Comparing a Pair of Paired Comparison Experiments: Examining the Validity of Web-Based Psychophysics

Michael D. Harris and Graham D. Finlayson, School of Computing Sciences, University of East Anglia, Norwich, UK

Abstract

Paired comparison experiments are frequently used to gather observer preference data in many areas of image enhancement. However, due to the large quantity of comparisons each individual must complete, these experiments are typically carried out with few observers. Taking this method onto the web is a quick way of gaining a larger number of observers and preference judgements. This work examines the validity of web based paired comparisons and whether the loss of control over viewing conditions causes significantly different results.

Introduction

The method of pairwise comparisons is often used to collect observer preference data for differing image reproductions. Indeed, pairwise comparisons, along with Thurstone's [21] method of ranking paired comparisons, has become a widely adopted standard for performing and analysing preference experiments. Some of the problems with this approach, however, include the time required to set up and administer such a well-controlled study, as well as a lack of participants to partake in the often lengthy and sometimes laborious sessions in which each observer must participate. Thus, typically, these experiments are carried out for only a small number of observers. To address this, there exists a hypothesis that pairwise comparison studies can be successfully carried out over the internet using an interface implemented in a regular web browser.

Web based paired comparison experiments provide a quick and easy method of gaining a very large number of participants in exchange for a minimal amount of time and effort on the part of the researcher. But do these benefits come at the cost of reliable data? Web based experiments lack the control over confounding variables and as such it is not obvious they will deliver data that are useful. However, it can be argued that having no control over these confounding variables gives a more 'real-world' representation of observers, and that the effects of the variance in these conditions will become minimised as the numbers of observers and differing viewing environments increase.

The aim of this work is to take an empirical approach to evaluating the validity of data acquired by web-based paired comparison experiments. Specifically, Mei [12] evaluated a variety of tone mapping algorithms by carrying out a pairwise experiment on the web. It should be noted that, at the time of publication, this experiment is still running and so the results are subject to change; the results discussed in this paper represent a snapshot in time of the web results. In this work, we replicated the pre-existing online paired comparison experiment under controlled laboratory conditions. The results from the laboratory experiment were analysed in the same way as [12], and a direct comparison of the two sets of results is presented in this work.

In our experiments we find that observer judgements made in the web based experiment differ markedly from those made by observers in our lab (under controlled standard viewing conditions). Therefore, web-based preference experiments cannot always substitute for controlled lab-based observer judgements.

Background

Web Based Experiments

There have been many web-based tools developed to gather data from participants on the web, a selection of which are examined in detail by Birnbaum [1]. However, the majority of these have historically been examples of survey-based data collection, implying that the presentation of the experiment itself has little impact on the response of the participant. In colour science however, the environment around the participant, the screen upon which they are observing any displayed images, and ambient lighting conditions, along with numerous other factors, can all play a significant role in the participant's responses.

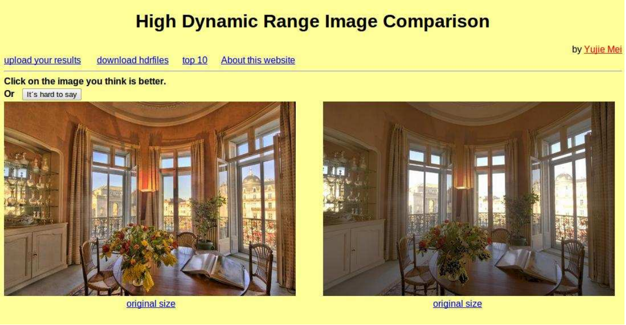
In recent years, there have been several web-based experiments in the field of colour science, *e.g.* the colour naming experiment by Moroney [14]. This example, among others, has been extremely successful in exploiting the power of the internet to collect data at a large scale, and arguably, due to the requirement of such a large range of participants, could not have been done without the use of the internet. Non-academic projects, such as Munroe's colour naming experiment [15], which attracted over 220,000 participants[1], show the huge potential for mass data collection and the public interest in scientific research performed in this way. This concept of 'crowd-sourcing' data is not new to the internet, but it has recently undergone a rise in popularity largely due to the surge in adoption of social networking sites and their integration with third-party services. Thanks to services such as Gravatar, Disqus and Facebook Connect, web users are much more involved in content creation as well as consumption. Because of this shift toward greater user-engagement on the web, casual browsers are now more inclined to participate in web-based experiments.

