Fidelity of Graphics Reconstructions : A Psychophysical Investigation

Ann McNamara[1] Alan Chalmers[2] Tom Troscianko[3] Erik Reinhard[4]
[1,2,4]Department of Computer Science,
[3]Department of Experimental Psychology,
University of Bristol, United Kingdom

Abstract. In this paper we develop a technique for measuring the perceptual equivalence of a graphical scene to a real scene. Ability to compare images is valuable in computer graphics for a number of reasons but the main motivation is to enable us to compare different rendering algorithms and to bring us closer to a system for validating lighting simulation algorithms against measurements.
In this study we conduct a series of psychophysical experiments to assess the fidelity of graphical reconstruction of real scenes. Methods developed for the study of human visual perception are used to provide evidence for a perceptual, rather than a mere physical, match between the original scene and its computer representation. Results show that the rendered scene has high perceptual fidelity compared to the original scene, which implies that a rendered image can convey albedo. This investigation is a step toward providing a quantitative answer to the question of just how "real" photo-realism actually is.

1 Introduction

Realistic image synthesis focuses on generating images that emulate the impression of real scenes. It is now possible to accurately simulate the distribution of light energy in a scene, however, physical accuracy in rendering does not ensure that the displayed images will have authentic visual appearance. Even if we assume the lighting simulation is correct to within a given tolerance, problems exist with the manner in which human observers perceive the resulting images. Reasons for this include the limited range of intensities that can be displayed on a display device, and the fact that in general, most state of the art renderers don't compensate for these limitations. Some encouraging research, which models the parameters of perceptual response [4, 11], has begun, but this poses some new questions; for example how can we compare these images which claim to be realistic, to real scenes. Furthermore, many of these techniques are based on algorithms which make assumptions about the functioning of the human visual system; such assumptions often apply to restricted viewing conditions rather than complex scenes. In general, the question of how to compare images with other images to determine how alike they are, remains largely unanswered.

Several computational methods for assessing the quality of such computer generated images have been proposed. The most effective of these methods compare images based on perceptual appearance rather than photometric accuracy. Rushmeier *et al.* explored a number of such perceptually based metrics and concluded that perceptual metrics may be used to numerically compare renderings and captured images in a man-

[1]E-mail: mcnamara@cs.bris.ac.uk

ner that approximately corresponds to human contrast perception[9]. An overview of these metrics is given in the next section.

There are many ways in which perceptual fidelity could be measured; in fact, it is likely that no single measure can fully deal with what is a complex issue. For this reason we use actual human response rather than perceptual data when attempting to compare images. We chose a particular task - that of matching materials in the scene against a display of originals - because the task has a number of attractive features. First, Gilchrist [6, 7] has shown that the perception of lightness (the perceptual correlate of reflectance) is strongly dependent on the human visual system's rendition of both illumination and 3-D geometry. These are key features of perception of any scene and are in themselves complex attributes. However, the simple matching procedure used here depends critically on the correct representation of the above parameters. Therefore, the task should be sensitive to any mismatch between the original and the rendered scene. Secondly, the matching procedure is a standard psychophysical task and allows excellent control over the stimulus and the subject's response. The task chosen here corresponds closely to the methodology of Gilchrist [2, 6, 7] which permits simple measures (of lightness) to be made at locations in complex scenes. Ultimately, the task was chosen to be simple while also being sensitive to perceptual distortions in the scene. In addition, this measure of perceived contrast may be used to provide a simple way to predict object visibility within a scene [2, 7].

2 Previous Work

The following image comparison metrics were derived from [3, 5, 8] in a study which compared real and synthetic images by Rushmeier *et al* [9]. Each is based on ideas taken from image compression techniques. The goal of this work was to obtain results from comparing two images using these models that were large if large differences between the images exist, and small when they are almost the same.These suggested metrics include some basic characteristics of human vision described in image compression literature. First, within a broad band of luminances, the eye senses relative rather than absolute luminances. For this reason a metric should account for luminance variations, not absolute values. Second, the response of the eye is non-linear. The perceived "brightness" or "lightness" is a non-linear function of luminance. The particular non-linear relationship is not well established and is likely to depend on complex issues such as perceived lighting and 3-D geometry. Third, the sensitivity of the eye depends on the spatial frequency of luminance variations. The following methods attempt to model these three effects. Each model uses a different Contrast Sensitivity Function (CSF) to model the sensitivity to spatial frequencies.

