Evaluating Tone Mapping Algorithms for Rendering Non-Pictorial (Scientific) High-Dynamic-Range Images

SUNG HO PARK and ETHAN D. MONTAG
Rochester Institute of Technology

Nine algorithms were implemented to overcome the problem associated with rendering high-dynamic-range scientific imagery to low-dynamic-range display devices. The algorithms were evaluated using two paired-comparison psychophysical experiments judging preference and "scientific usefulness". The results showed that, on average, the Zone System algorithm performed best and the Local Color Correction method performed the worst. However, the performance of the algorithms depended on the type of data being visualized. The low correlation between the preference and scientific usefulness judgments ($R^2 = 0.31$) indicated that observers used different criteria when judging the image preference versus scientific usefulness. The experiment was repeated using expert observers (radiologists) for an MR scan (Medical image). The results showed that the radiologists used similar criteria as the non-expert observers when judging the usefulness of the rendered images. A target detection experiment was conducted to measure the detectability of an embedded target in the Medical image. The result of the target detection experiment illustrated that the detectability of targets in the image is greatly influenced by the rendering algorithms due to the inherent difference in tone mapping among the algorithms.

Categories and Subject Descriptors: I.3.3 [Computer Methodologies]: Computer Graphics—Picture/Image Generation; H.1.2 [Information Systems]: Models; Principles—User/machine systems

General Terms: Experimentation, Human Factors, Performance
Additional Key Words and Phrases: High-Dynamic-Range Image, Rendering, Psychophysics

1. INTRODUCTION

One aim of image rendering or reproduction is the creation of images that share the identical appearance attributes as the real scene. The world exhibits a wide range of luminance values. The human visual system is capable of perceiving this wide range of dynamic scenes spanning five orders of magnitude and adapting more gradually to over nine orders of magnitude, which is facilitated by local adaptation that allows regions of various luminance levels to be viewed essentially simultaneously. Recent advances in high dynamic range capturing systems [Debevec and Malik 1997;
Nayar and Mitsunaga 2000; Xiao et al. 2002] make it possible to capture a highly detailed dynamic range representation of the scene and later process the data in order to select the image that better fulfills the given requirements. Unfortunately, the dynamic range of image display devices and image display media have not kept up with the progress in digital image capture devices and methods. Since typical desktop displays, such as CRTs and LCDs, are only capable of displaying about two orders of magnitude of dynamic range, the question is then how can we reproduce and visualize such high-dynamic-range (HDR) images in a standard output device. This is fundamentally possible because the human eye is sensitive to relative luminance values, rather than absolute luminance.

More recently, concern has grown in the visualization and scientific communities over the use of scientific imagery, its interpretation, and the relation of the data to its interpretation. Novel techniques are also required for imagery captured from non-visual sources such as remote sensing, medical imaging, astronomical imaging, etc. The goal of this study is to integrate the techniques used for the display of HDR pictorial imagery for the display of non-pictorial imagery while searching for perceptually based schemes for encoding this imagery that facilitate its interpretation. By applying these same HDR processing techniques developed for pictorial imagery, it is hypothesized that more information can be conveyed because local perceptual contrast in a wider range of the scene will be preserved by automatically adjusting the luminance and chromatic contrast in the image based on the image content.

In the fields of medical imaging and radiology, emphasis has been placed on developing standards for displays that perceptually linearize the displayed data. An example of this is the Digital Imaging and Communications in Medicine (DICOM) Part 14: Grayscale Standard Display Function [NEMA 2004]. This standard is based on Barten’s model of the contrast sensitivity function [1992] so that changes in digital values of the display represent equal perceptual steps in lightness based on threshold differences. For a display of the luminance level used in the present experiment, the DICOM implementation of the model predicts that there are 556 jnd’s between black and white, so that a display would require between 10-bits to encode intensity gradients without contouring artifacts. Interestingly, Cowan et al. (2004) calculated and verified experimentally that the eye’s sensitivity to change, based on Barten’s model, requires between 10- and 12-bit encoding for a display with a luminance that was less than 25% of the luminance of our display to reduce contouring. These differences are likely due to differences in the implementation of the model parameters such as spatial frequency and adaptation level.

The grayscale function that best represents perceptual linearity depends on the nature of the stimuli being displayed. For 2° x 2°, 4 cycle per degree patches on a uniform background, the DICOM Part 14 standard may represent good linearity, however, for other stimuli, a variety of curves have been determined (see Wyszecki and Stiles, [2000], for example). Our experience [Montag 1999] has shown that in complex images, observers’ judgments of lightness corresponds to a gray scale curve with a gamma of approximately 1.8. We can compare this value to the gamma of 6.28 best fit to the DICOM standard based on the luminance or a gamma of 2.49 based on the 1976 CIELAB lightness function, L^*.