Tone Mapping Operators

Tone mapping operators (TMOs) are functions designed to map high dynamic range images such that they can be viewed on low dynamic range monitors or printers, while maintaining the colour, contrast and brightness of the original image. Many such operators exist, and several authors have psychophysically assessed them under laboratory conditions [10].

The web-based paired comparison experiment launched in

[1] This estimate is based on user sessions, it does not account for participants taking part more than once.



High Dynamic Range Image Comparison

by Yujie Mei

Click on the image you think is better.

Or It's hard to say

Figure 1. *Interface of the web experiment*

2010 by Mei [12] (hereafter referred to as the 'web' experiment) collects user preferences of images produced by TMOs as viewed through a visitor's web browser on their own computer, as reported in [17]. Upon arrival at the site the visitor is presented with two images of the same scene treated by two different TMOs, and can click on either one to submit a preference, as shown in Figure 1. Alternatively the visitor may click a button to indicate a lack of preference, or a 'tie' situation. The results of these preference choices are collated, ranked and made available at [13].

The web experiment uses 13 different scenes, which are listed in the Appendix. The TMOs compared are[2]:

Drago
Adaptive Logarithmic Mapping For Displaying High Contrast Scenes Drago et al. [3]

LCIS
LCIS: A Boundary Hierarchy For Detail-Preserving Contrast Reduction Tumblin and Turk [22]

Mantiuk08
Display Adaptive Tone Mapping Mantiuk et al. [11]

Reinhard
Dynamic Range Reduction Inspired By Photoreceptor Physiology Reinhard and Devlin [18]

Filter
Fast Bilateral Filtering For The Display Of High-Dynamic-Range Images Durand and Dorsey [5]

GD
Gradient Domain High Dynamic Range Compression Fattal et al. [6]

Hier
Hierarchical Tone Mapping For High Dynamic Range Image Visualization Qiu and Duan [16]

LocalHA
Tone-Mapping High Dynamic Range Images By Novel Histogram Adjustment Duan et al. [4]

EMPJ
Photographic Tone Reproduction For Digital Images Reinhard et al. [19]

Ward
A Visibility Matching Tone Reproduction Operator For High Dynamic Range Scenes Larson et al. [9]

Experimental Design

To compare results with the web-based research, a controlled paired comparison experiment (hereafter referred to as the 'lab' experiment) was carried out with fourteen unpaid participants who were naïve to the objective of the experiment.

Viewing conditions were prepared in accordance with ISO standard 3664:2009, and images were displayed on a HP LP2480ZX monitor calibrated to sRGB standard [20]. The average image size subtended at the retina was approximately 6° visual angle, with approximately 1° of padding between the two images. Viewing time was not limited but was monitored. The average viewing time was 5.5 seconds per image pair.

The pairwise comparison was run using the same collection of scenes and TMOs as used in the web experiment. As in the web experiment, different subsets of the algorithms were used for each of the different scenes. There are 2 scenes for which 6 algorithms are evaluated (giving $(\frac{6\times5}{2})\times 2 = 30$ pairs), 5 scenes where 7 algorithms are tested (105 pairs), another 4 where 8 algorithms are tested (112 pairs) and 1 scene where respectively 9 and 10 algorithms are tested (36 and 45 pairs respectively). In grand total there are 328 pairs of images. Each pair is viewed as $[AB]$ and $[BA]$, where A and B are images for the same scene processed by two different tone mapping 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 thirty minutes each in order to minimise eye strain and loss of concentration among observers.

The images used in the experiment were taken directly from [12], and resized with bicubic resampling to fit within the intended observable angle at a standardised viewing distance of approximately one metre. Note that the images displayed to participants were exactly the same in each experiment (save for displayed size); it is the change in environment which is of interest.

