Photon Path Distribution in Inhomogeneous Scattering Media

Christian Schmaltz
30th September 2004

Diplomarbeit - Diploma Thesis
Universität des Saarlandes - Saarland University
Max-Planck-Institut für Informatik
AG 4: Computer Graphics Group

Abstract

In this thesis we will compare the dipole model by Jensen (and the result from a Monte Carlo simulation for the geometry used by Jensen) to simulation results for different geometries and to results from simulations in which inhomogeneous materials have been used. This will enable us to give hints when the dipole approximation yields feasible results and what can be done to improve them. Then we propose a simple model for inhomogeneous materials based on the dipole approximation and fit its parameters to simulation results.

Acknowledgments

Special thanks goes to my supervisor Michael Goesele for a lot of good ideas, corrections and remarks. Furthermore I’d like to thank Sabine Fischer for helping me with maple and the GIMP, Christian Rössl for finding a nasty bug in the simulation program and my other friends who helped me in one way or another.

Statement under Oath

Hereby I declare on oath that I have drafted the following thesis independently, that I have used the stated sources and means only, and that I have not presented this work at a different examination office previously.

Contents

Chapter 1

Introduction

In the past and especially during the last few years, it has become clear that there are virtually no scenes that cannot be rendered any more. Consequently, the question nowadays is not if a scene can be rendered but how long it takes to do so. Efficient hardware, rendering structures and algorithms (like BSP-trees, programmable shaders that run on graphic cards or efficient data structures to store photons in photon mapping) have been invented and improved.

There are, however, still some problems. Scenes with fluorescence, or phosphoresce are still very hard. These problems are usually not serious because these two effects are very rare. There is, however, a common effect that is often important and hard to compute. This phenomenon, which is called **subsurface scattering**, occurs in many non-metallic materials. Although the effect is in most cases small, it is clearly visible for certain materials like marble, skin or milk. A glass of milk, for example, will look like a block of chalk without subsurface scattering. The effects of subsurface scattering can also clearly be seen in alabaster, as seen in Figure 1.1.

Subsurface scattering happens when a photon that hits a solid or fluid object enters this object and traverses inside the object non-deterministically until it leaves it again, usually at a different position.

That is, these objects are not transparent and cannot be described by a BRDF [21] (and not by a mixture of these two approaches). Such materials are called **translucent**. Areas that are bright although they are not illuminated, rather dull highlights and similar effects can be the result from subsurface scattering. In general, features on such objects appear to be smoothed.

A wide spread dipole approximation for rendering subsurface scattering objects was given by Jensen [17]. It is, however, only valid under very special conditions that are usually not true in common scenes. This model is often used in these scenes nonetheless. One of

(a) including subsurface scattering (b) excluding subsurface scattering

Figure 1.1: Both figure shows a photograph of the head of an alabaster horse. However, the subsurface scattering was removed in the right figure by covering the model with fine dust. Comparing these two figures, it is obvious that Subsurface scattering can have a great influence on the appearance of objects.

the goal of this thesis is therefore to check whether the model introduced by Jensen still results in an acceptable solution if one of the preconditions if no longer valid. Moreover we try to generate a better model for a specific case.

1.1 Overview

To achieve these goals and understand the necessary concepts, we start by explaining the mathematical and physical background required in Chapter 2. The previous work is briefly covered in Chapter 3. In Chapter 4, we give the setup of our Monte Carlo photon-tracer, explain the possible outputs it can generate and give a short summary over other useful tools we wrote. Furthermore, we verify that the results are reasonable. Chapter 5 deals with simulation results from different shift-invariant geometries. In there we will compare the results from four types of geometries to Jensen's dipole approximation and give hints what can be done to improve the result from the approximation. In addition to that, we examine a few factors that might be important such as the importance of non-scattered photons. Shift-variant materials are dealt with in Chapter 6 in which discrete and continuous spatially varying materials are used for simulations. In Chapter 7 we fit a model to a generated marble block, trying to improve the result of Jensen's model in this special instance. The last chapter summarizes the results of the previous chapters and gives an outlook on future work.

Chapter 2

Basics

This chapter describes the basic definitions necessary to understand and solve the problems occurring when photons interact with participating media. Basic mathematical concepts are explained first, followed by the physic necessary to understand how light interacts with scattering and non-scattering media. Details are given in [1] or [14].

2.1 Introduction to Probability Calculus

In this section an introduction to probability calculus (i.e. the concept of expected value, variance etc. of random variables) is given. Starting with a formal definition of random variable, the concepts of probability density function, expected value ... will be explained for continuous and discrete random variables of one or more dimensions.

The following formal definitions are required to introduce random variables:

Let S be a non-empty set. A family of subsets F is called a **sigma-algebra** over S if the following three properties hold:

- $\emptyset \in F$
- $E \in F \Rightarrow \overline{E} \in F$
- $E_1, E_2 \ldots E_n \in F \Rightarrow \bigcup_{i=1}^{n} E_i \in F$

Examples include the full power set of S. The smallest sigma-algebra which contains all closed subsets of a **topological space** [1] T is called the **Borel sigma-algebra**.

[1] A topological space is a sigma-algebra S where every union and intersection of (not necessarily finite number of) sets from S is also in S.

A function $\mu : S \to \mathbb{R}_0^+ \cup \{\infty\}$ from a sigma-algebra to the non-negative real numbers (including infinity) is called a **measure** if and only if:

- $\mu(\emptyset) = 0$
- The function is countably additive:
  $E_1, E_2, \ldots E_n \in F, \quad E_i \cap E_j = \emptyset \; \forall 1 \leq i, j \leq n \quad \Rightarrow \quad \mu(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} \mu(E_i)$

Let Ω be a non-empty set, F a sigma-algebra and P a measure such that $P(\Omega) = 1$. The triple $PS = (\Omega, F, P)$ is called a **probability space**, and a **random variable** is defined as a measurable non-deterministic function from a probability space PS to a measurable space S with measure μ. The measure P is called **probability measure**. If the real numbers with the Borel sigma-algebra are chosen as S, the probability measure is also called **probability distribution**. This will always be the case in this thesis.

Loosely spoken, however, a **random variable** is nothing else but the numeric result of a non-deterministic process or experiment. The results from flipping a coin, rolling a die (or summing the spots of several dice) or drawing a bingo number are typical examples of random variables. Another common example is the **canonical random variable**. It results in a number from the interval $[0, 1)$, where every number in this interval has equal probability to be chosen.

In practice, the same random variable can be used to get a random number from the closed interval $[0, 1]$, because the probability that the value 1 is chosen is zero anyway[2].

For each random variable X, its **cumulative distribution function** $F(X) := Pr[X \leq x]$ can be defined.

Each random variable X gives rise to a **probability distribution**, which contains most of the important information about X. The probability distribution corresponding to X assigns to the interval $[a, b]$ the probability $Pr[a \leq X \leq b]$, where $Pr[a \leq X \leq b]$ is the probability that the random variable X will take a value from the interval from a and b. Similarly, $Pr[X = a]$ is the probability that the random variable X will take the value a and $Pr[X \in S]$ stands for: The random variable X will take a value from the set S.

If the set Ω (i.e., the set of values the random variable can attain) is finite or countable finite, the random variable is called **discrete random variable**. Its cumulative distribution function consists then of a series of discrete steps.

Unfortunately, there are two conventions what a **continuous random variable** is. The first convention defines that a random variable is continuous if and only if its cumulative distribution function is continuous, which is equivalent to $Pr[X = a] = 0 \; \forall a \in \mathbb{R}$. In

[2] The probability that 1 is chosen is zero because the random variable is continuous. The term continuous is explained later in this section.

this paper, however, the second convention is used: A random variable X is called (absolutely) continuous if the **probability density function** (or short **PDF**) exists, i.e., there is a function $p : \Omega \to \mathbb{R}$ such that

$$Pr[X \in S] = \int_S p(X)dX \quad \forall\, S \subset \Omega.$$

In this case, this relationship is denoted with $X \sim p$.

Note that these two conventions match for "normal" functions. However, some functions like the Cantor function [10] are continuous but not absolutely continuous. Furthermore, it is obvious that

$$p(S) \geq 0 \;\forall\, S \in \Omega$$

and

$$\int_\Omega p(X)dX = 1.$$

Every function that fulfills these two conditions is a PDF for a random variable and it should be clear that a random variable X can be uniquely described by both its cumulative distribution function and its PDF.

2.1.1 Expected Value, Variance and Standard Derivation

If X is a random variable over a probability space (Ω, F, P), the **expected value** $E[X]$ is defined as the Lebesgue integral of X over Ω:

$$E[X] := \int_\Omega X\,dP.$$

The random variable X has no expected value if this integral does not exist.

For a discrete random variable X with values $x_1, x_2, \ldots x_n$, this formula simplifies to

$$E[X] := \sum_{i=1}^{n} Pr[X = i] \cdot x_i.$$

If X is (absolutely) continuous and p is the corresponding PDF, $E[X]$ can be computed as

$$E[X] = \int_\Omega X \cdot p(X)dX.$$

Intuitively, the expected value is the "average" result of a random variable, but the name expected value must not lead to the wrong intuition that the result of a random variable will most of the time be the expected value. In reality, the expected value is often a value which is not possible as value of the random variable at all, e.g. the expected value when

throwing a die and noting the number is $\frac{1}{6}(1+2+3+4+5+6) = 3.5$, a result which will definitely never occur.

A very useful property of the expected value is its linearity: If X and Y are two random variables, the expected value of the sum of X and Y is the sum of the expected values of X and Y, i.e., $E[X+Y] = E[X] + E[Y]$. Thus, it follows easily that the expected value of the sum of the spots of n dice is $n \cdot 3.5$.

