Journal of Imaging Science and Technology® 57(2): 020502-1–020502-12, 2013.

Web-based Image Preference

Michael D. Harris and Graham D. Finlayson▲
School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom
E-mail: Michael.harris@uea.ac.uk

James Tauber
Eldarion, Inc., Burlington, MA, 01803

Abstract. *Some experimenters have begun to carry out image preference experiments over the web, with observers completing the task in their own time and using their own display devices. This reduces the administrative overhead, and opens the possibility to huge numbers of potential observers. However, we have to surrender some control over viewing conditions. In previous work, we evaluated an existing web-based paired comparison experiment against a lab-based counterpart and found that, generally, the two variants did not correlate to a significantly high degree. In this work we extend that study with the development of our own web-based research platform with greater control over viewing conditions and much larger quantities of observers (over 1,000, with more than 26,000 individual observations). With this, we show much more positive correlation between the web- and lab-based variants. We also show the similarity or otherwise between the two variants as a function of time, which reveals how many web-based observations are required to achieve stable results.* © *2013 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2013.57.2.020502]

INTRODUCTION

The images we see around us are often the end result of a long chain of image processing algorithms. The difference between the "raw" image recorded by a camera and the output of a processing pipeline can be very large, as shown in the example in Figure 1. The common aim of the majority of photographic pipelines is to construct images that look as "good" as possible. While "goodness" is a rather fuzzy and personal notion it is precisely this judgment that camera manufacturers and purveyors of image manipulation software must address.

This question can be evaluated systematically in a preference experiment, where some processing pipeline, A, is evaluated in concert with a second pipeline, B (where, for the purposes of illustration, A and B might be the same except B has a putatively improved white balance method). In a *paired comparison* preference experiment, images are processed with each of the two pipelines and presented in pairs. An observer (one of many) is asked to choose the image they prefer while, crucially, they are unaware of which image maps to which algorithm. Figure 2 shows a pair of images, that differ by their color balance, in a format that might be used in a preference experiment. Assuming the number favoring B is significant (above a criterion amount: perhaps 75%) then B is the preferred algorithm.

Care must be taken to undertake preference experiments correctly, for we do not wish the results of our experiment to depend on how the images are viewed. The display should be calibrated (to a standard like sRGB)[1] and the observer should view the image pairs in a dimly lit room (ideally with walls of a neutral gray surround). The pair of images should be shown on an average neutral background and the sequence of pairs shown should be random. Each observer should ideally see the same pair several times. Further details of an appropriate experimental set-up are encoded in the standard ISO 3664.[2]

Often we are interested in evaluating many algorithms, or pipelines, simultaneously. We might run five, six, seven or more different algorithms against each other. While in principle the preference experiment remains the same—the observer is presented with a pair of images at a time—the number of image judgments that need to be made increases rapidly. Assuming we are processing 10 raw images with four, six or eight algorithms and we present each unique pair of processed images, then there are respectively 60, 150 and 280 pairwise comparisons. For even a modest number of observers and a small number of repetitions it takes a long experimental session to obtain complete image preference data.

The main contribution of this article is the design of a web-based pairwise image preference experimental platform. We take inspiration from the color naming work of Moroney[3] and also the "typewar" platform.[4] Our idea is, simply, to meet the challenge of the need for large numbers of pairwise comparisons by crowd-sourcing via the internet. Indeed, we can achieve a far greater number of repetitions than could ever reasonably be found in the lab if only the smallest fraction of web users took part in the preference experiment. Given a greater set of preference data we would like to be able to arrive at stronger conclusions (and so make stronger recommendations about which algorithms perform most favorably).

Of course controlled, lab-based, image viewing was adopted for a reason; we cannot, for example, calibrate a remote observer's monitor. However, it is still possible to use the same presentation of image pairs. Images are entirely visible side by side and are viewed on a variegated

▲ IS&T Member.

Received Jan. 30, 2013; accepted for publication May 31, 2013; published online Mar. 1, 2013. Associate Editor: Susan Farnand.
1062-3701/2013/57(2)/020502/12/$20.00

(a) Camera raw image (b) Rendered JPEG

Figure 1. Camera captured image before and after application of processing pipeline.

Figure 2. Typical interface of a paired comparison experiment.

background that averages to gray. This simple insight, which is key to our work, proves to be crucial in making the web application work. Moreover, we hope that there is a law of large numbers that works in our favor: specifically that, given enough observations, the "web preference" will, on average, be the same as the lab preference.

In our first experiment we evaluate a large set of algorithms using a conventional lab-based methodology and also by crowd-sourcing the web. Broadly, our experiment shows that the two different experimental approaches do converge to similar algorithm rankings. This is, however, not an obvious result: a previous preference experiment run on the web, using identical images, but where care was *not* taken to standardize image presentation delivered a quite different ranking. That study, combined with our own, indicates (at least on a prima facie basis) that image preference studies can be successfully transplanted to the web so long as sufficient care is taken over image presentation. Significantly, in a second contribution of this article we track the similarity or otherwise of the preference results as a function of the number of observations. We do obtain convergence between lab- and web-based preference data but only after *sufficient* preference judgments are made.

BACKGROUND

Web-based Paired Comparisons

Several attempts have been made to gather data from participants over the web, many of which are introduced and examined by Birnbaum[5]. However, a large amount of the successful among these studies have followed survey-based formats, suggesting that the presentation and viewing conditions of the experiment have little or no impact on the results gathered. Work by Rasmussen[6] investigated defect detection over the web; observers were presented with two duplicates of the same image, one of which had been modified to exhibit some "defect", or noise, and the time taken for observers to identify which of the two images was defective was recorded. The results of this experiment were not compared to any lab-based alternative, but as every comparison had a correct answer, the authors could quantify the level of correctness of the observers, which was generally positive. The results were manipulated by discarding data points according to some filtering steps, such as removing user sessions below a certain accuracy level, or excluding observers who did not complete a minimum of 100 observations. This particular study also required a calibration stage by observers, and so represents a more restrictive kind of experiment than the one we envisage here. Observer engagement was encouraged by presenting the experiment in a game-like format: observers were challenged to identify the defects within the quickest time possible. Engagement was further incentivized by the inclusion of a monetary reward for top performers.

Zuffi et al.[7] carried out a web-based readability test, attempting to isolate the thresholds for lightness differences between text and background color on web pages. This was compared to a lab-based control experiment. Similar results were indeed found. Interestingly, in this experiment there were actually fewer web-based participants than those in the lab, but, that the two experimental formats produced similar results is encouraging.