Fundamentally, there is a trade-off between the number of levels that can be represented and the intensity of the display. Muka and Reiker [2002] argue that 8-bits are sufficient for reproducing certain radiological imagery given the viewing conditions and the limits of the visual system [Muka and Whiting 2002]. Despite the cost involved, however, an advantage may be achieved using displays with higher bit-depth and corresponding higher luminance levels. However, HDR algorithms attempt to bypass this problem by recreating the appearance of a scene at a lower luminance level and bit-depth. Therefore the goal here is to increase the amount of information displayed on common 8-bit displays using algorithms developed for the display of pictorial imagery. In this paper, we consider a variety of scientific imagery from a variety of fields. Further research is needed to determine the specific utility of these techniques in the various fields.

Much research has been done to develop algorithms that are capable of recreating a truthful rendition of high-dynamic-range image onto lower-dynamic-range displays [Reinhard et al. 2002; Durand and Dorsey 2002; Johnson and Fairchild 2003]. For pictorial imagery, the truthfulness or accuracy of the display lies in the ability to recreate the appearance qualities of the original scene. However, for non-pictorial imagery, the truthfulness of the display cannot be evaluated by comparison with the original scene. Instead, the usefulness of the display lies in the ability of the user to visually interpret and use the data. The term, non-pictorial, refers to scientific imagery captured outside the visible wavelength region or of objects not accessible to the human eye, such as hyperspectral data captured by spacecraft or aircraft, astronomical images captured using non-visible wavelengths, or characteristics of human tissue obtained in medical imaging. Since main focus of this project is to test algorithms for the display of non-pictorial HDR imagery that is univariate, the visualization of multidimensional data is not concerned in this study.

There are three aspects of this study: 1) The development and implementation of HDR algorithms including some used for HDR pictorial imagery 2) The psychophysical evaluation of these algorithms in rendering this non-pictorial imagery, and 3) The psychophysical measurement of the effect of tone and contrast mapping on target detection. The results from the evaluation aspect will be used as feedback to help improve the algorithms used to encode the data. Two psychophysical experiments were conducted to evaluate these algorithms. The goal of the psychophysical testing was to determine which algorithms were preferred and which algorithms were judged as being more scientifically useful.

This paper should be considered a starting point for the application of these HDR algorithms to imagery from sources distinct from pictorial imagery. It is but a survey of the possible application of these algorithms; more research will be required in order to fully understand the generality and applicability of these algorithms. Here we are demonstrating the distinction between image preference and the usefulness of these algorithms for the scientific evaluation of different types of imagery. The specific utility of an algorithm must take into account the specific way the imagery is used and therefore rely on the judgments of expert users for evaluation.
2. TONE MAPPING OPERATORS

The goal of tone reproduction (mapping) is to produce realistic renderings of captured scenes, showing no more and no less visual content than would be visible if actually present to see the original scene, and to produce such rendering while facing the limitations presented by output devices. In our daily environment, the Human Visual System (HVS) copes with the large variations of the luminance input to the eyes through a complex local adaptation process that allows regions of various luminance levels to be viewed effectively simultaneously. When the eye moves around in a natural environment, the luminance input to the eyes changes continually. The sensitivity of the HVS is continually adjusted in order to allow efficient transfer of information about the visual input to the brain. Without such an adjustment, small signals will drown in neuronal noise, and large signals will saturate the system. The purpose of luminance adaptation is thus to keep the response to rapidly varying visual input within the dynamic range of the neurons in the retina [Shapley and Enroth-Cugell 1984]. These adaptation mechanisms have been studied extensively, both psychophysically [Hayhoe et al. 1992; Foley and Boynton 1993; Kortum and Geisler 1995; Hood et al. 1997] and physiologically [Yeh et al. 1996; Shapley 1997; Victor et al. 1997]. Although these studies are giving an increasingly detailed view of early processes of adaptation, it is not clear how these processes act in performance under natural luminance conditions.

Adaptation is the term used for the process that changes the sensitivity of the visual system to different light levels. The problem of adapting to increase and decrease in illumination is best understood by considering the variety of situations confronting the human visual system. The human observer experiences a range of naturally occurring ambient light levels of nearly 14 log units and must be able to discriminate objects in the environment over 8 log units [Hood and Finkelstein 1986]. However, the differences in intensity reflected by those objects at any single light level are very small, spanning at best 2 to 3 log units [Walraven et al. 1990].