The instructions given to the user in the web experiment are "Click on the image you think is better", with a tie option given as "Or *It's hard to say*" (emphasis indicates clickable button text). The instructions given in the lab conditions were modified slightly, as the user did not click on images to indicate preference, but had separate physical buttons to select either image or the tie option, as such the instructions given were "Choose the image you think is better, or press [the tie button] if it is hard to say".

Results

It has become commonplace to analyse paired comparison data of this kind by using Thurstone's law of comparative judgement [21]. However, a Thurstonian analysis of the web-based study was not compiled, nor is the original raw data available to create one. Instead, the authors used what they called the 'Image Quality Ranking Index' (or IQRI), detailed in [17]. This index for a particular reproduction t is defined as:

$$IQRI_t = \frac{v}{w_t + \frac{d_t}{2}} \qquad (1)$$

where w is the number of wins for reproduction t, d is the number of tie situations involving t, and v is the total number of votes cast across all comparisons involving t; a lower IQRI score indicates a more favourable ranking.

[2] Abbreviated TMO names have been kept consistent with those used in the web experiment [12].

Clearly, we wish to compare our experimental results with the web-based rankings (available at [13]). We do this by comparing the IQRI rankings of both experiments using the Kendall rank correlation coefficient, as defined in [7]. This is a measure of the level of correlation between two sets of ranked data, giving a score ranging from 1, indicating perfect correlation, to -1, indicating that one ranking is correlated with the inverse of the other. A score of 0 indicates that the two rankings are uncorrelated.

To compute this statistic, τ, from two rank orders, those rankings must first be rearranged so that one is considered as a 'correct', or objective, order. For example, consider the two rankings A and B:

$$A = (2,1,5,4,3)$$
$$B = (1,3,4,5,2)$$

To rearrange these rankings, considering A objectively, the elements of A are rewritten such that they are in increasing order, while maintaining the corresponding elements of B:

$$A' = (1,2,3,4,5)$$
$$B' = (3,1,2,5,4)$$

Once this reordering is completed, a measure, k, of the ordered pairs within the ranking B' can be calculated:

$$k = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \begin{cases} 1 & \text{if } B'_j > B'_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This can then be normalised to give the correlation coefficient τ.

$$\Sigma = 2k - \frac{n(n-1)}{2} \tag{3}$$

$$\tau = \frac{2\Sigma}{n(n-1)} \tag{4}$$

This correlation coefficient was computed for the IQRI rankings for all scenes. Table 1 shows, for each scene, the value for the Kendall rank correlation coefficient, τ, and where there is a significant similarity, the corresponding p-value.

As shown in Table 1, the 'Synagogue' and 'Tinterna' scenes both have very high rank correlation ($p < 0.01$ and $p < 0.05$ respectively); however, the rank correlations for the 'Clock Building' and 'Venice' scenes produce drastically different results. Overall, only 4 of the 13 scenes produced TMO rankings which were similar across the two experiments at the 95% level. Figure 2 provides a visual representation of the rankings for those scenes which have the highest and the lowest rank correlations.

In light of the discrepancy between the two sets of rankings, it is desirable to know to what level of confidence we can hold the data. Measures of the Kendall coefficient of agreement among observers, the χ^2 score, and the Kendall coefficient of intra-observer consistency (all described in Connah et al. [2], Kendall and Smith [8], Ledda et al. [10]) were calculated for the lab data.