There have been a comparatively small number of paired comparison experiments carried out on the web, with varying degrees of success, but there has been little effort in empirically comparing the results gathered to any "ground

truth'' lab-based data. Some notable attempts to date are studies by Jiang et al.[8] and Sprow et al.[9]

Jiang et al.[8] performed a web-based paired comparison experiment and contrasted it with two lab-based counterparts using the same dataset as part of a larger study on soft-copy reproductions of fine art images. The two lab-based variants were carried out with and without the original image present. While strong positive correlation was found between the web-based results and the lab results without the original present, only weak correlation was found when comparing to the variant with the original present. The web-based variant in this study received a relatively small number of observers—88—and the authors do not describe their recruitment process. On top of this, some statistical power was lost due to the adaptive paired comparison procedure that was employed.

A study by Sprow et al.[9] focused on web-based and lab-based variants of a paired comparison preference experiment concerning a gamut mapping task, presenting an sRGB reference image as well as two images mapped to various device gamuts by competing gamut mapping algorithms. This study attracted a larger number of participants—around 700—and, generally, showed very good correlation between the two sets of results. The importance of these results does come with some notable caveats, however: many observers were friends, relatives and coworkers of the authors and also those recruited by solicitation via the ECI (European Color Initiative) mailing list. Some observers also participated in both variants, with 43 of the 70 observers in the lab variant contributing to the approximately 700 total for the web variant. This particular study utilized a questionnaire and adjustment/characterization images to gather extra data about observers' display devices. This extra intrusion was kept as minimal as possible, but would likely still drive away a substantial amount of possible observers had they not been recruited directly from the color community.

Here lies a significant problem with web-based research—attracting participants and maintaining engagement. Recruitment through mailing lists and pre-existing contacts is effective, but it carries the problem of introducing a sampling error in that the participants already have a vested interest in the results and/or are ''expert observers''. Casual web users have little or no commitment to the study in which they are voluntarily participating, and the task of keeping them engaged and entertained without introducing bias into the results is problematic. The offer of a material reward for participation, or for top contributors, has been used in the past but it introduces the problem of participants manipulating the system for their own reward, without taking any care over their responses.

Our own previous work[10] examined an established, long-running, web-based paired comparison experiment by Mei[11] (comparing the outputs of different tone mapping algorithms operating on high dynamic range images), and compared the results to those produced by a highly controlled lab-based variant performed using the same images and image treatments. We found that, largely, the correlation between these two sets of results was unsatisfactory. However, we suggest that this can be, at least partially, explained by the relatively small numbers of participants attracted to the web-based experiment (no large-scale recruitment was performed), as well as some presentation and implementation issues which may have affected a significant portion of the participants that were attracted—the images were displayed against a yellow background and occasionally the formatting of the web page led them to be displayed stacked atop each other.

The new contribution of this work is to compare the same lab-based data from our previous work[10] to a new set of web-based results gathered from our own, more highly controlled, web-based platform.

METHODOLOGY

Control experiments

Mei[11] and, by extension, our own previous work[10] examined observer preference of tone mapping operators (TMOs). TMOs are functions designed to map pixel values of high dynamic range images into a low dynamic range space such that those images can be viewed on low dynamic range monitors or printed using a conventional printer, all the while attempting to preserve the color, contrast and brightness information present in the original image. We shall be continuing to use TMOs as a test subject in this work, however it is not the purpose of this article to explain the use of TMOs or to examine their relative merits—many authors have already carried out such evaluations, such as Ledda et al.[12] For reference, the TMOs under comparison are given in the Refs. 13–22. We will be using the same lab-based data as our previous work,[10] and comparing that to new data from our own web-based platform.

In our lab-based experiment,[10] we carried out a controlled paired comparison experiment with 14 unpaid participants who were naïve to the objective of the experiment. The pairwise comparison was run using the same collection of scenes and operators as used in the existing web-based experiment by Mei.[11] Note that, for consistency with Mei,[11] different subsets of the algorithms were used for each of the different scenes. There were two scenes for which six algorithms were evaluated (giving $(\frac{6\times5}{2}) \times 2 = 30$ pairs), five scenes where seven algorithms were tested (105 pairs), another four where eight algorithms were tested (112 pairs) and one scene where respectively nine and 10 algorithms were tested (36 and 45 pairs respectively). In grand total there were 328 pairs of images. Each pair was viewed as [*AB*] and [*BA*], where *A* and *B* are images for the same scene processed by two different algorithms, making a total of $328 \times 2 = 656$ comparisons per observer. Due to this large amount of comparisons undertaken, the average observer completed the experiment in one hour, however this was split into sessions lasting no more than 30 minutes each in order to minimize eye strain and loss of concentration among observers.

Viewing conditions were prepared in accordance with ISO standard 3664:2009, and images were displayed on an HP LP2480ZX monitor calibrated to sRGB standard.[1] The average image size subtended at the retina was approximately 6° observable angle, with approximately 1° of padding between the two images. The observed experimental interface resembled that of Fig. 2. Viewing time was not limited but was monitored. The average viewing time was 5.5 s per image pair.

The images used in our lab-based replicate were taken directly from the existing web-based experiment,[11] and resized with bicubic resampling to fit within the intended observable angle at a standardized viewing distance of approximately 1 meter. Note that the images displayed to participants were exactly the same in the web- and lab-based variants (save for displayed size); it is the change in environment which is of interest.

To corroborate results from our TMO experiment, and to alleviate the problem of having results that are task dependent, we also carry out a second experiment using color-to-grayscale (C2G) algorithms. These are algorithms designed to reduce color images, usually three-dimensional RGB, into one-dimensional grayscale images. There are many existing approaches to solving this problem, a collection of which are reviewed by Connah et al.[23]—again, it is not the purpose of this article to compare the different techniques. We shall be using the preference data gathered by Connah et al.[23] as the lab-based variant in this second experiment. The C2G algorithms under comparison are detailed in Refs. 24–28 as well as simple luminance-channel grayscale which, assuming an image color space of sRGB, is given by:

$$lum = 0.2172 \times R + 0.7152 \times G + 0.0722 \times B. \quad (1)$$

The control conditions for the second lab-based experiment are summarized by Connah et al.[23]

Our web-based platform

Extending from our previous work,[10] we observe that the web-based experiment by Mei[11] suffered from fairly low numbers of participants and did not control for some factors which could still plausibly be controlled and/or monitored even in a web-based scenario. In light of this we opted to implement our own web-based research platform[29] so that we could gain greater control over the web-based data collection. We have also expanded our datasets to not only compare observer preferences for tone mapping operators, but also for color-to-grayscale algorithms. For the first experiment, we use observer preferences from our own web-based platform (hereafter referred to as the "web" variant) and from a controlled experiment carried out in our own lab (hereafter referred to as the "lab" variant). For the second experiment, we use observer preferences from Connah et al.[23] for the lab-based data, and data gathered from our own web-based variant of the same study for the web-based data.

