Verification of Physically Based Rendering Algorithms

Christiane Ulbricht[§] and Alexander Wilkie[§] and Werner Purgathofer[§]

Institute of Computer Graphics and Algorithms, Vienna University of Technology, Austria

Abstract

Within computer graphics, the field of predictive rendering *is concerned with those methods of image synthesis which yield results that do not only* look *real, but are also radiometrically correct renditions of nature, i.e. which are accurate predictions of what a real scene would look like under given lighting conditions.*

In order to guarantee the correctness of the results obtained by such techniques, three stages of such a rendering system have to be verified with particular care: the light reflection models, the light transport simulation and the perceptually based calculations used at display time.

In this report, we will concentrate on the state of the art with respect to the second step in this chain. Various approaches for experimental verification of the implementation of a physically based rendering system have been proposed so far. However, the problem of proving that the results are correct is not fully solved yet, and no standardized methodology is available. We give an overview of existing literature, discuss the strengths and weaknesses of the described methods and illustrate the unsolved problems. We also briefly discuss the related issue of image quality metrics.

Categories and Subject Descriptors (according to ACM CCS): I.6.4 [Simulation and Modeling]: Model Validation and Analysis

1. Introduction

In recent years, a lot of work has been published in the field of photorealistic computer graphics concerning more accurate rendering algorithms, more detailed descriptions of surfaces, more realistic tone mapping algorithms and many other improvements for the process of rendering photorealistic images. Unfortunately, there have been comparatively few attempts on verifying all these algorithms in practice. In the field of photorealistic rendering, it should be common practice to compare a rendered image to measurements obtained from a real-world scene. This is currently the only practicable way to prove the correctness of the implementation of a true global illumination algorithm. However, most rendering algorithms are just verified by visual inspection of their results.

The process of verifying a photorealistic renderer can be divided into three steps ([GTS*97]). First, one has to prove the correctness of the light reflection models through comparisons with measured physical experiments. One way to do this is to use a gonioreflectometer to measure the full bidirectional reflectance distribution functions (BRDFs) of all surfaces. The second step is to verify the light transport simulation, or – in other words – the actual rendering algorithm. This can be done by comparing the rendered image to a measurement of a real scene, or through analytic approaches. The final step in image synthesis generates an image from radiometric data provided by the rendering algorithm; it has to take the properties of the output device and the actual viewing conditions into account. In this stage psychophysical models are used to achieve convincing results, therefore perceptual experiments are used to evaluate how believable such an image is.

This state of the art report will focus on the second step: the verification of physically based rendering algorithms. Although this step is crucial in photorealistic image synthesis, comparatively little work has been published about this topic so far. The problem is far from being solved, though. Most of the available rendering systems are not verified. The questions of how this could happen, why most photorealistic images nevertheless look very good and why it is nonethe-

§ {ulbricht|wilkie|purgathofer}@cg.tuwien.ac.at

less very important to do research on this topic will be answered during this report.

This report is organized as follows: Section 2 points out the difference between believable and correct photorealistic rendering. Important applications of predictive rendering are illustrated in section 3. Section 4 provides an overview of existing verification approaches that are categorized by the device that was used to obtain the measurement of the real scene. Several attempts on defining scenes where the solution can be obtained analytically are described in section 5. Section 6 briefly overviews some approaches in the direction of image quality metrics. In the final section we discuss future challenges and draw our conclusions. Appendix A gives an overview of photometric quantities that are used throughout this report, whereas appendix B describes the corresponding radiometric quantities. The different kinds of measuring devices that are mentioned as being used in various verification approaches are characterized in appendix C.

2. Believable vs. Correct

Photorealistic image synthesis has made outstanding progress in the last decades. While the pioneers of this field tried to achieve realism with comparatively simple algorithms like raytracing and radiosity, a lot of sophisticated algorithms – such as path or photon tracing, as well as numerous hybrid techniques – have been developed since then.

Nowadays it is possible to render images that can hardly be distinguished from photographs. Consider for instance the Alias website “Is it Fake or Foto” [Ali05a]. There you can attempt to guess the origin of ten images depicting various scenes, some of them real and some of them rendered with Maya [Ali05b]. If you know what you have to look for, it is indeed possible to choose the right answers, but at first sight the renderings look believably real.

The results are convincing, but we cannot assume that the rendering algorithm is implemented correctly as far as physical accuracy is concerned. However, for many applications this is good enough. For movies, computer games and related areas it is essential that the images look appealing and – above all – believable. It is not necessary and sometimes even counterproductive that an image is a physically correct representation of the light transport in a scene.

However, for *predictive rendering* (as the field has become to be known), this is absolutely crucial. Its applications are much more restricted than the largely artistic realm of ordinary, “just believable” rendering techniques; the effort involved is far greater – e.g. because one has to use measured surface reflectance data or complicated analytical models for the scene description – and the creativity of the user is severely restricted by the constraints of physical correctness (e.g. the inability to fine-tune the lighting in the scene for artistic reasons on a per-object basis as it is common practice with current commercial systems).

3. Applications of Predictive Rendering

Even though it is not to be expected that predictive rendering will ever replace the existing artistic image synthesis tools for the abovementioned reasons (although methods from this field will probably continue to make inroads as optional enhancements, such as selective raytracing of reflective objects), it still commands a small but highly viable niche market for all those who need accurate imagery of virtual scenes.

3.1. Virtual Prototyping

Many appearance-sensitive branches of industry, for example the car industry, have to build expensive prototypes of their products before they actually go into serial production. In the last years, a lot of effort was put into reducing the costs by using computer generated images of the products. But this can only be done if the rendering software is proven to accurately simulate the light propagation within a scene.

The “RealReflect” project [Rea05] was initiated in order to investigate and improve the process of measuring surface reflectance properties and generating high quality renderings out of the data. However, these renderings are worthless if it cannot be proven that the implementation of the rendering algorithm is accurate.

3.2. Architecture and Lighting Design

Architects and lighting designers are interested in a reliable simulation of the lighting conditions of a building. The more precise this lighting simulation is, the more accurate is the output of the design process. Beside artificial light, many natural lighting situations have to be considered, like for example direct sunlight effects, ambient skylight effects, different time conditions, and different weather conditions. This kind of calculations requires a validated physically based rendering system.

4. Different Approaches to Verification

4.1. Visual Comparisons

A practicable way of verifying a renderer is to compare the results of the rendering process to a real scene. Two representative approaches are discussed in this section.

4.1.1. The Original Cornell Box

The first approach in this direction was done in 1984 by Cindy M. Goral [GTGB84]. She showed that the radiosity algorithm behaves similar to light propagation in a real setup of the scene. One reason for choosing this method was that the rendering equation had not been defined in the form known today [Kaj86].

