlett Packard P1100 21 in. monitor (1280×1024 pixels) driven by a Hewlett Packard Kayak XU computer.

2.3.7 *Procedure*

On each trial of the experiment, observers were shown pairs of the same scene rendered via different tone-mapping methods and were asked to choose the image that they found to be the most attractive. To further explain this instruction, observers were asked to choose the image they would select to put into their own photo album. The observers were also asked to look around the images before making a decision rather than focus on just one aspect.

The experiment started after a 2 min adaptation period. Three seconds after each pair of images was presented, two selection boxes appeared under the images. This 3 s delay was to encourage the observers to carefully consider their decision. There was no upper limit on response time. The observers indicated their preference by using a mouse to move a cursor to the selection box under the preferred image and clicking. The observer could subsequently change his/her mind by clicking on the alternative box. When the observer was satisfied with his/her selection, he/she clicked on an enter button to move to the next trial.

The images were viewed from a distance of 60 cm. The images ranged in width from 17 to 19 cm (subtending visual angles 16.1° to 18.0°) and ranged in height from 13 to 25 cm (subtending visual angles from 12.4° to 23.5°). Images were shown in pairs on the monitor, one on the left and one on the right. Each image had a border of width 1 cm which was rendered as the brightest simulated D65 illuminant the monitor could produce (78 cd/m^2). The remaining area of the monitor emitted simulated D65 illuminant but at a luminance level of about 20% of the border region (measured at 14.9 cd/m^2).

Using four rendering methods gives six pairwise presentation combinations per image. For the 25 experimental images, this produces a stimulus set of 150 image pairs.

2.3.8 *Observers*

Twenty observers participated in the experiment (12 males and 8 females) with an average age of 31 (range 19–62). The experiment took place at Hewlett Packard Labs in Palo Alto and the observers were recruited by posting flyers around the building complex. The observers were a mixture of Hewlett Packard employees and outside friends and family. Only color normal observers participated. Color vision was tested using the Ishihara color plates.[36]

2.3.9 *Data analysis*

The aim of our analysis was to summarize image tone characteristics using simple tone variables, and to determine whether these variables predicted image preference. We hoped to identify systematic relationships between preference ratings and tone variables. Thus our data analysis has two important components: the procedure used to transform the pairwise image judgments to image preference ratings and the procedures used to extract variables that capture image tone-mapping characteristics.

2.4 *Image Ratings*

The raw data consisted of pairwise rankings between the four different renderings of each image. For each image, we used a regression based scaling method[23] to convert the pairwise rankings to *preference ratings* for each of the four

versions. Denote these ratings as π_j^i where the superscript i denotes the image ($1 \leqslant i \leqslant 25$) and the subscript j denotes the rendering version ($1 \leqslant j \leqslant 4$, algorithms as numbered above). Within image, these ratings for the four different versions of an image are directly comparable. Since no preference judgments were made across images, however, the ratings across images are not necessarily commensurate.

Although we cannot make comparisons of preference ratings across images, we can make such comparisons of differences in preference ratings. Under assumptions that we found reasonable,[23] the four ratings generated for each image lie on an interval scale. The unit of this scale corresponds to one standard deviation of Gaussian perceptual noise that observers are assumed to experience when making preference judgments, and the unit is thus common to the ratings generated for all 25 images. What differs across images is the origin of the scale, which is assigned arbitrarily by the regression method. To remove the effect of origin, we can compute *difference ratings* between the j'th and k'th renderings, $\pi_{jk}^i = \pi_j^i - \pi_k^i$ ($1 \leqslant j,k \leqslant 4$). Because the rating scale constructed for each image has a common unit, the difference ratings are commensurate across images. Thus we can explore whether there are image tone characteristics whose differences predict difference ratings.

From the four renderings for each image, we can take six pairwise differences. Only three of these are independent, however, in the sense that given any three pairwise differences the other three may be reconstructed. To avoid this redundancy, we used only the difference ratings π_{12}^i, π_{23}^i, and π_{34}^i ($1 \leqslant i \leqslant 25$) in the analysis.

2.5 *Image Tone Characteristics*

To describe image tone characteristics, we used the L* coordinate of the CIELAB uniform color space.[37] This measure of luminance is normalized to a white point, and the normalized values are transformed so that equal differences in L* represent approximately equal differences in the perception of brightness. The maximum monitor output (all three phosphors set at the maximum) was used as the white point for converting image luminance to L*. We considered two summary measures of the L* histogram: the mean L* value and the standard deviation of the L* values. For each image i, we denote the mean L* of the j'th rendering value by μ_j^i and the standard deviation of the L* values by σ_j^i. These are both global tone variables, computed from the entire image. Note that μ_j^i is in essence a measure of the overall luminance of the image, whereas σ_j^i is in essence a measure of image contrast.

A preliminary analysis indicated that to the extent image quality ratings depended on the tone characteristics μ_j^i and σ_j^i, this dependence was not monotonic. This observation makes intuitive sense. Consider the mean L* value μ_j^i. An image with a μ_j^i value equal to zero will be entirely black and not provide a satisfactory rendering. Similarly, an image with a very large μ_j^i value will be entirely white. Clearly a rendering with a μ_j^i value between the two extremes is indicated. Similar arguments apply to σ_j^i.

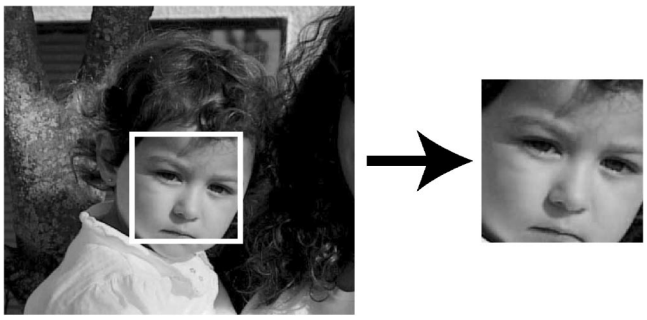

Fig. 2 Face subimages were created by cropping the faces out of the images as illustrated.