One of the limitations of the web-based experiment by Mei[11] was that, due to the design of the page and the size of the images compared, only an estimated 20%[30] of visitors to the site would be able to observe the entirety of both images in a pair on their screen without scrolling. Worse still, for an estimated 50% of observers the resolution of their display device would cause the page layout to display one image stacked atop the other, meaning that the observer would have to scroll between the two images, and would never be able to make a direct comparison of both images on the screen at the same time. In our system, the layout is fixed so that the images will always be shown side-by-side, and the statistics gathered show that 87% of our observers were able to see the entirety of both images on the screen at the same time without scrolling. To facilitate this, we resized the images to a smaller scale than that used by Mei.[11] All images were resized using bicubic resampling and, for the web experiment, we ensured that there would be no client-side rescaling of the images. It is worth noting that while the resolution of the displayed images can be controlled, the physical size of the displayed images cannot be reliably controlled or recorded. It is entirely feasible to record the resolution of observer's display devices and indeed this was done, but physical device size and pixel density cannot be recorded to a high level of accuracy or reliability with current browser support, and it is even less feasible to monitor the observable angle of the image, as viewing distances cannot be controlled. In principle some of this information could be obtained through a questionnaire, although in practice this approach can be unreliable—for example, to be useful, the questionnaire would need to include questions that require a high degree of technical competence to answer.

In previous similar experiments to this, authors have often recruited observers through friends and colleagues, or at conferences or through mailing lists etc. Obviously this can lead to an unrealistic sample of observer populations, as those recruited from within the community are likely to be expert observers, and anyone who is personally recruited is likely to feel an obligation to complete a large number of preference choices, or to spend more time scrutinizing their decisions in order to "get it right". We therefore opted against personal recruitment and targeted the wider online audience for our experiments. The project was publicized through social media and advertised through various other websites unrelated to color science. Participants were therefore attracted much more organically and represent a much better sample of internet users "in the wild".

Observers were free to complete as many or as few preference choices as they wished. If an observer submitted only a handful of preference choices these were added to the pool of data with equal weighting to those submitted by an observer who submitted hundreds. To date, the mean number of comparisons per observer is 18.9, with a standard deviation of 35.3.

Also in opposition to some previous approaches, we opted to have no calibration process, questionnaires or adjustment images. Observers visiting the site were immedi-

ately presented with their first preference choice. Primarily it was thought that immediate presentation of the task at hand would be more likely to engage observers and encourage them to partake; presentation of welcome pages, splash screens, or anything of the sort are well known to increase the "bounce rate" on websites. It is also noted that even if a calibration process were implemented, it would likely be of little value: observers' viewing conditions are likely to change with time, especially on mobile devices. Furthermore, observers could be employing multiple displays, returning to the site on multiple devices, or they could be using a device with an auto-dimming or auto-adjusting display.

Statistical tools

Thurstone's law of comparative judgment. When seeking a preference metric of the perceived quality of several differing image treatments, an intuitive approach is to compare every treatment with every other in a pairwise fashion, resulting in a "tournament" of comparisons where the image that receives the greater preference "wins" each comparison. The problem then, is aggregating the results from each comparison in the tournament into a definitive collection of preference scores. A common approach to this problem, which is still an active area of recreational mathematics,[31] is the application of Thurstone's[32] law of comparative judgment.

Thurstone proposes that a discriminatory process between two stimuli, causing responses S_A and S_B, can be modeled as a normally distributed random variable, where the distribution represents the value of $S_A - S_B$ over many observations, under the assumption that S_A and S_B are themselves normally distributed. The mean of this distribution should give a good approximation of the true value of $S_A - S_B$. This approach allows us to make an estimate of the scale of $S_A - S_B$, even though observers do not make any explicit assertions of that scale, rather they are only ever asked to judge which of the two stimuli produces the "greater" response. To accomplish this, Thurstone adopts some sets of assumptions, grouped by various cases which may apply to the experimental design. We shall use case V, which is the most commonly applied case in the imaging science literature.

Mosteller's test. As described above, Thurstone's case V solution makes several assumptions about the data being analyzed: specifically that the variances for the underlying discriminal processes are equal and that the coefficient of the correlation between observer responses is zero. However, there are occasions when these assumptions do not hold and the case V solution is inadequate. To detect these situations, Mosteller[33] put forth a chi-square test to evaluate the goodness-of-fit of the model to the data. When the χ^2 value obtained from this test is lower than the χ^2 value at some significance level p (with degrees of freedom $(t-1)(t-2)/2$, where $t =$ the number of algorithms, or *treatments*), we accept that the case V solution is suitable for this data.

Kendall coefficient of consistency. We would hope that, in general, observers are *consistent* when they make their preference choices. An inconsistency, in this case, refers to the situation where an observer prefers image A over B, and image B over C, but then prefers image C over A. Kendall and Smith[34] define such an occurrence as a *circular triad*, and they can occur in situations where the compared stimuli do not elicit a hugely different response, meaning that the observer has a hard time differentiating between them or, specifically to cases in image preference, in situations where different image treatments perform well in some image regions but not others, and the observer then chooses different image regions on which to base their preference for one comparison than they do for another.

When only a small collection of stimuli are being compared, it is simple to count these violations of consistency directly. However, for larger quantities, a process for calculating the frequency of the inconsistencies is described by Kendall and Smith.[34] Once the total number of inconsistencies has been calculated, this number is compared to the maximum possible number of inconsistencies for the given number of competing algorithms. This normalized measure of consistency, Ω, has a maximum value of one in the case where there are no violations of consistency, and decreases to zero as the observed inconsistencies increase.

Low values for Ω can be interpreted as an indicator that a particular observer was poor at making consistent preference choices. Alternatively, if Ω is low across many observers, it is an indicator that the stimuli being judged were too similar for the observers to make consistent choices.

It is important to note that, when giving summary statistics for an experiment, Ω is calculated separately for each observer and then averaged across all observers.

Kendall coefficient of agreement. If two observers make the same preference judgment on a pair of images, we denote this as one agreement. Kendall and Smith[34] give a method to calculate the number of pairs of observers in agreement over each pair of images, which is then normalized by the number of observers and the total number of pairs of images. This gives a measure, u, of observer agreement, which can range from 1 in the case of perfect agreement, to $-1/(n-1)$ when n is even, and $-1/n$ when n is odd, where n is the number of observers.

To gain some significance measure of the coefficient of agreement, we can use the χ^2 test described by Ledda et al.[12] to test the null hypothesis that all observers made their preference judgments entirely at random. A significantly high value for u suggests that there are differences among the images being compared, but we cannot necessarily tell where those differences are.

Score difference test. Upon compilation of a Thurstonian analysis, the outcome is a collection of assignments of scores to image treatments. From these scores it is possible to generate an ordinal ranking. However if the scores for two different treatments only differ by a small amount, we may

be hesitant to assign a definitive ranking. To quantify this uncertainty, we can use the *score difference test*, described by Ledda et al.[12]

This test groups a collection of scores such that two scores within the same group cannot be declared significantly different at a given significance level. Formally, we are grouping the scores so that the variance-normalized range of the scores within each group is less than or equal to some value R_α^+.