The question that models of adaptation must answer is how the visual system remains sensitive to such small differences over such a wide ambient range. Moreover, the problem is how to scale such wide luminance values to the limited displayable range of a standard output device. The pixel values in most output devices, such as CRTs and LCDs, are limited to a useful dynamic range of about two orders of magnitude represented by eight bits per pixel with values between 0–255, which falls far short of the range of real world luminance values. This is where tone reproduction operators come in to play their important role. Since the human visual system is more sensitive to relative rather than absolute luminance values, visualizing the high dynamic range of world luminance value in a low dynamic range of output device is fundamentally possible. In order to create realistic rendering of a scene, tone reproduction should provide not only a method of compressing the range of luminance values that mathematically transform scene luminances into output device with limited capabilities, but also prediction of a various visual phenomena that mimics perceptual qualities such as threshold visibility, contrast, brightness, and fine detail – all the visual sensations experienced by a human observer viewing the scene in the real world [Ferwerda et al. 1996; Pattanaik et al. 1998]. A more complete understanding of both early and higher levels of HVS is essential for
advances in both the efficiency and the effectiveness of realistic image rendering, especially at suprathreshold level. Although complete models of the visual system are still mysterious to a certain extent, enough information has been accumulated to develop tone reproduction operators to display HDR imagery. Reviews of such tone reproduction operators can be found in McNamara [2001], Devlin et al. [2002], and Park [2004].

Since the issues of realistic tone mapping were introduced, many operators have been proposed to overcome the problem of displaying HDR images. In order to simulate the realistic perception of world luminance levels on a standard output device, some operators utilize perceptual data based on psychophysical experiments while others exploit a mathematical approach to simply compress the luminance range with the aim of obtaining the maximum visibility on the display device and without considering the perceptual aspects of visual system. In those cases where tone reproduction attempts to simulate reality, one of the most important factors for rescaling the high dynamic range to fit into the smaller output dynamic range is that the final image maintains the lightness integrity of the original scene.

Tone reproduction operators can be classified in two main categories: spatially uniform (global) and spatially varying (local) operators. Spatially uniform operators do not imitate local adaptation processes of the HVS but use an implicit normalizing factor in order to scale the scene luminance to fall within the limited range of display device. These operators handle the images as a whole and apply the same transformation to every pixel discarding the original intensities of the scene, which may cause perceptual differences. On the other hand, spatially varying operators mimic the local adaptation process in the retina by applying different scaling factors to different parts of an image. These operators reduce scene contrast locally, relative to neighborhood intensities, and convert the original intensities to the displayable intensities of the low-dynamic-range device.

Nine methods primarily proposed for the display of HDR pictorial imagery were implemented for the display of non-pictorial imagery. (See Kuang et al.[2004] for a comparison of HDR algorithms applied to pictorial imagery.) The nine proposed operators varied from a simple linear scaling factor to more complete high end solutions, which take into account complex perceptual human attributes, in other words, from simple global (spatially uniform) mapping to complex multi-scale local (spatially varying) mapping to imitate the visual system. In all cases, an inverse display characterization was applied at the end of each operator to account for inherent device nonlinearity before displaying the image. That is, the digital values in the original values were considered as values of linear magnitude. The electro-optical transfer function (EOTF) of the display was used to render the output of the algorithm so that it presented the appropriate intensities related to these magnitudes specified by the algorithm for those algorithms where it was appropriate. Controllable parameters for each operator were set as stated and recommended in its reference. Each of the nine algorithms used here are briefly described as follows (see each reference and Park [2004] for more detailed description of each algorithm):

**Linear Mapping.** Linear mapping is the most common approach that simply linear scale to fit the high-dynamic-range image data to low-dynamic-range display device. This linear method was used as a baseline algorithm.
Spiral Rendering (curved color path)?. The Spiral Rendering uses an intensity scale based on the CIELAB uniform color space with an added hue and chroma. The algorithm basically extends the path length of the univariate scale by adding color to the path.

Sigmoid-Lightness Rescaling [Braun and Fairchild 1999]. This function was originally developed for gamut mapping to remap lightness based on a discrete cumulative normal function. By utilizing a sigmoid function, both the highlight and the shadow detail are compressed to enhance the image contrast in the low dynamic range.

Localized Sigmoid Mapping [Park 2004]. This algorithm was developed to locally control the contrast of the HDR image. Gaussian blur in the frequency domain was utilized to set the parameter for every pixel to perform this localized adaptation.

Photoshop (Auto-levels). As part of one of the most popular graphic software application with the capability of handling 16-bit images, the performance of the Photoshop tool called “auto-levels” was evaluated with the other algorithms.

Local Color Correction [Moroney 2000]. This process basically performs a pixel-by-pixel gamma correction by utilizing an inverted low-pass filtered version of the input image.

iCAM (Image Color Appearance Model) [Johnson and Fairchild 2003]. iCAM was originally designed for truthful rendition of overall color appearance in images. However, the iCAM framework can be tuned for the prediction of the appearance of high-dynamic-range images because iCAM includes spatially localized adaptation and spatially localized contrast control that can be applied to the problem of displaying HDR images.

Fast Bilateral Filtering [Durand and Dorsey 2002]. The Fast Bilateral Filtering is based on a two-scale decomposition of the image into a base layer and a detail layer. It compresses the contrast of the based layer by bilateral filtering while preserving the details of the original image.

