A Concept for Evaluating the Accuracy of Computer Generated Images

Konrad F. Karner and Manfred Prantl

Institute for Computer Graphics
Technical University Graz
Münzgrabenstr. 11
A-8010 Graz, AUSTRIA
email: karner | mprantl@icg.tu-graz.ac.at

Abstract. We present a method to compare computer generated images of a real scene to the corresponding photograph. The approach is based on a numerical comparison of luminance values. Ground-truth data was obtained by mapping the pixel values of the photograph to luminance values via point measurements obtained from an accurate luminance measurement device. We present results of this comparison performed on a very complex indoor scene and conclude with a critical discussion of the problems encountered with our approach and the task of image comparison in general.

Keywords: image comparison, lighting simulation, physically based, reflectance, illumination, surface properties, rendering.

1 Introduction

Algorithms and software from computer graphics for the photo-realistic rendering of complex scenes have reached a remarkable level of maturity over the last years. While a lot of effort is spent on making the visual results slowly approach photo-realism, little work has been done on evaluating the quality of such rendering algorithms on an objective quantitative rather then on a subjective qualitative basis. Such a test, however, would be very useful in order to rank rendering algorithms depending on the application domain and subsequently to automatically choose the algorithms most appropriate for the problem at hand.

Here we describe one possible approach towards such a metric based on the comparison of a photograph and an image of the rendered scene in luminance space. Luminance was chosen because it is essentially what the eye perceives and, hence, the differences in luminance will give a good indication on the achieved rendering quality. We present a general concept to perform such a test and highlight sources of possible problems when doing it for complex environments.

2 Main contributions of this paper

We report on the application of a global illumination software for the rendering of a complex indoor scene. The test scene is an accurate model of a real world bureau room with modeled geometry and material characteristics. The rendered scene in luminance space was compared to ground-truth data obtained from the real scene by the use of a luminance measurement device.

In order to improve the descriptiveness of our comparison the point measurements of the luminance measurement device were used to calculate ground-truth or near ground-truth data over the whole scene. We describe a simple method for accomplishing this.

However, the most important contribution of this paper will be a report on the experiences and problems encountered when trying to photo-realistically render very complex scenes using ray-tracing and an interpretation of the results obtained hereby.

3 Previous work

In the last years little work has been done evaluating the accuracy of rendering algorithms. Meyer et al. [4] described a method where both numerical as well as perceptual based comparisons were done. However, he only demonstrated his method on a very simple scene, basically only consisting of two small cubes inside a larger cube with its front side open (Cornell box). Furthermore, his measurements are based on illuminance instead of luminance.

Takagi et al. [7] compared color and luminance values for the use in car design at a small number of selected points. The aim was to render a car as realistically as possible under different weather conditions.

Rushmeier et al. [6] proposed a method to calculate a metric based on the perceptual characteristics of the human eye in order to compare computer generated images.

There was also some work done in the lighting design community where rendering algorithms, especially the radiosity algorithm, were used to simulate lighting conditions [5] [1]. Their comparisons tried to evaluate the accuracy of the simulated values against measured values, however, only the illuminance values in the working planes were taken into account.

We introduce a comparison method for computer generated images which uses both point to point as well as surface to surface comparisons and is applicable for general scenes.

4 A numerical evaluation method

The goal of our proposed method is to evaluate the accuracy of rendering algorithms. We want to come up with a quantitative measure of the rendering quality which allows for classification of different algorithms.

4.1 Obtaining Ground-Truth

In order to obtain ground-truth data for our comparison we needed to measure the luminance values present at the observed scene. Lacking a measurement device capable of obtaining accurate luminance values for every point in the scene we resorted to taking measurements at selected points which we subsequently extended to the rest of the scene by the following procedure.

1. **Luminance point measurements**
   Luminance point measurements were taken in a grid-like fashion over the whole area under investigation using a standard luminance measurement device. Positioning the device at a convenient place within the scene, we sequentially pointed it at the selected grid points and, thus, obtained a set of luminance values. Using the same viewpoint as for the point measurements we also took a photograph of the test scene. This photograph was then scanned into the computer for further processing (see Fig. 1 for the measuring setup).