Model 1 After Mannos and Sakrison [8]: First, all the luminance values are normalised by the mean luminance. The non linearity in perception is accounted for by taking the cubed root of each normalised luminance. A Fast Fourier Transform (FFT) is computed of the resulting values, and the magnitude of the resulting values are filtered with a CSF to an array of values. Finally the distance between the two images is computed by finding the Mean Square Error (MSE) of the values for each of the two images. This technique therefore measures similarity in Fourier amplitude between images.

Model 2 After Gervais et al [5]: This model includes the effect of phase as well as magnitude in the frequency space representation of the image. Once again the luminances are normalised by dividing by the mean luminance. An FFT is computed producing an array of phases and magnitudes. These magnitudes are then filtered with an anisotropic CSF filter function constructed by fitting splines to psychophysical data. The distance between two images is computed using methods described in [5].

Model 3 After Daly:adapted from [3]: In this model the effects of adaptation and non-linearity are combined in one transformation, which acts on each pixel individually. In the first two models each pixel has significant global effect in the normalisation by contributing to the image mean. Each luminance is transformed by an amplitude non-linearity value. An FFT is applied to each transformed luminance and then they are filtered by a CSF (computed for a level of 50 cd/m^2). The distance between the two images is then computed using MSE as in model 1.

One of the goals of our psychophysical investigation is to validate or refute existing metrics. In general, metrics of perceptual fidelity, such as the afore mentioned models [9], depend on factors such as the Fourier composition of the scene. We do not expect a moderate change in Fourier composition to affect the perceived illumination and 3-D properties of a scene. Thus, we seek to show a dissociation between our method and the other, algorithmic, approaches. In the case of moderate Fourier filtering (eg low-pass) we would expect "perfect" results from our method, but low apparent fidelity from the other metrics. A reverse situation is also possible, in which the 3-D geometry is corrupted without a change in Fourier content; here, our method would signal poor fidelity whereas existing metrics would indicate good fidelity. Such a double dissociation would clearly show that perceptual fidelity cannot be described by a single number; rather, psychophysical measures must be made to establish its degree.

Our technique looks for a similar set of matching data in the original and the rendered scene. If such an equality is obtained, we can conclude that the illumination and 3-D qualities of both scenes are perceptually equivalent. The psychophysical experiments we are performing aim to validate or refute existing metrics for image comparison. Our tests make no assumption about the functioning of the human visual system other than its ability to judge equality. This lack of prior assumptions makes it a potentially very robust technique. The disadvantage is the lack of a single algorithm to account for the psychophysical data.

3 Method

This study required an experimental set-up comprising of a real environment and a computer representation of that environment. Here we describe the equipment used to construct the real world test environment, along with the physical measurements performed to attain the necessary input for the synthetic representation.

3.1 The Real Scene

The test environment was a small box of 557 mm high, 408 mm wide and 507 mm deep, with an opening on one side, figure 1. All interior surfaces of the box were painted with white matt house paint.

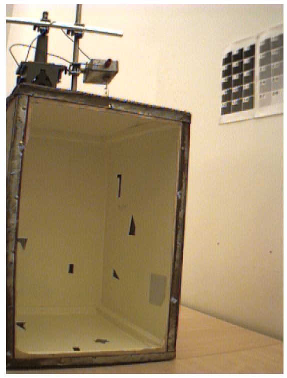

Fig. 1 Experimental Set up

To the right of this enclosure a chart showing thirty gray level patches, labelled as in figure 2, were positioned on the wall to act as reference. The thirty patches were chosen to provide perceptually spaced levels of reflectance from black to white, according to the Munsell Renotation System [13].