Zone System [Reinhard et al. 2002]. The Zone System utilizes an automatic dodging-and-burning technique used in traditional photography to accomplish dynamic range compression.

3. NON-PICTORIAL (SCIENTIFIC) IMAGERY

As stated earlier, the term, non-pictorial, refers to scientific imagery captured outside the visible wavelength region or of objects not accessible to the human eye, such as hyperspectral data captured by spacecraft or aircraft, astronomical images captured using non-visible wavelengths, or characteristics of human tissue obtained in medical imaging. In principle, any data arranged in a two dimensional matrix which can be displayed as an image can be considered.

Five different sources of scientific imagery were utilized in this study, and one pictorial image was also included for comparison. They are briefly described in Table I. Histograms of original images and thumbnails of each image processed by iCAM are shown in Figure 1. The processed radar image was cropped to 930(rows)×800(columns) in order to display the image in true size.

The astronomical image is a 16-bit binary image of a dying star captured by the Hubble Space Telescope. The image holds a peak pixel values at around 200 ADU.
Table I. Information of the imagery exploited in the study.

<table>
<thead>
<tr>
<th>Image Type</th>
<th>Source</th>
<th>Bit Depth</th>
<th>Size (pixels, deg, v x h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomical</td>
<td>Hubble Space Telescope</td>
<td>16-bits</td>
<td>1000x650, 14.7x9.6</td>
</tr>
<tr>
<td>Medical</td>
<td>Magnetic Resonance Image</td>
<td>11-bits</td>
<td>256x256, 3.8x3.8</td>
</tr>
<tr>
<td>Hyperspectral</td>
<td>AVIRIS</td>
<td>15-bits</td>
<td>614x512, 9.1x7.6</td>
</tr>
<tr>
<td>Radar</td>
<td>AIRSAR</td>
<td>14-bits</td>
<td>930x800, 13.7x11.8</td>
</tr>
<tr>
<td>Infrared</td>
<td>WASP</td>
<td>11-bits</td>
<td>640x510, 9.5x7.5</td>
</tr>
<tr>
<td>Pictorial</td>
<td><a href="http://www.debevec.org">http://www.debevec.org</a></td>
<td>16-bits</td>
<td>768x512, 11.3x7.6</td>
</tr>
</tbody>
</table>

(Analog-to-Digital Units), but there is real information at the level of 0.03 ADU. The medical image is a magnetic resonance image of a human spine that with a maximum digital value of 1655. The hyperspectral image is the 910 nanometer (nm) plane of a multispectral image of the Rochester, New York area (specifically, around RIT campus) captured by AVIRIS (Airborne Visible Infrared Imaging Spectrometer). This image delivers near 15-bits of information about the surface and atmosphere based on molecular absorption and particle scattering signatures. The radar image is a scene of Nabesca, Alaska captured by AIRSAR (Airborne Synthetic Aperture Radar), designed and built by the Jet Propulsion Laboratory (JPL) NASA, which contains 14-bits of data collected by penetrating through the clouds at night. The Infrared image is 11-bit imagery from long-wave band of the electromagnetic spectrum taken by WASP (Wildfire Airborne Sensor Program). The pictorial image is the 16-bit Memorial Church image from http://www.debevec.org.

In the case of an MR image, Hornak [2005] describes the typical image viewing process:

The image is typically displayed with an eight bit video display. This means there are 256 possible gray levels with which to display the 32768 possible data values from the 15 bits of magnitude information. A linear look-up table (LUT) is typically used. Here the video intensity between 0 and 255 is set by a linear relationship to the data value. The width of the data values set to the 256 possible gray levels is called the width or contrast. The data value assigned the center of the gray scale is referred to as the level or brightness. Adjusting the width and level allow the viewer to set the image attributes which best display the anatomy and pathology.

In this case, no special processing is involved in displaying the image other than the interaction of the viewer to select the desired range. However, spatial operators are used and have been evaluated in the processing of radiographic imagery. Flynn, et al. [2001] determined two global parameters associated with the latitude and degree of spatial equalization for chest radiographs. This technique bares similarities to the local operators used here however the algorithms used here dynamically adjust the latitude and contrast based on the local structure of the image as opposed to relying on global parameters.

4. PSYCHOPHYSICAL EXPERIMENTS

The psychophysical experiments were conducted on a 23” Apple Cinema HD flat-panel LCD display connected to an Apple Power Mac G4 dual 1GHz processor.
A MATLAB program was used for processing the images and running the experimental GUI. The display device was characterized using a LMT C1210 Colorimeter following the method described in Day et al. [2004]. The measured chromaticity of the display white point was $x = 0.30$ and $y = 0.34$ (CIE, 1931 2° obs.) and the luminance was 179 cd/m$^2$. The native gray scale curve (OETF) for this monitor can be well fit with a gain, offset, gamma (GOG) model with values of 1.08, -0.06, and 1.50, respectively. The experiment was performed in a darkened room with a viewing distance of 1 m. Table I gives the size of the images in degrees of visual angle. The size of the display was 27.8° x 17.6°. The images were surrounded by a neutral gray with the same chromaticity as the white point and a luminance of 33 cd/m$^2$. This gray was used as the background between image presentations as well.

