A Proposed Methodology for Evaluating HDR False Color Maps

AHMET OĞUZ AKYÜZ and OSMAN KAYA, Middle East Technical University

Color mapping, which involves assigning colors to the individual elements of an underlying data distribution, is a commonly used method for data visualization. Although color maps are used in many disciplines and for a variety of tasks, in this study we focus on its usage for visualizing luminance maps. Specifically, we ask ourselves the question of how to best visualize a luminance distribution encoded in a high-dynamic-range (HDR) image using false colors such that the resulting visualization is the most *descriptive*. To this end, we first propose a definition for descriptiveness. We then propose a methodology to evaluate it subjectively. Then, we propose an objective metric that correlates well with the subjective evaluation results. Using this metric, we evaluate several false coloring strategies using a large number of HDR images. Finally, we conduct a second psychophysical experiment using images representing a diverse set of scenes. Our results indicate that the luminance compression method has a significant effect and the commonly used logarithmic compression is inferior to histogram equalization. Furthermore, we find that the default color scale of the Radiance global illumination software consistently performs well when combined with histogram equalization. On the other hand, the commonly used rainbow color scale was found to be inferior. We believe that the proposed methodology is suitable for evaluating future color mapping strategies as well.

CCS Concepts: • **Human-centered computing → Visualization techniques**; • **Computing methodologies → Image processing;**

Additional Key Words and Phrases: HDR imaging, false color, visualization

ACM Reference Format:

Ahmet Oğuz Akyüz and Osman Kaya. 2016. A proposed methodology for evaluating HDR false color maps. ACM Trans. Appl. Percept. 14, 1, Article 2 (June 2016), 18 pages.
DOI: http://dx.doi.org/10.1145/2911986

1. INTRODUCTION

In image processing and computer graphics, false (or pseudo) colors are commonly used to visualize intensity distributions. If the underlying signal to be visualized is in the photometric domain, then they can be used to visualize luminances by assigning different colors to different degrees of the signal. Alternatively, if the signal is in the radiometric domain, as in infrared imaging, then they enable us to visually observe the radiance distributions that would otherwise be invisible. Regardless of the application domain, there are many degrees of freedom as to how to represent a given signal using false colors. As a mini-experiment, we invite the reader to try to judge which of the 12 different visualizations shown in Figure 1 most accurately conveys the luminance distribution in the scene? In this

Q1 Author's address: A. O. Akyüz, Middle East Technical University, Department of Computer Engineering, Ankara, 06800, Turkey; email: akyuz@ceng.metu.edu.tr.

DOI: http://dx.doi.org/10.1145/2911986

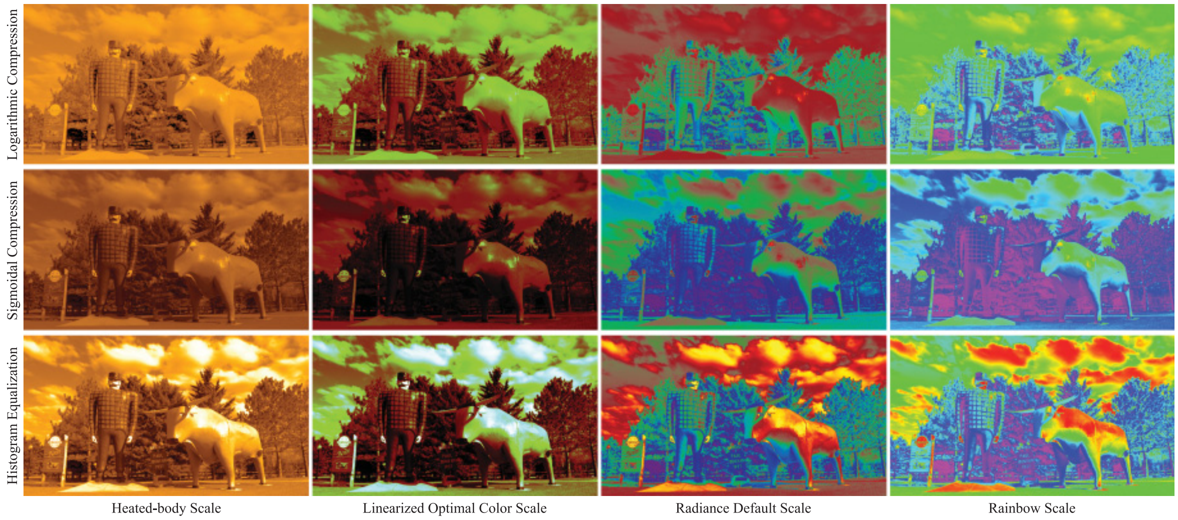


Fig. 1. The same high-dynamic-range image can be visualized in false color in different ways. One can change the type of compression for luminances and the color scale from which the colors are sampled. Here, 12 different combinations are shown. Which one is more informative? Our study aims to answer this question through subjective and objective evaluations using a large number of images.

article, we argue that the answer of this question is not obvious, and research is needed to determine the best way to visualize a luminance map using false colors.

The use of color maps in visualization is prevalent. It involves assigning different colors to different degrees of a modality with the goal that the color map conveys extra information that is not directly visible in the original signal. However, injudicious use of color maps can cause confusion rather than facilitating understanding [MacDonald 1999]. Therefore, it is critical to make the right choice for the task at hand.

There are many tasks that require studying luminance distributions to extract meaningful scene information. For example, Theodor and Furr apply the techniques of high-dynamic-range (HDR) imaging to study fossils [2009]. Cultural heritage and archeology also benefit from working directly with luminance maps as opposed to low-dynamic-range (LDR) footage [Happa et al. 2010]. Similarly, many other fields such as structural engineering [Grinzato et al. 2009], architecture [Cai 2013], medical imaging, and forensics [Brown et al. 2010] also use HDR luminance data for the purpose of scene and/or object analysis. It is clear that an enhanced depiction of luminance and/or radiance data using false colors can be beneficial to these disciplines.

Beltran et al. [2005] evaluate HDR photography as an alternative to taking precise measurements using spot meters. The authors found that the difference between actual luminance measurements and those obtained from HDR images are minimal. They therefore argue that HDR photographs can be used to rapidly assess the lighting requirements of various environments. False coloring, to this end, may facilitate quick inspection of lighting values in such photographs.

Despite its importance, we often observe that false color visualization is performed in an ad hoc manner. The typical workflow involves compressing the luminance data using a logarithmic function followed by mapping of the colors into an arbitrary color scale—and often the rainbow color scale. We argue that both of these choices, namely the compression function and the color scale, are critical, and ad hoc choices can impair the quality of the visualization. Our goal in this article is to allow making these choices in a more principled manner by studying which compression functions and color scales produce more descriptive visualizations.