Furthermore, it should be noted that functions of random-variables are random variables themselves. Therefore, if X and Y are random variables and f and g are functions of X and Y, respectively, the following formula holds:

$$E[f(X) + g(Y)] = E[f(X)] + E[g(Y)]$$

If X is continuous with $X \sim p$, the expected value of $f(X)$ is given by

$$E[f(X)] = \int_{\Omega} f(X) \cdot p(X) dx.$$

Other important concepts linked to random variables are the **variance** $V[X]$ and the **standard derivation** σ of a one-dimensional random variable X:

$$V[X] := E[(X - E[X])^2] = E[X^2] - (E[X])^2$$
$$\sigma := \sqrt{V[X]}$$

For multi-dimensional random variables, variance and expected value are not defined. Intuitively, the greater the standard derivation, the further the result of one experiment will be from the expected value (on the average). This is illustrated by the following example: The random variables X and Y that uniformly take one of the numbers $\{3,4,5\}$ and $\{-1,1,12\}$, respectively, both have an expected value of 4, but the variance and the standard derivation of the second variable are much bigger:

$$V[X] = \quad E[(X - E[X])^2] = \quad \frac{1}{3}\left(\sum_{i=1}^{3}(X - E[X])^2\right) = \qquad \frac{1}{3}(1+0+1) = \frac{2}{3}$$
$$V[Y] = \quad E[(Y - E[Y])^2] = \quad \frac{1}{3}\left(\sum_{i=1}^{3}(Y - E[Y])^2\right) = \quad \frac{1}{3}(25+9+64) = \frac{98}{3} = 32\frac{2}{3}$$

If X and Y are two random variables and the property

$$Pr((X \in S) \wedge (Y \in T)) = Pr(X \in S) \cdot Pr(Y \in T) \quad \forall S, T \in \Omega$$

holds, X and Y are called **independent random variables**. Then,

$$E[X \cdot Y] = E[X] \cdot E[Y] \quad \text{and} \quad V[X+Y] = V[X] + V[Y].$$

Two or more (absolutely) continuous random variables that have the same probability distribution function are called **independent identically distributed** random variables.

Especially important for Monte Carlo Integration (which will be discussed in the next section) are the weak and the strong laws of large numbers: Let $X_1, X_2 \ldots$ be an infinite sequence of independent identically distributed random variables with expected value $E[X]$ and $\overline{X}_n = \sum_{i=1}^{n} X_i$ the average of the first n random variables (the so called **estimated mean**). The **weak law of large numbers** state that, for every $\varepsilon > 0$

$$\lim_{n\to\infty} Pr\left(\left|\overline{X}_n - E[X]\right| < \varepsilon\right) = 1$$

Even better, the **strong law of large numbers** claims:

$$Pr\left(\lim_{n\to\infty} \overline{X}_n = E[X]\right) = 1$$

These laws basically mean that the estimated means is an approximation for the expected value:

$$E[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{for } n \to \infty.$$

That is, if a random experiment is done several times, the expected value of the experiment is approximated by the estimated means, and this approximations yields better results if more experiments are used in the approximation.

2.2 Monte Carlo Integration

In ray tracing it is often not possible to calculate exact results, especially in complex scenes. Using probability calculus, however, even integrals that can not be solved analytically can be approximated by a method called **Monte Carlo Integration**. When using photon mapping, for example, a certain number of photons are shot from the light sources and traversed through the scene. Those places where the photons are diffusely reflected are stored and used later to approximate the amount of light (and its color) of the areas shown. To obtain an exact result, however, billions and billions of photons would have to be shot in every possible direction, which is not practicable to do (and perhaps will never be) in a reasonable amount of time. When dealing with subsurface scattering, the

situation is even worse, as will be shown in the next section. More general, this problems always occurs when complicated integrals, e.g. the rendering equation, have to be solved.

Let X be a (absolutely) continuous random variable with $X \sim p$ and f be a function $f : \Omega \to \mathbb{R}$. By applying the definition of the expected value and using the approximation of the expected value with estimated means, it follows that:

$$\int_\Omega f(X)p(X)dX = E[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Ignoring the middle term and setting $g := f \cdot p$ yields

$$\int_\Omega g(X)dX \approx \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{p(X_i)}.$$

Thus, the integral $I := \int_\Omega g(X)dX$ can be approximated by choosing any PDF $p : \Omega \to \mathbb{R}^+$ and calculating the sum on the right side through random variables X_i with $X_i \sim p$. Choosing a PDF $p : \Omega \to \mathbb{R}$ is possible if and only if $p(a) \neq 0$ where $g(a) \neq 0$.

Thus, there are a huge number of possible approximations which will all converge towards the correct solution. The variance of these approximations, however, are very different: Choosing the right PDF can vastly reduce the variance and therefore increase the speed of convergence. This is illustrated in Table 2.1, which shows several sampling functions and the variance each function has when it is used in a Monte Carlo estimation of the integral described above. One of the calculations done for this table is illustrated in the following example:

Suppose the Monte Carlo solution (i.e. an approximation using Monte Carlo Integration) of the integral from 0 to 3 over the function $g(x) = x^2$ shall be computed:

$$I = \int_0^3 g(x)dx = \int_0^3 x^2 dx = \frac{1}{3}x^3\Big|_0^3 = 9$$

First, the sampling function p must be chosen. One valid choice is $p(x) = \frac{2}{9}x$, because:

- $\frac{2}{9}x$ only has one zero at $x = 0$, but $g(0) = 0$, i.e., $p(a) \neq 0$ where $g(a) \neq 0$.
- $p(x) \geq 0 \quad \forall\, x \in [0,3]$.
- $\int_0^3 p(x) = \frac{1}{9}x^2\big|_0^3 = 1$.

sampling function	variance	Number
$\frac{8}{10}(4-x)^3$	$\frac{1231,246}{n}$	1
$\frac{e^{-x}}{1-e^{-3}}$	$\frac{526,018}{n}$	2
$\frac{12-2x}{27}$	$\frac{108,928}{n}$	3
$\frac{1}{3}$	$\frac{64,8}{n}$	4
$\frac{x(4-x)}{9}$	$\frac{42,506}{n}$	5
$\frac{4}{81}x^3$	$\frac{10,125}{n}$	6
$\frac{2}{9}x$	$\frac{10,125}{n}$	7
$\frac{e^x}{e^3-1}$	$\frac{3,619}{n}$	8
$\frac{x^2}{9}$	0	9

Table 2.1: In this table, several possible estimators and their corresponding variances are shown when trying to find the solution of the integral $\int_0^3 x^2 dx$ using Monte Carlo methods. The last column indicates which plot in the Figures 2.1 and 2.2 is the plot of the given function. (Figure 2.1 and 2.2 are on Page 12).

The expected value of $\frac{g(x)}{p(x)}$ is:

$$E\left[\frac{g(x)}{p(x)}\right] = \int_0^3 \frac{g(x)}{p(x)} \cdot p(x)dx = \int_0^3 g(x)dx = I,$$

which should be no surprise. The variance is given by:

$$\begin{aligned} V\left[\frac{g(x)}{p(x)}\right] &= E\left[\left(\frac{g(x)}{p(x)}\right)^2\right] - \left(E\left[\frac{g(x)}{p(x)}\right]\right)^2 \\ &= \int_0^3 \left(\frac{g(x)}{p(x)}\right)^2 p(x)dx - I^2 \\ &= \int_0^3 \frac{9}{2}x^3 dx - 81 \\ &= 10.125 \end{aligned}$$

The variance of the chosen estimator is therefore $V\left[\frac{1}{n}\sum_{i=1}^n \frac{g(X_i)}{p(X_i)}\right] = \sum_{i=1}^n V\left[\frac{1}{n}\frac{g(X_i)}{p(X_i)}\right] = \frac{10.125}{n}$, since the random variables are identically independent distributed, and thus independent.

The variance for other possible sampling functions are shown in Table 2.1. Comparing the variance of these functions with their shapes, which are shown in the Figures 2.1 and 2.2,

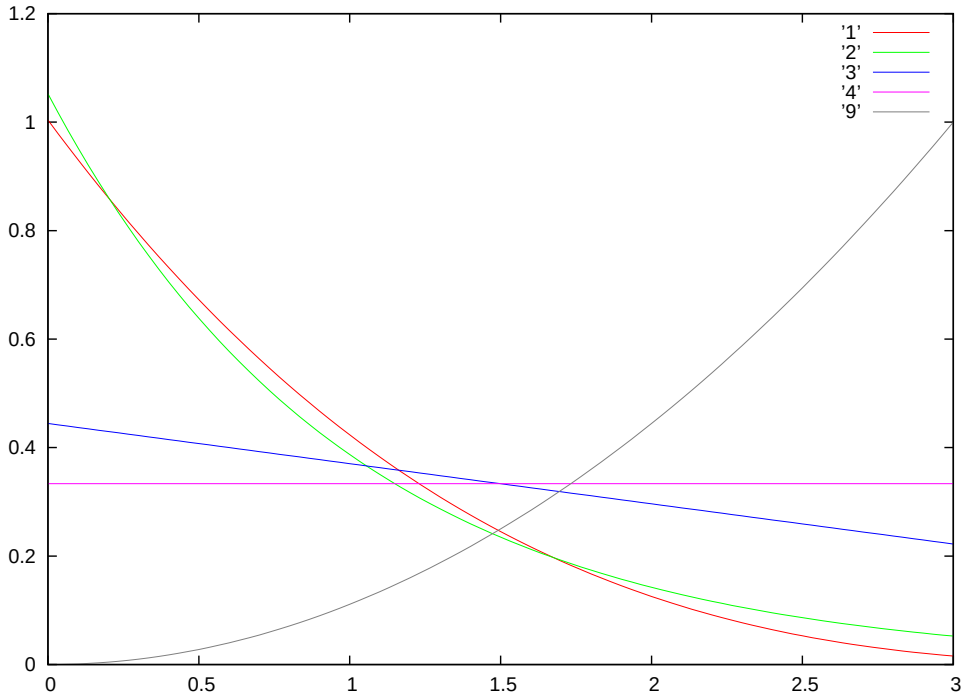


Figure 2.1: This figure shows the plots of four possible (but bad) estimators (1...4) together with the optimal estimator 9, which is nothing else but the scaled down shape of the initial function to be evaluated $g(x) = x^2$. It is quite obvious that the shapes of these four estimators are very different from the shape of $g(x)$. This results in a huge variance as shown in Table 2.1