To account for a possible nonmonotonicity of the relation between image quality and the tone characteristics μ_j^i and σ_j^i, we considered transforms of these variables:

$$\tilde{\mu}_j^i = |\mu_j^i - \mu_0|, \qquad \tilde{\sigma}_j^i = |\sigma_j^i - \sigma_0|. \tag{1}$$

Here the parameter μ_0 represents the *optimal value* for μ_j^i, that is the value that leads to the highest image quality across all images and renderings, and thus deviations of μ_j^i from μ_0 should correlate with reduced image quality. Similarly, the parameter σ_0 is the optimal value for σ_j^i.

2.6 *Analysis*

As noted previously, our data set does not provide us with direct access to image quality, but rather to quality difference ratings π_{jk}^i between pairs of images. To ask whether image tone characteristics predict image quality, we investigated whether differences between the tone variables $\tilde{\mu}_j^i$ and $\tilde{\sigma}_j^i$ predict the difference ratings π_{jk}^i. Specifically, we defined the tone variables differences $\tilde{\mu}_{jk}^i = \tilde{\mu}_j^i - \tilde{\mu}_k^i$ and $\tilde{\sigma}_{jk}^i = \tilde{\sigma}_j^i - \tilde{\sigma}_k^i$ and examined the linear dependence of π_{jk}^i on each of these differences. Since each transformed variable depends on its corresponding optimal value, numerical parameter search over the optimal value was used to maximize the predictive value (R^2) of $\tilde{\mu}_{jk}^i$ and $\tilde{\sigma}_{jk}^i$.

2.7 *Face Images*

In follow-up questioning conducted at the end of the experiment, many observers commented that for images containing people, the appearance of faces was an important factor in their decision making. For images containing faces (17 of 25) we examined the face regions in more detail and defined face subimages so the tone characteristics of these regions could be extracted. The subimages were defined by hand: an example of how a face subimage was defined is shown in Fig. 2. (One image had two faces; only the foreground face was used for this analysis.) The faces were of various ethnicities (8 Caucasian, 4 African-American, 3 Asian, 1 Hispanic, and 1 Polynesian).

For the images containing faces, we repeated our analysis of difference ratings when the tone characteristics depended only on the pixels in the face subimage. We denote

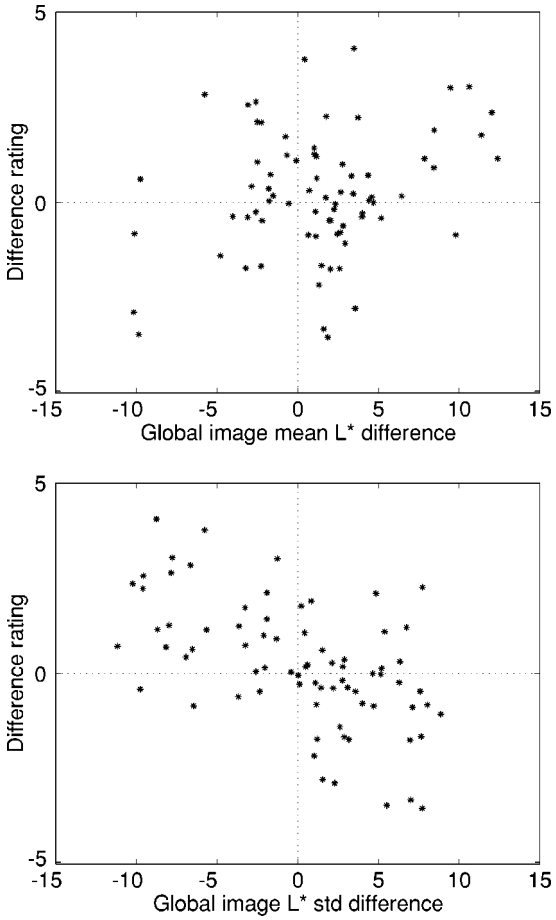

Fig. 3 Prediction of difference ratings from global tone characteristics. The figure plots the difference ratings obtained for all images in Experiment 1 against $\tilde{u}^i_{jk}$ (top panel) and $\tilde{\sigma}^i_{jk}$ (bottom panel).

these difference ratings by $\tilde{\mu}^i_{face\_jk}$ and $\tilde{\sigma}^i_{face\_jk}$. Note that these are local tone variables, in that they depend only on a subregion of the entire image.

3 Results

Figure 3 shows the difference ratings π^i_{jk} plotted against tone characteristic differences $\tilde{\mu}^i_{jk}$ (top panel) and $\tilde{\sigma}^i_{jk}$ (bottom panel) for our entire data set. From the figure, we can see that any systematic dependence of difference ratings on $\tilde{\mu}^i_{jk}$ is weak at best, but that there is a clear dependence of the difference ratings on $\tilde{\sigma}^i_{jk}$. Note that the negative slope of the dependence shown in the bottom panel of Fig. 3 makes sense: if a rendering j is preferred to image k (positive difference rating π^i_{jk}), then the deviation of image j's L* standard deviation from its optimal value is smaller than the corresponding deviation for image k (negative $\tilde{\sigma}^i_{jk}$). These conclusions are confirmed by statistical tests on the significance of the linear relation between the π^i_{jk} and each independent variable. The R^2 value for $\tilde{\mu}^i_{jk}$ is small (0.07) but significant ($p<0.05$), whereas $\tilde{\sigma}^i_{jk}$ explains a substantial fraction of the variance ($R^2=0.31$, $p<0.001$). The optimal value found for μ_0 was 46.6, whereas that found for σ_0 was 17.8.