Calculating R_α^+ is equivalent to finding some R' such that $P(R \geq R') \leq \alpha$. The distribution of the range R is asymptotically the same as the distribution of a variance-normalized range, W_t, of a set of normal random variables with variance $= 1$ and t samples.[35] This gives us

$$P\left(W_{t,\alpha} \geq \frac{2R - \frac{1}{2}}{\sqrt{nt}}\right), \tag{2}$$

where $W_{t,\alpha}$ is the value of the upper percentage point of W_t at significance level α, which is tabulated in many statistics texts, e.g. Pearson and Hartley.[36] From here we can directly calculate the value of R_α^+ given the value of $W_{t,\alpha}$.

$$R_\alpha^+ = \left\lceil \frac{1}{2} W_{t,\alpha}\sqrt{nt} + \frac{1}{4} \right\rceil. \tag{3}$$

To this resultant integer value, R_α^+, we ascribe the following quality: if the score difference between two image treatments is less than R_α^+, those two treatments cannot be described as perceptibly different at the chosen significance level, α.

Kendall rank correlation coefficient. To compare the results of our lab-based and web-based variants, we need a measure of computing the correlation between the two. Given the ordinal nature of the ranking derived from the scores output from a Thurstonian analysis, it follows to use a rank correlation statistic such as Kendall's τ.[37]

Kendall[37] gives a method for computing a significance measure, p, for τ. This measure is based on the likelihood of the observed correlation occurring given two independent variables. A low value for p indicates that a correlation to the extent of τ is unlikely to occur and so we reject the null hypothesis that the two variables are independent.

Sprow et al. chi-squared goodness-of-fit. From the Thurstonian analysis, we have access to more than just ordinal rank data. The scores give scale values as well as a rank ordering. In light of this, there may be some situations where a rank correlation statistic does not tell the whole story. Given a scenario with three treatments A, B, C with scores -1, 0.9, and 1 respectively, if in another experiment, the score for B was 1.1, a rank correlation statistic would penalize this small change just as heavily as if the score became -1.1, as both changes produce an equal rank order swap despite the differing magnitudes of the score difference.

To address this, Sprow et al.[9] devised a χ^2 statistic, similar in construction to Mosteller's test.[33] Instead of comparing the observed results of the experiment to an expected distribution, this test treats one experiment as the "observed" data, and the other as the "expected" data. This statistic is defined as:

$$\chi^2 = \sum_{j<l} \left(\frac{n_{jl} \cdot n'_{jl}}{n_{jl} + n'_{jl}}\right) \cdot (\arcsin(2p_{jl} - 1) - \arcsin(2p'_{jl} - 1))^2, \tag{4}$$

where P and P' are the proportion matrices (matrices where P_{AB} = the proportion of times algorithm A is preferred over algorithm B) of the "expected" and "observed" data respectively, and N and N' are matrices representing the total number of comparisons per pair in each of the experiments. This statistic accommodates for differing numbers of observers (and thus differing variance) between the two experiments and, due to its formulation, allows for unbalanced experiments, where each image pair is not necessarily viewed an equal number of times to every other pair.

Much like Mosteller's, this test is examining at what significance level can we assert that p_{ij} and p'_{ij} are from two different distributions. As such, and in juxtaposition with the significance measure for Kendall's τ, a low p-value from this statistic indicates a poor correlation.

RESULTS

Tone mapping operators

In our first experiment, we are seeking to re-evaluate our web- against lab-based comparison from our work in Harris and Finlayson[10] with a new, more highly controlled, web-based variant. Before doing so, however, it is important to evaluate the quality metrics of the lab data in isolation, so that we can uncover any statistical artifacts which may later impact our lab-to-web comparison: Table I shows the summary statistics described earlier for the lab variant of the TMO experiment.

The columns under the "Agreement" and "Consistency" headings show that, remarkably, all scenes showed significantly high inter-observer agreement ($p < 0.001$ for all scenes) and also high levels of intra-observer consistency. The columns under "Mosteller" show the χ^2 score and corresponding significance level (p-values greater than 0.05 are omitted for clarity) for the Mosteller test, which shows that, for the majority of the tone-mapped scenes, the case V solution adequately describes the preference data. However, the significantly high scores for the "Synagogue" and "Tahoe" scenes should be noted at this point—these suggest that, for these scenes, the assumptions of the case V solution may not hold and that these scenes should be treated with some caution when we later compare the web-based results to these lab-based results.

The data in this table convey several important messages. First, that the Thurstonian case V solution is, in most

Table I. TMO experiment: summary statistics for lab data.

Scene	Mosteller		Agreement			Consistency
	χ^2	Significance	u	χ^2	Significance	Ω
Atrium night	16.438		0.280	179.643	$p < 0.001$	0.691
Belgium	42.425		0.239	335.571	$p < 0.001$	0.592
Bristol bridge	15.052		0.222	195.821	$p < 0.001$	0.719
Clock building	24.126		0.433	355.357	$p < 0.001$	0.800
Fog	13.727		0.229	258.393	$p < 0.001$	0.694
Foyer	6.313		0.155	108.679	$p < 0.001$	0.577
Indoor	13.883		0.194	130.857	$p < 0.001$	0.707
Memorial	18.268		0.252	218.286	$p < 0.001$	0.646
Synagogue	36.019	$p < 0.05$	0.252	218.536	$p < 0.001$	0.815
Tahoe	19.535	$p < 0.05$	0.225	105.929	$p < 0.001$	0.633
Tinterna	18.058		0.274	126.000	$p < 0.001$	0.718
Tree	18.532		0.287	183.679	$p < 0.001$	0.719
Venice	4.668		0.227	149.429	$p < 0.001$	0.694

cases, sufficient for the task of analyzing preference data for tone mapping operators (although notably not in all cases). Second, that the observers in our lab made consistent preference judgments, and, finally, that the observers agreed with each other on image preference choices to a significantly high degree.

Now that we have some understanding of our lab-based results, we can begin to consider the data from the web-based variant. The data under consideration are taken from the first year of operation of our web-based platform,[29] during which time over 26,000 preference judgments were submitted by more than 1,000 observers. Unfortunately, due to its unbalanced nature, we cannot complete the same summary statistics as above for the web-based variant. Expecting web observers to complete every possible combination of images, in order to facilitate the balanced paradigm, is unreasonable. Indeed if we omitted all unbalanced sessions from our data we would be left with only two complete, balanced, sessions.

Table II shows how the web data compare to the lab data—we are considering how the Thurstonian analysis of one variant correlates with the other. The results of both the Kendall rank correlation coefficient and the Sprow goodness-of-fit test (as described earlier) are shown. Recall the disparity in the significance measures for the Kendall and Sprow statistics—a low p-value for the Kendall rank correlation coefficient suggests a strong correlation, while a low p-value for the Sprow goodness-of-fit test suggests a weak correlation. We can see that eight of the 13 scenes give significantly correlated rank orderings. However, for the "Clock building", "Fog", "Foyer", "Tahoe" and "Venice" scenes, both of the Kendall and Sprow measures agree that those scenes showed weak correlation, although we should bear in mind the results of the Mosteller test which suggest that the "Synagogue" and "Tahoe" scenes are ill-suited for the case V solution.