The setup – also known as the Cornell Box – consisted of

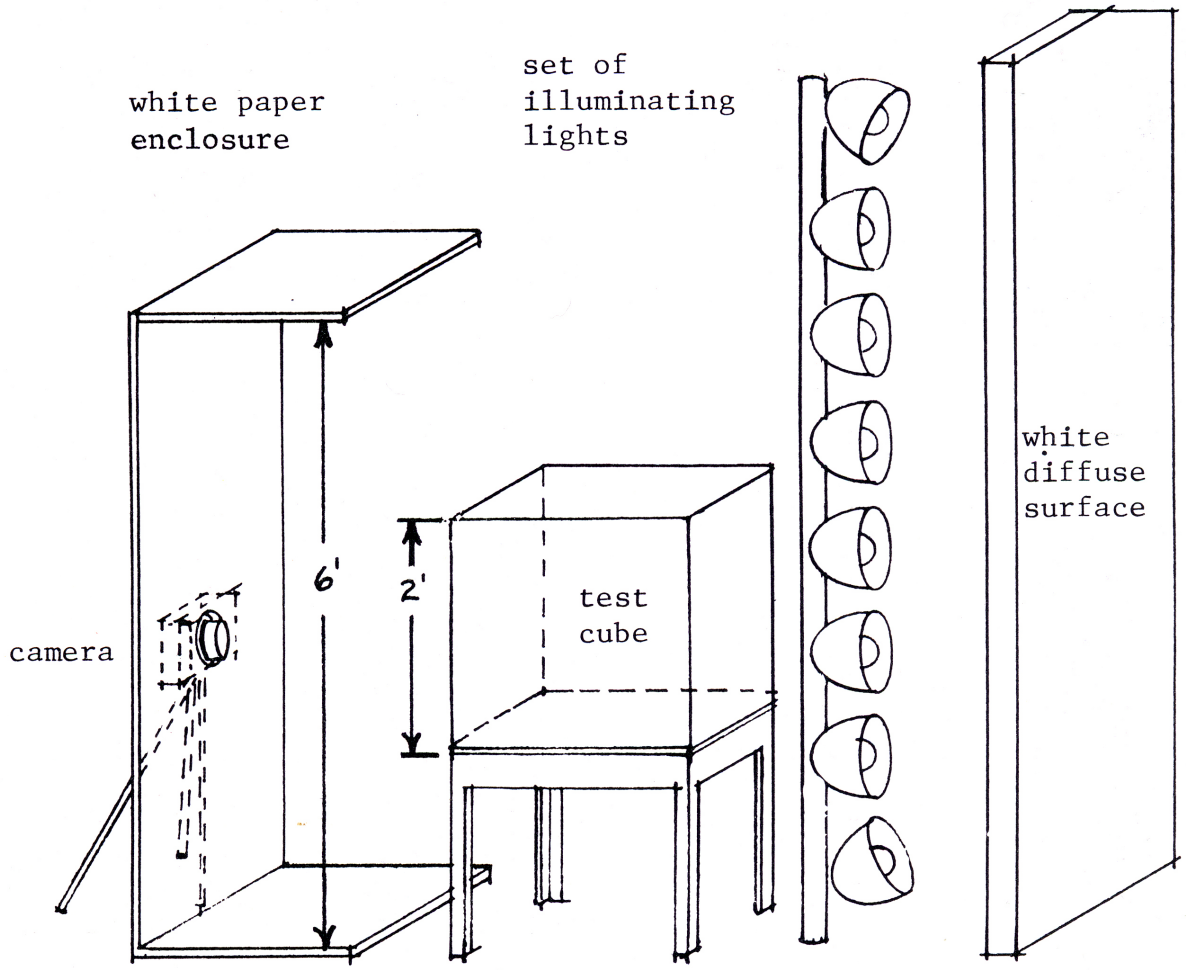

Figure 1: *A schematic of the first Cornell Box setup [GTGB84].*

a cube made out of fiber board panels, colored with flat latex paints to minimize specular reflections. The panels were painted with different colors to evoke the effect of color-bleeding. The cube had just five sides to be able to take photographs of the scene and to illuminate the inside with diffuse white light. A schematic of the setup can be seen in Figure 1. In order to simulate a diffuse and uniform area light source at the front side of the cube, the point light sources did not face the cube but a diffuse white surface. In front of the cube, there was another white surface that contained a small hole for the camera. Due to the multiple reflections, the lighting can therefore be considered diffuse.

Due to the lack of color measurement devices, no quantitative but only visual comparisons could be made. Hence, the result of the comparison is very superficial. It can only be said that the color-bleeding that is visible on a photograph of the cube (see Figure 2) is also present on a rendered image of the scene (see Figure 3). So, the structure of the light distribution can be verified but not the colors. They do not correspond to each other, e.g. the white surfaces of the photograph look much redder than the ones of the simulation. This is probably caused by using light bulbs that emit reddish light.

4.1.2. Recent developments

In 2000, McNamara et al. [MCTG00] again picked up the idea of visual comparisons. In their paper they present a way to quantify these subjective impressions. Like in the previous attempt, they also built a five sided box as a test environment. The interior of the box was painted with diffuse white paint, whereas the objects that were put inside the box were painted with different kinds of gray. The spectral reflectances of the paints were measured and converted to RGB values. Beside the box a computer monitor was placed to present simulated

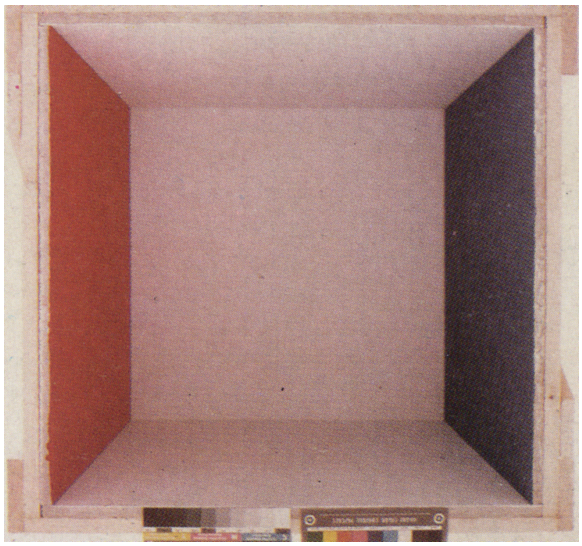

Figure 2: *A photograph of the real cube [GTGB84].*

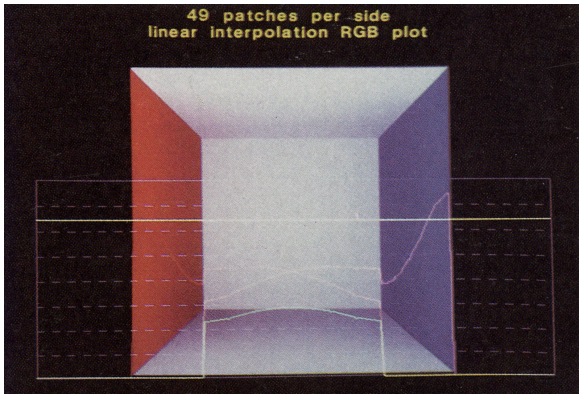

Figure 3: *Radiosity with linear interpolation [GTGB84].*

images of the scene. In addition, a mirror was used to facilitate alternation between the two settings.

Ten different types of images were selected to be compared to the real scene:

- A photograph,
- three images rendered with Radiance [War94] at different quality levels,
- a brightened high quality Radiance image,
- two Radiance simulations with controlled errors – one with inaccurate RGB values for the materials and one with inaccurate RGB values for the light source,
- a tone-mapped Radiance image and
- two images generated with RenderPark [Ren05], one using raytracing, the other using the radiosity algorithm.

Eighteen observers were asked to match the gray levels of the presented setting to a predefined set of samples, i.e. to judge the lightness of the objects. For this purpose, they have been trained to do this by recalling the different gray levels from memory before the actual experiment. The 11 different settings – the 10 images and the real scene – were presented in random order. Afterwards, the gray levels chosen by each participant in an image were compared with the values chosen in the real scene. The closer these values are, the closer the image is to the real scene.

In summary, the results of this study showed that high quality Radiance images, tone-mapped images and even one of the defective images were good representations of the actual scene. Though, the low quality Radiance simulations, the Controlled Error Materials image, the raytraced image and the radiosity image differ considerably from the real setting.

4.2. Radiometer

One year after the first experiments in Cornell (see section 4.1.1), the radiosity algorithm was improved by projecting onto an imaginary cube instead of a sphere. This hemi-cube radiosity [CG85] was again verified using the Cornell Box approach. Now, radiometric measurements of the cube were taken and compared with the results of the algorithm. Moreover, a visual side-by-side comparison was accomplished.

In the context of the development of the hemi-cube radiosity algorithm, the importance of verifying the implementation of a rendering algorithm was recognized. In 1986, Meyer *et al.* [MRC*86] investigated the procedure of experimental evaluation of rendering systems in detail. The following quotation emphasizes the necessity of experimental verification:

> "If a scientific basis for the generation of images is to be established, it is necessary to conduct experimental verification on both the component steps and the final simulation." [MRC*86]

The assembly of the Cornell Box had slightly changed since the first experiments had been made. The light bulbs and the diffuse white surfaces were replaced by a light source on top of the cube. A small rectangular opening was cut into the top panel and covered with a piece of opal glass to provide diffuse light. An incandescent flood light was mounted 15 inches above on top of a metal cone whose interior was painted white. In order to avoid interreflections with the surrounding, the box was placed on a black table and the walls were covered with black fabric. Moreover, the panels of the box could be exchanged by other panels with different colors.

In order to render an image that can be compared to the real world scene, the properties of the light source and the surfaces have to be measured. The spectral energy distribution of the light that shone through the opal glass was

acquired by a method described by Imhoff [Imh83]. Moreover, the intensity and the direction of the light were needed for the calculations. They were measured by using a photometer combined with an infrared filter. The inaccuracies of the device have been adjusted by a correction factor. Another piece of equipment – a spectrophotometer – was employed to measure the spectral reflectances of the surfaces that where painted in different colors.