The predictive value of $\tilde{\mu}^i_{jk}$ and $\tilde{\sigma}^i_{jk}$ is greater for images containing faces than for nonface images. The four panels of Fig. 4 show the difference ratings plotted against the $\tilde{\mu}^i_{jk}$ (top panels) and $\tilde{\sigma}^i_{jk}$ (bottom panels) for the face (left panels) and nonface (right panels) images separately. The linear predictive value of $\tilde{\mu}^i_{jk}$ and $\tilde{\sigma}^i_{jk}$ is significant only for the face images, and again only the global L* standard deviation accounts for a substantial proportion of variance. (Face images, $\tilde{\mu}^i_{jk}$: $R^2=0.09$, $p<0.05$; face images, $\tilde{\sigma}^i_{jk}$: $R^2=0.58$, $p<0.001$; nonface images, $R^2=0.04$, $\tilde{\mu}^i_{jk}$: $p=0.34$; nonface images, $\tilde{\sigma}^i_{jk}$: $R^2=0.00$, $p=0.83$.)

We focused on the face images for further analysis and considered whether the local tone variable differences $\tilde{\mu}^i_{face\_jk}$ and $\tilde{\sigma}^i_{face\_jk}$ extracted from the face region provided additional predictive value. Figure 5 plots the difference ratings for the face images against these two additional variables. Both local tone characteristics are predictive of the difference ratings ($\tilde{\mu}^i_{face\_jk}$: $R^2=0.59$, $p<0.001$; $\tilde{\sigma}^i_{face\_jk}$: $R^2=0.29$, $p<0.001$).

The analysis previously presented shows that both our global and local (face region) tone characteristics were predictive of image quality: differences in each variable separately are significantly correlated with the difference ratings. We used multiple regression to ask how well all four variables could jointly predict image quality. The overall R^2 when the difference ratings were regressed on $\tilde{u}^i_{jk}$, $\tilde{\sigma}^i_{jk}$, $\tilde{\mu}^i_{face\_jk}$, and $\tilde{\sigma}^i_{face\_jk}$ was 75%. Stepwise regression showed that almost all of the explanatory power was carried by two of the four variables: $\tilde{\sigma}^i_{jk}$ and $\tilde{\mu}^i_{face\_jk}$. These two variables alone provided an R^2 of 0.72. Figure 6 shows the measured difference ratings for the face images plotted against the predictions based on $\tilde{\sigma}^i_{jk}$ and $\tilde{\mu}^i_{face\_jk}$. If the two variables were perfect predictors of image quality, the data would fall along the diagonal line.

Recall that the data analysis involves finding the optimal values for the tone variables $\tilde{\sigma}^i_{jk}$ and $\tilde{\mu}^i_{face\_jk}$. Figure 7 shows a plot of how the R^2 measure for the face images varies with the optimal values σ_0 and μ_{face_0}. The optimal value σ_0 was 17.8, whereas that for μ_{face_0} was 48.7.

To test if the optimal values varied across ethnicities, we divided the images into two groups (8 Caucasian images and 9 non-Caucasian images) and then re-ran the analysis. The results for the two groups were very similar for face mean and standard deviation L* values (μ_{face_0} values were, Caucasian images: 48.6, non-Caucasian images: 48.8, and σ_{face_0} values were, Caucasian images: 19.2, non-Caucasian images: 18.4) but differed somewhat for global L* standard deviation (σ_0 values were, Caucasian images: 15.4, non-Caucasian images: 20.7). Although the performance of each of the four algorithms was not of primary concern in this paper, a summary of the preferences is shown in Table 1 for completeness. Note that Holm's method performed particularly well overall.

The data from Experiment 1 support the following conclusions: (i) We were unable to find a tone variable that predicted perceptual image quality for nonface images. (ii) For face images, a number of tone variables were significantly correlated with the difference ratings. Two variables accounted for the majority of the variance in the data that