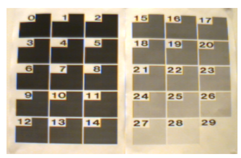

Fig. 2 Reference patches

A series of fifteen of these gray level patches were chosen at random, reshaped, and placed in no particular order within the physical environment. A small front-silvered, high quality mirror was incorporated into the set up to allow the viewing conditions to be changed to two settings, viewing of the original scene or viewing of the modelled scene on the computer monitor. When the optical mirror was in position, subjects viewed the original scene. In the absence of the optical mirror the computer representation of the original scene was viewed. The angular subtenses of the two displays were equalised, and the fact that the display monitor had to be closer to the subject for this to occur, was allowed for by the inclusion of a +2 diopter lens in its optical path; the lens equated the optical distances of the two displays.

The light source consisted of a 24 volt quartz halogen bulb mounted on optical bench fittings at the top of the test environment. This was supplied by a stabilised 10 amp DC power supply, stable to 30 parts per million in current. The light shone through a 70 mm by 115 mm opening at the top of the enclosure. Black masks, constructed of matt cardboard sheets, were placed framing the screen and the open wall of the enclosure, a separate black cardboard sheet was used to define the eye position. An aperture in this mask was used to enforce monocular vision, since the VDU display did not permit stereoscopic viewing.

The entire experimental set-up resides in an enclosed dark laboratory in which the only light sources are the DC bulb (shielded from direct view) or illumination from the monitor. Gilchrist [2, 6, 7] has shown that such an experimental environment is sufficient for our purposes.

3.2 The Graphical Representation

The geometric model of the real environment was created using Alias Wavefront [1]. The photometric instrument used throughout the course of the experiments was the Minolta Spot Chroma meter CS-100. The Minolta chroma meter is a compact, tristimulus colorimeter for non contact measurements of light sources or reflective surfaces. The one degree acceptance angle and through the lens viewing system enables accurate targeting of the subject. The chroma meter was used to measure the chromaticity and luminance values of the materials in the original scene and from the screen simulation. The luminance meter was also used to take similar readings of the thirty reference patches. For input into the graphical modelling process the following measurements were taken.

Geometry: A tape measure was used to measure the geometry of the test environment. Length measurements were made with an accuracy of the order of one millimetre.

Materials: The chroma meter was used for material chromaticity measurements. To ensure accuracy of the measurements five measurements were recorded for each material, the highest and lowest luminance magnitude recorded for each material discarded and an average was taken of the remaining three values. The CIE (1931) x, y chromaticity co-ordinates of each primary were obtained and the relative luminances for each phosphor were recorded using the chroma meter. Following Travis [10], these values were then transformed to screen RGB tristimulus values as input to the renderer by applying a matrix based on the chromaticity coordinates of the monitor phosphors and a monitor white point.

Illumination: The illuminant was measured by illuminating an Eastman Kodak Standard White powder, pressed into a circular cavity, which reflects 99% of incident light in a diffuse manner. The chroma meter was then used to determine the illuminant tristimulus values.

The rendered image was created using the Radiance Lighting simulation package [12] to generate our graphical representation of the real scene. Radiance is a physically based renderer, which means that physically meaningful results may be expected, provided the input to the renderer is meaningful.

4 Image Comparison Experiment

In this work we inspect perceptual, as opposed to physical, correspondence between a real and graphical scene by performing tests of visual function in both situations and establishing that both sets of results are identical (allowing for measurement noise). Such an identity will provide strong evidence that no significantly different set of perceptual parameters exist in each situation.

The subjects' task was to match gray level patches within a physical environment to a set of control patches. Then subjects were asked to repeat the same task with the original environment replaced by its computer representation, including some slight variations of the computer representation, such as changes in Fourier composition (blurring), see figure 3.

Fig. 3 Rendered Image (left) with blurring (right)

In the *Original Scene*, physical stimuli were presented in the test environment, described in the previous section. Subjects viewed the screen monocularly through a fixed viewing position. The experiment was undertaken under constant and controlled illumination conditions.

While viewing the *Computer Simulated Scene*, representation of the stimuli, rendered using Radiance, were presented on the monitor of a Silicon Graphics 02 machine. Again, subjects viewed the screen monocularly through a fixed viewing position.