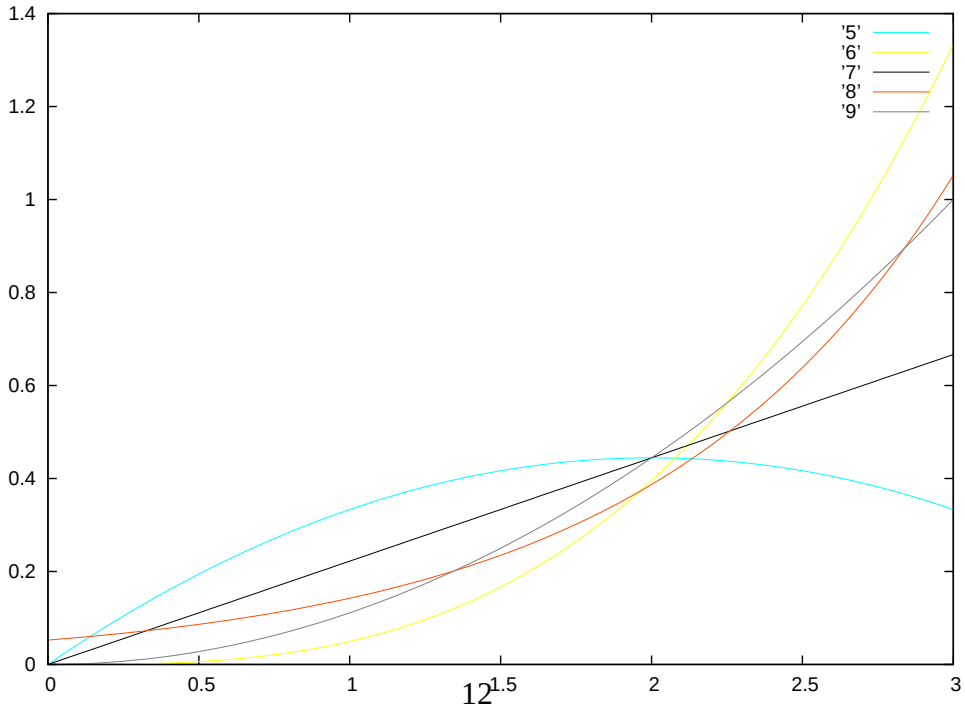


Figure 2.2: This time four different estimators (5...8) have been plotted whose shape is closes to the shape of $g(x)$ which is again the gray curve 9. As expected, their variance (given in Table 2.1 is far smaller than the variances of the first four estimators in the picture above.

it is noticeable that the variance is lower if the shape of the sampling function is similar to the shape of g.

Choosing a function p whose shape is similar to the initial function g results in less variance because then, more samples are taken in the "important" areas, i.e., these areas where $g(x)$ is large. The technique to try to find a "good" sampling function before the sampling process is started (or even when changing the sampling function in between) is called **importance sampling**. There are other possible sampling strategies, some one which will be covered briefly in the following section.

The best choice for p would be to choose $p(x) = g(x)/I$, because the variance will then be zero:

$$\begin{aligned} V\left[\frac{g(x)}{p(x)}\right] &= E\left[\left(\frac{g(x)}{p(x)}\right)^2\right] - \left(E\left[\frac{g(x)}{p(x)}\right]\right)^2 \\ &= \int_a^b \left(\frac{g(x)}{p(x)}\right)^2 p(x)dx - I^2 \\ &= \int_a^b \frac{(g(x))^2}{p(x)}dx - I^2 \\ &= \int_a^b \frac{(g(x))^2}{\frac{g(x)}{I}}dx - I^2 \\ &= \int_a^b I \cdot g(x)dx - I^2 = 0 \end{aligned}$$

To set p as $p(x) = g(x)/I$, however, I would have to be known. Then, the initial problem (approximating I) would already have been solved.

2.2.1 Quasi Monte Carlo Integration

Instead of using random samples it is possible to use more uniformly distributed samples. Equidistant and stratified sampling are some easy examples. While the term equidistant should be self-explanatory, stratified sampling might require an explanation:

When using stratified sampling, the interval (area, volume, ...) that should be sampled is divided into intervals (areas, volumes, ...) of equal size. The boundary (or boundaries) of these new objects are usually axis-aligned. Then, a random point is chosen in each interval. A simple example for stratified sampling can be seen in Figure 2.3.

More advanced methods like Poisson disc sampling and using Halton sequences are also feasible. When using one of these sampling methods, one speaks about quasi Monte Carlo

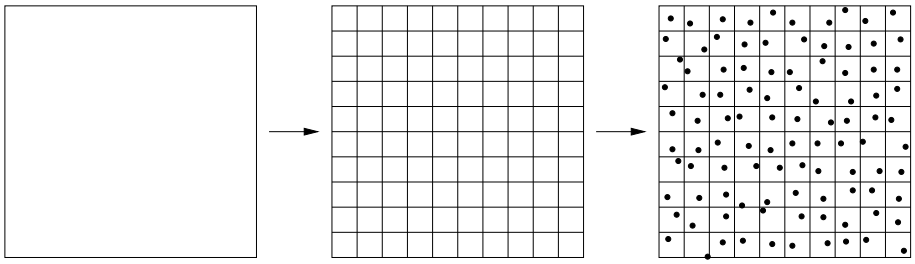


Figure 2.3: This figure illustrates stratified sampling of a square: First, the square is divided into a ten by ten square grid. Then a random sample is chosen in each cell.

integration. Details for both using and validating this approaches, however, would lead too far here and can, for example, be found in [19]. In short, one can say that using quasi Monte Carlo integration usually leads to better results, especially under certain conditions. But these conditions are not always true in graphics. It has been shown that, when using stratified sampling, the variance is not higher than with random samples, if all strata have equal measure. On the other side, however, using quasi Monte Carlo Integration methods can produce aliasing.

2.3 Uniform Sampling of Multidimensional Objects

In order to evaluate multi-dimensional integrals using Monte Carlo methods it is necessary to find random samples on multi-dimensional surfaces. Moreover, several samples should be equally distributed on the surface, i.e., in every subarea of the original surface, there should be (in the limit) a number of samples proportional to the area of the subarea. Samples chosen this way are called **uniform samples**.

Other examples in which finding uniform samples is important include sampling the light direction of area and volume light sources when using photon mapping or estimating the radiance at a given point in ray-tracing.

Choosing uniform samples for rectangles is easy: Let ξ_1 and ξ_2 be two independent canonical random variables. Then, a random variable that uniformly covers a rectangle with size (a,b) is given by $(a \cdot \xi_1, b \cdot \xi_2)$. In general, however, the situation is more complex. A random point on the unit disc can, for example, be calculated by choosing a random "angle" α from the interval $[0,1)$ and a "distance" d from the interval $[0,1]$. A random point on the disc is then given by $(d \cdot \cos(2\pi \cdot \alpha), d \cdot \sin(2\pi \cdot \alpha))$. However, not every point on the disc has equal probability to be chosen by this algorithm. Thus, the samples are not uniformly distributed, as can be seen in the Figures 2.4 up to 2.7.

Thus, the goal of this section is find a mapping from the unit square $[0,1] \times [0,1]$ to a two

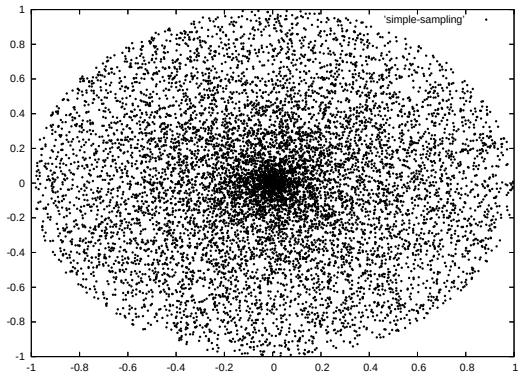


Figure 2.4: Here, 10000 random samples on a disc have been created using the simple formula $(\xi_1 \cdot \cos(2\pi\xi_2), \xi_1 \cdot \sin(2\pi\xi_2))$, where ξ_1 and ξ_2 are two (uniformly distributed) random numbers. Obviously, more samples have been mapped to the interior of the disc, resulting in non-uniform samples.

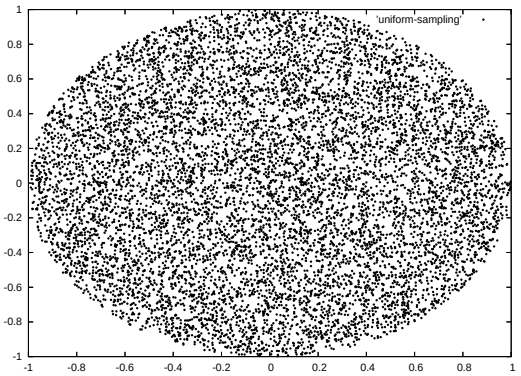


Figure 2.5: Like in Figure 2.4, 10000 samples have been created on a disc. This time, function inversion has been used. Thus, the samples are uniformly distributed. Similar results occur when using rejection sampling.

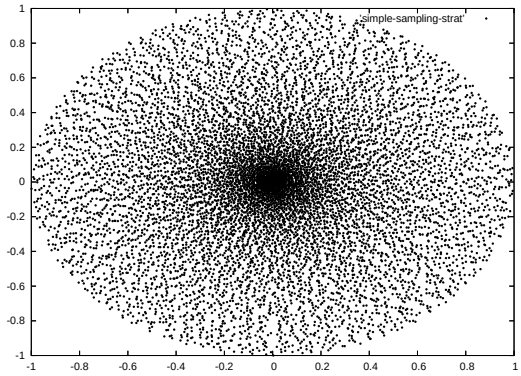


Figure 2.6: This figure shows samples generated with the same formula as the samples from Figure 2.4. This time, however, stratified random points have been used instead of uniformly distributed ones.

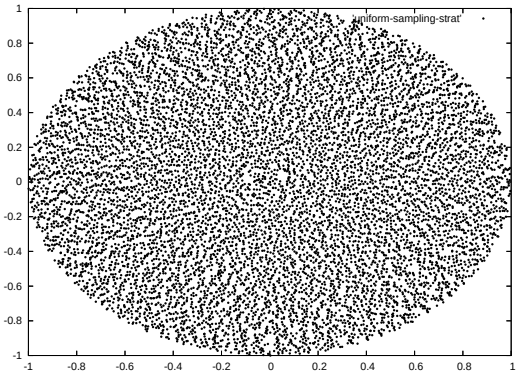


Figure 2.7: This is the result of samples being generated using function inversion and stratified points on the unit disc. Again, rejection sampling will achieve similar results.

dimensional area that preserves uniform sampling. It should be noted that such mappings will also preserve stratification.

The two dimensional area might be embedded in a coordinate system with more dimensions. The classical example is the (two-dimensional) surface of a (three-dimensional) sphere. In general, such an two-dimensional object is called a **2-manifold**. If the area has dimension *k*, it is called a *k*-**manifold**.