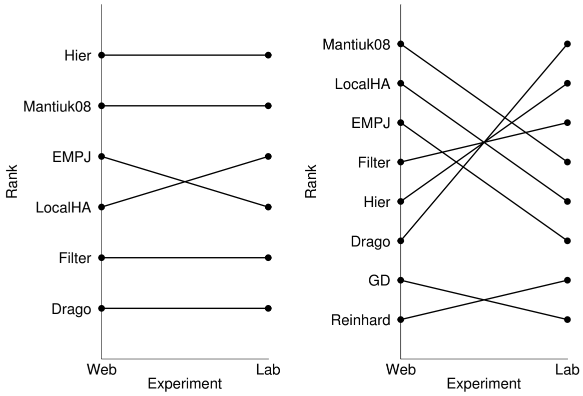
Table 2 shows the summary statistics for all scenes; the second column, u, gives the Kendall coefficient of agreement, the next column gives the χ^2 score, and the fourth column gives the confidence level at which we can accept that observers were in agreement in their preference decisions. Remarkably, $p < 0.001$ for all scenes, suggesting that observers were generally in agreement. The fifth column of Table 2, Ω, gives the average coefficient of consistency across all observers. This figure, which is high across all scenes except for 'Belgium' and 'Foyer', shows that intra-observer consistency was high, and so suggests that the compared images were perceptibly different to such a degree that an observer could make confident preference choices. The low score for the 'Belgium' and 'Foyer' scenes may suggest that observers were basing their decisions on different image features depending on the image pair presented. Upon inspection of the different reproductions of those scenes, it is evident that some operators perform well in the highlights but fail in the shadows, while some others perform conversely. Observers may have chosen to favour highlight performance for some image pairs, and shadow performance for others.

Table 1. Rank correlation metrics for all scenes

Scene	τ	Significance
Atrium Night	0.4286	
Belgium	0.5111	$p < 0.05$
Bristol Bridge	0.7143	$p < 0.05$
Clock Building	0.0714	
Fog	0.4444	
Foyer	0.3333	
Indoor	0.5238	
Memorial	0.5000	
Synagogue	0.7857	$p < 0.01$
Tahoe	0.4667	
Tinterna	0.8667	$p < 0.05$
Tree	0.2381	
Venice	0.1429	

Discussion

These results compare the outcomes of two very different experiments. Although both are paired comparison experiments and both are comparing the same collection of images, the levels of control in the lab experiment contrast greatly with the almost total lack of control in the web experiment. It is not the intention



(a) Tinterna (b) Clock Building

Figure 2. Rank correlations for 'Tinterna' and 'Clock Building' scenes

Table 2. Summary statistics for all scenes in the lab experiment

Scene	u	χ^2	Significance	Ω
Atrium Night	0.2798	179.6429	$p < 0.001$	0.6707
Belgium	0.2392	335.5714	$p < 0.001$	0.5920
Bristol Bridge	0.2220	195.8214	$p < 0.001$	0.7188
Clock Building	0.4330	355.3571	$p < 0.001$	0.7996
Fog	0.2288	258.3929	$p < 0.001$	0.6821
Foyer	0.1546	108.6786	$p < 0.001$	0.5493
Indoor	0.1938	130.8571	$p < 0.001$	0.6874
Memorial	0.2517	218.2857	$p < 0.001$	0.6464
Synagogue	0.2520	218.5357	$p < 0.001$	0.8147
Tahoe	0.2245	105.9286	$p < 0.001$	0.6334
Tinterna	0.2741	126.0000	$p < 0.001$	0.7176
Tree	0.2869	183.6786	$p < 0.001$	0.7000
Venice	0.2265	149.4286	$p < 0.001$	0.6735

of this work to examine why one TMO is preferred over another in each particular experiment; this is not a paper about TMOs. Rather, the data of interest is the extent of the similarity between the two sets of rankings, and what factors can account for any differences.

The interface of the web experiment breaks many conventions of displaying images to a participant. Aside from the many aspects of the environment which are beyond the feasible control of any web-based interface (such as ambient lighting, viewing angle, viewing distance, and screen resolution), that of the web experiment introduces some complications of its own. Images are displayed against a bright yellow background, bordered by other colourful interface elements. For several scenes many participants will have to scroll to see the whole image, and the degradation of display capabilities for those with a low screen resolution may mean that a participant has to scroll to see one image stacked atop the other, meaning that they would not be viewing both images on the screen at the same time and so could not make a direct comparison.