Since this project deals with non-pictorial imagery, the fidelity of the processed
images cannot be judged by comparison with the original scene. Instead, three psychophysical experiments were carried out to measure the effect of the different operators on the perception of the various images. Figure 2 shows the Medical image rendered with nine different operators.

4.1 Two Paired-Comparison Experiments

Two paired-comparison experiments were conducted to judge both the observers’ preference and the scientific usefulness of the images. The goal of the first experiment was to determine which encoding schemes rendered the high dynamic range images in a more preferable way. In this task, observers were instructed to choose the image that they preferred in each pair. In the second experiment, the same stimuli were used but the observers were instructed to choose the image in each pair that they considered to be “more scientifically useful.” Observers were allowed to use their own criteria for making these judgments.

Twenty-five observers (6 female and 19 male ranging in age from 23 to 43 with a mean of 33) with normal color vision participated in the experiment. (Color vision was only a factor for the images encoded by the Spiral Rendering method.) Eighteen of the observers were students, staff, and faculty of the lab. Although these observers have experience in performing visual experiments of this nature and judging image quality using pictorial images, they should be considered naïve in regard to evaluating the scientific imagery used here.

The six images shown in Figure 1 were processed through the nine operators,
producing 36 possible pairs for each image and 216 pairs for each experiment. These images were then randomly displayed side by side. The images were displayed until the observer made his judgment. There was a half-second gray interval between trials. The preference task was performed first and then the scientific usefulness of images was judged. On average, it took 40 minutes to complete both experiments.

To further elucidate the success of HDR rendering, the experiment was repeated in an online version for observers with expertise for a particular image type. Twenty-one radiologists participated in evaluating the success of rendering the Medical image. Since the experiment was performed on the web for convenience, the Medical image was processed by the nine different operators adjusted for a typical output device characteristics (sRGB with a gamma of 2.2) and again evaluated in a paired-comparison paradigm. Every possible pair of processed images (total of 36 pairs) was presented and the expert opinions from 21 radiologists were collected by asking them to choose which image from each pair would be a more useful based on their expertise.

It is recognized that image preference may not be a metric of image quality that will determine the usefulness of the various algorithms. In studies of image quality using pictorial imagery, indices of image quality can be determined based on preference or accuracy. For the types of imagery often encountered in science, the imagery has no visual equivalent to compare the accuracy of the reproduction. Therefore in order to compare these results to other studies, image quality based on preference was include. Image quality based on "scientific usefulness" will depend in the specific tasks for which the images are utilized. Therefore the results of these experiments should be considered preliminary. In this regard, however, it is interesting to compare the results from the naive observers to the radiologists.

4.2 Target Detection Experiment

A third psychophysical experiment was performed to measure how the change in contrast tone mapping due to the various operators affected the detection of a target as measured by the amplitude of the target in the raw image data. This experiment used a two-alternative forced-choice method of constant stimuli paradigm to find the threshold for detecting embedded noise in the Medical image. The task of target detection can be considered as a way of determining the change in detectability of a “tumor" embedded in the image. It should be pointed out that we are using the term "noise" to describe the spatial characteristics of the target used in this experiment. Typically, however, the “tumor" is considered the signal that is embedded in the noise of the anatomical background. Because we are measuring a sensory threshold, rather than actually evaluating the ability of the diagnostician to detect a tumor, only two observers participate in this experiment, as is typical in lower-level sensory experiments of this type.

Thresholds, in terms of the original digital values in the image data, were measured for three different targets in the image as shown in Figure 3. Each of the targets had a different spatial size and location in the image. The targets were constructed by creating normally distributed random noise and then multiplying this with a Gaussian envelope to reduce the sharp onset of the noise. The size of the Gaussian filter was set to 5, 8, and 10 pixels on 15×15, 20×20, and 30×30 noise patches for high-, mid-, and dark-tone regions, respectively. The range of the target
amplitudes varied depending on the rendering algorithm and was determined based on pilot experiments. The targets subtended 0.31°, 0.37°, and 0.46° of visual angle for the small, medium, and large target, respectively, at a fixed viewing distance of 1m where the image subtended 3.8°. The targets were placed at a fixed location for each of the three different lightness areas, dark- (large), mid- (medium), and high-tone (small) areas separately. For each target, a series of images were precomputed with different amplitudes of noise added to the original image data. These images were, in turn, processed through the nine operators. These images were displayed side by side with an image processed with the corresponding operator without the target, creating 189 pairs. Each target image was presented randomly with the corresponding rendered operator without the target 60 times. The observer’s task was to choose which image had the target. The images were displayed until the observer made a decision. The images were presented on the neutral gray background in a darkened room. A half second blank field was presented between trials. When the target amplitude was not detected, the observer would have a 50% chance of correctly guessing the image that contained the target. The experiment was divided into several sessions to complete the task. This within-subject design was analyzed using Probit analysis for each subject participated in the experiment to determine the corrected-for-chance 50% threshold for target detection. Two subjects participated in the experiment.