Interestingly, for "Bristol bridge" and "Tree", significant rank correlation is achieved but the Sprow test indicates a poor goodness-of-fit. This is examined in further detail later, with the aid of data from the color-to-grayscale experiment.

Also included in Table II (and Table IV), are approximate values for the quantity of observations for each scene in the web-based variants. These are not equal due to the unbalanced nature of the experimental design, but are of the same order of magnitude owing to the random assignment of observers.

Color to grayscale

To corroborate the TMO experiment, we ran a second experiment examining observer preference for color-to-grayscale algorithms. For the lab-based variant of this experiment, we are using existing data published by Connah et al.[23] The summary statistics for these data are recapitulated in Table III, with the addition of the results of the Mosteller test.

We can see that, as in the TMO experiment, there were high levels of intra-observer consistency for all scenes. However, for the "Girl" and "Hats" scenes, the inter-observer agreement was slightly lower—it is still significantly high ($p < 0.05$ and 0.01 respectively) but it is not at the $p < 0.001$ level as in the other scenes. The reasons for the poorer performance for these scenes are discussed by Connah et al.;[23] it is suggested that, for these scenes in particular, the compared algorithms all perform similarly and different observers may be selecting different criteria to judge the minor differences in these images.

The Mosteller test shows positive results for five of the six scenes but, as with "Synagogue" and "Tahoe" from the TMO experiment, we should be wary when considering the "Monet" scene due to its significantly high χ^2 score.

Table II. TMO experiment: correlations between lab and web results.

Scene	Approx. web observations	Kendall rank correlation		Sprow goodness-of-fit	
		τ	Significance	χ^2	Significance
Atrium night	1100	0.905	$p < 0.01$	23.123	
Belgium	2200	0.733	$p < 0.01$	48.589	
Bristol bridge	1500	0.571	$p < 0.05$	72.106	$p < 0.001$
Clock building	1400	0.357		150.678	$p < 0.001$
Fog	1800	0.333		98.427	$p < 0.001$
Foyer	1100	0.333		83.700	$p < 0.001$
Indoor	1000	0.714	$p < 0.05$	17.081	
Memorial	1400	0.643	$p < 0.05$	34.599	
Synagogue	1400	0.857	$p < 0.01$	26.182	
Tahoe	700	0.467		48.377	$p < 0.001$
Tinterna	800	0.867	$p < 0.05$	20.508	
Tree	1000	0.810	$p < 0.05$	62.485	$p < 0.001$
Venice	1000	0.619		41.896	$p < 0.01$

Table III. C2G experiment: summary statistics for lab data.

Scene	Mosteller		Agreement			Consistency
	χ^2	Significance	u	χ^2	Significance	Ω
Girl	4.362		0.040	28.833	$p < 0.05$	0.714
Hats	3.026		0.061	36.000	$p < 0.01$	0.604
Heron	11.569		0.521	194.833	$p < 0.001$	0.885
Monet	24.199	$p < 0.01$	0.435	165.167	$p < 0.001$	0.807
Parrot	13.172		0.386	148.000	$p < 0.001$	0.818
Poppies	9.070		0.226	92.833	$p < 0.001$	0.755

Table IV. C2G experiment: correlations between lab and web results.

Scene	Approx. web observations	Kendall rank correlation		Sprow goodness-of-fit	
		τ	Significance	χ^2	Significance
Girl	1600	0.333		17.970	
Hats	1600	0.867	$p < 0.05$	15.422	
Heron	1600	0.867	$p < 0.05$	99.281	$p < 0.001$
Monet	1700	0.600		48.534	$p < 0.001$
Parrot	1700	0.867	$p < 0.05$	29.811	
Poppies	1600	0.733	$p < 0.05$	27.162	

Our web-based variant of the C2G experiment ran parallel to the TMO experiment on our web-based research platform.[29] Observers were randomly assigned to one of the two experiments on their first visit to the site, but could opt-in to a different experiment if they so wished. Similarly, if an observer completed all the comparisons for a particular experiment (a feat managed by only two observers), they would be assigned to the other upon their next visit. Table IV shows how the web data compare to the lab data for the C2G experiment—much like Table II, these data represent the first year of data collection.

Four out of six scenes give significantly correlated rank orderings, while "Monet" exhibits weak correlation according to both the Kendall and Sprow measures—although we should once again bear in mind the results of the Mosteller

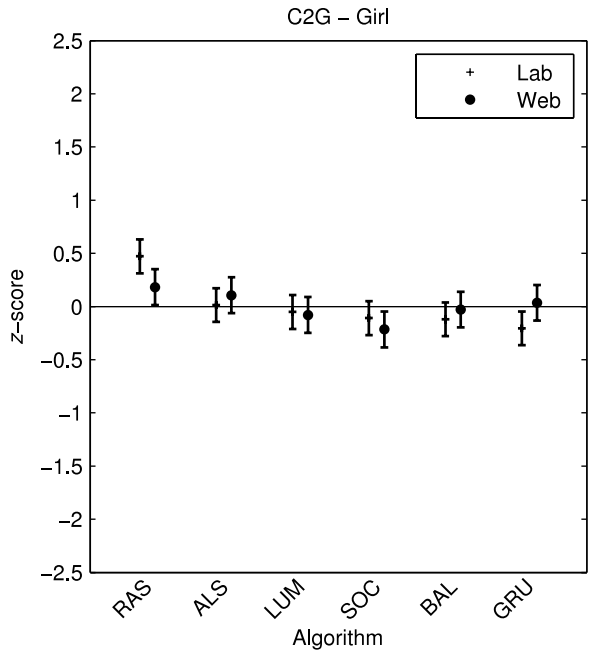

Figure 3. Thurstone scores for "Girl" scene.