The web experiment exposes the possibility of a sampling error, or a confirmation bias, by allowing multiple completions of comparisons by the same observer. Conversely, the average web user has little incentive to complete all comparisons, as they are not under monitored conditions, and so may become bored with the web experiment fairly quickly and only submit a small number of preference choices. The visitors who are more likely to donate a larger number of comparisons are those who are already interested in such studies, such as other researchers and photographers. These expert observers will likely have inherently different preference choices to the general population.

After completing the lab experiment, observers were consulted about the factors which influenced their preference decisions. Observers noted that some scenes had recurring artefacts generated by some TMOs but not others, and would intentionally seek out these artefacts to inform their preference choice. 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. While the lab experiment was balanced (every participant observed every image pair), it is unlikely that observers of the web experiment would complete a large number of comparisons before becoming bored and ceasing their participation. This suggests that the rankings of the web experiment 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 experiment is more likely to have been made by a participant who is still unaware of these image features.

In both experiments, the option for observers to opt out of a preference choice by submitting a 'tie' response may lead to loss of data in situations where two image versions are very similar. If observers were forced to make a choice, they may take more time and consideration in choosing an image version which outperforms the other. However if they are given the ability to opt out they may quickly decide that the two versions are too similar to make a preference judgement. Generally this should not incur too much penalty, if one image obviously outperforms another then the observer is unlikely to choose the 'tie' option (in the lab experiment only 2.7% of preference choices were ties). However, if two algorithms perform very similarly, then the lack of these detailed preference choices could create differences between rankings.

Many observers in the lab experiment mentioned the ambiguity in the instructions given. These were chosen to be as similar as possible to those in the web experiment, 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', 'choose the image you think has more artistic merit' or 'choose the image you would prefer to hang on your wall', all of which could produce different results. Observers noted that, because they were partaking in the experiment under laboratory conditions, they felt that they should choose images which looked more natural. Observers of the web experiment may have interpreted the prompt as in the latter interpretations above, considering that the sort of images traditionally associated with 'HDR photography', especially among online photo sharing websites such as Flickr, are those over-saturated, extremely crisp images that are seen to be more artistic. If 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 judgement 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).

Conclusions and Future Work

In this paper we replicated a web-based paired comparison experiment designed to determine tone mapping algorithm preference in a classical 'lab-based' experiment (where viewing conditions were controlled). We found that, at the 95% level, only 4 of the 13 scenes produced similar rankings. This is a strong indication that the web observers made different preference judgements compared with those in a controlled laboratory assessment. The high confidence measures derived from the statistical methods applied to the lab results, and the generally low correlation between the web and lab experiments, do not infer that the web results are erroneous. The different results from the experiments should be considered as valuable insights into how observer preferences can shift depending on the context of their decisions. However, if we take the view that the ISO recommended procedure for photographic preference assessment is the 'correct' way to judge between tone mapping algorithms, then our work indicates that a naïve web-based version of this experiment can deliver (surprisingly) different results.

There are not enough data available from these experiments to infer, in any definitive way, which of the numerous differences in the web-based and lab-based scenarios are the major influences in the discordance of the results. To facilitate a deeper insight into these discrepancies, and to identify any imaging topics which may lend themselves to web-based research better than TMOs, future work will include the development of a web-based platform upon which many different experiments can be carried out.

Acknowledgements

The authors would like to thank Jakkarin Singnoo[3] for help in administering the lab-based experiment discussed in this paper, as well as Yujie Mei[4] and Guoping Qiu[4] for access to data pertaining to the web experiment.

The support of EPSRC grant number EP/H022236/1 is gratefully acknowledged.

References

[1] M.H. Birnbaum. Human research and data collection via the internet. *Psychology*, 55(1):803, 2004.

[2] D. Connah, G.D. Finlayson, and M. Bloj. Seeing beyond luminance: A psychophysical comparison of techniques for converting colour images to greyscale. In *15th Color Imaging Conference: Color, Science, Systems and Applications*, pages 336–341, 2007.

[3] F. Drago, K. Myszkowski, T. Annen, and N. Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer Graphics Forum*, volume 22, pages 419–426, 2003.