Our use of the psychophysical lightness-matching procedure is chosen because it is sensitive to errors in perceived depth. Lightness constancy depends on a correct representation of the three dimensional structure of the scene [6, 7]. Any errors in depth perception, when viewing the computer model, will result in errors of constancy, and thus a poor psychophysical matching performance.

5 Results

Data were obtained for 15 subjects; 14 of these were naive as to the purpose of the experiment, and one was an author. Subjects had either normal or corrected-to-normal vision. Each subject performed a number of conditions, in random order, and within each condition the subject's task was to match the fifteen gray test patches to a reference chart on the wall. Each patch was matched once only and the order in which each subject performed the matches was varied between subjects and conditions.

Figure 4 shows the results obtained for comparison of a rendered and a blurred scene to the real environment. The x-axis gives the *actual* Munsell value of each patch, the y-axis gives the *matched* Munsell value, averaged across the subjects.

A perfect set of data would lie along a 45^0 diagonal line. The experimental data for the real environment lie close to this line, with some small but systematic deviations for specific test patches. These deviations show that lightness constance is not perfect for the original scene. What this means is as follows: when observing a given scene, small (but significant) errors of lightness perception are likely to occur. A perceptually-perfect reconstruction of the scene should produce a very similar pattern of errors if it

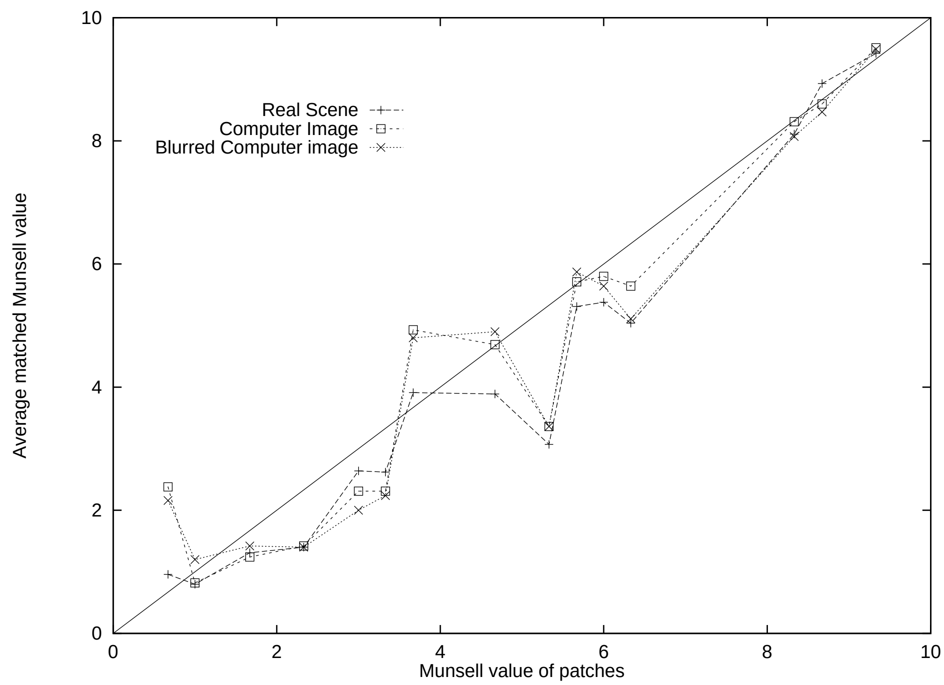

Fig. 4 Comparison of average matchings of Munsell values

is perceptually similar to the original.

The two other graphs relating to the rendered and the blurred rendered images are plotted on the same axes. In general, it can be seen that the matched values are very similar to those of the original scene, in other words, the same (small) failures of constancy apply both to the rendered and the blurred rendered images.

This, in turn suggests that there is no significant perceptual difference between the original scene and both the rendered version and the blurred rendered version. This is in spite of the fact that the mean luminance of the rendered versions was lower by a factor of about 30 compared to the original; also, under our conditions the blurred version looked very different subjectively, but again similar data were obtained.