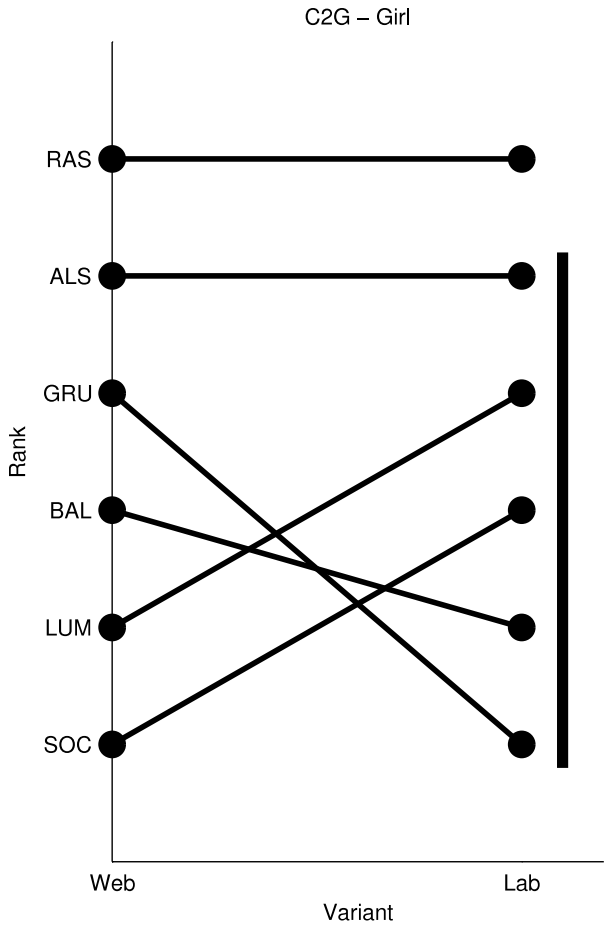

Figure 4. Rank position swaps for "Girl" scene.

test which suggest that the "Monet" scene is ill-suited for the case V solution.

The "Girl" scene presents an interesting situation: it exhibits weak rank correlation according to the Kendall measure, but favorable goodness-of-fit according to the Sprow measure. Figure 3 shows the results of the Thurstonian analysis of the "Girl" scene for both the lab and web variants plotted on the same axes. It is evident that the scores are very similar in both experiments, but the minor fluctuations happen to cause significant rank differences. Figure 4 shows the rankings for both variants, with the vertical bar to the right grouping algorithms that are, according to the score difference test (described earlier), not perceptibly dissimilar at the $\alpha = 0.05$ significance level. The rankings produce many rank position swaps, but they are all within the bounds of the perceptibly similar. This highlights the danger of relying solely on a rank correlation measure to quantify the similarities or otherwise of our lab- and web-based variants.

Notably, if we carry out the score difference test for all scenes (across both the TMO and C2G experiments), this same explanation holds true for every scene that does not exhibit significantly strong rank correlation—the rank position swaps are always among those algorithms which are, according to the score difference test applied to the lab data, not significantly dissimilar. This is an important point to underline—for every scene that does not exhibit strong rank correlation, the rank position swaps causing that weak correlation are all among algorithms which are not perceptibly dissimilar.

Another interesting situation arises for the "Heron" scene, as well as "Bristol bridge" and "Tree" from the TMO experiment: significantly strong rank correlation is achieved but the Sprow test indicates a poor goodness-of-fit. Figure 5 shows how this can be the case for the "Heron" scene—the rank orderings are very similar, with only one position swap between the "BAL" and "LUM" algorithms, however the web results are somewhat muted in comparison to the lab results. This could be due to the larger number of observers for the web experiment. The results for "Bristol bridge" and "Tree" show similar properties.

Correlation over time

A feature of our web-based platform is the ability to compute all the statistics used above in real time. This means that we can examine the correlation between the lab-based and web-based variants as a function of time or, equivalently, the number of comparisons completed. In so doing, we will consider the TMO and C2G experiments in unison.

Figures 6 and 7 show, for the two most strongly correlated scenes, rank correlation between the lab- and web-based variants as a function of the number of comparisons made in the web variant. The horizontal lines show the value of τ required to be significant at the 95 and 99% levels. Both of these scenes suggest that significantly strong correlation can be achieved after approximately 500 comparisons have been completed ($\approx$27 observers).

Conversely, Figure 8 shows correlation over number of comparisons for the weakly correlated "Fog" scene. We can see that the results do not correlate to any significant degree, but they do stabilize after approximately 500 comparisons.

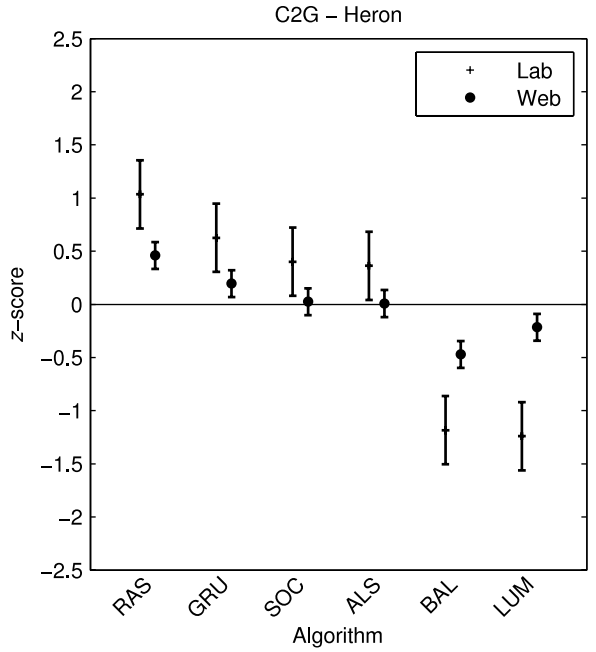

Figure 5. Thurstone scores for "Heron" scene.

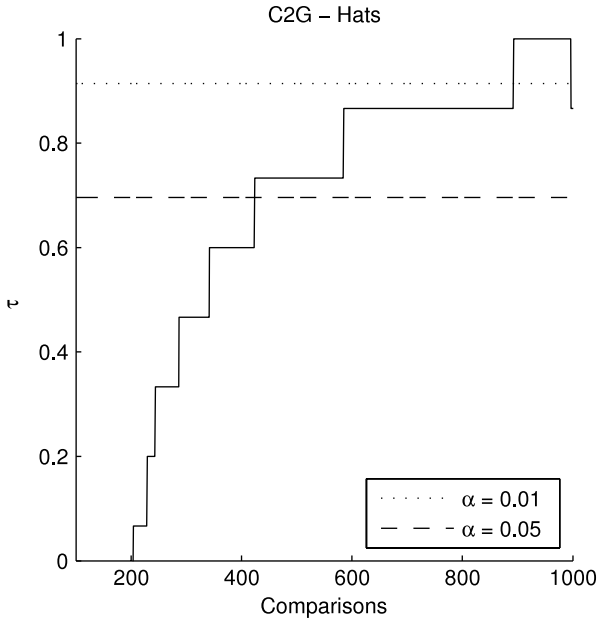

Figure 7. Rank correlation as a function of number of observations for the strongly correlated "Hats" scene.