2. **Extension from point to surface measurements**
   Provided the measured luminance values are approximately evenly distributed over the range of values occurring in the scene, we can estimate a mapping from pixel values (obtained from the photograph) to luminance values by a simple curve fitting procedure. We used the equation

$$l(p) = a + b * p^c \tag{1}$$

   for this mapping with p denoting the pixel value (usually in the range from 0 to 255) and a, b, c are constants to be determined by the fitting process. This method is commonly used with remote sensing imagery in order to estimate ground-truth values. Figure 2 depicts the fitted curves for our two test scenes.

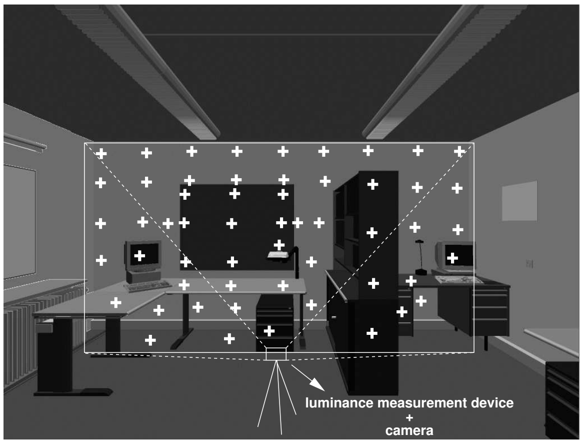
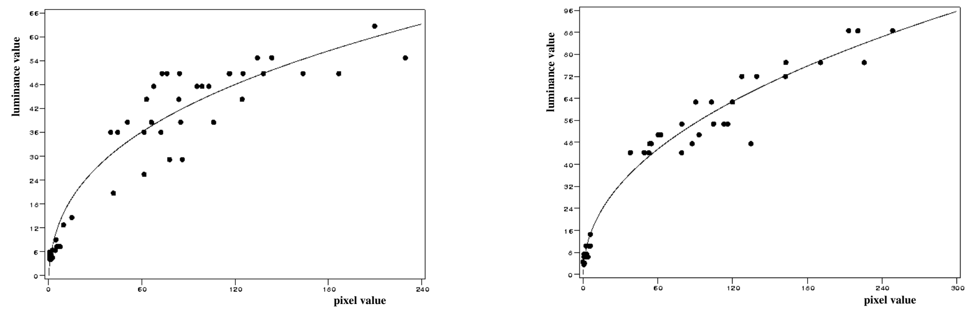


Figure 1: Measurement setup for obtaining the ground-truth values.

4.2 Rendering

As eventually we want to compare the scene generated by a specific rendering algorithm with the photograph taken from the real world, the parameters for the camera



(a) $l(p) = -7.59 + 10.45 * p^{0.35}$ (b) $l(p) = -2.62 + 7.83 * p^{0.44}$

Figure 2: The pixel to luminance value conversion

model of the rendering software have to be estimated, so that the resultant scene selection resembles the one from the real camera as closely as possible. The parameters to be chosen are the viewpoint, the viewing direction, the focal length and the field of view. We used a well known technique from photogrammetry, namely the so-called bundle adjustment algorithm, in order to estimate these parameters. Our estimates were based on a set of 10 pixel-to-world coordinate pairs, well distributed over the observed scene (see [3] for a detailed description of the procedure).

In addition to the camera specification we also need to know the geometry of the scene, the characteristics of the light sources and the reflectance functions of all observed objects, in order to do an accurate rendering. The problems that are involved in determining these quantities are discussed in section 6.

4.3 Comparison

The comparison between the luminance values of the rendered image and those of the mapped photograph was performed in two different ways:

1. **Point to point comparison**
   Here we simply compared corresponding values obtained from the luminance measurement device with those calculated by the rendering software. The RMS (root-mean-square) error and the average relative error are reported. This eliminates possible sources of bias due to the fitting function. However, this procedure does not allow for a quick visual identification of areas of large discrepancy. If the points of examination are chosen so that they are positioned inside a reasonable uniform image area, it will not be necessary to correct for slight misalignments between the rendered image and the photograph. Hence, this method of comparison can be understood as relatively robust against deviations in geometry and camera parameterization. Figure 3 illustrates the procedure.