It is possible to reduce the pattern of results to a single value as follows :

- taking the matches to the original scene as reference, calculate the mean signed deviation for the rendered and blurred rendered functions.
- Compute the mean and standard deviation of these

Table 1 shows the results obtained. A value of zero in this table would indicate *perceptually perfect* match; the actual values given come close to this and are statistically not significantly different from zero. This, therefore, again indicates high perceptual fidelity in both versions of the rendered scene

How do these values compare to other methods? Using the algorithm of Daly [3], we found a 5.04% difference between the rendered and blurred rendered images. As

Compared to Real	Mean Munsell Value Deviation
Rendered Scene	-0.37 ($\sigma = 0.44$)
Blurred Scene	-0.23 ($\sigma = 0.57$)

Table 1 Comparison of Rendered and Blurred Scene to Real Environment

a comparison, a left-right reversal of the image gives a difference value of 3.71%; and a comparison of the image with a white noise grey level image results in a difference value of 72%. Thus, the algorithm suggests that there is a marked difference between the rendered image and blurred rendered image; for example this is a 36% greater difference than that with a left-right reversed image. (This difference increases for less symmetrical images). However, our method suggests that these two scenes are perceptually equivalent in terms of our task.

It *may* therefore be that there is a dissociation between our method and that of the Daly technique. In addition, the algorithmic method cannot give a direct comparison between the original scene and the rendered version; this could only be achieved by frame grabbing the original which is a process likely to introduce errors due to the non-linear nature of the capture process. Further work is planned to attempt to capture a scene without errors of reproduction.

Figure 5 shows the results obtained for comparison of the real scene and rendered images of the environment after the depth has been altered by 50% and the patches specularity increased from 0% to 50%.

As can be seen from the graph, for simple scenes lightness constancy is extremely robust against changes in depth, specularity and blurring.

In summary: the results show that the rendered scenes used in this study have high perceptual fidelity compared to the original scene, and that other methods of assessing image fidelity yield results which are markedly different from ours. The results also imply that a rendered image can convey albedo.

6 Conclusions and Future Work

We have introduced a method for measuring the perceptual equivalence between a real scene and a computer simulation of the same scene. Because this model is based on psychophysical experiments, results are produced through study of vision from a human rather than a machine vision point of view. We have presented a method for modelling a real scene, then validating that model using the response of the human visual system.

By conducting a series of experiments, based on psychophysics, we can estimate how much alike a rendered image is to the original scene. Preliminary results show that given a real scene and a faithful representation of that scene, the visual response function in both cases is similar.

There is still much work to be done in this area, and numerous avenues for further research exist. While we have compared a real scene to the computer model we have yet to attempt the experiments using a captured image of our real scene. This has proved prohibitive in attempting to compare our method to those currently available. In particular the perceptual metrics should be used to compare our rendered image to