For radiometric comparisons, a device is needed that accurately simulates the behavior of the human eye, like the ability to capture multiple samples at each wavelength band of light. Since such a device was not available, Meyer *et al.* obtained a radiometer instead. This device was able to measure over the range of the radiometric spectrum but provided just a single reading, which is not enough to represent a whole scene. Therefore, 25 evenly distributed measurements were taken (see Figure 4). The sample points were chosen that way to avoid shadows and to maximize the amount of light that hits the probe.

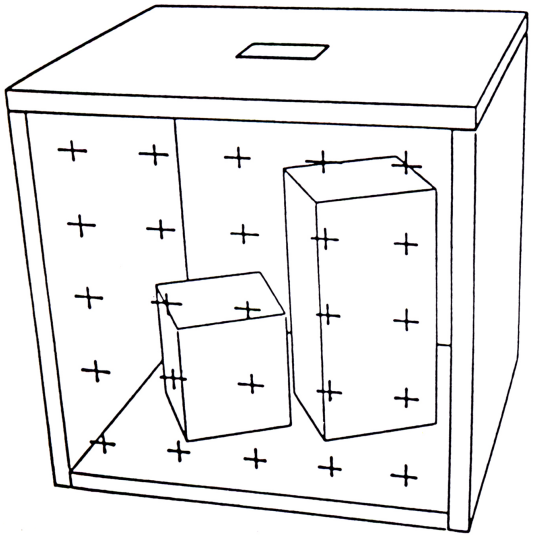

Figure 4: *The positions of the 25 radiometric measurements [MRC*86].*

Three different test scenes were created to analyze the verification procedure. Two of them consisted of an empty cube – the first had only white panels, the second contained one blue panel. The third test scene was a white cube with a white box inside it. Figure 5 shows the results of the third test scene. The cube is tilted in a way that the front panel is on top and the top panel is facing the left side. The irradiation H is shown on the vertical axis. The red lines show the result calculated by the radiosity algorithm whereas the blue lines represent the actual measurements. The correspondence between the values of the computer generated image and the real scene is clearly visible. They have a root mean square difference of about four percent.

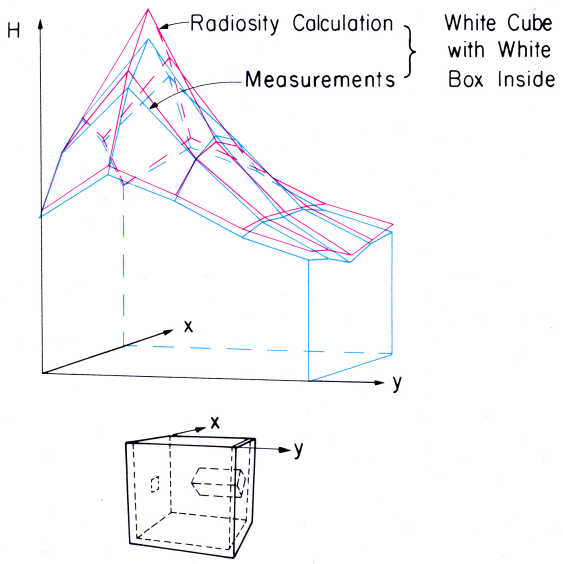

Figure 5: *Results of the test scene with white panels and a white box inside it [MRC*86].*

It has to be mentioned that the described method only works with diffuse environments. None of the devices is able to measure specular or anisotropic surfaces. Moreover, 25 measurements are not representative for more complex scenes. But using more samples makes the procedure even more time-consuming.

4.3. Incident Color Meter

In 1990, Tagaki *et al.* [TTOO90] wanted to verify the results of a global illumination algorithm they developed to render realistic previews of car models. They reduced the comparisons to a few relevant feature points. A car was analyzed under different weather conditions and the most critical sections were selected (see Figure 6).

A Minolta CS-100 incident color meter was used to measure the chromaticity and the luminance of these points. According to Tagaki *et al.*, the measured and the calculated values were almost equal. Unfortunately, the verification process is not described in detail. Therefore it is not clear how the measurements were acquired and why just six samples were taken. Figure 7 shows a photograph and a rendering of one of the car models. It can easily be seen that those images are not equal. A difference image (see Figure 8) reveals that they are not geometrically equal and that there are big differences in color values on some parts of the car, in the shadows and in the background. Figure 9 points out the color differences even more.

Figure 6: *Points A to F were selected for measurements [TTOO90].*

4.4. Scanner

Because of the disadvantages mentioned in section 4.2, the Cornell Box setup was further enhanced. The radiometer was replaced by a scanner to verify a global illumination algorithm based on spherical harmonics [SAWG91] in 1991. Three colored filters were used to retrieve information about the spectral distribution of the light. The filters transmitted a wide range of wavelengths, whereas the algorithm calculated values for three specific monochromatic channels. Thus, only visual comparisons of the structure of the illumination were possible – similar to the very first approach (see section 4.1).

4.5. Photometer

The following sections describe four different verification approaches that use a photometer to gather measurements of a real scene.

4.5.1. Model of an Office Room

Karner and Prantl [KP96] developed a method to verify complex scenes. They created a model of an office room including material characteristics and used the Radiance software package for rendering. Several different approaches for comparing a rendering of this model to measurements of the real scene were made. A photometer (Minolta Spotmeter F) was used to obtain luminance values at selected points of the scene. No color measurements were taken.

The first approach consisted in just comparing the measured values to the calculated results. In order to avoid problems that arose from misalignment of the two scenes, the positions of the samples were chosen in a way that they did not lie near edges (see Figure 10). For testing purposes, the model was rendered from two different viewpoints and in high and low quality. Figure 11 shows a high quality rendering, a low quality rendering and a photograph of the scene. The root mean square error lay within a margin of 18.2% to 21.8% while the average relative error lay between 44% and

Figure 7: *Top: A photograph of a car. Bottom: A rendered image of this car [TTOO90].*

Figure 8: *A difference image of photograph and rendered image (see Figure 7).*

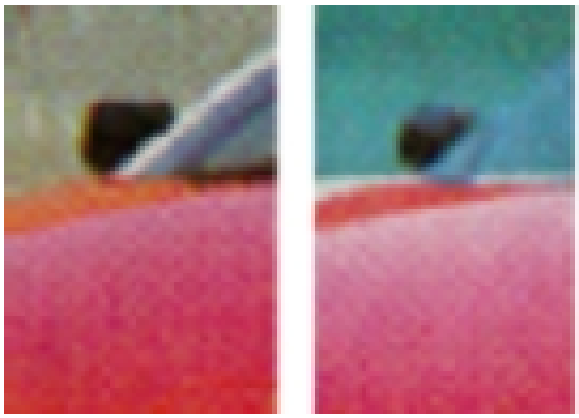

Figure 9: *A comparison of a small section of the front part of the car (see Figure 7).*

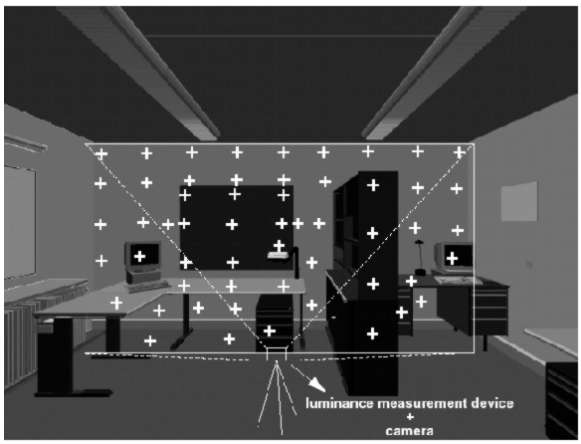

Figure 10: *The sample points do not lie near edges [KP96].*

Figure 11: *From top to bottom: a high quality rendering, a low quality rendering, and a photograph of the office scene [KP96].*

59%. The low quality renderings achieved similar results as the ones that were rendered in high quality although they obviously should perform worse. The authors' explanation for this is that the eye is more sensitive to changes than to absolute values. From this follows that the combination of point to point comparisons and the root mean square error is not sufficient for quantitatively verifying a rendering system.