To this end, we first propose a definition of *descriptiveness*. This definition involves maximizing the number of just-noticeably-different colors within a false color image. We then carry out a psychophysical experiment to compare the effectiveness of 12 sample false coloring strategies commensurate with this definition. Next, we develop an objective metric that correlates well with the results of the subjective experiment. Then, by using this metric, we evaluate these sample false coloring strategies on more than 100 images. Finally, we conduct a second psychophysical experiment involving images of various scenes using the best performing methods in the first two evaluations. The results indicate that histogram equalization outperforms all other compression methods. As for the color scale, the default color scale used in the Radiance ray tracer is found to be superior, especially when combined with histogram equalization as the compression function.

The rest of this article is organized as follows. In Section 2 we review the related work for using color in visualization in general and then more specifically for HDR image visualization. In Section 3, we outline a framework for producing false color maps from HDR photographs. Next, in Sections 4, 5, and 6, we describe our subjective and objective evaluations. Finally, we conclude our article with a discussion, conclusions, and ideas for future research directions.

2. RELATED WORK

2.1 Tone Mapping

Tone mapping or tone reproduction is the process of reducing the dynamic range of an HDR image to prepare it for display on LDR display devices [Reinhard et al. 2010]. Many tone mapping operators (TMOs) have been developed since the introduction of the problem into computer graphics [Tumblin and Rushmeier 1991]. These are generally classified as global and local operators. Global operators preserve the monotonicity of the luminance values in that higher luminances in the input image get mapped to higher (or equal) luminances in the compressed image. Thus, their compression function can be represented as a curve. The histogram adjustment method by Ward et al. [1997] and the global photographic operator by Reinhard et al. [2002] are two examples of notable global operators. Local operators, on the other hand, can alter the luminance values such that monotonicity is not preserved. This often leads to better visibility in high-contrast image regions but makes it impossible to represent their compression using a single curve. Among a large number of local operators, fast bilateral filtering [Durand and Dorsey 2002], gradient domain compression [Fattal et al. 2002], and the local photographic operator [Reinhard et al. 2002] are representative examples. There are a large number of TMOs that are beyond the scope of our review. We refer the reader to Reinhard et al. [2010] for a detailed coverage of the subject.

2.2 Color in Visualization

Color is an indispensable element of data visualization. Although its correct use can greatly enhance the effectiveness of the visualization, its ad hoc or incorrect use can cause further confusion [MacDonald 1999]. When using color for visualization, one of the first choices that needs to be made is the selection of a color scale. For univariate data, an appropriate color scale should preferably satisfy the following properties [Trumbo 1981; Levkowitz and Herman 1992]:

Order: The colors chosen to represent a set of data values must exhibit a perceived order that is congruent with the order of the data values themselves.

Uniformity: The perceived difference between the colors should correspond to the difference in magnitude of data values.

Continuity: The color scale should not create artificial boundaries that do not exist in the data. In other words, the color scale should be (perceived as) continuous.

Satisfying all three properties does not necessarily imply that a color scale is ideal for any given task. For instance while the linearized grayscale is continuous, uniform, and has a natural order, it displays a low contrast between its colors and suffers from visibility problems in dark regions, limiting its use for visualizing high-contrast information. We refer the reader to an excellent survey by Silva et al. [2011] for comprehensive guidelines of using color in visualization.

2.3 Color Maps for HDR Images

Although color maps for HDR images are commonly used for visualizing luminance distributions, to our knowledge there is no scientifically validated way to represent an HDR image in false colors. The most well-known tool that accomplishes this task is the Radiance software [Larson and Shakespeare 1998], which contains several color palettes and two methods of compression, namely linear and logarithmic. Based on the information provided to us by the developers,[1] the *SPEC* color scale represents spectral colors (i.e., the rainbow scale). *HOT* is a heated-body (thermal) scale that goes from black to white by passing through red and yellow. The *ECO* scale is borrowed from Ecotech, an environmental simulation software. *PM3D* is borrowed from Gnuplot [Williams et al. 2010], and the default scale, *DEF*, is a mixture of thermal and spectral scales.

In a more recent study, Akyüz [2013] performed a preliminary experiment to evaluate the performance of several compression functions, namely linear, logarithmic, and sigmoidal scaling, when used together with the rainbow scale. The participants were asked to rank these methods based on how well they represent the luminance distribution in three HDR images. For all three images, sigmoidal scaling was found to outperform the other methods. However, this experiment was limited in the sense that it only involved a single color scale and was based on purely subjective opinion.

3. COLOR MAPPING FRAMEWORK

In this section, we describe our framework that we used to convert an HDR image into false colors. We assume that the HDR image represents a luminance distribution stored in a linear RGB color space. Therefore, we first compute the luminance values by using an appropriate linear transformation that depends on the actual color space. For instance, if the HDR image is stored in the sRGB color space, its luminance values (Y) can be computed by the following formula [ITU (International Telecommunication Union) 2002]:

$$Y = 0.2126R + 0.7152G + 0.0722B. \tag{1}$$

The remaining process involves two stages. The first one is the compression of the luminance values and the second one is the mapping of the compressed values to color values ($\mathbf{C}$) in a given color scale.

3.1 Compression Stage

In the compression stage, one can apply an initial transformation to the luminance data to reduce its dynamic range. Otherwise, the ensuing color mapping would simply yield large regions of uniform color resulting in a flat visualization. While any TMO can be used to compress the luminance data, local operators may be unsuitable, as they do not preserve the monotonicity of luminance values. Here, we describe three global compression strategies:

[1]Personal communication with Greg J. Ward and Axel Jacobs.

Logarithmic Scaling (LOG): Logarithmic scaling, which approximates the human visual response to light [Drago et al. 2003], is defined as

$$f_{\log}(Y) = \frac{\log(Y+\epsilon) - \log(Y_{min}+\epsilon)}{\log(Y_{max}+\epsilon) - \log(Y_{min}+\epsilon)}, \tag{2}$$

where a small epsilon value (ϵ) is introduced to avoid singularity for black pixels. In this article, we used $\epsilon = 10^{-6}$.

Sigmoidal Compression (SIG): Sigmoidal compression was originally proposed by Naka and Rushton [1966] as a model of biological systems and was later used in a well-known tone mapping operator due to its simplicity and ability to produce natural looking images [Reinhard et al. 2002]. Its compression curve also mimics the S-shaped curves used in traditional photography when plotted on a logarithmic luminance axis:

$$f_{\text{sig}}(Y) = \frac{\alpha Y/\bar{Y}}{1+\alpha Y/\bar{Y}}, \tag{3}$$

where α denotes a user-defined key value and $\bar{Y}$ is the log-average luminance:

$$\bar{Y} = \exp\left(\frac{1}{N}\sum_{x,y}\log(Y(x,y)+\epsilon)\right), \tag{4}$$

with N representing the number of pixels in the image. We set $\alpha = 0.18$ as a generally used default value [Reinhard et al. 2002].

Histogram Equalization (HIS): Histogram equalization redistributes the luminance values such that each bin contains equal number of pixels [Gonzalez and Woods 1992]. In a false coloring framework, this means that each color value will be used for approximately equal number of times. Histogram equalization can be represented by using the following formula assuming an 8-bit output range:

$$f_{\text{his}}(Y) = \text{round}\left(255\frac{\text{cdf}(Y)-\text{cdf}_{\min}}{N-\text{cdf}_{\min}}\right). \tag{5}$$

Here, cdf(.) represents the cumulative distribution function of luminance values and $\text{cdf}_{\min}$ is the minimum non-zero value of the cdf.