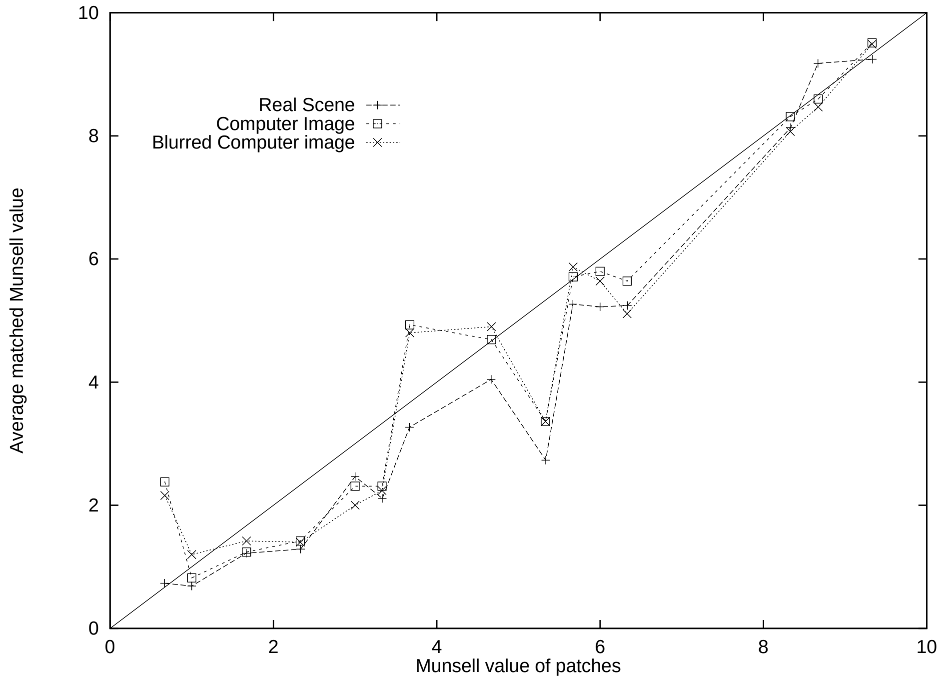

Fig. 5 Comparison of average matchings of Munsell values

a captured image of our real scene which does not suffer from distortions resulting from the capture process. This prevented us from being able to compare our method to those currently available. In an attempt to include comparison to existing metrics such as those outlined in section 2, we have compared the simulated scene to a blurred simulation of the measured scene. Results show that our method gives different results from the metric in the case of the blurred image. This is due to the likely lack of importance of high spatial frequency contents of the image to the task of matching lightness values.

Results from these experiments demonstrate how psychophysics may be used to compare real and synthetic images and in particular, how visual response in the original scene corresponds to visual response to the rendered scene. Our long range goal is to define a technique capable of identifying "perceptual" parameter space so that distances between images in a parameter space are close when humans perceive the images to be similar, and are not close otherwise. The utility being the capacity to evaluate the quality of photo-realistic rendering software, and develop techniques to improve such renderer's ability to produce high fidelity images.

Because the complexity of human perception and the compute expensive rendering algorithms that exist today, future work should focus on developing efficient methods from which resultant graphical representations of scenes yield the same perceptual effects as the original scene. To achieve this the full gamut of colour perception, as opposed to simply lightness, must be considered by introducing scenes of increasing complexity such as those including indirect illumination.

Acknowledgements

We would like to thank our colleagues who donated their time to do experiments.

References

1. Alias/Wavefront. *OpenAlias Manual*, 1996

2. J. Cataliotti & A. Gilchrist. Local and global processes in lightness perception. *Perception and Psychophysics, 57(2), 125-135, 1995*.

3. S. Daly. The visible difference predictor: an algorithm for the assessment of image fidelity. *In A. B. Watson Editor, Digital Images and Human Vision, pp 179-206, MIT Press, 1993*

4. J. Ferwerda, S. Pattanaik, P. Shirley, D. P. Greenberg. A Model of Visual Adaptation for Realistic Image Synthesis. *SIGGRAPH '97, pp 143-152, 1997*.

5. M.J. Gervais, L.O. Harvey,Jr., and J.O. Roberts. Identification confusions among letters of the alphabet. *Journal of Experimental Psychology: Human Perception and Performance, 10(5), pp 655-666, 1984*.

6. A. Gilchrist. Lightness, Brightness and Transparency. *Hillsdale: Lawerence Erlibaum Associates*.

7. A. Gilchrist. Lightness contrast and failures of lightness constancy: a common explanation. *Perception and Psychophysics, 43(5), 125-135, 1988*.

8. J.l. Mannos and D.J Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions on Information Theory, IT-20(4):525-536, 1974*.

9. H. Rushmeier et al. Comparing Real and Synthetic Images: Some Ideas About Metrics. *6th Eurographics Workshop on Rendering, pp 213-222, 1995*

10. D. Travis. Effective Colour Displays, Theory and Practice. *Academic Press, 1991*

11. G. Meyer, H. Rushmeier, M. Cohen, D. Greenberg, K. Torrence. An Experimental Evaluation of Computer Graphics Imagery *ACM Transactions on Graphics, Vol.5, No. 1, pp 30-50, 1986*

12. G. L. Ward. The RADIANCE lighting simulation and rendering system. *Computer Graphics, 1994*.

13. Wyszecki & Stiles. Colour Science: concepts and methods, quantitative data and formulae (2nd edition) *New York: Wiley*