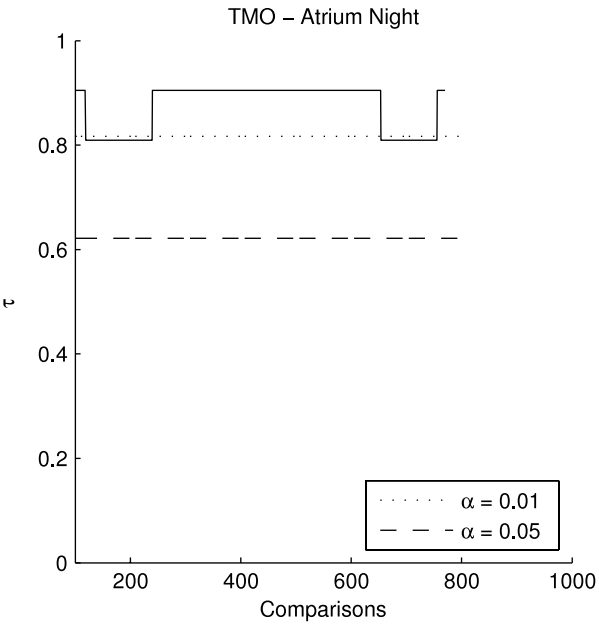

Figure 6. Rank correlation as a function of number of observations for the strongly correlated "Atrium night" scene.

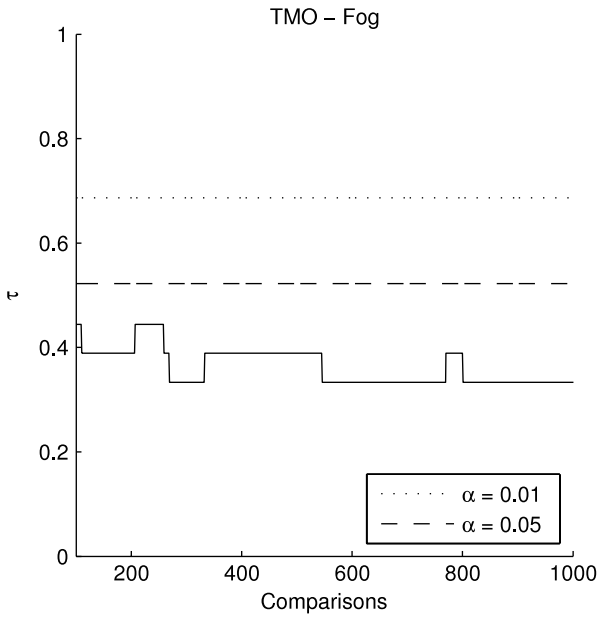

Figure 8. Rank correlation as a function of observations for weakly correlated "Fog" scene.

This suggests that, for the experiments detailed in this article, 500 comparisons is sufficient to obtain a stable result from a cohort of generic web users, but that this result cannot necessarily be relied upon to correlate with a lab-based experiment. This stability despite lack of correlation may also suggest a deeper underlying difference in preference metric for observers on the web.

It is important to note that, while 500 has emerged as the threshold for reliable results for the experiments discussed here, it should not be treated as a general criterion for web-based experiments. Experiments with differing numbers of treatments under scrutiny, with differently matched treatments, using different populations of observers, or employing varying experimental paradigms, may require a greater or lesser quantity of comparisons to meet this threshold.

DISCUSSION

It is apparent that often the results from web-based paired comparisons closely correlate with those carried out under laboratory conditions. It is also shown that, when the

results do not correlate, this can be attributed to lack of discriminatory power among the images being compared.

After completing the lab experiment, observers were consulted about the factors which influenced their preference decisions. Many revealed that they used different image features to inform their decision about different scenes; rather than taking the image as a whole they used specific regions or features of each scene to influence their decision. Further to this, observers noted that certain images had certain recurring artifacts generated by some image treatments but not others, and would intentionally seek these artifacts out upon being presented with an image pair of a certain scene. These cues to decision making are learned as the observer completes more comparisons. An observer beginning the experiment may take more time considering the image as a whole before making their decision, but as they continue they learn which salient image features to look for. This could be an important factor separating the lab and web variants. It is known that the observers in the web variants did not all complete large numbers of comparisons before ceasing their participation. This implies that the rankings of the web variants are likely to be made up of a greater number of observers each undertaking a smaller number of comparisons, which in turn means that each comparison in the web variants is more likely to have been made by a participant who is still unaware of these image features.

During consultation, the majority of observers in the lab-based TMO experiment mentioned the ambiguity in the instructions given. These were chosen to be as similar as possible to those in the web experiment we investigated in our previous work,[10] and it is easy to see how differences of interpretation could arise. The prompt ''choose the image you think is better'' could be interpreted as ''choose the image you think most represents a natural scene'' or ''choose the image you think has more artistic merit'' or even ''choose the image you would prefer to hang on your wall'', all of which could produce vastly different results. Observers noted that, because they were participating in the experiment under laboratory conditions, they felt that they should choose images which looked more natural. It is plausible that observers of the web variant may have interpreted the prompt as in the latter interpretations above, considering that the sort of images traditionally associated with ''HDR photography'' and ''tone mapping'', especially among online photo-sharing websites such as Flickr, are those oversaturated, extremely crisp images that are seen to be more artistic. If we suggest that the lab observers were choosing images which appeared more natural, while the web observers were choosing images which were more artistic (usually distinctly unnatural), then the two sets of observers were deriving completely different judgment metrics from similar instructions, due to the context in which the instructions were given (a formal, laboratory environment, or the informal environment of the internet). This may go some way to explaining the stability in some of the web results despite lack of correlation with the lab results, which was noted at the end of the previous section.

It is clear the question being asked of the observer is important. Prompts can easily be interpreted in many different ways depending on their environment. However, often in these kinds of experiment, we are seeking general observer preference. In both the TMO and C2G cases (and in many more like them), we are not looking for observer opinion on a specific metric such as ''which image appears more saturated?'', but we are seeking to quantify a quality as broad and expansive as general observer *preference*. Instructional context, and clearly separating quality metrics via precise instruction will be an intriguing area of further research in the future.

Our previous work[10] compared an existing web-based preference experiment to a lab-based replicate, and in this work we carry out the same task using the same lab-based data except with our own web-based experiment. Given the similarity of the experiments, it is surprising that we do not find similar results. Comparing our lab-based results to the web experiment by Mei,[11] we find only four of 13 scenes show significantly strong rank correlation, but comparing those same lab results to the results gathered from our web-based experiment we achieve significant rank correlation for eight of the same 13 scenes. Our experiment attracted a larger number of participants, indeed Mei[11] did not reach the 500 comparisons level which, according to our results, seems to be the point at which stable results are achieved. As well as this, perhaps some differences can be attributed to the small advances in control we implemented that Mei did not: namely ensuring consistent side-by-side display at appropriate resolutions, and displaying against a neutral gray background.

All of these points share a common theme: transplanting paired comparison experiments onto the web does not, necessarily, mean the complete surrender of all control over the experiment. With consideration over presentation, and large numbers of observers, it is entirely possible to achieve reliable results.