3.2 Color Mapping Stage

In this stage, a false color image is produced by mapping the compressed luminance values into color values from a given color scale. Algorithm 1 is used for this purpose:

ALGORITHM 1: Color Selection Algorithm for Logarithmic and Sigmoidal Compression

$Y' = f(Y)$
for $i = 0 \to 255$ **do**
    $bin[i] = \frac{i}{255}(Y'_{max} - Y'_{min})$
end
for ***each*** pixel $Y'_{m,n} \in Y'$ **do**
    find k **where** $|Y'_{m,n} - bin[k]|$ *is minimum*
    $\mathbf{C_{m,n}} = PALETTE[k]$
end

In this algorithm, $f(.)$ can be substituted with $f_{\log}(.)$, $f_{\text{sig}}(.)$, or $f_{\text{his}}(.)$ for different compression functions. $\mathbf{C_{m,n}}$ represents the false color value selected from the given color scale.

Fig. 2. Color scales evaluated in our study.

3.3 Color Scale Selection

Selection of a good color scale is critical for visualization. As described in Section 2.3, several color scales are commonly used for visualizing HDR images in false color. Among these, the rainbow scale and the heated-body scale are commonly used for other visualization tasks as well. Therefore, in this study we chose to include the following four color scales. Each scale is represented by a palette of 256 distinct colors (Figure 2).

Rainbow scale (RBS): This scale is one of the most commonly used scales in the literature. As the ordering of the colors is roughly based on their wavelength, it is also called the spectral scale. The palette of this scale is generally produced by varying the hue attribute in a color space such as HSV while keeping the other attributes constant. We have used hue angles between 0° (red) to 270° (magenta) to represent high and low luminances, respectively.

Heated-body scale (HBS): This scale represents a progression of colors going from black to white while passing through orange and yellow. The hue angle varies approximately between 15° and 60°. The advantage of this scale is attributed to the fact that the human visual system is most sensitive to luminance changes in that portion of the spectrum. We have used the perceptually linearized version of this scale, in which luminance difference between different color values correspond to roughly equal brightness differences.

Radiance default color scale (DEF): This is the default false color scale used in the Radiance software [Larson and Shakespeare 1998]. The scale is developed by Larson to represent a mix between the heated-body and rainbow scales. It was designed to maximize the number of named colors while still depicting a progression from cold to hot [Larson 2013].

Linearized optimal color scale (LOCS): This scale is designed to create a maximum number of just noticeable differences (JNDs) while preserving a natural order [Levkowitz and Herman 1992]. To our knowledge, this scale has not been used for visualizing HDR images in false color. This scale is also perceptually linearized.

Each of these four color scales satisfy the desired properties discussed in Section 2.2 to different extents. The HBS and LOCS satisfy all three of the order, uniformity, and continuity properties. RBS satisfies the continuity and order property, but the latter requires observers to be familiar with the spectral progression of colors. The DEF color scale, on the other hand, only satisfies the continuity property.

The 3 compression functions and 4 color scales gives rise to 12 false coloring strategies. In the following, we will refer to these strategies using the abbreviations shown in parenthesis. For instance, histogram equalization with the Radiance default scale will be identified as HIS-DEF. Other methods will be denoted similarly.

4. PSYCHOPHYSICAL EXPERIMENT ONE

We conducted a psychophysical experiment to evaluate the effectiveness of different false coloring strategies. Our experiment was was aimed to answer the following two questions: (1) Which of the aforementioned strategies is better for visualizing an HDR image in false color and (2) whether a quantitative metric can be derived that correlates well with the human observers' responses so any

Fig. 3. A tone mapped visualization of the HDR image used in the subjective evaluation. Image is retrieved from the HDR photographic survey [Fairchild 2007].

future strategy can be objectively evaluated using this metric. To this end, we first need to define the characteristics of a good false coloring strategy.

4.1 Criterion of Evaluation

First, it should be kept in mind that which false coloring strategy is the best depends on the application at hand. Certain compression functions and color scales may be more appropriate for certain applications. This is similar to the issue faced when evaluating tone mapping operators in that which TMO is the best depends on the purpose of tone mapping. For instance, a TMO used for medical imaging is likely to be desirable if it preserves visibility of small scale details. On the other hand, a TMO used for entertainment is likely to be desirable if it preserves, and even exaggerates, contrast at the cost of losing small details [Akyüz and Reinhard 2008]. Therefore, the studies that evaluate TMOs usually define the criteria according to which the tone mapping quality should be judged [Drago et al. 2002].

We adopt a similar approach in the current study. We define our criteria as if the luminance of two regions perceivably differ in the original image; they should be mapped to perceivably different colors in the false color image. In other words, we expect a false color visualization to convey noticeable luminance differences. This may be compared to preserving visibility during tone mapping. However, we also expect the order of luminances to be preserved. That is a lower luminance pixel should not be represented by a color that suggests a higher luminance than another pixel which actually has higher luminance.

4.2 Stimuli

In our experiment, we used a single calibrated HDR image depicting a scene of extremely high dynamic range (Figure 3) taken from a public HDR image database [Fairchild 2007]. The actual scene had a contrast ratio of a 1,000,000:1 and even when recorded the resulting HDR image retained a contrast ratio of 800,000:1 (the drop was due to flare). Eighteen exposures that are one f-stop apart were used to capture the scene. As shown in Figure 3, the image contains two sets of colored checkers, one in the dark region that receives no direct illumination, and the other directly illuminated by a bright light source. In addition to having an extremely high dynamic range, this image contains 48 uniform patches that can be used as test stimuli. To this end, we selected pairs of patches that are closest in luminance giving rise to 24 pairs of stimuli.

Table I. Mean Percentages of Correct Answer for Each Compression-Color Scale Combination Averaged over All Participants

	LOG	SIG	HIS	Average
RBS	61%	55%	71.1%	62.4%
HBS	65.5%	69%	76.8%	70.4%
DEF	67%	65.2%	72%	68.1%
LOCS	70%	63.7%	**81**%	71.6%
Average	65.9%	63.2%	75.2%	

4.3 Experimental Process

During the experiment, we asked the participants to indicate which of the two patches in a randomly selected pair from a false color image has a higher luminance. To prevent other factors, such as the proximity of a patch to the light source, from affecting participants' decisions, all parts of the image were masked out with a neutral gray color except the patches being compared. On the top-left corner of the screen the current color scale was shown to remind the participants of the progression of the colors with luminance. The participants indicated their responses by clicking on the patch that appears to have a higher luminance. After each response, a new random pair was automatically shown from the remaining stimuli. To avoid confusing the participants by rapidly switching between different color scales, all pairs from one scale were first consumed before proceeding with the next scale. The order of the compression functions was randomized within each color scale. The scales were randomized for each participant. The duration of the experiment took about 30 minutes for each participant.