[4] J. Duan, M. Bressan, C. Dance, and G. Qiu. Tone-mapping high dynamic range images by novel histogram adjustment. *Pattern Recognition*, 43(5):1847–1862, 2010.

[5] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 257–266. ACM, 2002.

[6] R. Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. *ACM Transactions on Graphics*, 21 (3):249–256, 2002.

[7] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30 (1/2):81–93, 1938.

[8] M.G. Kendall and B.B. Smith. On the method of paired comparisons. *Biometrika*, 31(3/4):324–345, 1940.

[9] G.W. Larson, H. Rushmeier, and C. Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *Visualization and Computer Graphics, IEEE Transactions on*, 3(4):291–306, 1997.

[10] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen. Evaluation of tone mapping operators using a high dynamic range display. *ACM Transactions on Graphics (TOG)*, 24(3):640–648, 2005.

[11] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. *ACM Transactions on Graphics (TOG)*, 27(3):68, 2008.

[12] Y. Mei. High dynamic range image comparison. `http://hdri.cs.nott.ac.uk/v1/index.php`, . Retrieved 26 August 2011.

[13] Y. Mei. High dynamic range image comparison - top 10. `http://hdri.cs.nott.ac.uk/v1/top10.php`, . Retrieved 26 August 2011.

[14] N. Moroney. Unconstrained web-based color naming experiment. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5008, pages 36–46, 2003.

[15] R. Munroe. Color survey results. `http://blog.xkcd.com/2010/05/03/color-survey-results/`. Retrieved 26 March 2011.

[16] G. Qiu and J. Duan. Hierarchical tone mapping for high dynamic range image visualization. In *Proc. SPIE*, volume 5960, pages 2058–2066, 2005.

[17] G. Qiu, Y. Mei, and J. Duan. Evaluating hdr photos using web 2.0 technology. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7867, page 24, 2011.

[18] E. Reinhard and K. Devlin. Dynamic range reduction inspired by photoreceptor physiology. *Visualization and Computer Graphics, IEEE Transactions on*, 11(1):13–24, 2005.

[19] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002.

[20] 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.

[21] L.L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273–286, 1927.

[22] J. Tumblin and G. Turk. LCIS: A boundary hierarchy for detail-preserving contrast reduction. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 83–90. ACM Press/Addison-Wesley Publishing Co., 1999.