Generating equally distributed points on a 2-manifold can be done in several ways, two of which will be explained in the following sections.

2.3.1 Rejection Sampling

A rather simple method to generate uniform samples is **rejection sampling**, i.e., generate uniform (or stratified ...) samples for a simpler object and discard all those that are not valid for a more complicated object.

Again, suppose the positions of a unit disc should be sampled. When using rejection sampling, uniform samples are generated for the unit square. Then, every sample that is not inside the unit disc is thrown away. Thus, pseudo code for creating uniform samples for the unit disc is given by:

- Generate a sample (ξ_1, ξ_2) on the square $[-1,1]^2$.
- Check if $(\xi_1^2 + \xi_2^2) \leq 1$, i.e., check if the sample is inside the unit circle.
- If the sample is not in the disc, discard it and restart the algorithm.
- Otherwise, (ξ_1, ξ_2) is a new sample for the unit disc.

Unfortunately, there is a major drawbacks in this algorithm: there are no runtime guarantees. Although it is very unlikely, the code given above might run forever without finding a single sample. This observation will be true for all algorithms given in this section.

Moreover, the algorithm is that simple only for relative simple areas. While discs and spheres can be sampled very easily, sampling the surface of a sphere is already more tricky. Doing so can be done by first sampling the sphere itself and mapping the found samples to the surface afterwards:

- Generate a sample (ξ_1, ξ_2, ξ_3) on the cube $[-1,1]^3$.
- Calculate the distance of the sample from the midpoint of the sphere/cube: $\mathsf{len} = \sqrt{\xi_1^2 + \xi_2^2 + \xi_3^2}$.
- Discard the sample and restart if $\mathsf{len} > 1$, i.e., restart if the sample is outside the unit sphere.

- Discard the sample and restart if $\mathsf{len} = 0$, i.e., restart if $(\xi_1, \xi_2, \xi_3) = (0, 0, 0)$.
- Map the sample to the surface of the sphere by dividing each sample through len. The point $\left(\frac{\xi_1}{\mathsf{len}}, \frac{\xi_2}{\mathsf{len}}, \frac{\xi_3}{\mathsf{len}}\right)$ is now a sample on the surface of the sphere.

Using rejection sampling, It is also possible to create random numbers with a given underlying PDF $p : [a, b] \to [0, c]$. Note that here, the PDF is required to be bounded[3]. Then, the random variable can be sampled in the following way:

- Generate a random point (ξ_1, ξ_2) in the rectangle $[a, b] \times [0, c]$.
- If $p(\xi_1) > \xi_2$, discard the sample and restart.
- Otherwise, the sample ξ_1 is valid.

2.3.2 Function Inversion

A more advanced approach to generate uniform samples is to use **function inversion**. This technique requires more mathematical background but leads to a deterministic algorithm that is usually quicker than using rejection sampling.

The drawback of this algorithm is that it requires a **parametrization** of the k-manifold M to work, where a parametrization is a **diffeomorphism** (i.e., smooth bijection) from the unit hypercube $[0, 1]^k$ to the k-manifold for which samples should be generated. The parametrization of the manifold M is denoted as Φ_M.

To use function inversion, Φ_M does not need to be bijective everywhere, but only **nearly everywhere**, i.e., the integral of 1 over the set T that does not satisfy the condition (here: Φ_M is bijective) is zero[4]. Especially, this condition holds if T is a part of the boundary of $[0, 1]^k$. Thus, finding a diffeomorphism from $[0, 1)^k$ (or $(0, 1] \times [0, 1)^{k-1}$ or similar constructs) to M is sufficient.

Function Inversion Algorithm for 2-Manifolds

If $M \subset \mathbb{R}^n$ is a 2-manifold with arbitrary parametrization $\Phi_M = \Phi$, then the following algorithm can be used to create an area-preserving parametrization of the 2-manifold M:

- Set

$$\begin{aligned} \sigma : [0, 1]^2 &\to \mathbb{R} \\ (s, t) &\mapsto \sqrt{(\Phi_s \cdot \Phi_s)(\Phi_t \cdot \Phi_t) - (\Phi_s \cdot \Phi_t)^2} \end{aligned}$$

[3]The requirement $p(x) \geq 0$ is no limitation, since this is common to all PDFs as discussed earlier.
[4]Other equivalent formulations are: "T is a null set" and "the measure of T is zero".

where $\Phi_s = \left(\frac{\delta\Phi_1}{\delta s}, \cdots, \frac{\delta\Phi_n}{\delta s}\right)$ and $\Phi_t = \left(\frac{\delta\Phi_1}{\delta t}, \cdots, \frac{\delta\Phi_n}{\delta t}\right)$ are the vectors of partial derivatives of Φ.

For $n = 2$, the formula simplifies to $\sigma(s,t) = \det(D_{(s,t)}\Phi)$. Here $D_{(s,t)}$ is the 2×2 Jacobian matrix of $\Phi = (\Phi_1, \Phi_2)$ at the point (s,t):

$$D_{(s,t)}\Phi = \begin{pmatrix} \frac{\delta\Phi_1}{\delta s} & \frac{\delta\Phi_1}{\delta t} \\ \frac{\delta\Phi_2}{\delta s} & \frac{\delta\Phi_2}{\delta t} \end{pmatrix}$$

For $n = 3$, the formula can be written as the length of the dot-product of the two vectors of partial derivatives:

$$\sigma(s,t) = |\Phi_s(s,t) \times \Phi_t(s,t)|$$

This function satisfies the condition that, for every subset A of $[0,1]^2$, the integral of σ over A is equal to the area of the part of the 2-manifold M that Φ maps A to:

$$\int_A \sigma = \mathsf{area}(\Phi(A))$$

- Furthermore, define the functions F and G, which are both defined as function from $[0,1]$ to $\mathbb{R}$:

$$F(s) := \frac{\int_0^1 \int_0^s \sigma(u,v)\, du\, dv}{\int_0^1 \int_0^1 \sigma(u,v)\, du\, dv}$$
$$G_s(t) := \frac{\int_0^t \sigma(s,v)\, dv}{\int_0^1 \sigma(s,v)\, dv}$$

Note that F and G are cumulative distribution functions of σ along different paths.

- Calculate the inverse functions f and g of F and G:

$$f(z) := F^{-1}(z)$$
$$g(z_1, z_2) := G^{-1}_{f(z_1)}(z_2)$$

- An area-preserving parametrization φ of the 2-manifold is then given by:

$$\varphi(z_1, z_2) := \Phi(f(z_1), g(z_1, z_2))$$

This parametrization can therefore be used to map uniform (or stratified ...) samples from the unit square to the 2-manifold M.

The only difficult step in the algorithm given above is the function inversion step. Depending on the parametrization or the 2-manifold, it might not be possible to find a explicit, closed-form solution. In this case, the inverse can be approximated numerically.

Using a weighting function to generate non-uniform samples is easy, too: Simply replace the function σ by $\sigma \cdot w(s,t)$, where $w(s,t)$ is the weighting function desired. As an example of this technique, a area-preserving parametrization for a phong lobe will be generated on Page 22.

2.3.3 Examples of Uniform Sampling Algorithms using Function Inversion

To clarify the algorithm explained in the last section and to show how easy it is to use this algorithm, the formulas for uniform sampling of the unit disc, the unit hemisphere, the unit sphere and a phong lobe will be derived.

Sampling the Unit Disc

Once again, a area-preserving parametrization of the unit disc is to be computed. Using the non-uniform parametrization $\Phi(d,\alpha) = (d \cdot \cos(2\pi \cdot \alpha), d \cdot \sin(2\pi \cdot \alpha)) \subset \mathbb{R}^2$, which has already been explained in Section 2.3, the function σ is given by:

$$\begin{aligned}
\det(D_{(d,\alpha)}\Phi) &= \det \begin{pmatrix} \cos(2\pi \cdot \alpha) & -2\pi d \cdot \sin(2\pi \cdot \alpha) \\ \sin(2\pi \cdot \alpha) & 2\pi d \cdot \cos(2\pi \cdot \alpha) \end{pmatrix} \\
&= 2\pi d \cdot \cos^2(2\pi \cdot \alpha) + 2\pi d \cdot \sin^2(2\pi \cdot \alpha) \\
&= 2\pi d
\end{aligned}$$

The functions F and G_d are:

$$\begin{aligned} F(d) &= \frac{\int_0^1 \int_0^d 2\pi u \; du \; dv}{\int_0^1 \int_0^1 2\pi u \; du \; dv} \\ &= \frac{\int_0^1 \pi d^2 \; dv}{\int_0^1 \pi \; dv} \\ &= d^2 \\ G_d(\alpha) &= \frac{\int_0^\alpha 2\pi u dv}{\int_0^1 2\pi u dv} \\ &= \alpha \end{aligned}$$

The inverses of F and G_d are obviously $f(z) = \sqrt{z}$ and $g(z_1, z_2) = z_2$. Thus, an area-preserving parametrization of the unit disc is given by:

$$\varphi(z_1, z_2) = \Phi(\sqrt{z_1}, z_2) = (\sqrt{z_1} \cdot \cos(2\pi \cdot z_2), \sqrt{z_1} \cdot \sin(2\pi \cdot z_2))$$

In fact, this function has been used to generate the Figures 2.7 and 2.5 mentioned earlier.