CONCLUSIONS

The results in this article compare the outcomes of two differing experimental techniques. Lab-based paired comparisons, with all the control and standardization they are typically carried out under, are seen by many as the ''correct'' way of performing visual psychophysics, while web-based techniques are criticized and often disregarded for their lack of traditional control. We have shown that, largely, similar results can be gathered by performing these experiments on the web and that, when the results are not similar, it is indicative of an underlying problem with the images under comparison that may suggest that they are ill-suited for this type of experiment in general.

We observe that convergence in results can be met, so long as careful consideration is given to image presentation, the phrasing of the prompt given to the observer and whether or not general web users may have a predisposition to favor certain images that a lab observer may not. We also note that many previous studies in this area have exhibited poor results

which may be attributable to small numbers of observers, or to samples of web users that are not generally representative of the observers on the web at large.

ACKNOWLEDGMENTS

We would like to thank Eldarion, Inc. for the generous provision of web hosting and support for the web-based aspects of this work. Thanks to David Connah for providing raw data from his color-to-grayscale experiments.[23] The support of EPSRC grant number EP/H022236/1 is gratefully acknowledged.

REFERENCES

[1] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, "A standard default color space for the internet-srgb". Microsoft and Hewlett-Packard Joint Report (1996).
[2] ISO 3664:2009 graphic technology and photography: Viewing conditions. URL www.iso.org.
[3] N. Moroney, "Unconstrained web-based color naming experiment," Proc. SPIE **5008**, 26–46 (2003).
[4] J. Tauber, typewar. http://typewar.com/. Accessed 28 January 2013.
[5] M. H. Birnbaum, "Human research and data collection via the internet," Psychology **55**, 803 (2004).
[6] D. R. Rasmussen, "Online image quality surveys based on response time," Proc. SPIE **6808**, 68080K (2008) http://dx.doi.org/10.1117/12.757895.
[7] S. Zuffi, C. Brambilla, R. Eschbach, and A. Rizzi, "Controlled and uncontrolled viewing conditions in the evaluation of prints," Proc. SPIE **6807**, 680714 (2008).
[8] J. Jiang, J. Frey, and S. Farnand, "Evaluating the perceived quality of soft-copy reproductions of fine art images with and without the original present," *Proc. IS&T/SID Nineteenth Color and Imaging Conf.* (IS&T, Springfield, VA, 2011) pp. 276–284.
[9] I. Sprow, Z. Baranczuk, T. Stamm, and P. Zolliker, "Web-based psychometric evaluation of image quality," Proc. SPIE **7242**, 724201 (2009).
[10] M. D. Harris and G. D. Finlayson, "Comparing a pair of paired comparison experiments: Examining the validity of web-based psychophysics", *Proc. IS&T/SID Nineteenth Color and Imaging Conf.* (IS&T, Springfield, VA, 2011) pp. 29–34.
[11] Y. Mei, High dynamic range image comparison. http://hdri.cs.nott.ac.uk/v1/index.php. Accessed 28 January 2013 .
[12] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," ACM Trans. Graphics (TOG) **24**, 640–648 (2005) ISSN 0730-0301.
[13] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," Computer Graphics Forum **22**, 419–426 (2003) (Wiley Online Library).
[14] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," ACM Trans. Graphics **21**, 267–276 (2002) ISSN 0730-0301.
[15] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," ACM Trans. Graphics (TOG) **21**, 257–266 (2002) ISBN 1581135211.
[16] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," ACM Trans. Graphics **21**, 249–256 (2002) ISSN 0730-0301.
[17] G. Qiu and J. Duan, "Hierarchical tone mapping for high dynamic range image visualization," Proc. SPIE **5960**, 2058–2066 (2005).
[18] J. Tumblin and G. Turk, "LCIS: A boundary hierarchy for detail-preserving contrast reduction," *Proc. 26th Annual Conf. on Computer Graphics and Interactive Techniques* (ACM Press/Addison-Wesley Publishing Co., 1999), pp. 83–90, ISBN 0201485605.
[19] J. Duan, M. Bressan, C. Dance, and G. Qiu, "Tone-mapping high dynamic range images by novel histogram adjustment," Pattern Recognit. **43**, 1847–1862 (2010), ISSN 0031-3203.
[20] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping," ACM Trans. Graphics (TOG) **27**, 68 (2008), ISSN 0730-0301.
[21] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," IEEE Trans. Vis. Comput. Graphics **11**, 13–24 (2005), ISSN 1077-2626.
[22] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," IEEE Trans. Vis. Comput. Graphics **3**, 291–306 (1997), ISSN 1077-2626.
[23] D. Connah, G. D. Finlayson, and M. Bloj, "Seeing beyond luminance: A psychophysical comparison of techniques for converting colour images to greyscale," *Proc. IS&T/SID Fifteenth Color Imaging Conf.* (IS&T, Springfield, VA, 2007) pp. 336–341.
[24] A. Alsam and Ø. Kolås, "Grey colour sharpening", *Proc. IS&T/SID Fourteenth Color Imaging Conf.* (IS&T, Springfield, VA, 2006) pp. 263–267.
[25] R. Bala and R. Eschbach, "Spatial color-to-grayscale transform preserving chrominance edge information," Signal **100**, 4 (2004).
[26] M. Grundland and N. A. Dodgson, "The decolorize algorithm for contrast enhancing, color to grayscale conversion. Technical Report, University of Cambridge, Computer Laboratory, XX, XX, vol. UCAM-CL-TR-649 (2005), pp. 1–15.
[27] K. Rasche, R. Geist, and J. Westall, "Detail preserving reproduction of color images for monochromats and dichromats," IEEE Trans. Comput. Graphics Appl. **25**, 22–30 (2005).
[28] D. A. Socolinsky and L. B. Wolff, "Multispectral image visualization through first-order fusion," IEEE Trans. Image Process. **11**, 923–931 (2002).
[29] M. D. Harris, Colourwar. http://colourwar.com/. Accessed 28 January 2013 .
[30] Google browser size. http://browsersize.googlelabs.com/. Accessed 28 January 2013.
[31] J. P. Keener, "The perron-frobenius theorem and the ranking of football teams," SIAM Rev. **35**, 80–93 (1993).
[32] L. L. Thurstone, "A law of comparative judgment," Psychol. Rev. **34**, 273–286 (1927), ISSN 0033-295X.
[33] F. Mosteller, "Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," Psychometrika **16**, 207–218 (1951).
[34] M. G. Kendall and B. B. Smith, "On the method of paired comparisons," Biometrika **31**, 324–345 (1940), ISSN 0006-3444.
[35] H. A. David, *The Method of Paired Comparisons* (Charles Griffin and Company, London, 1988).
[36] E. S. Pearson and H. Hartley, *Biometrika Tables for Statisticians*, 3rd ed. (Cambridge University Press, 1966), Vol. 1.
[37] M. G. Kendall, "A new measure of rank correlation," Biometrika **30**, 81–93 (1938), ISSN 0006-3444.