Comparing Real and Synthetic Images: Some Ideas About Metrics *

H. Rushmeier[1] and G. Ward[2] and C. Piatko[1] and P. Sanders[3] and B. Rust[1]

[1] National Institute of Standards and Technology, Gaithersburg MD 20899, USA
[2] Lawrence Berkeley Laboratory, Berkeley CA 94720, USA
[3] GE NELA Park, Cleveland OH 44112, USA

Abstract. This paper explores numerical techniques for comparing real and synthetic luminance images. We introduce components of a perceptually based metric using ideas from the image compression literature. We apply a series of metrics to a set of real and synthetic images, and discuss their performance. Finally, we conclude with suggestions for future work in formulating image metrics and incorporating them into new image synthesis methods.

1 Introduction and Data from Physical Experiments

Research in realistic image synthesis suffers from a very basic gap in knowledge: no one knows how to evaluate a visual simulation. We need good image metrics to compare simulation methods, to validate simulations against measurements, and to guide progressive image synthesis calculations more efficiently. In this paper, we develop measures based on appearance rather than photometric accuracy. These perceptual metrics can be used to compare and to validate different rendering methods. Furthermore, they can be used to identify appropriate error quantities and error bounds for developing more efficient rendering algorithms. The goal of this paper is not to examine the accuracy of the particular program used to generate the synthetic images, but to consider how metrics for image comparison should be constructed.

There is bound to be some argument over which perceptual metric to use, since agreeing on one is so important to comparing rendering algorithms. Our goal in this paper is to get the conversation started, in hopes that one day we can settle on a satisfactory basis for comparing our visual simulation results. We intend to test these perceptual metrics on more complicated environments under more varied lighting conditions, and propose that others do the same before passing final judgement.

* Certain commercial equipment and instruments are identified in this paper in order to specify the experimental procedures adequately. Such identification is not intented to imply recommendation or endorsement by the National Institute of Standards and Technology or by the Lawrence Berkeley Laboratory, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

1.1 Input Data and Captured Images

We begin with the physical measurements performed to obtain the input for the synthetic images and to capture the "real" images. Understanding the limitations in the measurement process is important to understanding the construction of a comparison metric.

In our study we use captured and synthetic images obtained in the course of a study of metrics for lighting design. The goal of that study, which is still in progress, is to relate subjective impressions of an environment to values computed from measured luminance images, along the lines suggested by Flynn [5]. A secondary goal is to determine if current lighting simulations are capable of predicting luminance images that are accurate enough to use the new metrics in lighting design. This study required an experimental setup with a room whose lighting systems could be changed to present different qualities of illumination to human observers and to a calibrated electronic camera. The geometry, materials and light sources were also measured so that lighting simulations (renderings) of the space could be produced. The test space was a conference room, shown

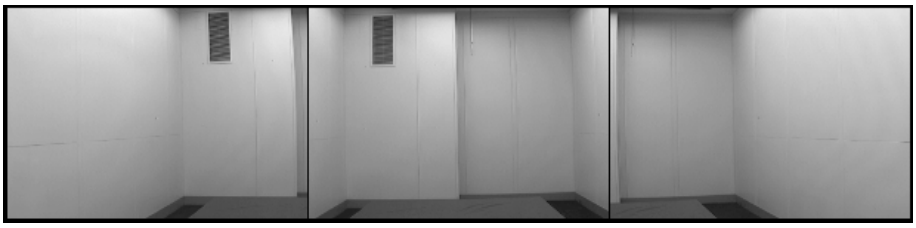

Fig. 1. These three images, obtained with a calibrated CCD camera, show left, middle and right views of the conference room with indirect fluorescent lighting.

in Fig. 1, at the National Institute of Standards and Technology. The room was modified to install downlights, wall washers, and to make adjustments in the direct and indirect lighting effects of the existing fluorescent fixtures. The room was deliberately uncluttered to isolate the effects of lighting changes on occupant impressions. As a real architectural space however, some complexity remained, such as the vent in the corner and seams in the walls.