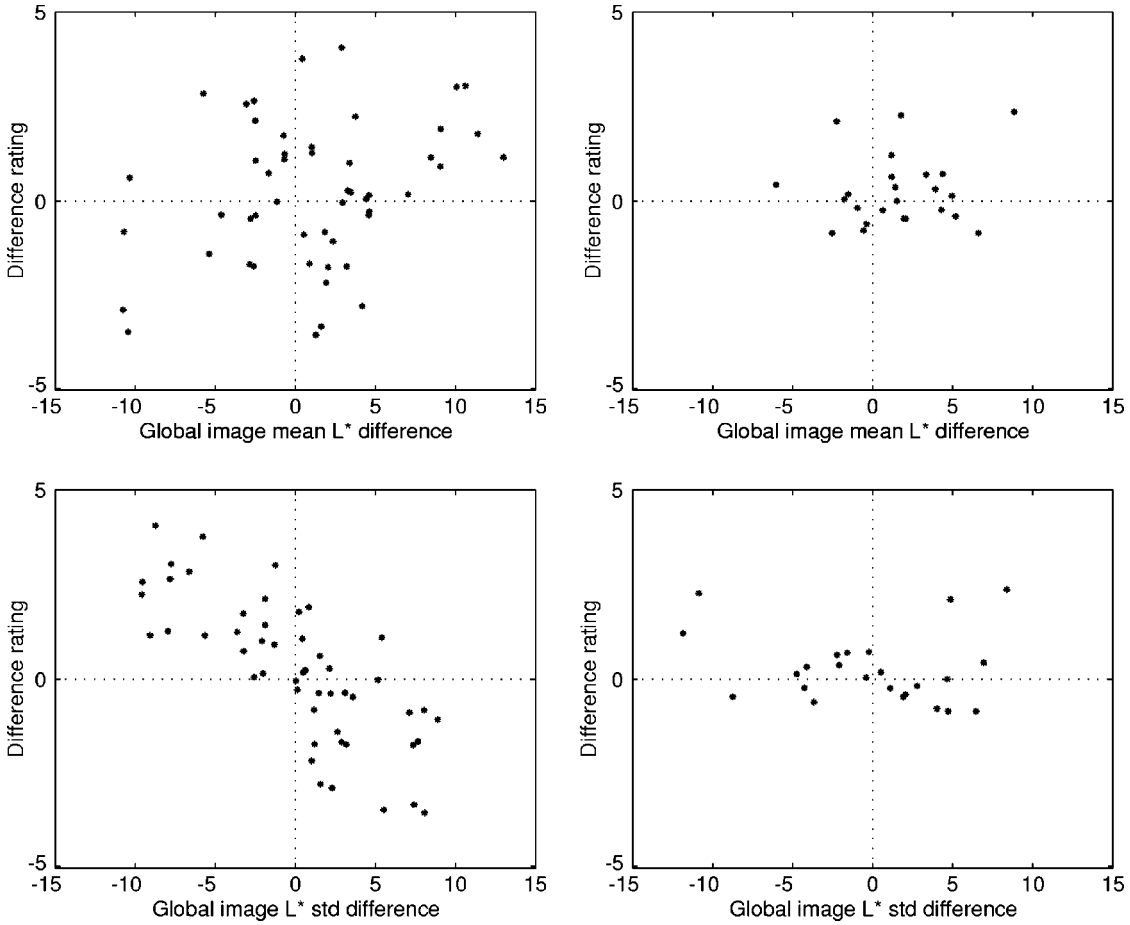

Fig. 4 Prediction of difference ratings from global tone characteristics, face images (left panels) and nonface images (right panels) shown separately. The difference ratings obtained for all images in Experiment 1 against $\tilde{u}^i_{jk}$ (top panels) and $\tilde{\sigma}^i_{jk}$ (bottom panels) are plotted.

we could explain. These were the difference in L* standard deviations across the entire image ($\tilde{\sigma}^i_{jk}$) and the mean L* value difference for the face subimage ($\tilde{\mu}^i_{face\_jk}$). (iii) The data allowed identification of optimal values for each of these variables.

4 Experiment 2

The results from Experiment 1 suggest that for images containing a face, the standard deviation of image luminance values and the mean luminance level of the face itself do a good job of predicting predictive image quality. In Experiment 2, we explored the effect of face mean luminance in more detail. We used a diverse set of face images that included people with a wide range of skin tones and images with multiple faces.

4.1 *Methods*

The methods were the same as for Experiment 1 except for the following.

4.1.1 *Image acquisition*

Images were acquired using the Kodak DCS-420 digital camera. Fifteen images were selected, all of which were portraits taken under daylight. Face subregions were again identified by hand. Ten contained only one subject (5 Caucasian, 3 African-American, 2 Asian) and five contained multiple subjects (1 of Caucasians only, 2 with African Americans only, and 3 with a mixture of ethnicities). For the images containing multiple faces, the identified face subregions included all faces. The dynamic range of the images, computed as described for Experiment 1, varied between 37 and 245.

4.1.2 *Image processing*

We wanted to generate rendered images with different face luminance levels with minimal changes to the L* standard deviation. This was done by applying a smooth global tone-mapping curve to the images, with the curve parameters chosen so that the output images had the desired face region mean L* and L* standard deviation tone characteristics. Face subimages were selected by hand and 5 versions of each image were created with different mean face L* target values (42, 48, 52, 56, and 62) and with the L* standard deviation value held fixed at approximately 18.8. Five different renderings per image produced ten possible pairwise presentations for each of the fifteen images. Difference ratings π^i_{12}, π^i_{23}, π^i_{34}, π^i_{45} and corresponding differences in tone variables were used in the analysis.

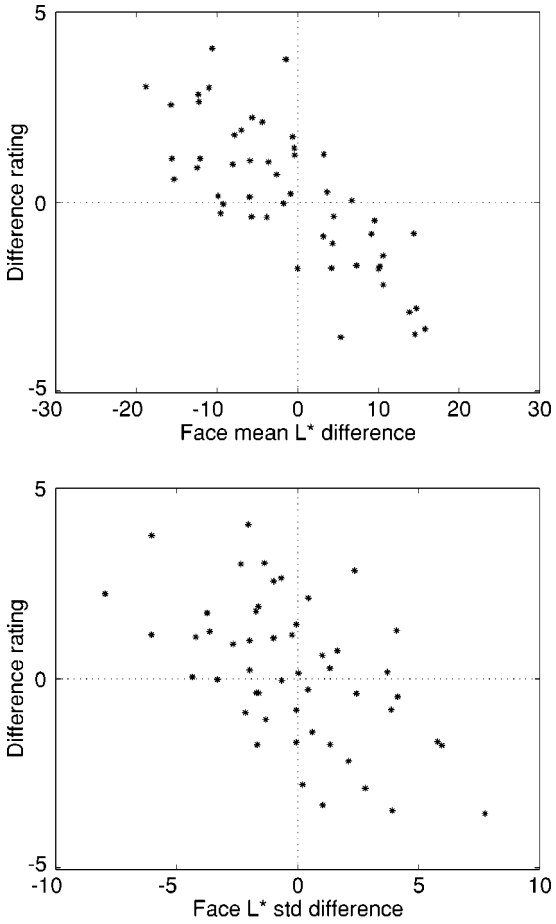

Fig. 5 Prediction of difference ratings from face-region tone characteristics, face images only. The difference ratings obtained for the face images in Experiment 1 against $\tilde{u}^i_{face\_jk}$ (top panel) and $\tilde{\sigma}^i_{face\_jk}$ (bottom panel) are plotted.