Sampling the Unit (Hemi-)Sphere

A little more complex but still very simple is the derivation of an area-preserving parametrization of the unit hemisphere, i.e. the part of a unit sphere[5] whose third component is greater than zero. A common parametrization of this object can be obtained by using polar coordinates:

$$\Phi(s,t) = \begin{pmatrix} \sin(\frac{\pi s}{2}) \cos(2\pi t) \\ \sin(\frac{\pi s}{2}) \sin(2\pi t) \\ \cos(\frac{\pi s}{2}) \end{pmatrix}$$

The codomain of Φ is $\mathbb{R}^3$. Thus, σ is defined as:

[5]The sphere centered at the origin with radius 1.

$$
\begin{aligned}
\sigma(s,t) &= |\Phi_s(s,t) \times \Phi_t(s,t)| \\
&= \left| \begin{pmatrix} \frac{\pi}{2} \cdot \cos(\frac{\pi s}{2}) \cos(2\pi t) \\ \frac{\pi}{2} \cdot \cos(\frac{\pi s}{2}) \sin(2\pi t) \\ -\frac{\pi}{2} \cdot \sin(\frac{\pi s}{2}) \end{pmatrix} \times \begin{pmatrix} -2\pi \cdot \sin(\frac{\pi s}{2}) \sin(2\pi t) \\ 2\pi \cdot \sin(\frac{\pi s}{2}) \cos(2\pi t) \\ 0 \end{pmatrix} \right| \\
&= \pi^2 \left| \begin{pmatrix} \sin^2(\frac{\pi s}{2}) \cos(2\pi t) \\ \sin^2(\frac{\pi s}{2}) \sin(2\pi t) \\ \cos(\frac{\pi s}{2}) \sin(\frac{\pi s}{2}) (\sin^2(2\pi t) + \cos^2(2\pi t)) \end{pmatrix} \right| \\
&= \pi^2 \sqrt{\sin^4\left(\frac{\pi s}{2}\right) \underbrace{(\cos^2(2\pi t) + \sin^2(2\pi t))}_{=1} + \cos^2\left(\frac{\pi s}{2}\right) \sin^2\left(\frac{\pi s}{2}\right) \cdot 1} \\
&= \pi^2 \sqrt{\sin^2\left(\frac{\pi s}{2}\right) \left(\sin^2\left(\frac{\pi s}{2}\right) + \cos^2\left(\frac{\pi s}{2}\right)\right)} \\
&= \pi^2 \sin\left(\frac{\pi s}{2}\right)
\end{aligned}
$$

Again, the functions F and G, given by

$$
\begin{aligned}
F(s) &= \frac{\int_0^1 \int_0^s \pi^2 \sin\left(\frac{\pi u}{2}\right) \, du \, dv}{\int_0^1 \int_0^1 \pi^2 \sin\left(\frac{\pi u}{2}\right) \, du \, dv} &&= 1 - \cos\left(\frac{\pi s}{2}\right) \\
G_s(t) &= \frac{\int_0^t \pi^2 \sin\left(\frac{\pi u}{2}\right) \, dv}{\int_0^1 \pi^2 \sin\left(\frac{\pi u}{2}\right) \, dv} &&= t
\end{aligned}
$$

have to be inverted, which is trivial: $f(z) = \frac{2\cos^{-1}(1-z)}{\pi}$ and $g(z_1, z_2) = z_2$. Using the fact that $\sin(x) = \sqrt{1-\cos(x)}$, the final parametrization can be written as:

$$
\varphi(z_1, z_2) = \begin{pmatrix} \sqrt{z_1(2-z_1)} \cos(2\pi z_2) \\ \sqrt{z_1(2-z_1)} \sin(2\pi z_2) \\ 1 - z_1 \end{pmatrix}
$$

The costs of this formula can be sightly decreased by substituting $z\prime$ for $1 - z_1$.

Doing the same computations for the unit sphere with the parametrization given by spherical coordinates $((\sin(\pi s)\cos(2\pi t), \sin(\pi s)\sin(2\pi t), \cos(\pi s))^t)$ results in $\sigma(s,t) = 2\pi^2 \sin(\pi s)$ and the area-preserving parametrization $\left(\sqrt{4z_1(1-z_1)}\cos(2\pi z_2), \sqrt{4z_1(1-z_1)}\sin(2\pi z_2), 1-2z_1\right)^t$.

It should be noted that the choice of the initial parametrization is rather important. Choosing the parametrization

$$\Phi(x,y) = \begin{pmatrix} 1-2x \\ 1-2y \\ \sqrt{1-(1-2x)^2-(1-2y)^2} \end{pmatrix}$$

for the upper hemisphere, for example, the steps of the algorithm get far more complicated.

Sampling a phong lobe

Function inversion can easily be adapted to handle non-uniform PDF. The typical example is the sampling of a phong lobe, i.e. the sampling of the (upper) hemisphere with a density distribution that is proportional to the cosine of the polar angle. As mentioned above, all that have to be done is to replace the function σ by the product of σ and the desired weighting function. In the case of a phong lobe with phong exponent k, that is:

$$\sigma' := \sigma \cdot \cos^k\left(\frac{\pi s}{2}\right)$$

Similar calculations then lead to $F(s) = \cos^{k+1}\left(\frac{\pi s}{2}\right)$ and $G_s(t) = t$ and finally

$$\varphi(z_1,z_2) = \begin{pmatrix} \sqrt{1-z_1^{\frac{2}{k+1}}}\cos(2\pi z_2) \\ \sqrt{1-z_1^{\frac{2}{k+1}}}\sin(2\pi z_2) \\ z_1^{\frac{1}{k+1}} \end{pmatrix}.$$

More information about this algorithm can be found in [16], as well as formulas for uniform sampling of other primitives.

2.4 Reflection and Refraction

As already mentioned, the goal of this thesis is to analyze subsurface scattering, i.e. to see how photons interact with participating media. However, before a photon can interact with a material, it must first enter it. Whenever light (or another electromagnetic wave) hits the border between different materials, there are basically two things that can happen: The wave might get (partially) reflected instead of entering the next material and/or it might be refracted.

It is a well know fact that all electromagnetic waves traverse though vacuum with a constant velocity of $c := 299.792.458\frac{m}{s}$. However, in other media, the light is slowed down. The number n defined by the ratio between the observed speed v and the speed of light

$$n := \frac{v}{c}$$

is called the **refraction index** of the material. It depends on the material and the frequency of the wave.

Another aspect of photons that is sometimes important is their **polarization**: A photon can be seen as both a wave and a particle (Wave-particle duality [13]) at the same time, and the polarization corresponds to the direction in which the wave oscillates. Note that this oscillation direction is usually perpendicular to the propagation direction. Moreover it should be noted that this explanation of polarization has been strongly simplified. See [12] for a more detailed description.

For some media, the refraction index differs depending on the polarization and the direction of the wave. This is called **birefringence** or double refraction. Instead of the refraction index, the **dielectric constant** ε is then used to describe the effects mentioned above which is, in this general case, a rank 2 tensor, i.e. a 3 by 3 matrix. For non-metallic materials, however, the dielectric constant is simply the square root of the refraction index. In this thesis, however, birefringence will be ignored since it only appears in metallic materials, which have no subsurface scattering.

The directions of reflection and refraction and their magnitude depends on several factors: The angle between the incoming wave and the material normal, the frequency of the wave, the polarization of the wave and the refraction index of both materials. Details are given in the following sections that explain what percentage of light is reflected (**Fresnel Equations**) and how the direction of the refracted wave can be computed (**Snell's Law**).

2.4.1 Snell's Law

Snell's law, which describes how an electromagnetic wave is refracted when it crosses the border of two materials with different refraction indices, is named after Willebrord van Roijen Snell (1580-1626), a Dutch astronomer and mathematician. But it seems that he was not the first derive it. It appears to have been recovered by Ibn Sahl (tenth century) and Thomas Hariot (late sixteenth century) before Willebrord Snell and René Descartes discovered it again. It is, however, not clear which of the two was the first to do so. Thus, in French, Snell's law is still called "la loi de Descartes".

The statement of Snell's law is the following: Whenever an electromagnetic wave crosses the border between two materials with refraction indices n_1 and n_2, the following property holds:

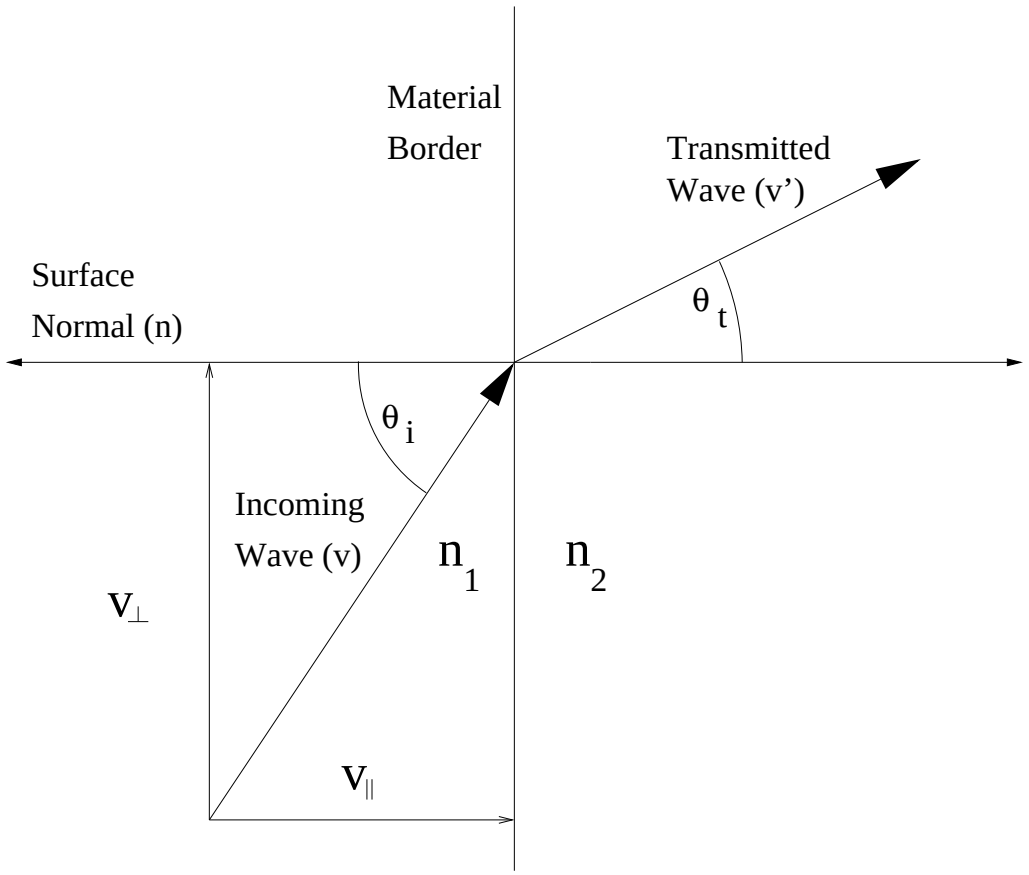


Figure 2.8: This figure illustrates the identifiers used in Snell's law: θ_i is the angle between the incoming wave and the normal (same for the transmitted wave and θ_t). In this picture, the transmitted ray is bent towards the normal with means that the refraction index n_1 is smaller than n_2.

$$n_1 \sin(\theta_i) = n_2 \sin(\theta_t)$$

Here, θ_i is the angle between the incoming wave v and the surface normal, and θ_t is the angle between the transmitted wave v' and the surface normal, as seen in Figure 2.8. Therefore, θ_t can be computed as $\theta_t = \arcsin(\frac{n_1}{n_2}\sin(\theta_i))$. It is easy to see that this equation does not always have a real solution if $n_1 > n_2$. In this case, there is no refracted ray and the whole wave will be reflected. Thus, this is called **total internal reflection**. The angle $\arcsin(\frac{n_1}{n_2})$ is called the **critical angle** because all waves arriving with a lower angle will split into a reflected and a transmitted wave, while the waves arriving with a higher angle will only be reflected.