For input into the synthetic imaging process the following data were taken:

Geometry: The room geometry was measured with a simple tape measure. In general, length measurements were made with an accuracy on the order of millimeters. Measurements were not made of the squareness of the room. The tilt of the vent blades and the angular position of the light fixtures were measured, but the data are accurate only to within a few degrees. The position of the table in the room was not measured precisely at the time luminance images were taken, so the table position is accurate only within a few centimeters.

Light Fixtures: No equipment was available for taking near-field photometry on the specific bulbs and light fixtures used in the experiment. Manufacturers' data were used for both goniometric and spectral data of the light bulbs. Since the fluorescent fixtures were modified, manufacturers' data were available for the tubes only, with the egg crate fixtures modeled as scene geometry. In general, variations on the order of 30% can be expected between the manufacturers' data and actual performance depending on supplied voltage, operating temperature and bulb to bulb variations.

Reflectances: As many of the surfaces as could be sampled and carried to a measurement device (i.e. wall paint, carpet and table covering) were measured with the goniophotometer described in [9] and were found to be nearly Lambertian. The spectral reflectances for these surfaces were measured using a Minolta CM-2002 spectrophotometer. Other surfaces' reflectances – such as the baseboards, ceiling, table legs and fluorescent fixture components were estimated. The baseboards and table legs were estimated by comparing colorimeter output for these surfaces to output for a white diffuse standard under the same lighting conditions. The ceiling and lighting fixture surface reflectances were estimated visually by comparing with a card painted with a white diffuse paint which had been measured to have a reflectance of 90%.

The reflectances measured with the goniophotometer and spectrophotometer are quite accurate for the small samples on which they were made. However, as this is a room in daily use significant variations exist in all of the surfaces due to texture, aging and dirt.

Luminance images of the room under various lighting conditions were captured using the CapCalc imaging system described in ref. [8]. CapCalc is a CCD based system developed several years ago at RPI. The system is no longer state-of-the-art, but was adequate for this initial study. A major limitation is that it stores only 8 bits per pixel. The luminance range for any particular image is adjusted by changing the camera aperture. Images requiring larger dynamic ranges can be obtained by carefully combining images. All of the images captured for this study were captured at a resolution of 512 x 480 with a pixel aspect ratio of .75 and a field of view of 35°. These images were resampled to 512 x 360 images with pixel aspect ratio of 1 for applying the various metrics being examined.

An obvious limitation of a luminance imaging system is that all of our comparisons are restricted to grey scale images. However, spectral data were used in the simulations to account for the impact of spectral variations in emission and reflectance on the measured luminances.

1.2 Impact of Measurement Uncertainty on Image Metrics

There are clearly very high levels of uncertainty in the measurements made in this experiment. The combination of uncertainties results in even larger uncertainties in the final simulated luminances. For example, the uncertainty in the directional distribution coming from one of the light bulbs aimed at the wall, coupled with a few degrees uncertainty in the orientation of the bulb and small uncertainties in the surface reflectance can easily result in uncertainties on the

order of a factor of 10 for the luminance at any specific point. However, the data we obtained are typical of what can be expected to be available in many (although not all) applications. Furthermore, the uncertainty in the data is typical of an environment in daily use. If appearance were closely dependent on precise photometry, the use of synthetic images for most design and training applications would be doomed. Fortunately, while appearance does depend on photometry, it is not a detailed linear dependence. Even with the data with high uncertainty obtained in this experiment, it is possible to synthesize images that look "more or less" like the real physical space. The purpose of the metrics we are looking for in this paper is to quantify "more or less".

2 How Should We Build an Image Metric?

One reason we want to compare real and synthetic images for specific applications is to validate software and measurement systems for input data. The level of "matching" required and the specific measurement accuracy required will always depend to some extent on the application. Another reason we want these metrics is to incorporate them into our global illumination calculations. We can compute images most effectively if we only compute image features that are visible in the final image.