We also used an eye-tracker, SMI Red 60/120Hz, to measure the duration that each participant looked at the color scale. This was used to rank the scales in terms of intuitiveness, as a less intuitive scale may require studying of the palette for a longer time.

All stimuli were shown on an NEC SpectraView Reference 241W monitor calibrated to the sRGB profile using an X-Rite i1Display Pro colorimeter. The peak display luminance was set to 80cd/m^2 for full sRGB compliance. The black level was measured as 0.5cd/m^2. The participants viewed the display in a dark room from a distance of approximately 70cm. No head mounting was used to avoid discomfort. At this distance, the angular size of a center pixel was approximately 0.0221° in both dimensions.

Fourteen participants (4F and 11M) contributed to the experiment. Each participant received a brief training about the color scales and the relationship between the colors and luminance values prior to taking the experiment.

4.4 Results

The mean percentages of correct answers for each compression-color scale combination is shown in Table I. As can be seen from this table, histogram equalization together with the linearized optimal color scale (LOCS) yields the highest percentage of correct answers (81%). Sigmoidal compression with the rainbow scale yields the lowest percentage (55%), which is slightly higher than what would be obtained by chance if subjects were making random decisions (50%). We can also observe that histogram equalization-based methods outperform logarithmic and sigmoidal compression for all color scales. Logarithmic compression surpasses sigmoidal compression except with the heated-body scale. When averaged across all color scales, histogram equalization clearly outperforms the other two compression types. When averaged across all compression types, LOCS marginally outperforms HBS and DEF. However, the rainbow color scale (RBS) clearly underperforms in this task.

These observations are supported with a two-way within-subjects ANOVA test that was conducted to understand whether these differences are statistically significant. These results are summarized

Table II. Statistical Result for the User Study

Factor	Statistical result
Compression	$F(2, 26) = 28.99$, $p < 0.001$
Color scale	$F(2, 39) = 5.284$, $p = 0.004$
Compression × Color scale	$F(6, 78) = 2.101$, $p = 0.062$

Compression: HIS LOG SIG **Color scale:** LOCS HBS DEF RBS

Fig. 4. Statistical similarity groups of the user study. Items underlined by the same line are statistically similar.

Table III. Total Times in Minutes and Seconds during Which Participants Looked at the Palettes Shown in the Top-Left Corner (Accumulated over All Compression Types and Participants)

Color scale	Total time (m:ss)
HBS	3:59
LOCS	4:50
RBS	4:54
DEF	5:31

in Table II. Based on these, we can observe that both the compression type and the color scale have a statistically significant effect on the quality of the visualization ($p < 0.05$ for both). However, the interaction between the compression type and color scale was found to be marginally insignificant ($p = 0.062$). Therefore, pairwise differences between compression type and color scale combinations were not computed.

Next, we performed pairwise t-tests with Bonferroni correction to identify which compression and color scales statistically differ from each other. We found that histogram equalization is significantly better than logarithmic and sigmoidal scaling, but the latter two are statistically equivalent. As for the color scale, LOCS, heated-body, and Radiance scales formed the first group, and Radiance and rainbow scales formed the second. Figure 4 illustrates the similarity groups.

Finally, in Table III, we report the total time elapsed when participants studied the color palettes shown in the top-left corner during the experiment. This duration was the shortest for the heated-body and linearized optimal color scales. This can be expected, as their color palettes are ordered in increasing order of luminance, which facilitates making decisions. The rainbow scale had a similar timing to that of LOCS. Radiance's default color scale took the longest time for the participants to interpret the relationship between colors and luminance. This could be expected, as this color scale has the least intuitive ordering.

4.5 Discussion

The subjective study reveals that histogram equalization outperforms other dynamic range reduction methods irrespective of the color scale that was being used along with it. Next, we set out to understand whether this result can be explained by examining some low level image statistics. For example, it could be hypothesized that histogram equalization was better in this task, as it produces images with higher entropy. To this end, we experimented with several image statistics such as variance and entropy but could not find a strong correlation. That is, a false color visualization strategy with high variance or entropy did not perform well in the user study.

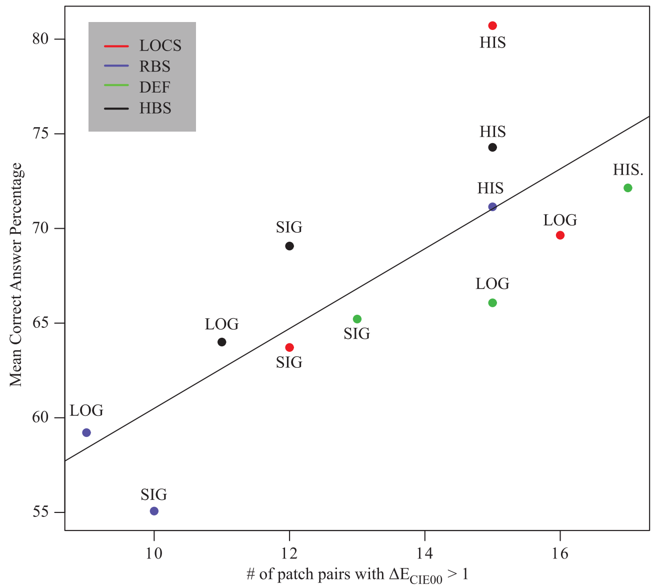


Fig. 5. Correlation between the number of patches for which $\Delta E_{CIE00} > 1$ and the mean correct answer percentage across all patches and participants. Dashed line shows the least-squares fit.

Next, we hypothesized that the best false color image must have the highest perceivable color difference between the patches that were compared. To test this hypothesis, we compared the CIEDE2000 [Sharma et al. 2005] color differences ($\Delta E_{\mathrm{CIE00}}$) between the compared patches and the mean number of correct answers given for those patches. Surprisingly, these two variables also did not have a strong correlation. Further investigation revealed that as $\Delta E_{\mathrm{CIE00}}$ values increased the number of correct answers also increased. However, as color difference grew, the number of correct answers could not go beyond 14 (the number of participants). Therefore, we slightly modified the correlation variables and compared the number of patches where $\Delta E_{\mathrm{CIE00}} > 1$ with the number of correct answers.[2] The Pearson product-moment correlation coefficient indicated a strong positive correlation between the two variables, $r = 0.77$, $n = 12$, $p = 0.0018$ (also see Figure 5). If HIS-LOCS is removed as an outlier, then the correlation coefficient increases to $r = 0.84$. This high correlation suggests that the total number of patch pairs with $\Delta E_{\mathrm{CIE00}} > 1$ in a false color image is a good indicator of perceived difference between those patches. It should be noted that the color difference value is not linearly correlated with the actual luminance difference of the patches due to the initial non-linear compression (see Mantiuk et al. [2009] for an experimental demonstration of this phenomenon). However, it can still be used to indicate the *presence* of a perceivable color difference. This observation enabled us to perform the objective evaluation explained in the next section.