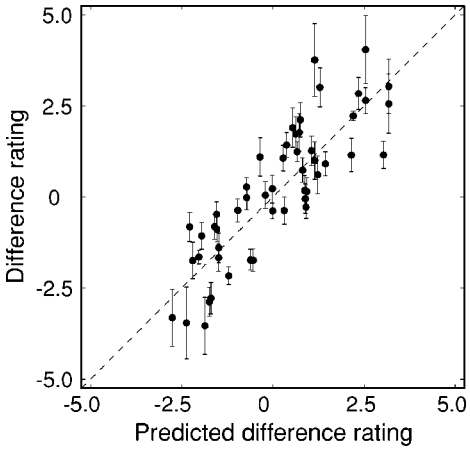

Fig. 6 Measured difference ratings plotted against difference ratings predicted as the best linear combination of $\tilde{\sigma}^i_{jk}$ and $\tilde{u}^i_{face\_jk}$ for the face images of Experiment 1. If the predictions were perfect, the points would fall on the diagonal line. The error bars show ±1 standard error of measurement for the difference ratings, computed using a resampling method (Ref. 41). The raw preference data were resampled 50 times. For each resampling, difference ratings were computed and the standard deviation of the resulting difference ratings was taken as the standard error.

4.1.3 *Observers*

Nineteen color normal observers participated in the experiment (12 males and 7 females) with an average age of 35 (range 23–62). Eight of the observers had previously participated in Experiment 1.

4.2 *Results*

The data were analyzed in the same fashion as were the data for Experiment 1 with respect to the predictive power of the $\tilde{\mu}^i_{face\_jk}$ variable. The top panel of Fig. 8 shows a scatter plot of the difference ratings against mean face-region L* value differences. For images with multiple faces, the mean face L* value was used. The regression results showed that this tone characteristic difference was significantly correlated with the difference ratings ($p < 0.001$) and that percent variance explained was $R^2 = 0.49$. This replicates and extends the results of Experiment 1 with respect to this tone characteristic.

After the experiment, observers were given a chance to provide comments and feedback. In Experiment 2, a number of observers noted that some renderings of three of the images contained visible artifacts in the facial regions, and that these artifacts had a strong negative influence on their preference for those images. Post-hoc examination of the images confirmed the observer reports. We believe the artifacts arose because the tone-mapping procedure amplified the noise in some of the darker image regions. Because our interest was in tone characteristics, not artifacts, it seemed of interest to repeat the analysis with the three problematic images excluded. This led to an increase in the percent of variance accounted for by the face L* mean difference variable, with $R^2 = 0.66$ rather than 0.49. The bottom panel of Fig. 8 shows the relation between difference ratings and this variable after the exclusion.

As part of the analysis, numerical search was again used to find value μ_{face_0} that optimized R^2. This value was 49.2 when the full data set was analyzed and 46.5 with the three images excluded, both very close to the value of 48.7 found in the first experiment. The dependence of the R^2 value on the optimal parameter is shown in Fig. 9 for the two cases.

We examined if the optimal μ_{face_0} value varied across ethnicities. The images were divided into two groups (4 images of Caucasians, 6 images of non-Caucasians). Five of the images were not included (2 had multiple faces of different ethnicities and three has visible artifacts in the face region as discussed above). We re-ran the analysis and the results for the two groups were very similar (μ_{face_0} values were, Caucasian images: 46.3, non-Caucasian images: 45.7).

5 Discussion

5.1 *Summary*

The paper presents experiments that explore whether a number of simple image tone characteristics are predictive of perceptual image quality. For the nonface images we studied, we were unable to identify any such variables. For images consisting primarily of faces, however, the results suggest that the best image quality results when the face L*

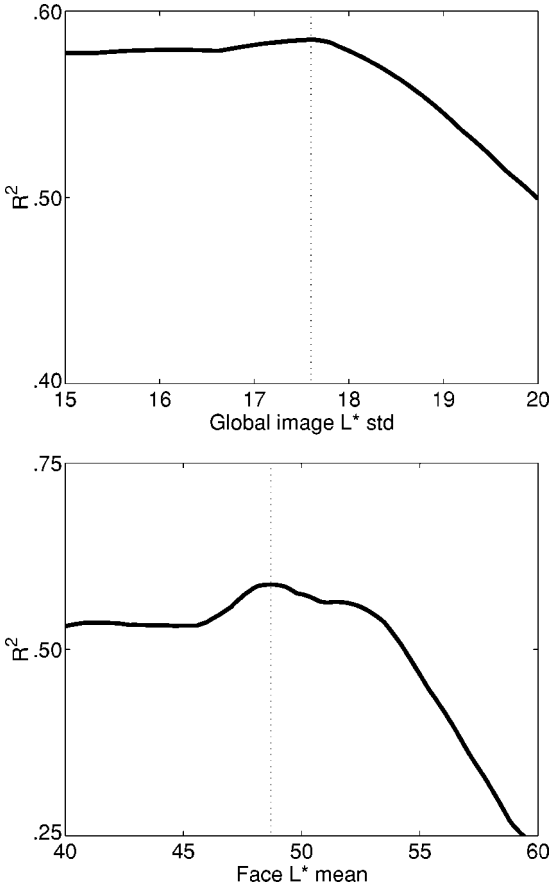

Fig. 7 Optimal values σ_0 and u_{face_0} for the face images used in Experiment 1. Each panel plots the percent variance explained by a single tone characteristic (top panel: $\tilde{\sigma}^i_{jk}$; bottom panel: $\tilde{u}^i_{face\_jk}$) as a function of the corresponding optimal value (top panel: σ_0; bottom panel: u_{face_0}).