A simple approach to comparing a synthetic and a captured luminance image of the same scene is to require that the luminances match at every pixel. If it were possible to obtain such a match, the resulting synthetic and captured images would appear identical to a human observer. However, matching exact luminances is sufficient, but not necessary. The human visual system is not a photometer. Ideally, we want a metric that measures how much alike two images *appear*.

How can we quantify the performance of such a metric? Here are some simple initial ideas for the properties of a metric $M(A;B)$ measuring the distance between images A and B. First, $M(A;A) = 0$. Second, $M(A;B) = M(B;A)$. Third, if images A and B appear similar, and A and C appear different, $M(A;C)/M(A;B) \gg 1$. Fourth, if images A, B, and C all look similar to one another, $M(A;B)/M(A;C) \simeq 1$. Fifth, if the differences between image A and image B are similar to the images between image C and image D, $M(A;B)/M(C;D) \simeq 1$.

In our study, all of the metrics we use meet the first two criteria. To examine the third and fourth criteria we use "random" images – i.e. images in which pixel values are assigned randomly with the same mean and standard deviation as the measured images – as images that are "obviously" different. We will use "flat" images – i.e. images with surfaces rendered assuming an arbitrary uniform ambient light – as an image that looks more like the measured image than the random image. Images rendered with the *Radiance* system [10] with varying levels of quality specified will be used as images which clearly look more like the measured images than the random or flat ones.

To examine the fifth criterion on the list we will look at the distance between two captured luminance images for two different lighting configurations, and the distance between two synthetic images for the same two lighting configurations.

The literature on human perception is voluminous, and deals with many aspects of vision outside of our area of interest. Most vision research in the area of imaging deals with detecting thresholds, that is, looking at image A and image B can you see any difference at all? For our metric, we are interested in suprathreshold effects, that is, a measure that is large when A and B are very different, and small when they are nearly the same. The vision research which would provide the basis for an entirely psychophysical model of this type does not yet exist. We have to rely on modeling a few basic characteristics of the human visual system to construct a practical system.

To get a start, we will look at how human vision models have been used in a discipline that deals specifically with images – digital image compression. Similar to the synthesis problem in which we want to save time by only computing what will be visible in the image, in image compression the goal is to save space by saving only what will be visible in the image.

For static, grey scale images, there are several basic characteristics of human vision described in the compression literature that should be included in a metric. First, within a broad band of luminances, relative luminances rather than absolute luminances are sensed by the eye. A model should account for luminance variations, not absolute values. Second, the response of the eye is non-linear. The "brightness" perceived is a non-linear function of luminance. The particular non-linear relationship is not well established. Third, the sensitivity of the eye depends on the spatial frequency of luminance variations. At least for isolated sine grating patterns, the sensitivity is low for very low spatial frequencies, increases to a maximum for a frequency on the order of a few cycles per visual degree, and then decreases again for high spatial frequencies [3].

There are many ways these three effects can be modeled and combined. To test how sensitive a metric is to the specific model used we compare three metrics "inspired" by the compression and vision literature. We use the term "inspired" because while we are using models from these other papers, our application is different, and we are not using them in the same way as in the original papers.

Each model has slightly different features. Each uses a different CSF (contrast sensitivity function) to model the sensitivity to spatial frequencies. We use the combinations of transformations as found in the literature, rather than using the same CSF for each model, since the transformations have been found successful by others only in combination.

Model 1: After Mannos and Sakrison: Following ideas presented in [7], we begin by normalizing all of the luminance values L_{ij} in the matrix $\|L\|$ by the mean luminance L_{m}. Next, the nonlinearity in perception is accounted for by taking the cube root of the normalized luminance. A fast Fourier transform (FFT) is computed of the resulting values, and the magnitude of the resulting transform at frequencies in the horizontal and vertical directions (u, v), (where u and v are expressed in terms of cycles per visual degree) is denoted $f_{uv}(\|(L/L_{\mathrm{m}})^{.333}\|)$.