5. RESULTS AND DISCUSSION

5.1 Paired-comparison Analysis

The paired-comparison data was analyzed using Thurstone’s Law of Comparative Judgments (Case V) [Engeldrum 2000] which results in interval scales of the judged attribute. For the preference task, observers were asked to choose which of the two images they preferred in terms of overall image quality. For scientific usefulness, no specific criteria were given to observers. They had to decide what is meant by “scientifically useful.” The image preference and judged scientific usefulness of all images are shown in Figure 4. The error bars on all plots were calculated in terms of interval scale units for a 95% confidence interval [Montag 2004]. Both figures indicate that performance of each operator depends on the image type. Comparison of these graphs also shows the different pattern of response between the two tasks.

This distinction is more apparent in average performance data shown in Figure 5. As shown in Figure 6, the low correlation ($R^2 = 0.31$) between the two sets of
results demonstrates that the observers were using different criteria for the two tasks. However, the images processed using the Zone System were judged high both in preference and scientific usefulness compared to the other operators. The Local Sigmoid function showed the most prominent changes between the two tasks. It can be noted that the Local Sigmoid method suffers from some halo artifacts, as is a common for spatially varying operators. Observers may have not preferred the images processed by the Local Sigmoid method due to these artifacts but may have found that they revealed data that were more scientifically useful.

Individual variability for the Astronomical image and the Pictorial image is plotted using diagrams that show the observer’s response patterns in Figure 7. Individual observer data is shown along the rows and the columns represent the operator types. A box with a lighter shade indicates that the operator in that column was chosen more frequently in the experiment than the other operators. Therefore, white boxes show often chosen operator types and black boxes show rarely chosen types. The appearance of vertical stripes indicates the consistency of responses.
Evaluating Tone Mapping Algorithms

Fig. 5. Averaged results of paired-comparison experiments over image type for Preference (left) and Scientific Usefulness (right).

among the observers. For the Astronomical image, Figure 7(a), illustrates that observers agreed well on their preference judgments but not on their judgments of scientific usefulness. By contrast, Figure 7(b), shows similar individual agreement on both preference and scientific usefulness task for the Pictorial image. Similar analysis (not shown) for the other images demonstrate that these results depend on image type.

The judgments of the scientific usefulness of the the Medical image of the non-radiologists is compared to the radiologists’ results in Figure 8 to determine if they have similar opinions on the success of the rendered images. This comparison
demonstrates the similar trends \((R^2 = 0.80)\) in opinion between the two groups; the best three operators are the Zone System, Bilateral Filtering, and Local Sigmoid and the worst two operators are the Spiral and Linear methods. However, the results of Radiologists, Figure 8 (left), show greater distinction between the operators with localized adaptation feature than the ones reproduced by the operators with global mapping. The diagrams shown in Figure 9 indicate that two groups share similar striped patterns. However, the results of Radiologists, Figure 9 (left), show a more apparent striped pattern exhibiting greater consistency among the experts. The dark vertical lines in these diagrams reveal that the Spiral and Linear methods are judged as less useful by the both groups. To the extent that the na"ıve observers results agree with the results from the radiologists, we can presume that the notion of "scientific usefulness" may have some degree of validity.

As mentioned above, the images for this experiment were rendered assuming typical output device characteristics (sRGB). We did not control for the actual display or viewing conditions used by the participants of this experiment. It is therefore possible that differences in the displays or viewing conditions could have introduced confounding influences on the results. Therefore, it is remarkable to note the consistency in the results from the radiologists. Although the experiment did not directly measure the usefulness of these renderings for the requisite task that radiologists view these types of images, namely detecting anomalies, the degree of consistency in the results would indicate that further study of the feasibility of these techniques is warranted.

The Zone System performed well for the majority of the tested images. However, the results for the Infrared image were an exception. As shown in Figure 10, the performance of the operators for the Infrared image can be divided into two groups.
Except for iCAM, operators with local contrast feature were judged worse than the ones with global operators so that simple linear mapping renders the image better. Figure 11 shows the plot of the Infrared image’s pixel values for the Linear rendering versus the Zone System (left) and versus iCAM (right). The Linear vs. Zone System and Linear vs. iCAM plots show how the operator with spatial filtering rendered the image compared to one without. For this particular image, the relationship between Linear and Zone System can be explained by a simple gamma curve. The shadows are more compressed and the highlights are more expanded in the Linear method compared to the Zone System. However, in iCAM, only the shadow regions are expanded while the other regions are relatively unaffected. These results depend on the spatial structure of the image.