In the second approach, the authors compared whole surface patches instead of points. They did not increase the number of measurements but took a photograph of the scene and scanned it for further processing. Hence the luminance values are known for some of the pixels of the photograph, the other values could be calculated by using a curve fitting function. The root mean square error was now between 16.4% and 18.5% and the average relative error was between 44% and 71%. Most of the errors occurred because of a misalignment of the images. Moreover it should be mentioned, that for one of the scenes the root mean square error for the low quality rendering was again lower than for the one with high quality.

In order to cope with the problem of misalignment, the edges were excluded from the evaluation for testing purposes. This reduced the root mean square error a value between 12.9% and 17.8% and the average relative error to a value between 32% and 52%. However, the margin of error was still very high. One reason for this was that the specifications of the light sources differed by up to 30% from the true value. Furthermore, the devices used to measure the material characteristics introduced an error in the range of 5% to 10%.

4.5.2. Validation of Daylight Simulations using Radiance

In 1999, Mardaljevic [Mar99] finished his PhD on the validation of daylight simulation using the Radiance rendering system. In the course of his work he investigated whether

Radiance yielded reliable results for daylight modeling of architectural scenes. The measurements required were accomplished within the scope of the International Daylight Measurement Programme (IDMP) organized by the CIE. These particular measurements were taken from July 1992 to July 1993. On the one hand, the sky was scanned by mounting a sky monitoring apparatus on top of a building. During daylight hours, 150 readings were made every 15 minutes. On the other hand, the illuminance inside two rooms of the same building was measured using photocells at the same time. Six photocells were positioned 0.7m above floor level along the center line of each room (see figure 12). The second room was used to test the effect of several different innovative glazings. The simulation was based on the recordings of the scan of the sky. The result of the simulation could therefore directly be compared to the values that were measured inside the building.

In order to achieve a wide range of different sky conditions, 754 representative sky measurements were selected for further processing. The internal illuminances at the six photocells were calculated using Radiance for all of the 754 skylight configurations. Figure 13 shows one scatter plot for each photocell, each containing a comparison of the predicted and the measured illuminances. It can be seen that most of the dots lie on or near the diagonal, i.e. the measured and the calculated values are equal or nearly equal. Though, for high illuminances the accuracy decreases noticeably. Especially the first photocell, placed in front of the window, yielded a high number of over and under predictions. So, for bright clear sky conditions the prediction is less reliable than for overcast skies. Still, 63.8% of the internal illuminance predictions were within a margin of $\pm 10\%$ of the measured values.

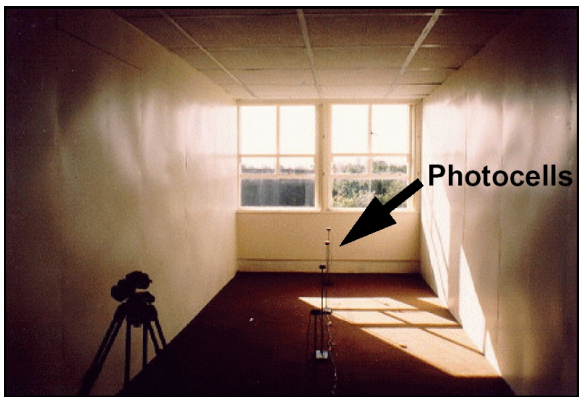

Figure 12: *Six photocells were positioned along the center line of each room [Mar99].*

An exhaustive analysis was done in order to find out whether the errors were related to measuring errors or misalignments in the model representation, or whether Radiance yielded inaccurate predictions. Although it was not possible to find a single cause of error, it could be shown that most likely multiple inaccuracies in the model representation were responsible for the errors rather than the Radiance program itself. Geometric misalignments as well as meteorologic phenomena such as small bright clouds, fast moving patchy clouds, rain, snow or heavy showers could have biased the model. Radiance was therefore capable of reliably predicting the illuminance inside a building for a wide range of sky types – especially overcast skies – but results that were generated when the circumsolar region was visible from the point of calculation were considered to be potentially inaccurate.

The PhD of Mardaljevic also contains an evaluation of existing skylight models (e.g. the CIE models and the Perez model [PSM93]) and a description of how Radiance can be used to predict the *daylight factor*, which describes the ratio of the internal illuminance at a point to the global horizontal illuminance under overcast sky conditions and which is commonly used by architects and design consultants to evaluate the lighting situation inside a building.

4.5.3. Model of the Atrium at the University of Aizu

Two years later, Drago and Myszkowski [DM01] used a photometer (more precisely a luxmeter, see photometer in appendix C) to acquire data about a real scene. Their goal was to provide a complete set of data which characterized a non-trivial existing environment for the test of physically based rendering engines. Unlike the Cornell Box scenes, which are of low complexity in terms of geometry and lighting, they wanted to build a scene that can be used to test the overall functionality of a rendering system. For this purpose, they created a model of the atrium at the university of Aizu based on the blueprints and on the actual scene. More than 80% of the surfaces in the atrium consisted of six materials, whose BRDFs were measured and included in the model. The reflectance properties of the remaining surfaces were estimated. For the luminaires, the goniometric diagrams were received from the manufacturer and corrected by a maintenance factor accounting for depreciation and aging. Figure 14 shows a rendering and a photograph of the atrium.

For the comparison, 84 sample points were chosen on the floor of the atrium (see figure 15). The illuminance values were obtained with the luxmeter and then compared to the output of a rendering engine that used the DEPT technique [VMKK00]. The calculated values and the measurements matched quite well. For a high quality rendering, the average simulation error was 10.5%.

Moreover, Drago and Myszkowski did a visual comparison of a rendered image, a photograph, the real scene and a tuned rendering, i.e. the Lambertian and mirror reflection coefficients were intuitively tuned by a skilled artist to get the best match of image appearance in respect to the real scene. They asked 25 subjects to rate how similar the im-

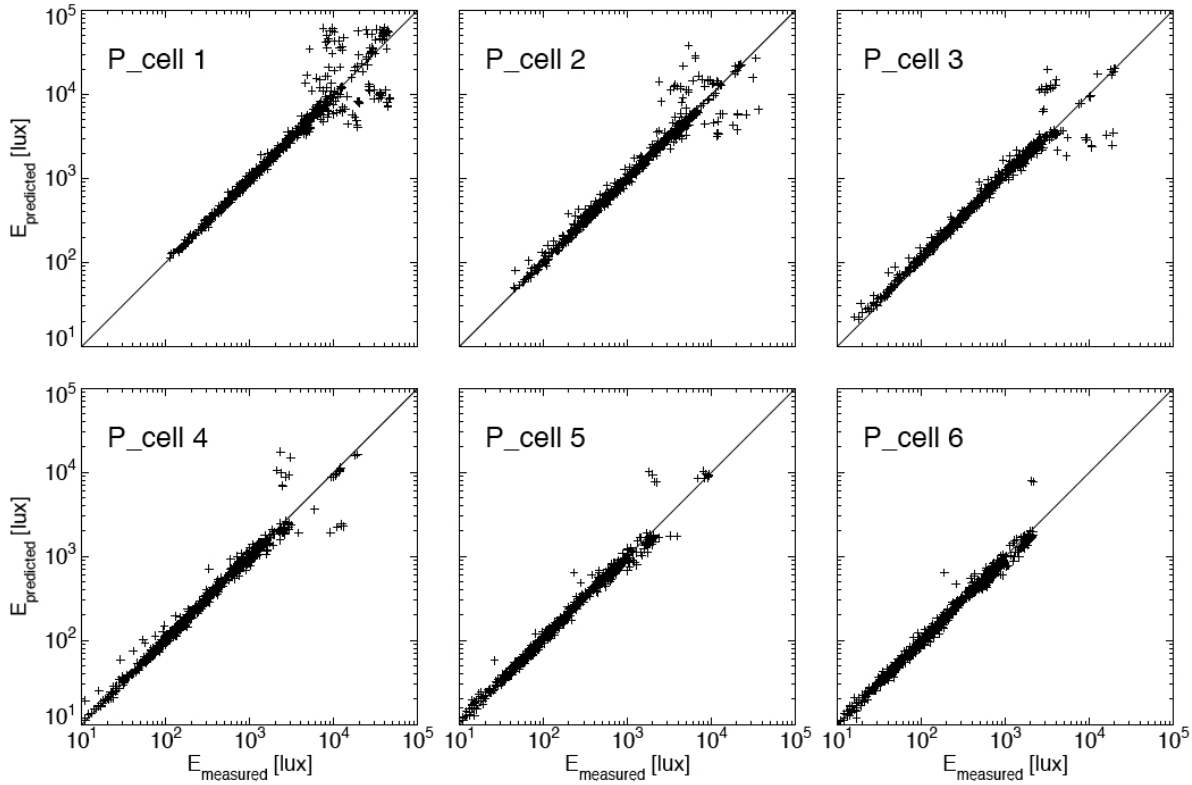

Figure 13: *Six scatterplots that compare the predicted and the measured illuminances for each photocell [Mar99].*

ages were in respect to the real atrium scene. In all cases, the photograph got the highest score. The tuned image was found to be more realistic than the rendered image in terms of overall lighting and tone, which is not surprising because it was post-processed. This might be part of the explanation why industry prefers tweaking rendering parameters instead of doing physically based renderings. But it has to be mentioned that the artistic approach cannot be used if the scene does not exist in reality, i.e. when predictions of a scene have to be generated.