5. OBJECTIVE EVALUATION

For the objective evaluation, we used the 105 images in the HDR photographic survey [Fairchild 2007]. This survey contains various images depicting different environments. To understand the diversity of the images in this survey we clustered them into six bins using the k-means algorithm. As features we have used the HSV and gradient magnitude histograms. That is, for the purpose of clustering, each

[2] $\Delta E_{\mathrm{CIE00}}$ value of 1 corresponds to the human detection threshold [Reinhard et al. 2008].

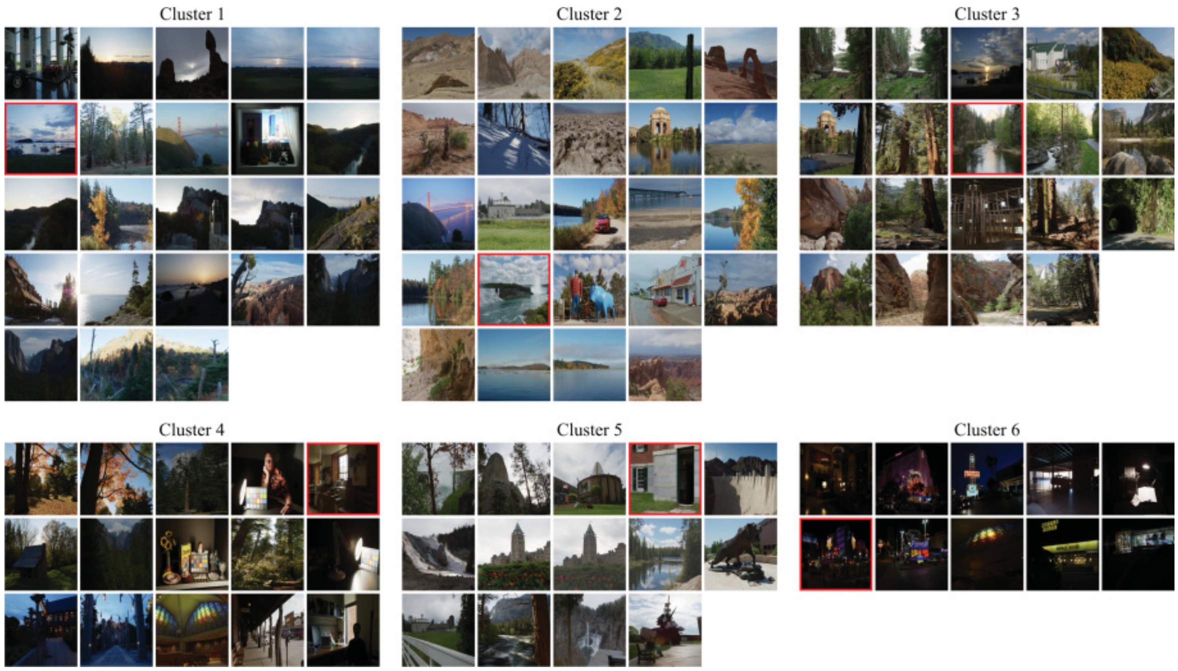


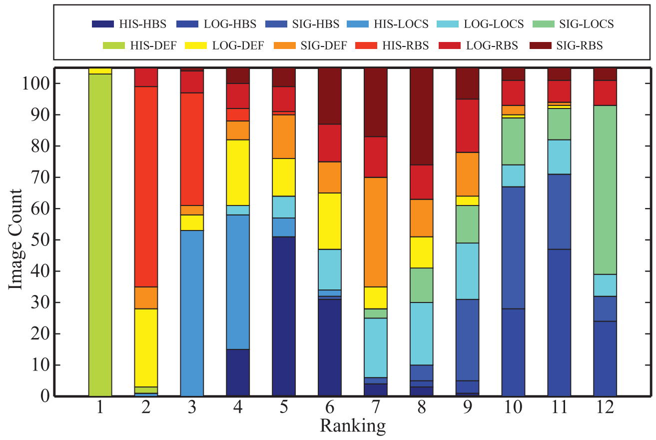
Fig. 6. The images in the HDR photographic survey [Fairchild 2007] are categorized into six clusters using a k-means algorithm according to their HSV and gradient magnitude histograms as feature vectors. See Table IV for a description of image characteristics in each cluster. The images with a red border are used for the second psychophysical experiment. In cluster order, their names are *BarHarbor*, *Niagara*, *Amikeus*, *HancockIn*, *HancockOut*, and *LasVegas*.

Table IV. Characteristics of Images in Different Clusters

Cluster 1	Mostly sunset and sunrise images with bimodal histogram distributions
Cluster 2	Daytime outdoor images with mostly blue tones
Cluster 3	Daytime outdoor images with mostly foliage
Cluster 4	Darker outdoor images and several indoor images
Cluster 5	Images containing buildings and man-made structures
Cluster 6	Night scenes

image was represented as a 60-dimensional feature vector with each component represented using 15-bin histograms [Ben-Haim et al. 2006]. The resulting clusters are depicted in Figure 6. We can see that each cluster contains images with different characteristics, although outdoor environments are more heavily represented than indoor ones (see Table IV). This clustering is performed to demonstrate the variability of the images in the objective evaluation dataset.

Each image was visualized in false color using the 12 compression type–color scale combinations discussed earlier. As our metric, we decided to use the number of pixel pairs where $\Delta E_{\text{CIE00}} > 1$. The decision to use pixel pairs instead of larger patches was motivated by the fact that the latter requires segmenting the input images into uniform patches—an operation that would be dependent on the



Mean Rankings		
1	HIS-DEF	1.0190
2	HIS-RBS	2.4476
3	HIS-LOCS	3.5714
4	LOG-DEF	4.7333
5	HIS-HBS	5.3429
6	SIG-DEF	6.6286
7	LOG-RBS	7.2952
8	SIG-RBS	7.5810
9	LOG-LOCS	8.0476
10	SIG-HBS	9.9619
11	LOG-HBS	10.6857
12	SIG-LOCS	10.6857

Fig. 7. Left: Bar plot summarizing the results of the objective evaluation. Each colored box shows the number of times each method was ranked the first, the second, and so on. Right: The mean rankings of each method.

segmentation algorithm used. To avoid such interaction effects, we opted to use pixel values directly. All visualizations are sorted according to this metric to create a ranking. We have repeated this process for each HDR image and obtained 105 rankings. The aggregate results are shown as a bar plot on the left of Figure 7 with mean rankings shown on the right of the same figure.

According to the results, histogram equalization combined with the Radiance color scale (HIS-DEF) produced the maximum number of pixel pairs with $\Delta E_{\mathbf{CIE00}} > 1$ for 103 of the 105 images in the database. As such, it was the clear winner according to the objective metric. The only two images where it came the second were *North Bubble* and *Delicate Arch*, which had relatively low dynamic ranges, 200:1 and 500:1, respectively. For those images, LOG-DEF was the winner.