f_{uv} is then filtered with $A_M(u,v) = A_M(r), r = u^2 + v^2$ to account for spatial frequency sensitivity to produce the array of values g_{uv}:

$$A_M(r(u,v)) = 2.6 * [0.0192 + 0.144\sqrt{r}] \exp[-(0.144\sqrt{r})^{1.1}]$$
$$g_{uv} = f_{uv}(\|(\frac{L}{L_{\mathrm{m}}})^{.333}\|) * A_M(r(u,v)) \quad (1)$$

Finally, the distance between the two images is computed by finding the MSE (mean squared error) of the values of g_{uv} for the two images X and Y:

$$M(X;Y) = \frac{1}{N} \sum_{all\ u,v} (g_{X,uv} - g_{Y,uv})^2 \quad (2)$$

Model 2: After Gervais et al.: In our second model, we adapt ideas from a study of confusion between letters of the alphabet [6]. Although our problem is quite different, we selected this model because it includes the effects of phase as well as magnitude in the frequency space representation of the image. Once again the luminances are normalized by dividing by the mean luminance. An FFT is computed for this normalized image, producing an array of magnitudes $f_{uv}(\|L/L_{\mathrm{m}}\|)$ and phases $p_{uv}(\|L/L_{\mathrm{m}}\|)$. The magnitudes are then filtered with an anisotropic filter function constructed by fitting splines to psychophysical data presented by Campbell et al. [2], producing the filtered values $g_{uv}(\|L/L_{\mathrm{m}}\|)$ The distance between images X and Y is then computed using:

$$M(X;Y) = \frac{1}{N} \sum_{all\ u,v} (((\log g_{X,uv} + 1) - (\log g_{Y,uv} + 1))(1 + p_{X,uv} - p_{Y,uv}))^2 \quad (3)$$

Model 3: After Daly: In the third model, we use ideas from a paper on visible difference prediction [4]. Because we are only interested in suprathreshold comparisons, we only use the first part of the algorithm described in the paper, and omit the components modeling detection mechanisms. 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 normalization by contributing to the image mean. In this model, each luminance is transformed to an "amplitude non-linearity value" b_{ij} defined by: $b_{ij} = L_{ij}/(L_{ij} + 12.6L_{ij}^{0.63})$ where the constants 12.6 and 0.63 apply when luminance is expressed in cd/m^2.

An FFT is then applied to the values b_{ij}. The resulting magnitudes, $f_{uv}(b)$ are filtered by a CSF which is a function of the image size in degrees, light adaptation level L_{m}. The dependence on the light adaptation level is very weak, and for this study the CSF was computed for a level of 50 cd/m^2. The resulting filter $A_D(r(u,v))$ is given by:

$$A_D(r(u,v)) = (.008/(r)^{1.5} + 1)^{-.2} 1.42\sqrt{r} \exp(-.3\sqrt{r}) \sqrt{1 + .06 \exp(.3\sqrt{r})} \quad (4)$$

The distance between the two images is then computed as the MSE for the filtered magnitudes of the amplitude nonlinearity values as in Eq. 2 for Model 1.

3 Results and Conclusions

In all of the tests we conducted, comparisons are made on luminance images – not on 0-255 images for which a tone reproduction operator has been used to transform the image to display. We want to find luminance images that match, with the assumption that whatever tone reproduction operator is used to represent the image in the final display medium will preserve the appearance of the luminance image. Also, comparing images before tone mapping allows us to say something about how well the simulation matches the measurements assuming an "ideal" display device.

3.1 Comparisons to Pixel by Pixel MSE

The first issue we want to address is that the three models we have proposed really do perform better with respect to our requirements for a metric than a MSE measure computed on untransformed images. Table 1 shows a comparison of the results of pixel by pixel MSE for two test cases – the left view of the room with down lights, and the left view of the room with wall washers. The results in Table 1 are typical for all of the various room configurations in our study.