4.5.4. Component Case Studies

Recently, Schregle and Wienold [SW04] presented another approach for using luxmeters to verify a photorealistic rendering system, which is also the topic of Schregle's PhD thesis. Unlike Drago and Myszkowski, they focused on a setup that can be used to test each effect of global illumination separately. Therefore, they built a box similar to the Cornell box (see figure 16).

On the top, the bottom, and the sides of the box, belts were mounted in order to guide the sensors. The belts were covered with the box's interior material. They were driven in parallel by a shaft, which was operated manually. Figure 17 shows a schematic of the sensor guidance mechanism.

Moreover, there were four additional sensors mounted on the front face of the box to measure the direct illuminance from the light source. So, they were able to correct the specifications of the manufacturer by a maintenance factor. The inside of the box was covered with heavy molleton, which was found to be nearly lambertian. In order to be able to validate caustics as well, a so-called sandblasted aluminium light shelf was attached at the open side of the box. The BRDFs of the molleton and the aluminium were obtained using a goniophotometer.

Schregle and Wienold did two different kinds of case study. First, *component* case studies were done, where individual components of the rendering system were tested. After this, *compound* case studies were performed which were combinations of different *component* case studies. Lighting simulations were performed using photon maps and the Radiance engine to be able to compare forward and backward raytracing methods.

Four different case studies are presented in this paper: diffuse patch reflection, light shelf caustics, diffuse interreflec-

Figure 14: *The left image shows a rendering of the atrium at the university of Aizu, while the right image shows a photograph of the real scene [DM01].*

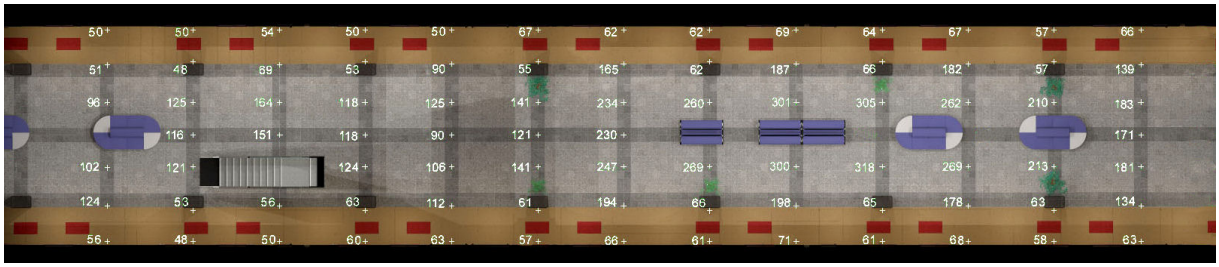

Figure 15: *Illuminance values measured with a luxmeter at sampled points on the floor of the atrium [DM01].*

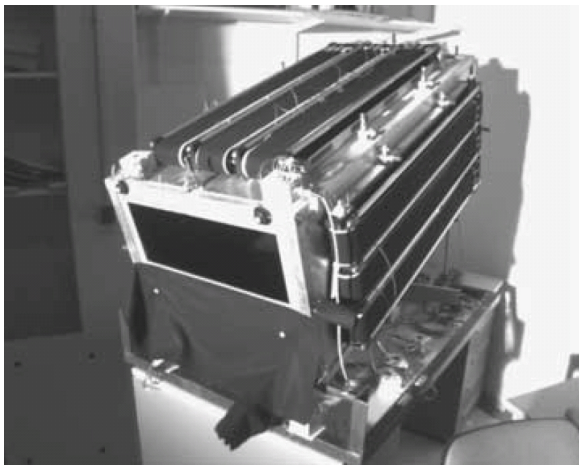

Figure 16: *The validation test box [SW04].*

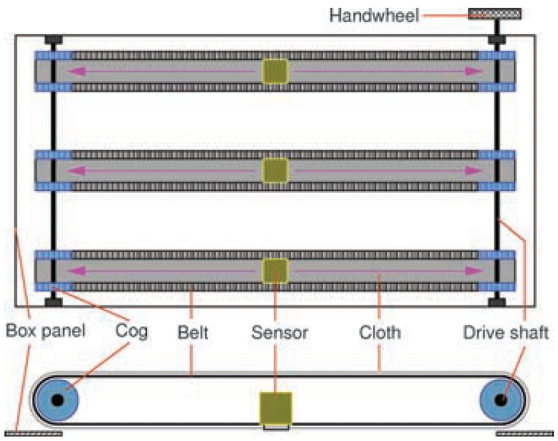

Figure 17: *A schematic of the sensor guidance mechanism [SW04].*

tion, and a combination of light shelf caustics and diffuse interreflection. For the first test case, two patches of light gray molleton where placed between the floor sensor tracks. The resulting illuminance was measured on the ceiling. A schematic of the inside of the box can be seen in figure 18a.

A comparison of the measured values and the results of the Radiance and photon map calculations is shown in figure 18b. The vertical bars in the measured data indicate the sensor tolerance. It can be seen that the curves fit quite well. The average deviation for Radiance is 3%, while it is 2% for the photon map.

For the case study involving caustics, the aluminium light shelf was mounted on the outside of the window in order to create a caustic directed toward the ceiling. The accurate simulation of this effect was possible with both rendering algorithms, though the calculations of Radiance were slightly more noisy than the ones of the photon map algorithm. This is also reflected in the average deviations, which are 7% for Radiance and only 2% for the photon map.

The third case study that is described in the paper is a generalization of the diffuse patch reflection case and therefore a compound case study. The whole interior of the box was covered with light gray molleton. The BRDF data had to be corrected because the molleton parts were from different consignments. After this adjustment, the average deviations were 1% for Radiance and 2% for photon map.

Similar results were achieved for the fourth test case, a combination of the light shelf caustic and the diffuse interreflection. Due to the correction of the BRDF data, deviations averaged 1% and 2%, for Radiance and the photon map respectively.

Both algorithms performed well and the relative deviations were consistently within the estimated error margins. Nonetheless, this setup can only be used for point samples but not for other purposes as for instance for verifying the exact shape of the caustic.

4.6. Charge-coupled Device (CCD)

In 1997, the Cornell Box approach was again used for verification purposes. Now a CCD camera was used for direct colorimetric comparisons of synthetic and real images. This was the first attempt where values for the whole image and not just a few samples were captured. Pattanaik *et al.* [PFTG97] describe the procedure of calibrating a CCD camera to reduce the error that is introduced by different forms of noise and the extraction of color values out of a monochromatic CCD. Seven narrow band filters were used to distinguish between the different ranges of wavelengths. Then, the CIE tristimulus values (X,Y,Z) were computed for each CCD element and for each pixel of the computer generated image. Figure 19 shows the results of those calculations. A difference in color values is clearly visible. The measured image is redder in the upper left corner whereas the computed image appears greener on the right side. A scaled difference image of the luminance values Y can be seen in Figure 20. It can be seen that most of the errors occur at the edges of the objects and on the light source.