Snell's law can be derived from **Fermat's Principle**: The actual path between two points taken by a beam of light is the one which is traversed in the least time. [6]

A classic and beautiful analogy is given by Richard P. **Feynman** (1918-1988), an American physicists[11]: Suppose a person on a beach want to rescue a drowning person in the sea. Since the rescuer is quicker on the beach, it can be shown that the fastest way to the drowning person is given by a path that follows Snell's law.

The value of the angle θ_t is sufficient to calculate the direction of the transmitted ray. In ray tracing, however, the value of θ_t is usually never computed because there are much quicker ways to compute the bend ray than to use θ_t directly. The first method was used by Whitted [29]. Here, however, the formula that appeared in [8] will be shown. The formula derived by Whitted (and an additional formula) are also summarized in [7].

Heckbert's idea was to write the refracted ray v' as sum of the normal vector and a second unit vector in the refraction plane that is perpendicular to the normal (compare Figure 2.8):

$$v' = \cos(\theta_t)n + \sin(\theta_t)\frac{v_\perp}{\|v_\perp\|}$$

Assuming n and v have unit length, this formula can be simplified in the following way:

[6] There also exists a modern version of this principle. It states that, for each material (whose optical density, i.e. the index of refraction, at the point x is given by $n(x)$) and each path γ, the **optical path length** $\int_\gamma n(x)dx$ must be extremal, i.e. either minimal or maximal.

$$
\begin{aligned}
v' &= -\cos(\theta_t)n + \sin(\theta_t)\frac{v_\perp}{\|v_\perp\|} \\
&= -\cos(\theta_t)n + \sin(\theta_t)\frac{v - v_\parallel}{\|\sin(\theta_i)v\|} \\
&= -\cos(\theta_t)n + \sin(\theta_t)\frac{v - (v\cdot n)n}{\sin(\theta_i)} \\
&= -\cos(\theta_t)n + \frac{n_1}{n_2}(v - (v\cdot n)n) \\
&= \frac{n_1}{n_2}v - \left(\cos(\theta_t) + \frac{n_1}{n_2}(v\cdot n)\right)n \\
&= \eta v - (\cos(\theta_t) + \eta(v\cdot n))\,n
\end{aligned}
$$

where $\eta = \frac{n_1}{n_2}$. The term $\cos(\theta_t)$ can be computed in the following way:

$$
\begin{aligned}
\cos(\theta_t) &= \sqrt{1 - (\sin(\theta_t))^2} = \sqrt{1 - \eta^2(\sin(\theta_i))^2} \\
&= \sqrt{1 - \eta^2(1 - \cos^2(\theta_i))} = \sqrt{1 - \eta^2(1 - (n\cdot v)^2)}
\end{aligned}
$$

Thus, the direction of the reflected wave is given by:

$$
v' = \eta v - \left(\sqrt{1 - \eta^2(1 - (n\cdot v)^2)} + \eta(n\cdot v)\right)n
$$

Note that v' is normalized.

2.4.2 Fresnel Equations

The Fresnel equations are named after **Augustin-Jean Fresnel** (May 10, 1788 - July 14, 1827). Whenever an electromagnetic wave hits the border of two materials with different refractive indices, they can be used to calculate the percentages of energy that are reflected and transmitted. These percentages do not only depend on the angle between the incoming wave and the material border θ_i and the transmitted wave and the material border θ_t (and therefore on the refractive indices of the materials) but also on the polarizations of the photons in the wave. For a wave with a polarization that is parallel to the **plane of**

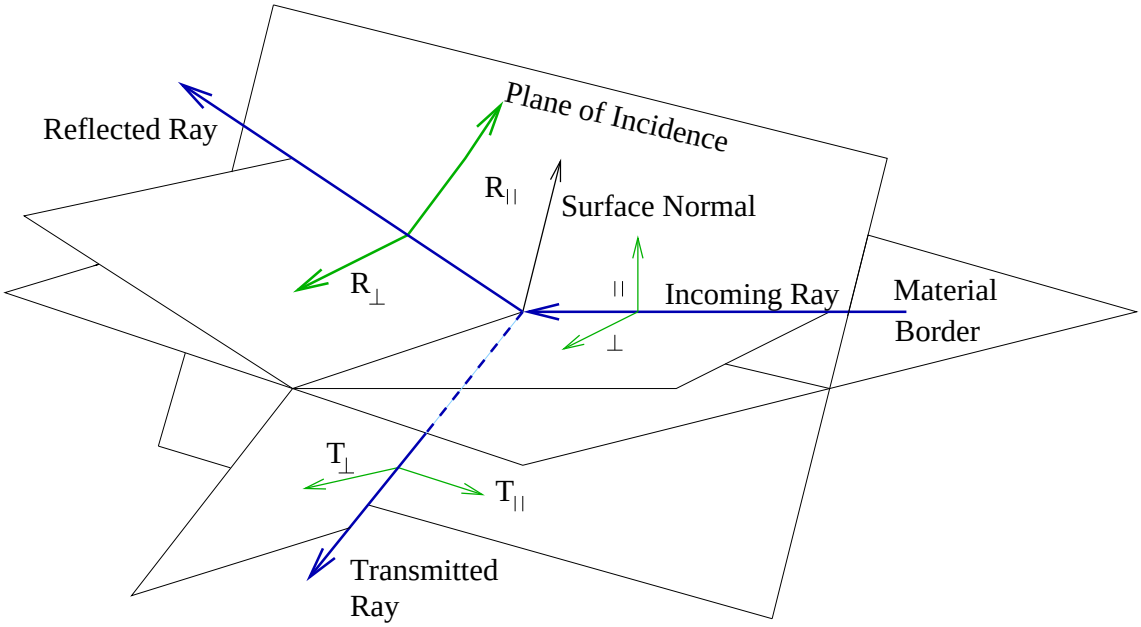


Figure 2.9: Here, the same situation as in Figure 2.8 is given: An electromagnetic wave hits the border between two materials with different refraction indices. This figure, however, has been done in 3D to illustrate the terms "plane of incidence" and to show the wave is polarized if the initial wave was *s* or *p*-polarized .

incidence of the wave[7], Fresnel's equations state that the following percentage of the wave is reflected:

$$R_{\parallel} := \left(\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}\right)^2$$

A wave that with this polarization is called *s*-**polarized**.

In case of a polarization perpendicular to the plane of incidence (*p*-**polarized**) the percentage of energy reflected is given by:

$$R_{\perp} := \left(\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}\right)^2$$

The Figure 2.9 illustrates what the plane of incidence is and how the polarization of the wave changes after reflection and refraction happened. Note that in Figure 2.8, only the plane of incidence can be seen.

For waves whose photons are unpolarized the amount of energy reflected is the average of those two cases:

$$R := \frac{R_{\parallel} + R_{\perp}}{2}$$

One should note, however, that the calculations done in this formula can become numerically unstable if $\theta_i \approx \theta_t \approx 0$ since the difference of two nearly identical numbers has to be computed. Luckily it is possible to transform the formulas to deal with this problem:

$$\begin{aligned}
R_{\perp} &:= \left(\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}\right)^2 \\
&= \left(\frac{\sin(\theta_i)\cos(\theta_t) - \sin(\theta_t)\cos(\theta_i)}{\sin(\theta_i)\cos(\theta_t) + \sin(\theta_t)\cos(\theta_i)}\right)^2 \\
&= \left(\frac{\frac{\sin(\theta_i)}{\sin(\theta_t)}\cos(\theta_t) - \cos(\theta_i)}{\frac{\sin(\theta_i)}{\sin(\theta_t)}\cos(\theta_t) + \cos(\theta_i)}\right)^2 \\
&= \left(\frac{\frac{n_2}{n_1}\cos(\theta_t) - \cos(\theta_i)}{\frac{n_2}{n_1}\cos(\theta_t) + \cos(\theta_i)}\right)^2 \\
&= \left(\frac{n_2\cos(\theta_t) - n_1\cos(\theta_i)}{n_2\cos(\theta_t) + n_1\cos(\theta_i)}\right)^2
\end{aligned}$$

A similar calculation leads to

[7] i.e. the photons are oscillating in a plane given by the wave direction and the material normal

$$R_{\parallel} := \left(\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}\right)^2 = \left(\frac{n_1\cos(\theta_t) - n_2\cos(\theta_i)}{n_1\cos(\theta_t) + n_2\cos(\theta_i)}\right)^2$$

Now it is obvious that, for $\theta_i \approx \theta_t \approx 0$, $R_{\parallel}$ and $R_{\perp}$ are both approximately $\left(\frac{n_1-n_2}{n_1+n_2}\right)^2$.

Another interesting fact is that $R_{\perp}$ will drop to zero if $n_1 < n_2$ at a certain angle known as **Brewster's angle**. For visible light hitting glass ($n \approx 1,5$) from air, this angle is approximately $56°$.