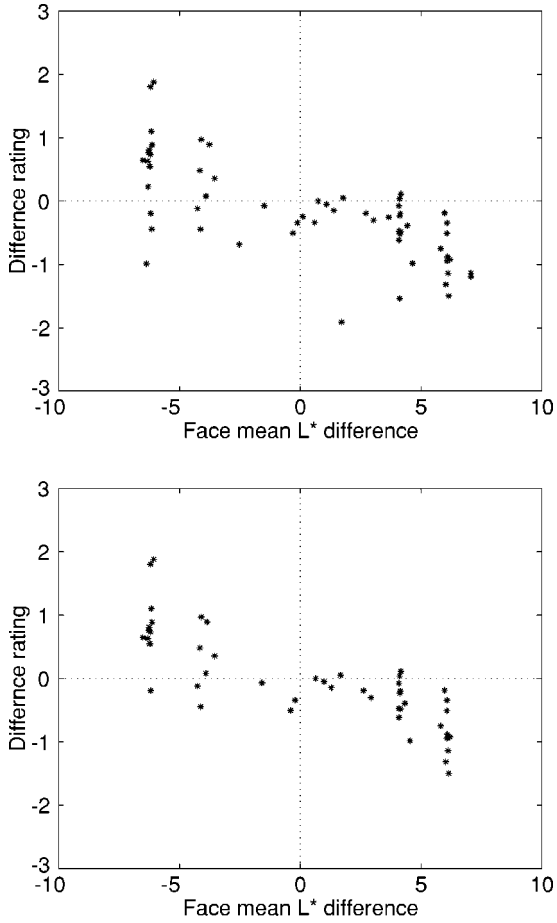

Fig. 8 Prediction of difference ratings from face-region tone characteristics, for Experiment 2. The difference ratings against $\tilde{u}^i_{face\_jk}$ are plotted. The top panel shows the full data set and the bottom panels shows the data when three images with artifacts were excluded.

luminance is in the range 46–49, and the standard deviation of the image L* luminances is approximately 18. This conclusion was suggested by the results of Experiment 1, and the conclusion concerning the optimal level of face L* was confirmed directly in Experiment 2.

The images used in our experiments contained faces with a wide variety of skin tones. Analysis of Caucasian and non-Caucasian subgroups suggest that the conclusions concerning optimal face L* level may generalize to a wide array of face images. We do note, however, that our image sample was relatively small and that follow-up work might profitably probe the generality of our results. For example, we do not know how sensitive the data are to the noise properties of the camera sensors. The analysis of the Experiment 1 data by ethnicity also suggests that the optimal global L* standard deviation for the rendered image may depend on ethnicity, although again the generality of this result is not clear.

Table 1 The overall percentage of times the output of each tone-mapping method was chosen as the preferred image in Experiment 1. Results for each algorithm were obtained by taking all of the pairwise comparisons involving the output of each algorithm and computing the percentage of times the output of that algorithm was chosen as preferred. Data were aggregated across all images and observers.

Images	Clipping (%)	Histogram (%)	Larson (%)	Holm (%)
All	26.4	19.0	19.0	35.7
Face	30.0	15.7	16.2	38.2
Nonface	18.8	25.9	24.9	30.4

5.2 *Other Image Statistics*

In addition to the image tone characteristics on which we previously reported in detail, we also examined other possible predictors of image quality. These included chromatic variables and a histogram difference measure. The histogram difference measure increased with the difference between the luminance histogram of the input and output of the tone-mapping algorithms. The chromatic variables did not provide predictive power. This is perhaps not surprising given that the images were all color balanced to a common illuminant and that the tone-mapping algorithms did not affect pixel chromaticities. The histogram difference measure was correlated with image quality for the face images.

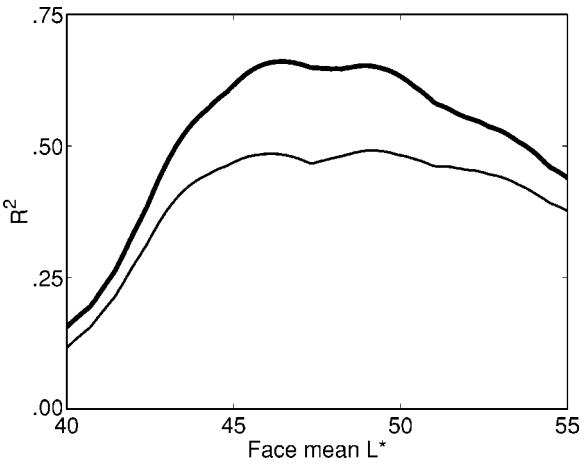

Fig. 9 Optimal value u_{face_0} for Experiment 2. The plot shows the percent variance explained by $\tilde{u}^{i}_{face\_jk}$ as a function of the optimal value u_{face_0}. Thin line: full data set. Thick line: data set when three images with artifacts were excluded.

A stepwise regression analysis, however, showed that adding the histogram difference measure to the face L* and image L* standard deviation did not explain substantial additional variance.