2. **Surface to surface comparison**
   Comparing whole surface patches instead of just single image points requires slightly more effort. It now is more critical to adjust for geometrical misalignment. If the approximation of the camera parameters and the geometry of the

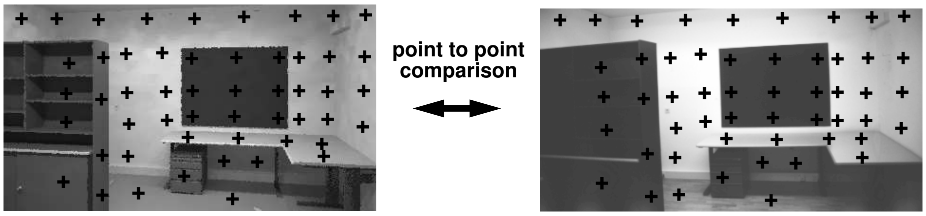


Figure 3: The point to point comparison in view B

scene, both input to the rendering algorithm, are close to the real values (as compared to the photograph), it often is sufficient to do a simple 6 parameter affine mapping between the two images. This can quickly be accomplished by selecting a number (> 3) of corresponding points from the images and then performing a least-squares approximation of the transformation parameters. In most cases this should give a result sufficiently accurate for the comparison. Misalignment might still be visible at the edges of objects and, therefore, those areas are excluded from the evaluation using a mask obtained by an edge detection algorithm which ignores points where edges occur. Again we calculate the RMS and the average relative error. Using the surface based evaluation method now offers also a visual way for a quick comparison and a spotting of problem zones. Furthermore, as the calculation of the error values now is based on a larger number of measurements, the accuracy of the test should increase. Figure 4 depicts the general outline of the method described above.

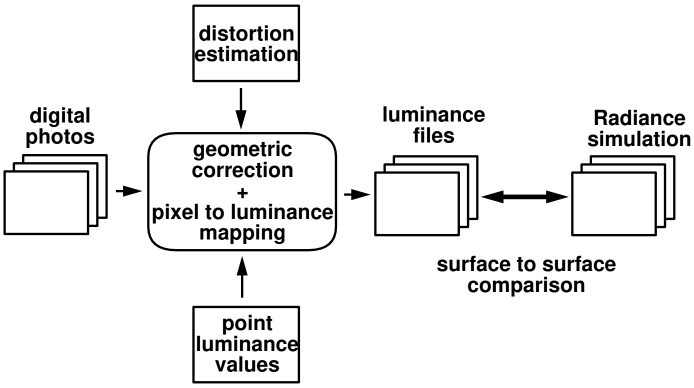


Figure 4: Work flow of geometric correction, pixel to luminance mapping and the surface to surface comparison

4.4 Experimental Setup

We tested our method on two views obtained from the same scene as depicted in Figure 3 and Figure 6. The office scene was accurately modeled using a common CAD package so that geometric deviations from real world are within a few millimeters. For the color measurements we used a Minolta Chroma Meter CR-200b and the reflectance functions were determined by the method described in [2]. The data from the light sources (luminous power and light distribution function) are obtained from accurate measurements supplied by the manufacturer. The images were rendered using the Radiance software package, a distributed ray tracing algorithm from the Lawrence Berkeley Laboratory [8] that returns luminance values. For each view a high and low quality rendering was performed. Finally, the luminance values at selected points were obtained by Minolta Spotmeter F and extended over the whole image area as described in section 4.1 using a scanned photo.