Once again there is a more efficient implementation possible in which there is no need to calculate the angle θ_t. Defining $A := \sin^2(\theta_i - \theta_t)$, $B := \sin^2(\theta_i + \theta_t)$, R can be computed as:

$$\begin{aligned}
2R =& R_{\parallel} + R_{\perp} = \left(\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}\right)^2 + \left(\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}\right)^2 \\
=& \frac{A^2}{B^2} + \left(\frac{\frac{A}{\cos^2(\theta_i - \theta_t)}}{\frac{B}{\cos^2(\theta_i + \theta_t)}}\right)^2 \\
=& \frac{A^2}{B^2} + \frac{A^2}{B^2}\left(\frac{\cos^2(\theta_i + \theta_t)}{\cos^2(\theta_i - \theta_t)}\right)^2 \\
=& \frac{A^2}{B^2}\left(1 + \left(\frac{1 - \sin^2(\theta_i + \theta_t)}{1 - \sin^2(\theta_i - \theta_t)}\right)^2\right) \\
=& \frac{A^2}{B^2}\left(1 + \frac{1 - B^2}{1 - A^2}\right)
\end{aligned}$$

Thus, knowing A^2 and B^2 is sufficient to use the Fresnel equations. Setting $C := \sin(\theta_i)\cos(\theta_t)$ and $D := \cos(\theta_i)\sin(\theta_t)$ yields $A = (C - D)^2 = C^2 + D^2 - 2CD$ and $B = (C + D)^2 = C^2 + D^2 + 2CD$. The terms C^2 and D^2 can easily be compute using the facts that

$$\sin^2(\theta_i) = 1 - (v \cdot n)^2, \quad \sin^2(\theta_t) = \sin^2(\theta_i)\left(\frac{n_1}{n_2}\right)^2$$

and $2CD$ is then given through $2CD = 2\sqrt{C^2D^2}$. Thus, R can be computed quite efficiently from $v \cdot n$ and the ratio $\eta = \frac{n_1}{n_2}$ without computing the angle θ_o, just like in the computations done for Snell's law[8].

[8] It is possible to further optimize this calculations. However, the additional benefi t from these further optimizations are rather small: One multiplication is removed for another addition. Details can be found in [22] and are thus omitted here.

2.5 Participating Media

This section will deal with materials that influence the light (i.e. the photons) that traverses through them. Two different photons-material interactions can be noted in such a material, which is then called **translucent** or **participating media**. The two possible interactions (light absorptions and light scattering) will be discussed first, followed by the derivation of an algorithm that can be used to sample the new direction of a photon after a scattering event. This section will, however, not give most of the physical background that explains why the formulas given here are valid because that would lead too far here. The justification of these formulas can be found in [14], however.

2.5.1 Absorption

The first kind of photon-material interaction is that it is possible for photons to be absorbed by the atoms or molecules the material consists of. While this might be a deterministic process in reality, this is usually seen as a random process. Thus, speaking of the percentage of photons (or light) that are transmitted (and not absorbed) while the light traverses from one point x to another point x' makes sense. This percentage is called the **beam transmittance** and is usually abbreviated by $T_r(x,x')$, were $r := \|x-x'\|$ is the length of the path from x to x'. Furthermore, $\omega := \frac{x'-x}{\|x'-x\|}$ is the unit vector in the direction from x to x'. Note that this function does not take into account if photons have been scattered into the wave. This effect is considered separately in the next section.

The beam transmittance is used to define the **volume attenuation function**

$$a_r(x,\omega) := \frac{1-T_r(x,x')}{r}$$

which is required for the definition of the **absorption coefficient**:

$$\sigma_a(x,\omega) := \lim_{r\to 0^+} a_r(x,\omega)$$

It is quite easy to see that the derivation of T_r with respect to r is $\frac{dT_r}{dr} = -\sigma_a T_r$, and this differential equation has the solution

$$T_r = e^{-\int_0^r \sigma_a(x)dx}$$

Most of the time, the absorption coefficient is independent from the direction of the wave and only depends on the position x. In most uniform materials like plastic or milk, it is even constant. The equation above then simplifies to **Beer's law**:

$$T_r = e^{-\sigma_a r}$$

For more complex materials like marble, however, the position dependency is important. Still, in the following section, the absorption coefficient is always written as σ_a, i.e. without its parameters, since its actual form is not important for the remaining discussion.

Obviously, the greater the value of σ_a, the more photons are absorbed on their way. Not so obvious is the fact that the average length a photon traverses before it is absorbed is given by $\frac{1}{\sigma_a}$.

2.5.2 Scattering

Photon scattering events are more complex to simulate than absorptions. Not only has to be found out how often a photon is scattered, moreover the direction the photon takes after scattering must be computed. Luckily the first question can be solved just like in the last section resulting in the definition of the **scattering coefficient** σ_s[9]. The **attenuation coefficient** σ_t is the sum of the absorption coefficient and the scattering coefficient:

$$\sigma_t := \sigma_a + \sigma_s$$

This is an important quantity since the distance d a photon can travel through a material with attenuation coefficient σ_t can be modeled by setting $d := \frac{-\log(\xi)}{\sigma_t}$, where ξ is a canonical random variable. The expected value of d is then given by: $\int_0^1 \frac{-\log(x)}{\sigma_t} dx = \frac{1}{\sigma_t}$.

Thus, the average distance a photon can travel through a material with attenuation coefficient σ_t is given by $\frac{1}{\sigma_t}$.

Note that physically, a scattering is nothing else but a collision of the photon with an atom or molecule of the media. By a scattering event, however, the photon is not absorbed but reflected (or it is absorbed but re-emitted). Thus, since both absorption and scattering take place when the photon hits a particle of the media, it makes sense to model absorption and scattering together. Then, the fraction of energy Λ that is re-emitted (or, more precise, the probability that a photon is re-emitted)

$$\Lambda := \frac{\sigma_s}{\sigma_t}$$

is called **albedo**.

It is now clear how to find out when a photon will be scattered, but the question in which direction the photon is scattered is still open. It can be answered by a **phase function** k.

The phase function is a probability density function taking the direction of the incoming light ω and outgoing ray ω' and whose value corresponds to the probability that light

[9] The parameters are omitted once more.

arriving from ω is scattered into the direction ω'. Often, however, the form of k is much simpler. For nearly all naturally-occurring media, the value only depend on the angle θ between ω and ω'. Then, the media is called an **isotropic media**. Sometimes ([28]), the phase function is even defined as: "The reflected intensity as a function of angle normalized by the intensity at normal incidence", i.e. in the definition of the phase function, it is assumed that the material is isotropic since this is usually the case in current research (and will be the case in this thesis). In an **anisotropic media**, however, more dimensions are needed. A typical example where all four dimension may be needed is inside a crystalline structure.

Care must be taken when using the terms isotropic and anisotropic since these terms are not only used for materials but also for the phase function itself: An **isotropic phase function** does not depend on its parameters. In other words, it is a constant function, usually with the value 1. On the other side, an **anisotropic phase function** is the general function described above.

A large range of different analytical phase functions have been developed. Some are derived from simplified assumptions, some are empirical and have a few additional parameters that can be used to fit the function to observed data. When dealing with anisotropic phase functions and an isotropic media, the situation can be simplified by introducing the **asymmetry parameter** g which is defined as the average of the product of the phase function and the cosine between the two directions θ over the sphere S^2. That is, for any ω

$$g := \int_{S^2} k(\omega,\omega')\cos(\theta)d\omega' = \int_{S^2} k(\omega,\omega')(\omega\cdot\omega')d\omega'$$

Note that g is always a value from the interval $[-1,1]$. The larger the value of g, the less the direction of most photons is usually changed by each scattering. A value of 1 means total forward scattering (i.e. no scattering at all) while a value of -1 is achieved in total back-scattering (i.e. every photon is reflected and travels back where it came from). Obviously, an isotropic phase function has an asymmetry parameter of $g = 0$.

Although the value of g is not enough to describe the phase function completely it is not unusual to use only g to describe the scattering behavior of a media. This is done because the loss of accuracy is considered rather small in comparison to the gain of being able to describe the phase function by a single parameter. One of several possible ways to estimate the phase function from the asymmetry parameter is to set

$$k(\omega,\omega') := \frac{1-g^2}{2(1+g^2-2g(\omega\cdot\omega'))^{\frac{3}{2}}} = \frac{1-g^2}{2(1+g^2-2g(\cos(\theta)))^{\frac{3}{2}}}$$

This formula was introduced by Henyey and Greenstein in [9] and is thus called **Henyey-Greenstein phase function**. It will be used here because it is widespread, cheap to evaluate and close to a lot of observed phase functions. For more complex phase functions,

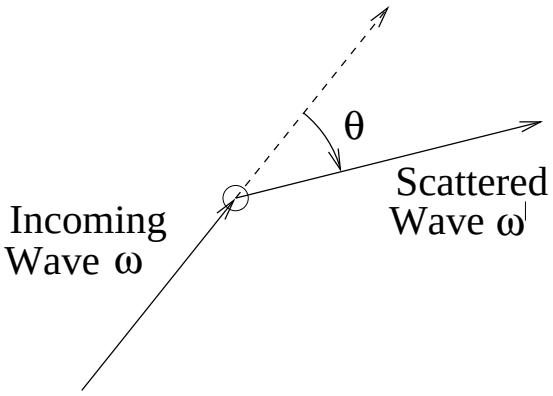


Figure 2.10: This figure shows the situation when an electromagnetic wave arriving from ω hits an atom and is thus scattered into a new direction ω'. The larger the angle θ, the more the wave is diffracted in the scattering event. As described in the text, this angle is often the only interesting value when regarding scattering events.

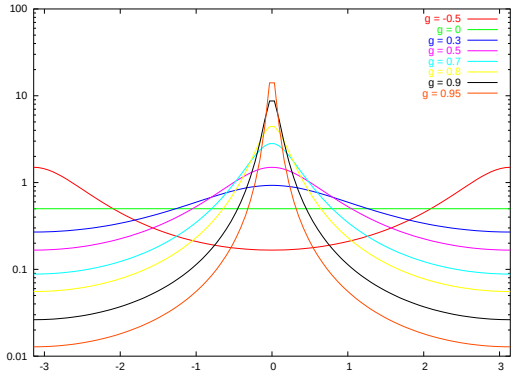


Figure 2.11: The Henyey-Greenstein phase function with different values for g is shown in this figure, where $\theta \in [-\pi, \pi]$

a typical approach is to use the weighted sum of several simple phase functions like the one described above. The function is shown in Figure 2.11 for different g's.

Some authors have an additional multiplicative factor in the definition of g and thus have an other factor than $\frac{1}{2}$ in the Henyey-Greenstein phase function. Here, the factor used is of no importance since it will cancel out in the next section.

2.5.3 Sampling Scattering Directions using the Henyey-Greenstein Approximation

In this section, the techniques explained in Section 2.3.2 are used to derive a uniform sampling algorithm of the sphere with the importance function given by the Henyey-Greenstein formula. In other words, a function that takes the incoming direction of a photon and two canonical random numbers and returns the new direction of the photon after a scattering event will be acquired.