Author Biography

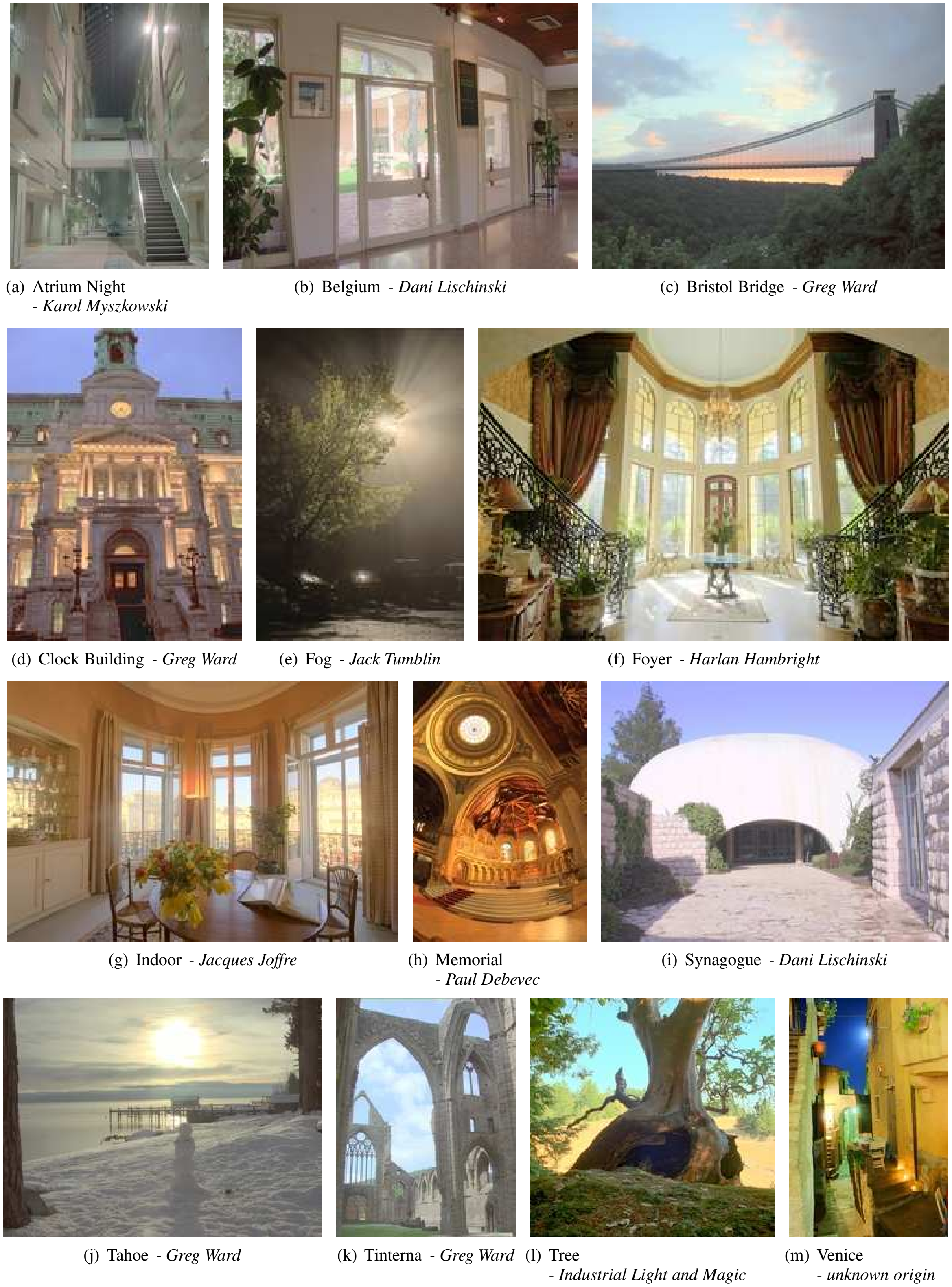
Michael D. Harris received his BSc in Computing Science from the University of East Anglia (Norwich, UK) in 2008. After spending some time in industry, he returned to the University of East Anglia in 2010 as a PhD student under the supervision of Prof. Graham Finlayson.

Graham D. Finlayson obtained his BSc in Computer Science from the University of Strathclyde (Glasgow, Scotland) in 1989. He then pursued his graduate education at Simon Fraser University (Vancouver, Canada) where he was awarded his MSc and PhD degrees in 1992 and 1995 respectively. From August 1995 until September 1997, Dr. Finlayson was a lecturer in Computer Science at the University of York (York, UK) and from October 1997 until August 1999 he was a Reader in Colour Imaging at the Colour & Imaging Institute, University of Derby (Derby, UK). In September 1999 he was appointed a Professor in the School of Computing Sciences, University of East Anglia (Norwich, UK).

[3] School of Computing Sciences, University of East Anglia, UK.

[4] School of Computer Science, University of Nottingham, UK.

Appendix



(a) Atrium Night - *Karol Myszkowski*

(b) Belgium - *Dani Lischinski*

(c) Bristol Bridge - *Greg Ward*

(d) Clock Building - *Greg Ward*

(e) Fog - *Jack Tumblin*

(f) Foyer - *Harlan Hambright*

(g) Indoor - *Jacques Joffre*

(h) Memorial - *Paul Debevec*

(i) Synagogue - *Dani Lischinski*

(j) Tahoe - *Greg Ward*

(k) Tinterna - *Greg Ward*

(l) Tree - *Industrial Light and Magic*

(m) Venice - *unknown origin*

Figure 3. *Scenes used in experiments. Scene names have been kept consistent with those used in the web experiment [12].*