In particular, we look at the ratio of the distance between the random image and the measured image (R) to the distance between a good simulation and the measured image (S), and at the ratio of the distance between a flat image and the measured image (F) to the distance between a good simulation and the measured image. Our goal is to have the first ratio be very large, since the simulation clearly looks much more like the measured image than the random image does. We also want the ratio for the flat image to be much smaller than the ratio for the random image, since the flat image does capture features of the measured image. In Table 1 we show the results of adding each of the transformations in Model 1 for the left image of the downlight and wall washer lighting configurations. First, the pixel by pixel MSE is clearly very poor – it essentially gives the random image as being as close to the measured image as the simulation, and closer than the flat image. As various transformations are added in – normalizing by average luminance (NORM), comparing cube roots (CUBE ROOT), comparing the magnitudes of the FFT of the cube root luminances (FFT), and comparing filtered magnitudes (filt FFT, Eq. 2) the measure of the relative distance to the random image increases, while the relative distance to the flat image goes to a value above one, but much less than the distance to the random image.

3.2 Considering Image Registration

One reason the MSE comparison is so poor for the measured and synthetic images is simple geometric misalignments. To test this effect, we used the "ip warp" module in AVS [1] to warp the synthetic images so that geometric features such as the vent and the table matched in the two images. The distance comparisons for the warped versus simulated images are shown in the rightmost

column of Table 1 for the wall washer lighting configuration. Aligning the geometries has minimal effect on improving pixel by pixel comparisons. This is because geometric misalignment of objects in the final image is only one type of misalignment problem. Another problem is small errors in the directionality of the light source orientation, which can cause registration problems in the light patterns. We conclude that rather than introduce more assumptions and alter images to compensate for geometric factors, it is better to compare images with metrics that measure appearance rather than point by point photometry.

3.3 Comparing the Three Perceptual Models

Given that we want to use metrics that incorporate vision models, we now test the three models using the criteria we outlined. We use three different types of comparisons.

First we look at the performance of the three models in comparing the relative distances to random and flat images, as discussed in the first subsection. Again we show results for the down lights and wall washers, with the results for other lighting configurations we studied showing similar trends. These results are presented in Table 2. The most striking feature of this comparison is that Model 2 does a much poorer job at measuring the relative distances for random and simulated images. A second observation is that Model 3 reports greater relative distances between the flat and simulated images. This is desirable performance in the metric, particularly in the case of the wall washers in which the bright spots on the walls are features that clearly differentiate the measured image from a flat rendering.

Next we look at the performance of the models at detecting more subtle effects. For this test we use the measured image, flat rendering, and the three increasingly better approximations of the images of the room with indirect fluorescent lighting. Ideally the metrics should detect the slight improvements in the decreasing spatial artifacts in the three simulations. In Table 3 we show results for the three models comparing the distances for the flat, very bad, and a little better rendering relative to the distance for the best rendering. In this case again, the most noticeable feature is that Model 2 does a poor job measuring improvements in the approximation.

Finally, we look at how well the models predict similar changes in pairs of images. We look at how consistently the models give similar results when the distance between a pair of measured images for two different lighting configurations is compared to the distance between a pair of synthetic images for the same two lighting configurations. These results are presented in Table 4. In this case Model 2 does a slightly better job. Since in this test there are no geometric misalignments, the phase data that is included in Model 2 does not detect any registration problems, which mask other parts of the comparisons.

3.4 Conclusions

We have demonstrated how perceptual metrics adapted from vision and image compression research may be used to numerically compare renderings and captured images in a way that roughly corresponds to our own subjective, human impressions. In particular, the model inspired by Daly [4] tested very well against our criteria for a good perceptual metric. The Daly model was also the only one we examined that considers human limits in dark adaptation, and we expect it to outperform the other models in low light situations which we have not yet tested. The Gervais model [6] was the only one that included phase (i.e. pixel position) information, and its performance thereby suffered due to subjectively minor registration problems between captured and simulated images in our tests. In situations where geometric alignment is not a problem, or is of critical importance for some other reason, this model may actually outperform the others.