Holm[16,38] has suggested that classifying images based on histogram properties and then applying different tone mapping depending on the classification can be effective. To explore this, we computed Holm's *key value* statistic from our input image histograms and divided the scenes into two sets, low key and high key, based on this statistic. Low-key scenes have luminance histograms that are skewed toward dark values, whereas high-key scenes have luminance histograms that are skewed toward light values. In Experiment 1, we found that the relation between global L* value and image quality was strong for the low-key scenes and not significant for the high-key scenes, whereas the relation between global L* standard deviation and image quality was significant for both low- and high-key scenes. There was a difference in optimal global L* standard deviation between the two sets, but this difference was not stable with respect to small perturbations of the criterion key value used to divide the data set. In Experiment 2, the dependence on face L* values was significant for both low- and high-key scenes with the optimal value varying between 52 (low-key) and 47 (high-key). Further experiments focused on the stability of scene key as a modulator of optimal tone characteristics, as well as on other potential higher-order histogram statistics (e.g., degree of bimodality), would be of interest.

5.3 *Relation to Other Work*

The work here emphasizes comparisons are among images displayed on a common output device, so that the dynamic range of the comparison set is constant. This is a reasonable choice for the goal of improving the appearance of images acquired with current digital cameras, whose image capture range is approximately matched to current display technology. In contrast, a number of papers have examined tone-mapping across large changes in dynamic range between input and output.[8,15,17,18,20,39] The experimental methods and analysis presented here are general and could be used to evaluate the efficacy of these methods for high-dynamic range imagery.

A second feature of our work is our focus on the tone characteristics of the displayed images, rather than on the functional form of the tone-mapping curve. The results presented here suggest that there is considerable utility in examining tone characteristics. Other recent experimental work[19,20] has focused on the efficacy of tone-mapping operators *per se*. These two approaches may be viewed as complementary. Also of note is the diverse set of psychophysical techniques that have been employed across studies.[19,20,39] Here we have focused on image preference, which is conceptually quite different from perceptual fidelity.

5.4 *Using the Results*

Although our positive results only apply to images that contain faces, such images probably form a large proportion of those acquired by the average camera user—many consumers take pictures of their friends and families. Thus our results have the potential for leading to useful practical algorithms.

Since our work shows how preference for images containing faces depends on tone variables, tone-mapping methods might profitably include algorithms to identify images that contain faces and to apply appropriate mapping parameters to these images. (Face recognition software has advanced greatly in recent years. See recent review by Pentland and Choudhury[40]). Indeed, the present work led directly to the development of a novel proprietary tone-mapping algorithm at Agilent Laboratories.[42] The idea that empirical image preference studies can enable development of effective image processing algorithms was also supported by our earlier study.[4] We believe further studies hold the promise of providing additional algorithmic insights.

Acknowledgments

The authors wish to thank Jerry Tietz for help with image acquisition and Jack Holm and Jeff DiCarlo for help with image processing and implementation of tone-mapping methods. Jack Holm also helped with the experiment room setup. Finally they would like to thank Russell limura, Amnon Silverstein, Joyce Farrell, and Yingmei Lavin for helpful suggestions.

References

1. C. J. Bartleson, "Memory colors of familiar objects," *J. Opt. Soc. Am.* **50**, 73–77 (1960).
2. S. M. Newhall, R. W. Burnham, and J. R. Clark, "Comparison of successive with simultaneous color matching," *J. Opt. Soc. Am.* **47**(1), 43–56 (1957).
3. R. M. Boynton, L. Fargo, C. X. Olson, and H. S. Smallman, "Category effects in color memory," *Color Res. Appl.* **14**, 229–234 (1989).
4. P. Longere, X. Zhang, P. B. Delahunt, and D. H. Brainard, "Perceptual assessment of demosaicing algorithm performance," *Proc. IEEE* **90**, 123–132 (2002).
5. H. de Ridder, "Naturalness and image quality: saturation and lightness variation in color images of natural scenes," *J. Imaging Technol.* **40**(6), 487–493 (1996).