5. Analytical Tests

Arbitrary images that were created by using a stochastic global illumination algorithm are often just verified by the programmer because the result includes random variables and can therefore not easily be compared with a reference solution. This leads to the fact that sometimes errors in the implementation are explained by random noise or other artifacts and are not further investigated. In order to avoid this

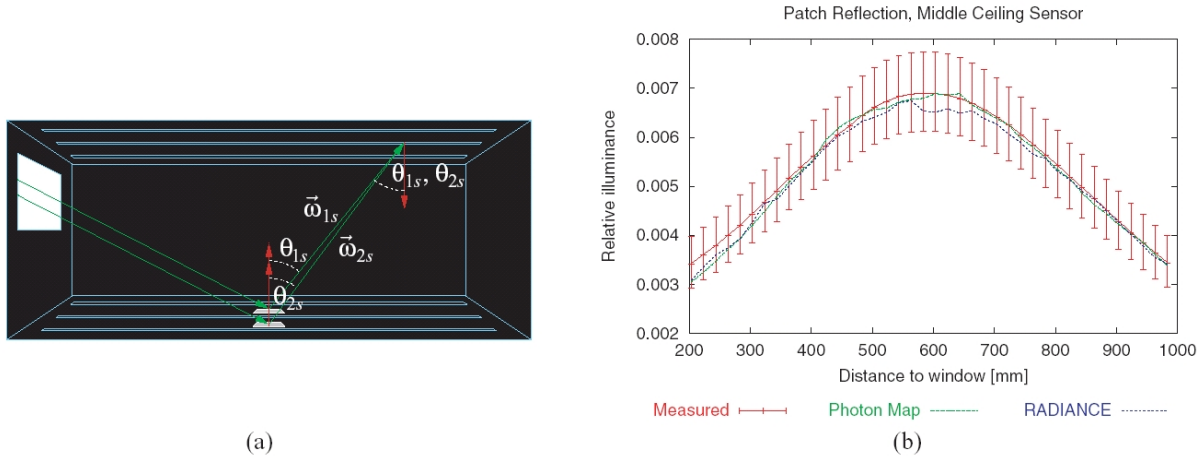

Figure 18: *(a) A schematic of the diffuse patch reflection test case; (b) Comparison of the measured illuminance with the results of the photon map and Radiance [SW04].*

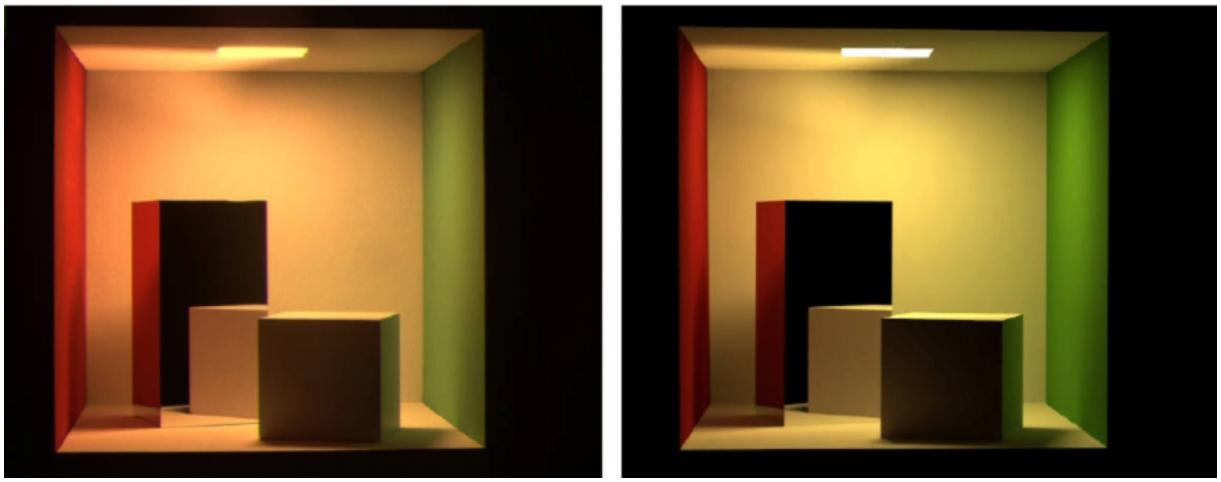

Figure 19: *Left side: Measured image. Right side: Computed image [PFTG97].*

kind of misinterpretation, another way of testing the correctness of the implementation of Monte Carlo global illumination algorithms is to create scenes where the solution can be determined analytically.

5.1. Scenes With a Constant Solution

In 2001, Szirmay-Kalos *et al.* [SKKA01] extended previous attempts in order to be able to test Monte Carlo global illumination algorithms and to use arbitrary BRDF models and light sources. Two different types of test scenes were described in this paper: scenes with constant radiance and scenes that represent the internal surface of a sphere. Scenes with constant radiance could contain arbitrary geometry with the restriction that it had to be a closed environment. As the radiance was constant everywhere and in every direction, the incoming radiance was also constant. To be able to use non-diffuse rendering algorithms, two approaches were presented. First, the BRDFs were given and the light source that was needed to provide constant radiance had to be calculated. In the second approach, the light sources were left unchanged and the BRDFs were determined in a way that the radiance was again constant.

In scenes that represent the internal surface of a sphere, the material models and the light sources could be defined more generally. Again Szirmay-Kalos *et al.* presented two different versions of this test scene. In one case, all the surfaces and light sources were supposed to be diffuse, in the other case, the surfaces were perfect mirrors.

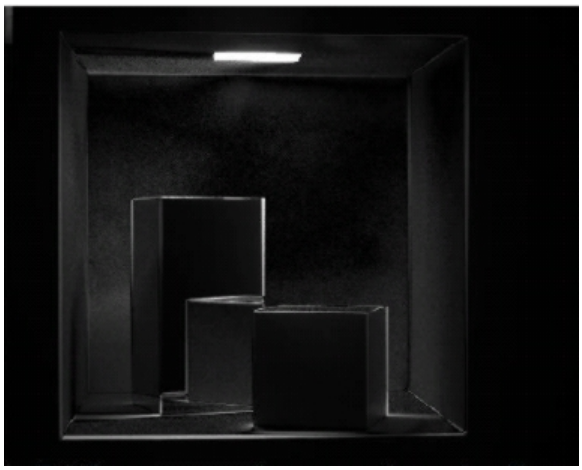

Figure 20: *Scaled difference image of the luminance values of the images shown in Figure 19 [PFTG97].*

In all cases, the fully converged solution was completely white. This allows the programmer to verify the basic implementation of the global illumination algorithm, because the correct solution of the calculation is already known. If the result of the rendering process contains areas that are not white, the programmer can be sure that the implementation is not correct.

5.2. A Representative Set of Test Cases

Another attempt on defining a set of representative test cases with analytical solutions was done by Maamari and Fontoynont [MF03] in 2003. They created six test cases where the solution is known and therefore can be compared to the results of a simulation. The tests were done with a radiosity solution only, but the authors state that the set of test cases will be expanded by scenes that include specular surfaces as well. This work is part of the main objectives of the CIE Technical Committee 3.33 [Cie05], that concentrates on "Test Cases for Assessment of Accuracy of Interior Lighting Computer Programs".

The first test case consists of a square room with either an opening at the roof or at one of the walls. The interior surfaces are black, i.e. they have a reflectance of 0%. A skylight model for cloudy sky and a sun at 90° elevation and zero luminance is used as light source. The total direct luminous flux arriving at the opening surface has to be equal to the flux reaching the interior surfaces. The radiosity algorithm produced a systematic error of about -16% for the roof opening and +13% for the wall openings.

To simulate windows, a second test case was defined where a perfectly specular glass material is positioned at the top of the opening. Several simulations were done where the incident angle of the light varied from 0° to 90° in 10° steps. However, the software that was used did not take into account the transmittance variation of incoming light with incidence angle. Therefore the internal illuminances were over-estimated.

In order to describe the intensity distribution of luminaires, the most commonly used file formats are the IESNA format and the Eulumdat format. To test the quality of the interpolation of the intensity, a third test case was created that consists of a horizontal surface with a point light source at 3 m height from the surface center. It could be shown that the radiosity algorithm solved this test case correctly because the error was below 0.2% compared with the analytical solution.

The shape factor formulae, which are used to define the fraction of luminous flux leaving a perfectly diffuse surface that reaches an elementary surface, were verified in the fourth test case. The geometry is again a square room as in the first two test cases, but instead of the opening there is a diffuse area light source at the center of the ceiling. Again, the error was very low (below 1.3%) and it could be proved that the software used was able to calculate this effect correctly.

The fifth test case dealt with the reflectance of the room surfaces. The surfaces of a square room are defined as uniform diffusers. A point light source at the center of the ceiling is used for illumination. The errors were within a negligible margin except for the 0 and 0.9 to 1 reflectance values. The authors assumed that these errors came from an epsilon value being affected at these extreme values, but also point out that these values rarely exist in reality.