Ultimately, the biggest challenge is to take insights into human perception and apply them to visual simulation directly, computing only as much as is needed to satisfy the observer. Recasting a perceptual model into a progressive calculation is not an easy task, and it is essential to start with the right model. The more reliable the model of human perception used in image synthesis, the more efficient and accurate rendering engines will be.

References

1. Advanced Visual Systems. *AVS User's Guide* AVS Inc., 300 Fifth Ave., Waltham, MA 02154, 1992.
2. F. W. Campbell, J. J. Kulikowski, and J. Levinson. The effect of orientation on the visual resolution of gratings. *Journal of Physiology*, 187:427–436, 1966.
3. T. N. Cornsweet. *Visual Perception*, Academic Press, New York, NY, 1970.
4. S. Daly. The visible difference predictor: an algorithm for the assessment of image fidelity. In A. B. Watson, editor, *Digital Images and Human Vision*, pages 179–206. MIT Press, 1993.
5. J. E. Flynn. A study of subjective responses to low energy and nonuniform lighting systems. *Lighting Design and Application*, pages 6–15, 1977.
6. 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):655–666, 1984.
7. 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.
8. M. S. Rea and I. G. Jeffrey. A new luminance and image analysis system for lighting and vision I. Equipment and calibration. *Journal of the Illuminating Engineering Society*, pages 64–72, Winter 1990.
9. G. J. Ward. Measuring and modeling anisotropic reflection. *Computer Graphics*, 26(2), 1992.
10. G. J. Ward. The RADIANCE lighting simulation and rendering system. *Computer Graphics*, July, 1994.

Table 1. Comparing the Performance of Several Metrics

	R/S		F/S		W/S
IDEAL	a big number		$R/S > F/S > 1$		—
	Downlights	Wall Washers	Downlights	Wall Washers	Wall Washers
MSE	1.2	1.0	1.8	1.2	0.98
NORM	1.8	4.5	0.8	2.4	0.93
CUBE ROOT	1.8	13.8	1.1	2.6	1.09
FFT	7.1	31.6	1.4	2.4	1.19
Filt FFT	44.6	599.4	1.7	1.9	0.77

R = M(measured; random), S = M(measured; simulated)
F = M(measured; flat), W =M(measured; warped)
All comparisons are for the left view of the room.

Table 2. Comparing models' sensitivities to similar and dissimilar images.

Model #	1		2		3	
	R/S	F/S	R/S	F/S	R/S	F/S
Downlights -left view	44.6	1.7	2.5	0.7	46.2	2.9
Downlights -middle view	78.7	1.5	2.7	0.8	98.4	2.6
Downlights -right view	67.1	0.9	2.0	1.5	84.3	4.1
Wall Washers -left view	599.4	1.9	7.1	1.2	768.3	3.5
Wall Washers -middle view	338.5	1.4	4.2	1.1	468.7	1.9
Wall Washers -right view	705.1	1.9	5.7	2.3	1041.3	4.4

Table 3. Comparing models' abilities to detect progressive rendering improvements.

Model #	1			2			3		
	F/S	VB/S	B/S	F/S	VB/S	B/S	F/S	VB/S	B/S
Indirect Fluorescents -left	1.41	0.99	1.00	0.8	0.95	0.95	4.6	1.05	1.02
Indirect Fluorescents -middle	1.55	1.03	1.02	0.9	0.96	0.98	3.7	1.05	1.02
Indirect Fluorescents -right	1.22	1.02	1.01	1.6	0.93	0.98	12.7	1.14	1.06

VB = M(measured; very bad simulation), B = M(measured; bad simulation)

Table 4. Comparing models with different lighting, but no geometric misalignments.

	$\frac{M(SDL;SWW)}{M(MDL;MWW)}$		
Model #	1	2	3
left view	.86	.97	1.41
middle view	1.05	1.08	0.84
right view	0.79	1.13	1.42

SDL = Simulated downlights, SWW = Simulated wall washers
MDL = Measured downlights, MWW = Measured wall washers