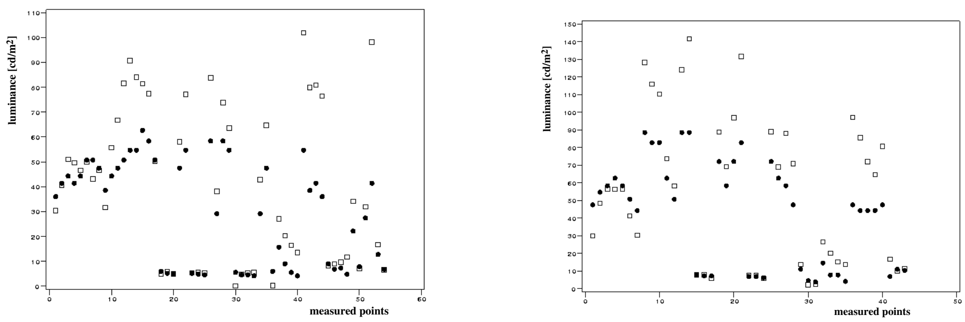
5 Results

1. **Point to point comparison**

   The luminance values measured at discrete grid points were compared to corresponding simulated luminance values. Figure 5 graphically illustrates the obtained values for the high quality rendered images. We then calculated the RMS (root-mean-square) error and the average relative error (F_{rel}) for the two different views.

$$
\begin{aligned}
RMS &= \sqrt{\frac{\sum_{i=1}^{n} (lum_{calc,i} - lum_{meas,i})^2}{n}} \\
F_{rel} &= \frac{1}{n} * \sum_{i=1}^{n} \frac{(lum_{calc,i} - lum_{meas,i})}{lum_{meas,i}} \\
n &\ \ldots\ \text{number of points}
\end{aligned}
$$

   The results are summarized in Table 1.



(a) Image A_{high} (b) Image B_{high}

Figure 5: Results from point comparison for the high quality renderings. Filled dots denote measured quantities whereas squares indicate calculated luminance values.

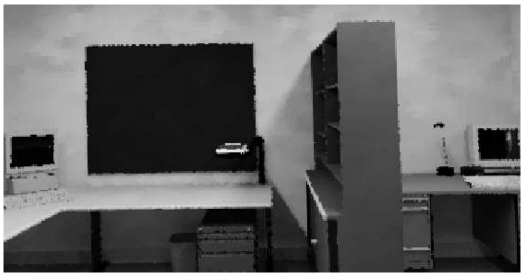
2. **Surface to surface comparison**

   Here we compare pixels stemming from a whole image rather then just selected points. The luminance values for the photograph were obtained by the procedure described in 4.1. Figures 6a-b show a high and low quality rendering

Table 1: RMS and average relative error of the point to point comparison

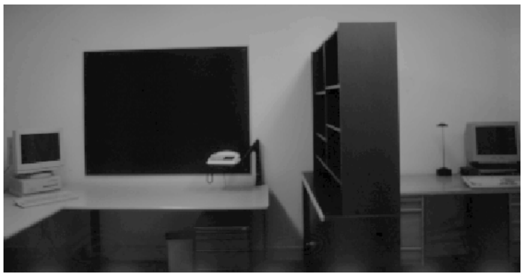
Image	RMS error $[cd/m^2]$	F_{rel} [%]
A_{high}	18.2	46
A_{low}	21.8	59
B_{high}	21.8	45
B_{low}	20.6	44

of one of the views selected for comparison, with both images being already geometrically corrected to the photograph.



(a) High quality computer generated image of view A

(b) Low quality computer generated image of view A



(c) Photograph of view A

Figure 6: Selected test scene (rendered with high and low quality) and corresponding photograph

Figure 7a shows the relative error in the image. It can be seen that most of the errors are due to geometric misalignment. Thus, we excluded those places in our error calculations. The 'mask' which can be seen in Figure 7c shows the ignored parts (white). Figure 7b depicts errors larger than 100%. Table 2 summarizes the RMS and the average relative error (F_{rel}) for the unmasked and masked luminance images for a high and low quality rendering.

(a) Relative error (b) Relative error larger than 100% (c) Mask

Figure 7: Relative error, relative error larger than 100% and the mask

Table 2: RMS and average relative error of the surface to surface comparison

Image	RMS error $[cd/m^2]$	RMS_{mask} error $[cd/m^2]$	F_{rel} [%]	$F_{rel,mask}$ [%]
A_{high}	16.5	12.9	44	32
A_{low}	18.5	17.8	56	52
B_{high}	17.8	14.8	71	36
B_{low}	16.4	16.3	55	39

6 Critical Discussion of the Method

Having a critical look at the results of Table 1 and Table 2 one will recognize several aspects. It sometimes happens that even though the image quality is subjectively poorer the values of the RMS and/or the average relative error suggest the opposite or do at least not clearly give a rate towards this observation. One reason for this is that the eye is more sensitive to changes than to absolute values, a fact that is not well captured in our metrics. Another problem lies in the fact that bright and dark image regions are not treated uniformly. The RMS error tends to be small on images with many dark regions whereas the F_{rel} is usually pretty large on this kind of test scenes. A possible remedy might be to adjust all test images to the same average luminance.