The last test case was specified to verify the calculation of the daylight factor, which is commonly used to evaluate daylighting inside buildings. The geometry is the same as in the fourth test case. For the analytical solution, a module developed and validated by Dumortier and Van Roy [DVR02] at the LASH laboratory of the ENTPE–France was used. Unfortunately, in this case the software yielded errors of over 13%.

The work of Maamari and Fontoynont shows that defining a set of test cases where the solution can be derived analytically is a promising approach to validate global illumination rendering engines. Having a reliable benchmark would help to classify existing software.

6. Image Quality Metrics

Section 4 described various approaches on how to gather measurements of a real scene and how to compare them to computer-generated images. All of them used their own methods to compare the results. Unfortunately, the results are therefore not directly comparable with one another. The question of how to quantify the actual difference of two images has not been solved satisfactorily so far. Basically, there

are two different approaches: either the values are compared analytically or the properties of the human visual system and human perception in general are taken into account as well. The following sections will present different methods of both approaches. Additional information can be found in [CMD*00] and [McN00].

6.1. Mean Square Error

The mean square error (MSE) and related purely numerical metrics of image difference give rise to some problems, when the task is to evaluate whether two images would appear similar to a human observer or not. If a new image is generated by randomly redistributing the pixels of a given picture (i.e. by scrambling it), the MSE would declare them to be equal, whereas a human observer would obviously report them to be completely different. Another example can be seen in figure 21, where the MSE yields a better result for the corrupted version than for the blurred one.

6.2. A Perception Based Metric

Neumann et al. [NMP98] proposed an algorithm that operated in image space and on color values, without any Fourier or wavelet transforms. The basic idea was to calculate the CIE LUV average color difference between corresponding subregions of the images. As subregions they used a number of various rectangles. The differences between the images were weighted according to the rectangle size and the contrast sensitivity function. The sum of this differences was divided by the number of rectangles, which gave the actual difference between the images. Elaborate tests showed that this approach behaved as requested.

6.3. A Meta-Metric

In 1995, Rushmeier *et al.* [RWP*95] performed a comparison of three different perceptual image metrics. They tested variations of methods previously described by Mannos and Sakrison [MS74], Gervais *et al.* [GHR84] and Daly [Dal93]. All three methods are performed in Fourier space. The tests were based on how human vision models are used in the field of digital image compression. This discipline tries to save only those features of an image that will actually be visible to a human observer. The same methods can be used in image comparison to detect those parts of two images that are visually different. Rushmeier *et al.* suggest that the following three particular characteristic properties of the human visual system should be included in a metric. First, the human eye is more sensitive to relative luminances than to absolute luminances. Second, the response of the eye is non-linear. Third, the eye is more sensitive to spatial frequencies on the order of a few cycles per visual degree than to lower and higher spatial frequencies.

Figure 21: *From top to bottom: the actual image, a blurred version and a corrupted version [NMP98].*

The comparisons were not only done between measurements and simulations, but also with two other kinds of images. In one case, the pixel values were generated randomly, but with the same mean and standard deviation as the measured image (“random image”). In the other case, the image was rendered with uniform ambient light, so that it looked similar to the original image, but was not equal (“flat image”). According to human vision, this type of image looks more like the original image than the random image, although it has a different mean value. A proper perceptual

metric should account for this fact. All tests were made before the tonemapping step and on luminance values only.

Comparing the mean square error (MSE) of those four types of images indicates that the randomly generated image is more close to the measured image and to the simulation than the flat image. This is obviously not the desired behavior of a perceptually based image metric. Rushmeier *et al.* point out that the proposed metrics perform significantly better than the MSE. The first and the third metric clearly identify the random image as completely different from the measured image and the flat image as very close. Though, the second metric does not perform that well because it uses information about pixel positions and is therefore vulnerable to geometric misalignment. A slightly shifted image yields practically the same result as a randomly generated one. To avoid this, they tried to compensate for this effect by warping the synthetic images so that the geometry matched the measurements. Though, the improvement was only minimal because there were also other misalignment problems, like errors of the direction of the light that lead to different light patterns. Rushmeier *et al.* propose to use metrics that measure appearance rather than metrics that do a point by point comparison.

7. Future Challenges and Conclusions

In predictive rendering, it is crucial that one can rely on the results of the rendering system one uses. In section 4 we have discussed several different approaches for gathering measurements of a real scene which have been published so far; none of them is what one could consider a truly workable, robust solution suitable for widespread use. While the sophistication of the published techniques has grown considerably over the years, even the latest contributions still have weaknesses.

One possible avenue for improvement over the state of the art is that in the past years new, stable, compact and robust radiometers have been introduced to the market; such devices have not been used in computer graphics verification experiments so far. Devices of this kind– such as for example the Minolta CS-1000 – are typically able to accurately capture absolute radiance values over the entire visual range in nanometer intervals. Even though it provides just one spot measurement at a time, use of such devices could offer a new quality to the accuracy of verification experiments.

However, as we point out in section 4.2, spot measurements are not sufficient for the verification of a rendering system. Advanced CCD sensors with waveband filters could be used for whole-frame captures, possibly in conjunction with one of the new high-accuracy spectroradiometers to provide integrated calibration and reference data to the experiment.

The fact that no standard procedures for this fundamental problem of computer graphics exist also raises the question: what open research challenges remain – apart from the possible use of improved measurement devices? The following list is probably incomplete, but should give an idea on just how far-spread the problems are that face anyone who wants to improve on the state of the art:

- Differences in scene and viewing geometry between real and synthetic images – mainly caused by misalignment and inaccurate measurements – are a source of error whose exact influence on the verification process has not been characterized in sufficient detail. The approach of Karner and Prantl [KP96] to simply skip the problematic areas of the images under consideration might be an improvement over not acknowledging the problem at all, but is something that should definitely be improved upon.
- Whole classes of physical effects like e.g. polarization and fluorescence have never been verified in practice yet. Since the problems for which predictive renderers are being used – e.g. glare prediction in the automotive industry, which needs polarization support if it is to be highly accurate – very often depend on just these capabilities, this is a grave omission. This is aggravated by the fact that the measurement procedures published in literature so far are usually not capable of characterizing such effects at all.
- Is it possible to introduce a ranking scheme for physically based rendering systems? What are the criteria for such a ranking scheme? Although the focus of this report mainly is on the acquisition of measurements of a real scene, these questions should not be disregarded.
- One reason why verification is not very popular in this field is because currently there is no easy way of performing it. Accurate measurements require very expensive measuring devices. It has to be investigated whether there is a possibility to use less costly devices (such as for instance commercially available digital cameras), and to characterize the error that is introduced by doing this.

As the capabilities of photorealistic rendering engines continue to grow and appearance–sensitive industries increasingly rely on predictive rendering technologies for their virtual prototyping processes, one can confidently expect that these problems will become active research areas in the foreseeable future, and that the field of image synthesis system verification will become more active as the demands of customers for accuracy guarantees in commercial rendering solutions grow.

As a concluding remark it can be said that the groundwork in this field has been well laid, but that the task of developing robust, practicable solutions is still mostly before us.

Appendix A: Photometric Quantities

Photometric quantities consider only the visible part of the electromagnetic spectrum and take the visibility factor $V(\lambda)$ into account, which is the photopic sensitivity curve of the human eye. There is an analogous series of units for radia-

tion measurements that consider the total amount of radiation (see appendix B).

Luminous intensity

Luminous intensity is a measure of the energy emitted by a light source in a particular direction. The SI unit of luminous intensity is *candela (cd)*.

Luminous Flux

Luminous flux is a measure of the energy emitted by a light source in all directions. The SI unit of Luminous flux is the *lumen (lm)* or *candela* · *steradian*. One lumen is defined as the amount of light that falls on a unit spherical area at unit distance from a light source of one candela.

Illuminance

Illuminance is the total luminous flux incident per unit area. It refers to the amount of incident light. The unit is *lux(lx)* or *lumen per square meter*.

Luminance

Luminance is the amount of visible light leaving a point on a surface in a given direction. The unit is *candela per square meter*.

Appendix B: Radiometric Quantities

The quantities of the radiometric measurements correspond to photometric quantities (see appendix A) with the *watt* replacing the *lumen*.