6. E. A. Fedorovskaya, H. D. Ridder, and F. J. J. Blommaert, ''Chroma variations and perceived quality of color images of natural scenes,'' *Color Res. Appl.* **22**(2), 96–110 (1997).
7. T. Tanaka, R. S. Berns, and M. D. Fairchild, ''Predicting the image quality of color overhead transparencies using a color-appearance model,'' *J. Electron. Imaging* **6**(2), 154–165 (1997).
8. G. W. Larson, H. Rushmeier, and C. Piatko, ''A visibility matching tone reproduction operator for high dynamic range scenes,'' *IEEE Transactions on Visualization and Computer Graphics*, available at http://radsite.lbl.gov/radiance/papers (1997).
9. C. N. Nelson, ''Tone and color reproduction Part 1: Tone Reproduction,'' *The Theory of the Photographic Process*, Macmillan, New York (1977).
10. C. J. Bartleson and E. J. Breneman, ''Brightness reproduction in the photographic process,'' *Photograph. Sci. Eng.* **11**(4), 254–262 (1967).
11. C. J. Bartleson and E. J. Breneman, ''Brightness perception in complex fields,'' *J. Opt. Soc. Am.* **57**, 953–957 (1967).
12. L. D. Clark, ''Mathematical prediction of photographic picture quality from tone-reproduction data,'' *Photograph. Sci. Eng.* **11**(5), 306–315 (1967).
13. R. G. W. Hunt, ''The effect of viewing conditions on required tone characteristics in colour photography,'' *Brit. Kinematogr. Sound Telev.* **51**, 268–275 (1969).
14. R. W. G. Hunt, I. T. Pitt, and P. C. Ward, ''The tone reproduction of colour photographic materials,'' *J. Photogr. Sci.* **17**, 198–204 (1969).
15. E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, ''Photographic tone reproduction for digital images,'' *ACM Trans. Graphics* **21**, 267–276 (2002).
16. J. Holm, ''Photographic tone and colour reproduction goals,'' *CIE Expert Symp. '96 on Colour Standards for Image Technology*, 51–56 (1996).
17. J. Tumblin, J. K. Hodgins, and B. K. Guenter, ''Two methods for display of high contrast images,'' *ACM Trans. Graphics* **18**, 56–94 (1999).
18. J. M. DiCarlo and B. A. Wandell, ''Rendering high dynamic range scenes,'' *Proc. SPIE* **3965**, 392–401 (2000).
19. F. Drago, W. L. Martens, K. Myszkowski, and H.-P. Seidel, ''Perceptual evaluation of tone mapping operators with regard to similarity and preference,'' Report No. MPI-I-2002-4-002, Max-Planck-Institut (2002).
20. G. J. Braun and M. D. Fairchild, ''Image lightness rescaling using sigmoidal contrast enhancement functions,'' *J. Electron. Imaging* **8**, 380–393 (1999).
21. J. L. Simonds, ''A quantitative study of the influence of tone-reproduction factors on picture quality,'' *Photograph. Sci. Eng.* **5**(5), 270–277 (1961).
22. S. B. Novick, ''Tone reproduction from colour telecine systems,'' *Brit. Kinematogr. Sound Telev.* **51**(10), 342–347 (1969).
23. D. A. Silverstein and J. E. Farrell, ''Efficient method for paired comparison,'' *J. Electron. Imaging* **10**(2), 394–398 (2001).
24. Kodak, ''Programmer's reference manual models: DCS-200ci, DCS-200mi, DCS-200c, DCS-200m,'' Report 2, Eastman Kodak Company (1992).
25. P. L. Vora, J. E. Farrell, J. D. Tietz, and D. H. Brainard, ''Image capture: simulation of sensor responses from hyperspectral images,'' *IEEE Trans. Image Process.* **10**, 307–316 (2001).
26. D. H. Brainard, ''An ideal observer for appearance: reconstruction from samples,'' Report No. 95-1, UCSB Vision Labs Technical Report, Santa Barbara, CA (1995), http://color.psych.upenn.edu/brainard/papers/bayessampling.pdf
27. D. H. Brainard, ''Bayesian method for reconstructing color images from trichromatic samples,'' *IS&T 47th Annual Meeting*, 375–379 (1994).
28. D. H. Brainard and D. Sherman, ''Reconstructing images from trichromatic samples: from basic research to practical applications,'' *IS&T/SID Color Imaging Conf.: Color Science, Systems, and Applications*, 4–10 (1995).
29. M. J. Vrhel, R. Gershon, and L. S. Iwan, ''Measurement and analysis of object reflectance spectra,'' *Color Res. Appl.* **19**(1), 4–9 (1994).
30. D. H. Brainard and W. T. Freeman, ''Bayesian color constancy,'' *J. Opt. Soc. Am. A* **14**(7), 1393–1411 (1997).
31. W. Frei and C. C. Chen, ''Fast boundary detection: a generalization and a new algorithm,'' *IEEE Trans. Comput.* **26**, 988–998 (1977).
32. D. H. Brainard, ''Calibration of a computer controlled color monitor,'' *Color Res. Appl.* **14**(1), 23–34 (1989).
33. ISO, ''Viewing conditions—for graphic technology and photography,'' *ISO* **3664**, 1998(E) (1998).
34. D. H. Brainard, ''The psychophysics toolbox,'' *Spatial Vis.* **10**(4), 433–436 (1997).
35. D. G. Pelli, ''The Video Toolbox software for visual psychophysics: transforming numbers into movies,'' *Spatial Vis.* **10**(4), 437–442 (1997).
36. S. Ishihara, *Tests for Colour-Blindness*, Kanehara Shuppen Company, Ltd., Tokyo, Japan (1977).
37. CIE, ''Recommendations on uniform color spaces, color-difference equations, psychometric color terms,'' Report No. Supplement No. 2 to CIE Publication No. 15, Bureau Central de la CIE (1978).
38. J. Holm, ''A strategy for pictorial digital image processing,'' *Proc. of the IS\&T/SID 4th Color Imaging Conf.*, 194–201 (1996).
39. A. MacNamara, ''Visual perception in realistic image synthesis,'' *Comput. Graphics Forum* **20**, 211–224 (2001).
40. A. Pentland and T. Choudhury, ''Face recognition for smart environments,'' *Computer* **33**(2), 50–55 (2000).
41. B. Efron and R. LePage, ''Introduction to bootstrap,'' *Exploring the Limits of Bootstrap*, L. Billard, Ed., Wiley & Sons, New York (1992).
42. X. Zhang, R. W. Jones, I. Baharav, and D. M. Reid, ''System and method for digital image tone mapping using an adaptive sigmoidal function based on perceptual preference guidelines.'' U.S. Patent No. EP02000691 (Jan. 2005).

Peter Delahunt received his BS degree in psychology from Lancaster University, England in 1996. He received his MA and PhD degrees in psychology from the University of California, Santa Barbara, in 1998 and 2001, respectively. He is currently working as a human factors scientist for Exponent Inc.

Xuemei Zhang received her bachelor's degree in psychology from Beijing University, master's degree in statistics, and PhD in psychology from Stanford University. She is currently a research scientist working in Agilent Technologies Laboratories.

David Brainard received his AB in Physics from Harvard University in 1982. He attended Stanford University for graduate school and received his MS degree in electrical engineering and PhD in psychology, both in 1989. He is currently professor of psychology at the University of Pennsylvania. His research interests include human vision, image processing, and visual neuroscience.