The second-best method was found to be HIS-RBS, which was followed by HIS-LOCS and HIS-HBS (determined by the mean rankings of the algorithms). Here, we can see that histogram equalization approach gives the best results regardless of the color scale being used. These results support the findings of the user study where histogram equalization outperformed logarithmic and sigmoidal compressions for all color scales. However, the rankings of color scales within this compression type has changed. Whereas HIS-LOCS was the winner in the user study, it was the third best in the objective evaluation. Also HIS-DEF was the third best in the user study, and it was found to be the winner in the objective evaluation. Finally, HIS-RBS had the worst performance in the user study within histogram equalization, although it came as the second best in the objective evaluation.

The rankings of the remaining methods were more intermixed. The methods with the worst ranking were SIG-LOCS and LOG-HBS, which were found to be the third and fourth methods from the last, respectively, in the user study as well. The differences between the rankings of the methods were confirmed by Friedman rank sum test, $\chi^2(11) = 905.31$, $p < 0.001$ [Hollander et al. 2013]. To understand which methods truly differ from each other, we performed Wilcoxon post hoc tests with Bonferroni correction applied. The similarity groups at 95% significance level are shown in Figure 8.

6. PSYCHOPHYSICAL EXPERIMENT TWO

The overall correlation between the rankings of the subjective evaluation and objective evaluation was found to be 0.586 according to Spearman's rank correlation. This moderate correlation suggests that further investigation could be needed to determine the most suitable approach for false color

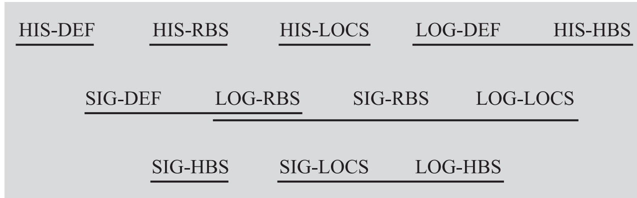


Fig. 8. Statistical similarity groups for the rankings of each compression type–color scale combination. Methods underlined by the same line are statistically similar. Rankings are given in increasing order from left to right and top to bottom.

visualization. To this end, we selected the best performing five methods, namely HIS-DEF, HIS-RBS, HIS-LOCS, LOG-DEF, and HIS-HBS, and compared them in a final experiment. Also, to help generalize this experiment for changing scene conditions, we selected one HDR image from each cluster shown in Figure 6, resulting in a total of six scenes (the selected images are indicated by a red border in the figure).

The experiment was designed as a paired comparison experiment due its increased reliability over rating, ranking, and similarity experiments [Mantiuk et al. 2012]. The participants were shown an HDR image in the middle of the screen with two different false color versions on either side. The HDR image was initially linearly mapped to the computer screen such that the mean luminance value was set to 127.5. The participants could change this scaling factor by pressing the UP and DOWN arrow keys to allow bringing different regions of the image into proper exposure. The scaled HDR image was shown after applying gamma correction by using the sRGB gamma. The display device, which was NEC Spectraview Reference 241W, was calibrated to the sRGB profile as in experiment one. The experiment was conducted in a dark room and the participants sat approximately 70cm from the display device.

The participants' task was to choose the false color image that best describes the distribution of luminance across the HDR image. On top of each false color image, the corresponding color palette was visualized to help participants interpret the meaning of the colors. The participants could choose the image they prefer by pressing the LEFT and RIGHT arrow keys, which drew a gray border around the selected the image. They could then finalize their decision and move on the next trial by pressing the ENTER key. The experiment started with a short warm-up session during which the responses were not recorded. The mean experimental duration was 21 minutes with a standard deviation of 11 minutes. A short break was given in the middle of the experiment. A total of 17 participants (7F and 10M) with normal or corrected-to-normal color vision participated in this experiment.

A complete block design was utilized in which each participant judged all stimuli. This amounted to $C(5, 2) = 10$ responses per each HDR image and $6 \times 10 = 60$ responses in total. The responses were collected in a preference for each HDR image. By summing up these individual matrices, the aggregate preference matrix was generated. The per-scene and overall results of the experiment are shown in Tables V and VI.

According to the overall results HIS-DEF was preferred the highest number of times (281). It was followed by HIS-HBS (256), HIS-RBS (204), and HIS-LOCS (195). The least-preferred method was LOG-DEF (84). This overall trend also exhibits itself in per-scene results as well. In four of six scenes, HIS-DEF was preferred the highest number of times, with HIS-HBS being the winner in the remaining two. LOG-DEF was the least preferred in all scenes as well. The preference counts of HIS-LOCS was more variable across scenes. HIS-RBS, on the other hand, was more stable but was preferred relatively fewer number of times.

Table V. Per-Scene Results Aggregated Over Participants. A: HIS-DEF, B: HIS-HBS, C: HIS-LOCS, D: HIS-RBS, E: LOG-DEF. Each Entry Shows the Number of Times the *Row* Method Was Preferred over the *Column* Method. Please Refer to Figure 6 for the Images. The Numbers in Parenthesis Show the Cluster Number of the Corresponding Image. The Algorithms Whose Total Scores Differ by at Least 14 Statistically Differ from Each Other

Amikeus (3)	A	B	C	D	E	Total	BarHarbor (1)	A	B	C	D	E	Total
A	0	11	11	12	15	**49**	A	0	7	11	8	15	41
B	6	0	7	8	13	34	**B**	10	0	10	12	14	**46**
C	6	10	0	10	14	40	C	6	7	0	7	15	35
D	5	9	7	0	12	33	D	9	5	10	0	13	37
E	2	4	3	5	0	14	E	2	3	2	4	0	11
HancockIn (4)	A	B	C	D	E	Total	HancockOut (5)	A	B	C	D	E	Total
A	0	8	5	13	11	37	**A**	0	8	16	15	17	**56**
B	9	0	11	11	14	**45**	B	9	0	13	11	13	46
C	12	6	0	12	11	41	C	1	4	0	3	7	15
D	4	6	5	0	13	28	D	2	6	14	0	16	38
E	6	3	6	4	0	19	E	0	4	10	1	0	15
LasVegas (6)	A	B	C	D	E	Total	Niagara (2)	A	B	C	D	E	Total
A	0	10	12	12	15	**49**	**A**	0	10	13	12	14	**49**
B	7	0	8	10	15	40	B	7	0	13	9	16	45
C	5	9	0	9	15	38	C	4	4	0	8	10	26
D	5	7	8	0	15	35	D	5	8	9	0	11	33
E	2	2	2	2	0	8	E	3	1	7	6	0	17

Table VI. The Aggregate Results Combined over All Scenes and Participants. The Algorithms Whose Total Scores Differ by at Least 32 Statistically Differ from Each Other

Aggregate	A	B	C	D	E	Total
A	0	54	68	72	87	281
B	48	0	62	61	85	256
C	34	40	0	49	72	195
D	30	41	53	0	80	204
E	15	17	30	22	0	84

To understand whether the results are significant, we performed the least significant difference test [Starks and David 1961]. This test computes a D value using the following formula:

$$D = 4\left[\sum_{i=1}^{t} a_i^2 - \tfrac{1}{4}tn^2(t-1)^2\right] \Big/ (nt), \tag{6}$$

where a_i denotes the total preference count of method i, t is the number of methods, and n is the number of participants. However, when aggregating the per-scene results, n must be set to the product of the number of participants and the number of scenes. The D value approaches zero if the preference counts are similar. Larger D values indicate higher confidence of a statistically significant result. In our experiment, we found that $D = 181.286$. This value is then compared with the upper $100p\%$ point of the χ^2 distribution with $(t-1)$ degrees of freedom, where p indicates the desired level of significance. In our experiment, we set $p = 0.001$, which corresponds to a χ^2 value of 18.465, allowing us to strongly reject the null hypothesis that all methods are equal.