Radiant Intensity

Radiant intensity corresponds to luminous intensity. The unit is *watts per steradian*.

Radiant Flux

Radiant flux corresponds to luminous flux. The unit is *watt*.

Irradiance

Irradiance corresponds to illuminance. The unit is *watts per square meter*.

Radiance

Radiance corresponds to Luminance. The unit is *watts per steradian per square meter*.

Appendix C: Measurement Devices

Incident Color Meter

An Incident Color Meter is used for the measurement of luminance and standard colour value properties of colored light sources.

Gonioreflectometer

A gonioreflectometer is a device to measure the reflectance properties of a material and is therefore commonly used to determine BRDFs. It consists of a light source, an object being measured and a detector.

Photometer

In the broadest sense, a photometer is any instrument used to measure light intensity. Different types of photometers are available. A *luxmeter* is calibrated to provide the measured values in lux. A *spectrophotometer* measures intensity as a function of the wavelength of light. A *goniophotometer* is a device for measuring the directional pattern of light distribution from a source.

Radiometer

Radiometer is a general term for a device that is used to measure the intensity of radiant energy. A photometer is a special type of radiometer.

References

[Ali05a] Alias: Is it Fake or Foto?, http://www.alias.com/eng/etc/fakeorfoto/, June 2005. 96

[Ali05b] Alias: http://www.alias.com/, June 2005. 96

[CG85] COHEN M., GREENBERG D. P.: The Hemi-Cube: A Radiosity Solution for Complex Environments. In *Computer Graphics (ACM SIGGRAPH '85 Proceedings)* (Aug. 1985), vol. 19, pp. 31–40. 98

[Cie05] CIE TC 3-33: Test Cases for Assessment of Accuracy of Lighting Computer Programs http://ciediv3.entpe.fr/indcie/indact333.htm, June 2005. 107

[CMD*00] CHALMERS A. G., MCNAMARA A., DALY S., MYSZKOWSKI K., TROSCIANKO T.: Image quality metrics. *SIGGRAPH 2000 Course 44* (July 2000). 108

[Dal93] DALY S.: The Visible Difference Predictor: An Algorithm for the Assessment of Image Fidelity. In *Digital Images and Human Vision* (1993), Watson A., (Ed.), MIT Press, pp. 179–206. 108

[DM01] DRAGO F., MYSZKOWSKI K.: Validation proposal for global illumination and rendering techniques. *Computers & Graphics 25*, 3 (2001), 511–518. 102, 104

[DVR02] DUMORTIER D., VAN ROY F.: SODA daylighting resource, 2002. 107

[GHR84] GERVAIS M., HARVEY JR. L. O., ROBERTS J. O.: Identification confusions among letters of the alphabet. *Journal of Experimental Psychology: Human Perception and Performance 10*, 5 (1984), 655–666. 108

[GTGB84] GORAL C. M., TORRANCE K. K., GREENBERG G., BATTAILE B.: Modelling the Interaction of Light Between Diffuse Surfaces. In *Computer Graphics (SIGGRAPH '84 Proceedings)* (July 1984), vol. 18, pp. 213–222. 96, 97, 98

[GTS*97] GREENBERG D. P., TORRANCE K. E., SHIRLEY S., ARVO J., FERWERDA J. A., PATTANAIK S., LAFORTUNE E. P. F., WALTER W., FOO S.-C., TRUMBORE B.: A Framework for Realistic Image Synthesis. In *SIGGRAPH 97 Conference Proceedings* (Aug. 1997), Whitted T., (Ed.), Annual Conference Series, ACM SIGGRAPH, Addison Wesley, pp. 477–494. ISBN 0-89791-896-7. 95

[Imh83] IMHOFF E. A.: *Raman Scattering and Luminescence in Polyacetylene During the Cis-trans Isomerization*. Ph.d. dissertation, Cornell Univ., Ithaca, N.Y., Physics Dept., May 1983. 99

[Kaj86] KAJIYA J. T.: The Rendering Equation. In *Proceedings of SIGGRAPH* (1986), ACM Press, pp. 143–150. 96

[KP96] KARNER K. F., PRANTL M.: A Concept for Evaluating the Accuracy of Computer-Generated Images. In *Proceedings of the Twelfth Spring Conference on Computer Graphics (SCCG' 96)* (Comenius University, Bratslavia, Slovakia, 1996). 100, 101, 109

[Mar99] MARDALJEVIC J.: *Daylight Simulation: Validation, Sky Models and Daylight Coefficients*. Ph.d. dissertation, De Montfort University, Leicester, Institute of Energy and Sustainable Development, Dec 1999. 101, 102, 103

[McN00] MCNAMARA A.: Star: Visual perception in realistic image synthesis,. In *Eurographics 2000* (August 2000), Eurographics. 108

[MCTG00] MCNAMARA A., CHALMERS A., TROSCIANKO T., GILCHRIST I.: Comparing Real and Synthetic Scenes using Human Judgements of Lightness. In *Proceedings of the Eurographics Workshop* (June 2000), Piroche B., Rushmeier H., (Eds.), Eurographics. 97

[MF03] MAAMARI F., FONTOYNONT M.: Analytical tests for investigating the accuracy of lighting programs. *Lighting Research and Technology 35*, 3 (2003), 225–242. 107

[MRC*86] MEYER G. W., RUSHMEIER H. E., COHEN M. F., GREENBERG D. P., TORRANCE K. E.: An Experimental Evaluation of Computer Graphics Imagery. *ACM Trans. Graph. 5*, 1 (1986), 30–50. 98, 99

[MS74] MANNOS J., SAKRISON D.: The Effects of a Visual Fidelity Criterion on the Encoding of Images. *IEEE Transactions on Information Theory 20*, 4 (1974), 525–536. 108

[NMP98] NEUMANN L., MATKOVIC K., PURGATHOFER W.: Perception based color image difference. In *Eurographics 1998* (1998), Ferreira N., Göbel M., (Eds.), vol. 17, Eurographics. 108

[PFTG97] PATTANAIK S. N., FERWERDA J. A., TORRANCE K. E., GREENBERG D. P.: Validation of Global Illumination Solutions Through CCD Camera Measurements. In *Proceedings of the Fifth Color Imaging Conference* (Nov. 1997), Society for Imaging Science and Technology, pp. 250–253. 105, 106, 107

[PSM93] PEREZ R., SEALS R., MICHALSKY J.: All-weather model for sky luminance distribution - preliminary configuration and validation. *Solar Energy 50*, 3 (1993), 235–245. 102

[Rea05] RealReflect Project `http://www.realreflect.org/`, June 2005. 96

[Ren05] RenderPark: A Test-Bed System for Global Illumination `http://www.renderpark.be/`, June 2005. 98

[RWP*95] RUSHMEIER H., WARD G., PIATKO C., SANDERS P., RUST B.: Comparing Real and Synthetic Images: Some Ideas About Metrics. In *Eurographics Rendering Workshop* (1995). 108

[SAWG91] SILLION F. X., ARVO J. A., WESTIN W., GREENBERG D. P.: A Global Illumination Solution for General Reflectance Distributions. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics (SIGGRAPH '91)* (Las Vegas, Nevada, USA, July 1991), Sederberg T. W., (Ed.), ACM Press, pp. 187–196. 100

[SKKA01] SZIRMAY-KALOS L., KOVÁS L., ABBAS A. M.: Testing Monte-Carlo Global Illumination Methods with Analytically Computable Scenes. In *Winter School of Computer Graphics Conf.* (2001). 106

[SW04] SCHREGLE R., WIENOLD J.: Physical Validation of Global Illumination Methods: Measurement and Error Analysis. *Computer Graphics Forum 23*, 4 (2004), 761–781. 103, 105, 106

[TTOO90] TAKAGI A., TAKAOKA H., OSHIMA T., OGATA Y.: Accurate Rendering Technique Based on Colorimetric Conception. In *Computer Graphics (SIGGRAPH Proceedings) 24* (1990), ACM Press, pp. 263–272. 99, 100

[VMKK00] VOLEVICH V., MYSZKOWSKI K., KHODULEV A., KOPYLOV E.: Using the Visible Differences Predictor to Improve Performance of Progressive Global Illumination Computations. *Transactions on Graphics 19*, 2 (2000), 122–161. 102

[War94] WARD G.: The RADIANCE Lighting Simulation and Rendering System. In *Proceedings of SIGGRAPH* (1994), ACM Press, pp. 459–472. 98