The Photoshop Autolevels method shows the worst performance for the pictorial image (see Figure 12), though it performs well on average (see Figure 5). This result
Fig. 10. Plot of paired-comparison result for Infrared image: Preference (circles) and Scientific usefulness (triangles).

Fig. 11. Comparing processed Infrared image; Linear vs. Zone System and Linear vs. iCAM.

might be explained by the histogram of the image (see Figure 1). The histogram shows that the majority of pixels are located at the extremely low end and only a relatively small number of pixels are dispersed over the complete range. Photoshop tends to produce better results with images that have a wider distribution of pixel values, such as the radar image. A simple method for rendering HDR pictorial imagery is to apply a 99 percentile clipping and a gamma correction. These techniques are simple but powerful enough to obtain acceptable reproduction.

The Spiral encoding is the only operator that adds color to the image. This operator can be treated as a linear $L^*$ mapping since the digital values were first mapped linearly to $L^*$ values, and then chroma and hue values were added to the monochrome image. Observers tend to favor color over monochrome image when tone mapping is acceptable. However, this tendency diminishes when judging the scientific usefulness, see Figure 5. Due to the limitation of tone mapping, spiral encoding can't reveal much hidden information. If other tone mapping techniques can be combined with color, the performance may show a possible increase.
The Local Color Correction operator was judged as the worst method to use for rendering the HDR images on average and especially for the Radar and Infrared image. However, this method performed well for the Pictorial and Astronomical image. It is better than operators with global mapping but not good enough to compare with operators with local contrast mapping function.

The performance of iCAM was neither excellent nor terrible. The results are somewhat expected since iCAM is intended to render a pictorial scene truthfully rather than enhancing it. The aim of iCAM is accurate prediction of a variety color appearance phenomena that mimic human perception. Experiments on accuracy, which is not possible for scientific imagery, can be conducted to support this hypothesis by comparing pictorial renderings with original scenes.

5.2 Target Detection Analysis

The target detection experiment was conducted to measure the detectability of an embedded target in the Medical image to demonstrate the effect of the algorithms on target detection. It is obvious that the spatial structure and tone scale mapping of the images and their resultant renderings will introduce distortions that will effect target detection. Therefore, the characteristics of the targets and their surrounding contrasts in the image should be taken into account when determining the appropriate rendering algorithm. In theory, better algorithms will allow detection of targets with low amplitude regardless of the surrounding local contrast.

Figure 13 shows the plot of psychometric function (frequency-of-seeing curve) with 95% fiducial limits for the analysis of the iCAM processed image. The results from the two subjects are similar to one another. Some of the plots show larger fiducial limits than the others indicating lower levels of precision about the estimated threshold from the Probit analysis.

The threshold results are shown in Table II. The threshold results of the two subjects are similar indicating that we are measuring a low-level threshold response. These results are not meant to indicate that one or another algorithm can nec-
Fig. 13. Plots of psychometric functions for the iCAM processed image: Subject one (top), Subject two (bottom). The asterisks are the observed proportions of detection at each target amplitude, the solid line is the psychometric function determined by Probit analysis, and the dashed lines are the 95% fiducial limits. From left to right, High-, Mid-, and Dark-tone targets.

Lower thresholds indicate detection of the target at lower amplitudes. As is illustrated by the Table II and Figure 14, the results differ depending on the target size and location. For the high-tone area, the Local Sigmoid method increase detectability and the Zone System has high threshold values, even though the Zone System was judged as being most useful. However, for the mid- and dark-tone area, the Zone System and the Local Sigmoid method shows the best detectability with the lowest threshold values and the Linear and the Spiral method shows the highest threshold values representing the worst detectability. These results coincide with the scientific usefulness paired-comparison results. In general, the target amplitude must be increased as the background increase which is consistent with Weber’s Law. The pattern of detectability across the different algorithms are more similar for the targets embedded in mid- and dark-tone areas than the high-tone area and correspond more closely with the results of the scientific usefulness paired-comparison experiment.

6. CONCLUSION

High-dynamic-range imaging is a very attractive way of capturing real world appearances since it allows the preservation of more complete information about the luminance values in the scene. Nine operators used for the display of HDR pictorial imagery were applied to the display of non-pictorial image from a variety of scientific disciplines. The underlying principle is that by applying these HDR operators, more information can be conveyed because local perceptual contrast in a wider range of the scene is preserved by automatically adjusting the intensity in the image based on the image content.