	25		52		9		111	
HIS-DEF		HIS-HBS		HIS-RBS		HIS-LOCS		LOG-DEF

Fig. 9. Statistical similarity groups of the second experiment. Items underlined by the same line are statistically similar. The numbers indicate the difference of preference counts between the methods.

Once the null hypothesis is rejected, one can proceed with identifying which algorithms statistically differ from each other. A suitable method for this task is the test of equality of two pre-assigned treatments [Starks and David 1961]. This test computes a critical difference value, m_c, as follows:

$$m_c = \lceil 1.96(0.5nt)^{0.5} + 0.5 \rceil. \tag{7}$$

The algorithms with preference counts greater than or equal to the m_c value can considered to statistically differ from each other. In our experiment, we found $m_c = 32$ for the aggregate results and $m_c = 14$ for the individual scene results.

The statistical similarity groups based on the aggregate results are shown in Figure 9. According to this, HIS-DEF and HIS-HBS emerged in the first similarity group, followed by HIS-RBS and HIS-LOCS. LOG-DEF was isolated in the third group.

The second psychophysical experiment reveals interesting findings, which can be deduced by comparing Figures 4, 8, and 9. First, all experiments establish histogram equalization as the most preferred method of visualization. As for the color scale, the first experiment found DEF, HBS, and LOCS to be in the same statistical similarity group. The objective metric isolated HIS-DEF into the first group. The second experiment placed HIS-DEF and HIS-HBS in the first group as well. In the light of all three experiments, we can confidently argue that HIS-DEF appears to be a more favorable method of visualizing HDR images in false color than the other evaluated methods. Irrespective of the color scale, it is also observed that histogram equalization appears to a more preferred method of luminance compression than logarithmic and sigmoidal compression for the task of displaying HDR images in false color.

7. DISCUSSION

Based on the results of the two psychophysical experiments and the objective evaluation, we can argue that histogram equalization-based luminance compression gives the best color mapping results regardless of the color scale being used. We believe that this finding is important because it is not common practice to use this method of compression for visualizing HDR luminance maps in false color. More often, logarithmic scaling is used—a method found to be inferior by our experiments. Furthermore, Akyüz had found that sigmoidal compression may outperform logarithmic scaling [2013]. But in our study, we found these two methods to have a very similar performance (logarithmic scaling was only marginally better in the first experiment). Because these finding are obtained from a diverse set of images and through subjective and objective evaluations, our findings are likely to generalize to other images as well.

What makes histogram equalization better in this task? We believe that this can be attributed to the more uniform distribution of the luminance values across the display range. Although histogram equalization may not be the best method for photographic tone mapping, it appears to produce more informative false color visualizations due to a more balanced use of colors. However, it is important to note that histogram equalization distorts the relationships between luminances. In general, it violates the uniformity principle discussed in Section 2.2. Therefore, in applications where preserving uniformity is important, logarithmic scaling may be a better choice (LOG-DEF was found to be the best among logarithmic compression methods).

Furthermore, the selection of parameters (such as ϵ and α) may affect the performance of logarithmic and sigmoidal compression methods. As histogram equalization does not need user parameters, it is likely to better adapt to the contents of each individual image, which may be one of the reasons for the overall superiority of this technique over the other compression methods.

The effect of color scale was less pronounced. Moreover, it shows some variability between the subjective and objective evaluations. The first subjective evaluation indicated that the rainbow color scale is inferior to the other color scales in this task. However, it performed relatively well in the objective evaluation. It is possible that some of the perceptual challenges presented by this color scale, such as the highlighting effect of saturated yellows, which may subjugate other hues [Rogowitz and Treinish 1998], are not captured by the objective metric. Second, yellow is known to have the least number of perceived saturation steps [Wang et al. 2008]. This may make it difficult for the observers to distinguish small saturation variations in yellow. Furthermore, the ordering of the colors in the rainbow scale is not necessarily intuitive for people—an issue that is irrelevant to the objective metric [Borland and Taylor 2007]. However, in the light of all three experiments, DEF appears to be a more preferred color scale than the other evaluated scales. In particular, the HIS-DEF combination was found to be in the first statistical similarity group in all of the evaluations performed.

An important visual phenomenon that is ignored by our metric is *visual masking*. According to this phenomenon, natural images may contain highly textured and high-contrast regions that may induce visual masking for the neighboring pixels. In such regions, it is known that the threshold of luminance discrimination is elevated [Daly 1993]. This means that a pixel pair with $\Delta E_{\mathrm{CIE00}} > 1$ in the false color map may not have visibly differed in the HDR image due to the masking effect. In other words, in regions affected by visual masking, this color difference could be too conservative. For these regions, pixels with visually unnoticeable luminance differences (due to masking) could be rendered with visually different colors. While the currently proposed metric does not capture this effect, modulating our metric output with the output of a visual masking model is feasible, and this can be an interesting future research direction.

8. CONCLUSIONS AND FUTURE WORK

In this work, we conducted a comprehensive evaluation of false color mapping strategies for luminance distributions (i.e., HDR images). Our study included a carefully designed psychophysical experiment that allowed us to extract a correlation between people's preferences and a metric based on color differences. Using this metric, we carried out an objective experiment using a large number of HDR images. To further build confidence on the results, we conducted a second psychophysical experiment using HDR images of different types of scenes. The findings of all experiments suggest that HIS-DEF is generally the most preferred method of false color visualization for HDR images.

Certainly, false color visualization strategies are not limited to the methods tested in this study. One can think of different compression functions and color scales. However, we believe that the experimental methodology described in this article can be useful for future studies that may perform similar evaluations.

Using local compression functions instead of global ones may be more effective in conveying the visibility of small luminance variations. However, it should be kept in mind that local mappings may distort the monotonicity of luminances. As such, we leave it as a future work to study their appropriateness for different applications.

It is important to keep in mind that our study approaches the problem of false color visualization from a specific perspective. In particular, we have only used natural photographic images in our evaluations, and even there, the scene type seemed to have some impact on which method of visualization is the best. This suggests that in other domains where HDR images are used, such as medical imaging and remote sensing, different false color visualization strategies may in fact be more appropriate. Further research is required to answer this question.

In all, the current study sheds light on the issue of false color visualization of HDR images, for which no systemic evaluation has hitherto been conducted and proposes several evaluation strategies which may benefit future studies.

ACKNOWLEDGMENTS

We are grateful for all the volunteered participants of our experiments. We especially thank Okan Tarhan Tursun for his help in conducting the second psychophysical experiment. Finally, we thank the reviewers for their insightful comments and suggestions.