Other problems connected to the method of a ranking metric are more difficult to alleviate. There is, for example, the problem in obtaining accurate input values. The data about the luminous power of the light sources as stated by the manufacturer often deviates by up to 30% from the true value. Additionally, the light distribution functions are normally only valid in a distance of about 12 meters, where a point light source can be assumed. Errors in this data, however, are fully visible in the result of the ray tracing algorithm. Other sources of errors are the geometry of the modeled scene and the accuracy of the reflectance function and the color of the observed materials. The error introduced by the former we tried to reduce by excluding locations of sharp reflectance discontinuities from our error calculations. Nevertheless, orientation inaccuracies of materials with specular peaks or strongly anisotropic reflectance function can clearly contribute a considerable amount of deviation to our result. The error in the reflectance function and the color of the materials stems from the devices used to measure these quantities and lies in the range of 5% for the chromaticity and 10% for the reflectance function. This all suggest that a ranking scheme should try to not rely on absolute values but rather try to mimic characteristics of the eye, which obviously compensates for many such inaccuracies. Furthermore, it might not be the most useful way of obtaining a ranking scheme to compensate the rendering algorithms to a photograph as ground-truth measure but to some standard 'high-quality' rendering algorithm with precisely known input data. Of course this does not give an answer of how much photorealism an examined rendering approach has achieved.

Concluding we can remark that image comparison still lies in its infancy and that the contribution of this paper should be seen as one small piece towards a solution to this problem. We presented an approach that obviously has shortcomings, however, the discussion of these problems might help to make the difficulties conscious to people working in this field.

7 Acknowledgments

The authors would like to thank Michael Gervautz for his encouragement and many fruitful discussions. The second author also gratefully acknowledges the support by the Austrian 'Fonds zur Förderung der wissenschaftlichen Forschung' under grant S7003.

References

[1] Andrew J. Anderson and Michael Grant. Visulux: a radiosity based lighting design tool. In P. Brunet and F.W. Jansen, editors, *Photorealistic Rendering in Computer Graphics (Proceedings of the Second Eurographics Workshop on Rendering)*, pages 227–239. Springer-Verlag, New York, 1994.

[2] Konrad F. Karner. Using images to estimate reflectance functions. In *International Conference in Central Europe on Computer Graphics and Visualization, Plzen, Czech Republic*, 1996.

[3] Karl Kraus. *Photogrammetrie, Band1, Grundlagen und Standardverfahren*. Ferd. Duemmlers Verlag, Bonn, 1990.

[4] Gary W. Meyer, Holly E. Rushmeier, Michael F. Cohen, Donald P. Greenberg, and Kenneth E. Torrance. An experimental evaluation of computer graphics imagery. *ACM Transactions on Graphics*, 5(1):30 – 50, January 1986.

[5] M. Nithisoontorn. Shooting algorithm for illuminance: Comparison of prediction with experiment. *Lighting Res. Technology*, 26(1):13 – 18, 1994.

[6] H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust. Comparing real and synthetic images: Some ideas about metrics. In P.M. Hanrahan and W. Purgathofer, editors, *Rendering Techniques '95 (Proceedings of the 6th Eurographics Workshop on Rendering)*. Springer-Verlag, 1995.

[7] Atsushi Takagi, Hitoshi Takaoka, Tetsuya Oshima, and Yoshinori Ogata. Accurate rendering technique based on colorimetric conception. In *Computer Graphics (Proc. Siggraph '90)*, volume 24, pages 263 – 272. ACM Press, August 1990.

[8] Gregory J. Ward. The radiance lighting simulation and rendering system. In *Computer Graphics, Siggraph Conference Proceedings*, July 1994.