Two paired-comparison psychophysical experiments were performed to evaluate which algorithms produced the most preferred images and images that were
Table II. Results of the target detection experiment. The thresholds, which are the target noise amplitudes in digital values for the original image data, were determined at the 50% level of detection.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Subject 1</th>
<th></th>
<th>Subject 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Mid</td>
<td>Dark</td>
<td>High</td>
</tr>
<tr>
<td>Linear</td>
<td>42.8</td>
<td>21.8</td>
<td>28.9</td>
<td>38.0</td>
</tr>
<tr>
<td>iCAM</td>
<td>70.8</td>
<td>14.4</td>
<td>4.8</td>
<td>57.4</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>41.9</td>
<td>17.0</td>
<td>13.1</td>
<td>42.4</td>
</tr>
<tr>
<td>Spiral</td>
<td>34.8</td>
<td>17.9</td>
<td>43.5</td>
<td>39.5</td>
</tr>
<tr>
<td>Local Sigmoid</td>
<td>16.6</td>
<td>8.7</td>
<td>4.7</td>
<td>21.4</td>
</tr>
<tr>
<td>Local Correction</td>
<td>90.6</td>
<td>15.1</td>
<td>4.2</td>
<td>78.2</td>
</tr>
<tr>
<td>Bilateral</td>
<td>71.8</td>
<td>12.0</td>
<td>3.8</td>
<td>53.9</td>
</tr>
<tr>
<td>Zone System</td>
<td>87.8</td>
<td>11.2</td>
<td>3.7</td>
<td>84.1</td>
</tr>
<tr>
<td>Photoshop</td>
<td>32.9</td>
<td>15.4</td>
<td>17.2</td>
<td>36.4</td>
</tr>
</tbody>
</table>

Fig. 14. Plots of the thresholds for the detection experiment; subject one (top), subject two (bottom).
considered scientifically useful. Although the Zone System has the best performance on both average preference and scientific usefulness, the results of the paired-comparison experiments suggest that different encoding schemes might be optimal depending on the image type. There was little correlation between preference and scientific usefulness indicating that observers used different criteria for the two tasks.

To further explicate the success of the HDR rendering on these data, an online experiment was performed with radiologists evaluating the rendered MR images and these results were compared to the data from the non-expert observers. Both radiologists and non-radiologists judged the Medical image rendered by the Zone System as having the highest quality for scientific usefulness when compared to the other operators, however, the radiologists exhibited greater consistency in their judgments.

The effects of image distortion introduced by the rendering operators were investigated using a target threshold detection paradigm. The results of the high-tone area target indicate that the detectability does not strictly correspond with the results of the paired-comparison experiment. However, the threshold results of the mid- and dark-tone area targets show a somewhat closer relationship with the scientific usefulness paired-comparison experiment. It is instructive to note that in applications where locating embedded targets is a goal of the visualization, such as finding tumors in medical images, the spatial structure and tone-mapping algorithm may affect the detectability of the targets. Therefore, the rendering algorithm should be properly matched or tuned for the specific type of imagery being used.

More research is needed to evaluate the effectiveness of these algorithms for the specific tasks in which the images are utilized, especially in the field of radiology. As mentioned above, in MR imaging, specifically, interactive windowing is used to present the relevant range of information linearly within the range of the display. More research is needed to determine whether algorithms such as the ones presented here can provide value in automating the presentation and increasing the amount of usable information. As radiological imaging is not a monolithic discipline, we cannot presume that the same utility can be derived in different areas such as X-ray radiology, ultrasound, or MRI, to name a few, nor can we determine that the same algorithms will provide such utility across these domains.

A main goal of this research is to develop perceptually based schemes for automating the encoding scientific imagery that facilitate its interpretation. Here we focused on imagery that was univariate along the two spatial dimensions. For this type of imagery, it is likely that the addition of color may facilitate interpretation by either extending the perceived range of the data or by identifying image characteristics. Additional research is therefore required for determining rules and techniques for the display of multidimensional graphical information, as well as how to use color appropriately as a tool.

The results of this experiment indicate that, despite the overall trends observed in the average results, the particular algorithm needed for optimal scientific image rendering is dependent on both the image type and the purpose of the visualization. Just as in pictorial imagery where the rendering intent may be image fidelity or image preference, for scientific imagery the intent may also vary. Because of
the dependence of image type on the results, it seems likely that expert observers may have different criteria for judging scientific usefulness than novices. Therefore more research using experts, as was done with the Medical image, is recommended. Using the results of this research as a starting point, future experiment can be conducted involving other operators (especially, those with spatially varying mapping functions) and additional images that sample the range of image types used within a scientific domain. Additional research and development is necessary for better models that mimic human perception and facilitates better data mining and interpretation.

ACKNOWLEDGMENTS
The authors are sincerely thankful to Garrett Johnson, Jiangtao (Willy) Kuang, and Hiro Yamaguchi for their help, valuable discussions, and comments throughout the research; and all the observers who participated in the psychophysical experiments for their time and support

REFERENCES