REFERENCES

Ahmet Oğuz Akyüz. 2013. False color visualization for HDR images. In *HDRi2013 - First International Conference and SME Workshop on HDR Imaging*, Bessa M. and Mantiuk R. (Eds.). Porto, Portugal.

Ahmet Oğuz Akyüz and Erik Reinhard. 2008. Perceptual evaluation of tone reproduction operators using the Cornsweet-Craik-O'Brien illusion. *ACM Trans. Appl. Percept.* 4, 4 (2008), 1–29.

Liliana O. Beltrán and Betina Martins Mogo. 2005. Assessment of luminance distribution using HDR photography. In *ISES Solar World Congress, ISES Solar World Congress*.

N. Ben-Haim, B. Babenko, and S. Belongie. 2006. Improving web-based image search via content based clustering. In *CVPR*. IEEE, 106–106.

David Borland and Russell M. Taylor. 2007. Rainbow color map (still) considered harmful. *IEEE Comput. Graph. Appl.* 27 (2007), 14–17. DOI:http://dx.doi.org/10.1109/MCG.2007.46

King C. Brown, Teresa Bryant, and M. Dawn Watkins. 2010. The forensic application of high dynamic range photography. *J. Forens. Identif.* 60, 4 (2010), 449–459.

H. Cai. 2013. High dynamic range photogrammetry for synchronous luminance and geometry measurement. *Light. Res. Technol.* 45, 2 (2013), 230–257.

S. Daly. 1993. The visible difference predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, A. B. Watson (Ed.). MIT Press, Cambridge, MA, 179–206.

Frédéric Drago, William L. Martens, Karol Myszkowski, and Hans-Peter Seidel. 2002. *Perceptual Evaluation of Tone Mapping Operators with Regard to Similarity and Preference*. Technical Report MPI-I-2002-4-002. Max Plank Institut für Informatik.

Frédéric Drago, Karol Myszkowski, Thomas Annen, and Norishige Chiba. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *Comput. Graph. Forum* 22, 3 (2003).

Frédo Durand and Julie Dorsey. 2002. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Trans. Graph.* 21, 3 (2002), 257–266.

Mark D. Fairchild. 2007. The HDR photographic survey. In *Color and Imaging Conference*, Vol. 2007. Society for Imaging Science and Technology, 233–238.

Raanan Fattal, Dani Lischinski, and Michael Werman. 2002. Gradient domain high dynamic range compression. *ACM Trans. Graph.* 21, 3 (2002), 249–256.

R. C. Gonzalez and R. C. Woods. 1992. *Digital Image Processing*. Addison Wesley, Reading, MA.

E. Grinzato, G. Cadelano, P. Bison, and A. Petracca. 2009. Seismic risk evaluation aided by IR thermography. In *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 72990C–72990C.

Jassim Happa, Alessandro Artusi, Silvester Czanner, and Alan Chalmers. 2010. High dynamic range video for cultural heritage documentation and experimental archaeology. In *Proceedings of the 11th International Conference on Virtual Reality, Archaeology and Cultural Heritage*. Eurographics Association, 17–24.

Myles Hollander, Douglas A. Wolfe, and Eric Chicken. 2013. *Nonparametric Statistical Methods*. Vol. 751. John Wiley & Sons.

ITU (International Telecommunication Union). 2002. *ITU-R Rec. BT.709-5, Parameter Values for the HDTV Standards for Production and for International Programme Exchange*. ITU (International Telecommunication Union), Geneva.

Greg Ward Larson. 2013. Personal communication. (2013).

Greg Ward Larson and Rob A. Shakespeare. 1998. *Rendering with Radiance*. Morgan Kaufmann Publishers, San Francisco, CA.

Haim Levkowitz and Gabor T. Herman. 1992. The design and evaluation of color scales for image data. *IEEE Comput. Graph. Appl.* 12, 1 (1992), 72–80.

L. W. MacDonald. 1999. Using color effectively in computer graphics. *IEEE Comput. Graph. Appl.* 19, 4 (Jul/Aug 1999), 20–35. DOI:http://dx.doi.org/10.1109/38.773961

Radoslaw Mantiuk, A. Tomaszewska, and W. Heidrich. 2009. Color correction for tone mapping. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 193–202.

Rafa K. Mantiuk, Anna Tomaszewska, and Radosaw Mantiuk. 2012. Comparison of four subjective methods for image quality assessment. *Comput. Graph. Forum* 31, 8 (2012), 2478–2491.

K. I. Naka and W. A. H. Rushton. 1966. S-potentials from luminosity units in the retina of fish (Cyprinidae). *J. Physiol.* 185 (1966), 587–599.

Erik Reinhard, Erum Arif Khan, Ahmet Oğuz Akyüz, and Garrett M. Johnson. 2008. *Color Imaging: Fundamentals and Applications*. AK Peters, Wellesley, MA.

Erik Reinhard, Michael Stark, Peter Shirley, and Jim Ferwerda. 2002. Photographic tone reproduction for digital images. *ACM Trans. Graph.* 21, 3 (2002), 267–276.

Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. 2010. *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting* (second ed.). Morgan Kaufmann, San Francisco, CA.

Bernice E. Rogowitz and Lloyd A. Treinish. 1998. Data visualization: The end of the rainbow. *IEEE Spectr.* 35, 12 (1998), 52–59.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. 2005. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Res. Appl.* 30, 1 (2005), 21–30.

Samuel Silva, Beatriz Sousa Santos, and Joaquim Madeira. 2011. Using color in visualization: A survey. *Comput. Graph.* 35, 2 (2011), 320–333. DOI:http://dx.doi.org/10.1016/j.cag.2010.11.015

T. H. Starks and H. A. David. 1961. Significance tests for paired-comparison experiments. *Biometrika* (1961), 95–108.

Jessica M. Theodor and Robin S. Furr. 2009. High dynamic range imaging as applied to paleontological specimen photography. *Palaeontol. Electron.* 12, 1 (2009).

Bruce E. Trumbo. 1981. A theory for coloring bivariate statistical maps. *Am. Stat.* 35, 4 (1981), 220–226.

Jack Tumblin and Holly Rushmeier. 1991. *Tone Reproduction for Realistic Computer Generated Images*. Technical Report GIT-GVU-91-13. Graphics, Visualization, and Useability Center, Georgia Institute of Technology.

Lujin Wang, Joachim Giesen, Kevin T. McDonnell, Peter Zolliker, and Klaus Mueller. 2008. Color design for illustrative visualization. *IEEE Trans. Vis. Comput. Graph.* 14, 6 (2008), 1739–1754.

Greg Ward, Holly Rushmeier, and Christine Piatko. 1997. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Trans. Vis. Comput. Graph.* 3, 4 (1997).

Thomas Williams, Colin Kelley, and many others. 2010. Gnuplot 4.4: an interactive plotting program. (March 2010).

Received November 2014; revised March 2016; accepted April 2016

Query

Q1: AU: Please provide full mailing and email addresses for all authors.