Using the same parametrization as before yields

$$\sigma'(s,t) = 2\pi^2 \sin(\pi s) \frac{1-g^2}{2(1+g^2+2g\cos(\pi s))^{\frac{3}{2}}}$$

Since σ' is independent of t, it follows easily that $G_u(v)$ (and therefore also its inverse function) is the identity (compare the Pages 20 and 21, where this has already been shown for the disc and the hemisphere).

The function $F(u)$ is given by:

$$\begin{aligned}
F(u) &= \frac{\int_0^1 \int_0^u \frac{2\pi^2 \sin(\pi s)(1-g^2)}{2(1+g^2+2g\cos(\pi s))^{\frac{3}{2}}} ds\, dt}{\int_0^1 \int_0^1 \frac{2\pi^2 \sin(\pi s)(1-g^2)}{2(1+g^2+2g\cos(\pi s))^{\frac{3}{2}}} ds\, dt} \\
&= \frac{\int_0^u \frac{\pi \sin(\pi s)(1-g^2)}{(1+g^2+2g\cos(\pi s))^{\frac{3}{2}}} ds}{\int_0^1 \frac{\pi \sin(\pi s)(1-g^2)}{(1+g^2+2g\cos(\pi s))^{\frac{3}{2}}} ds} \\
&= \frac{\left| \frac{(1-g^2)}{g\sqrt{(1+g^2+2g\cos(\pi s))}} \right|_0^u}{\left| \frac{(1-g^2)}{g\sqrt{(1+g^2+2g\cos(\pi s))}} \right|_0^1} \\
&= \frac{\frac{1-g^2}{g\sqrt{(1+g^2+2g\cos(\pi u))}} - \frac{1-g^2}{g(1+g)}}{\frac{1-g^2}{g(1-g)} - \frac{1-g^2}{g(1+g)}} \\
&= \frac{\frac{1-g^2}{g\sqrt{(1+g^2+2g\cos(\pi u))}} - \frac{1-g}{g}}{2}
\end{aligned}$$

Now, the inverse is calculated:

$$\begin{aligned}
&\frac{1-g^2}{2g\sqrt{(1+g^2+2g\cos(\pi u))}} - \frac{1-g}{2g} = y \\
\Leftrightarrow &\frac{1-g^2}{\sqrt{(1+g^2+2g\cos(\pi u))}} - 1 + g = 2gy \\
\Leftrightarrow &\frac{2gy+1-g}{1-g^2} = \frac{1}{\sqrt{(1+g^2+2g\cos(\pi u))}} \\
\Leftrightarrow &\left(\frac{1-g^2}{2gy+1-g}\right)^2 = 1+g^2+2g\cos(\pi u) \\
\Leftrightarrow &-\cos(\pi u) = \frac{1+g^2-\left(\frac{1-g^2}{2gy+1-g}\right)^2}{2g}
\end{aligned}$$

This formula can now be used to sample the photon direction after a scattering event.

Chapter 3

Previous Work

The topic of how waves interact with scattering media is important in several different areas. In optical tomography, for example, the scatterings done are used to determine how the interior of an object looks like. This requires a difficult inversion step. Optical tomography is used in medicine and nondestructive material testing. In this thesis, however, rendering translucent objects is the main goal.

When creating images with translucent objects there are basically two questions: How can such an object be described and how can it be rendered? In the past, different approaches have been made to answer these questions. We only describe some of them here. For a more detailed overview see Goesele et al. [5].

3.1 Physical Background and Modeling

One of the early important articles was published by Henyey and Greenstein in 1941 [9], in which the Henyey-Greenstein phase function was introduced and used to solve problems about diffuse radiation in the galaxy. Later, in 1978, Ishimaru published a summary about the physical process of scattering in two volumes[14]. Using this work, a translucent object can be described by the absorption coefficient σ_a, the scattering coefficient σ_s and possibly an asymmetry parameter g for every point inside its volume.

However, instead of giving σ_a, σ_s (and g) for each point, an object can also be described by the 8-dimensional BSSRDF (Bidirectional Scattering Surface Reflection Distribution Function) introduced by Nicodemus et al. [21], which relates incoming irradiance at a certain point from a certain direction to the outgoing radiance at a (possibly different) point in a certain direction. By ignoring the dependence on the directions, the function collapses to the 4D diffuse subsurface reflection function R_d, making it far easier to handle.

In this thesis, we will use R_d, mainly because the incoming and outgoing directions are very important for single scattering, but rather unimportant if multiple scattering occurs.

3.2 Acquisition

In order to represent an object with subsurface scattering properties, one needs either the BSSRDF, the scattering parameters of the object or some other representation of the object.

Feng et al. analyzed the photon paths in scattering media. In [4], they especially examined two cases: In the first case, the material is infinite and occupies the whole space while in the second, the material is semi-infinite and consists of a half space. They noted that, in the second case, most of the photons are inside a region that looks like a banana. The term **photon banana** comes from this paper.

In 2001, Jensen et al. solved the problem of subsurface scattering analytically for a shift-invariant material that has the form of an infinite half space [17]. Their model consists of two light sources (one of which emits "negative" energy) to model energies from multiple scattering and a rather complex single scattering term. Due to these two lights, this model is called the **dipole approximation**. This approach requires little data but the results are only exact for this special case.

The dipole approximation states that the energy traveling from x to y in an infinitely half space with a homogeneous material is given by:

$$R_d(x,y) = \frac{\alpha'}{4\pi}\left(z_r(1+\sigma_{tr}d_r)\frac{e^{-\sigma_{tr}d_r}}{d_r^3} + z_v(1+\sigma_{tr}d_v)\frac{e^{-\sigma_{tr}d_v}}{d_v^3}\right)$$

In this equation, $\alpha' = \frac{\sigma_s'}{\sigma_t'}$,$\sigma_{tr} = \sqrt{3\sigma_a\sigma_t'}$ with $\sigma_s' = \sigma_s(1-g)$ and $\sigma_t' = \sigma_s' + \sigma_a$. The variables z_r and z_v are the distances of the two light sources (one of which is inside the material) from x. These distances must be computed from the material parameters. d_r and d_v are the distances from the light sources to y.

Moreover, the dipole model allows to compute the absorption coefficient σ_a and the reduced scattering coefficient σ_s' from measurements. The parameters of some materials in which subsurface scattering is considered important (such as marble, skin and milk) are given in [17].

Another approach has been done by Goesele et al. in [6]. They explicitly acquired the amount of light that traverses between each pair of points on a given object. This enabled them to interactively render acquired objects with subsurface scattering under arbitrary lighting conditions without inversion [18].

3.3 Rendering

How translucent objects are rendered depends on the representation. If a BSSRDF is given, the rendering step is expensive using a naive implementation. If only the scattering parameters are given, methods like photon mapping [2], diffusion [27] or Monte Carlo simulations [25] can be used to feign the effects occurring in translucent objects.

In 2001, Hendrik Wann Jensen and his co-authors published severals works regarding subsurface scattering [15], [16], [17]. In [15], photon mapping is used to handle (among other things) scattering media. A similar approach has also been done by Dutré et al. in [3]. In [16], Jensen gives a summary about ray-tracing scenes with subsurface-scattering, including sampling 2-manifolds, Monte Carlo integration and other possible approaches like metropolis sampling.

Chapter 4

Monte Carlo Simulation and Verification

As described in the last chapter, the dipole model by Jensen et al. is easy to compute and yields very good results but only under certain assumptions. It is not a priori clear if it is still feasible to use the dipole approximation if one or several of these assumptions are not true.

Thus, one goal of this thesis is to compare the estimation from the dipole approximation with a Monte Carlo simulation. The results are used to decide under which conditions the dipole model can be used without loosing too much precision, or if it can (and should) be adapted to be able to handle the new situations more accurately.

We start by briefly explaining how the Monte Carlo simulation is done (as we will see, several decision have to be done in the simulation) and explain the different kinds of results that can be generated, depending on the geometry of the object in which the subsurface scattering will take place. This is the topic of the first section. The second section deals with verifying that the simulation yield correct results.

4.1 The Monte Carlo Simulator

The main idea of evaluating subsurface scattering using Monte Carlo ray tracing is not very complicated. The amount of radiance $L_0(p, \omega_0)$ that leaves a material with surface A at a specific point p with normal n into the direction ω_0 is given by

$$L_0(p, \omega_0) = \int_A \int_{2\pi} S(q_i, \omega_i, p, \omega_0) L_i(q_i, \omega_i)(n\omega_i) d\omega_i dA(x_i)$$

where $S(q_i, \omega_i, p, \omega_0)$ is the amount of the energy that enters the material at the point q_i with direction ω_i and leaves the material at p in the direction ω_o. The function S is called **Bidirectional Scattering Surface Reflection Distribution Function**, or short **BSSRDF** and is a generalization of the well known BRDF [21],[17]. Integrating $L_0(p, \omega_0)$ over all possible directions for ω_0 yields the total amount of energy leaving the material at p.

If the BSSRDF is known this integral can be calculated rather easily. Unfortunately this is usually not the case. As already mentioned, an analytical solution of S has been computed for only one very simple geometry (infinite half space), a shift-invariant material and under several other assumptions. In the general case, computing the BSSRDF would require to integrate the incoming energy over all incoming directions at q_i. But, to get the energy arriving from a certain direction, the integral of the energy scattered into the direction of q (and not scattered again (or absorbed)) over all possible distances from q has to be computed. Even then, only photons that have been scattered at most once would be considered. To get the influence of photons that have been scattered more often, more and more integrations are required. Since it is not uncommon that photons are scattered a few hundred times (or even several thousand times), it is no longer feasible to compute this integral[1].

Thus, we use Monte Carlo integration to estimate this function. A short overview how the photons are traced in our simulator is given in Figure 4.1. All steps shown there (and some more) are explained next in this chapter.

We start with a photon with a 'random' starting direction (uniformly distributed from the hemisphere to the point q, using the formulas from Section 2.3.3, Page 21).

This photon enters the material at the point q, changing it's direction according to Snell's Law (compare Section 2.4.1).

Furthermore Fresnel's equations must be taken into account. However, there are two possible approaches how to do this. One approach would be to decide if the photon enters the object at all or if it is reflected. This approach is called **Russian Roulette**, just like any other approach where two or more possibilities exist and exactly one is taken.

Another possibility is to let all photons pass, but give each photon a 'virtual energy' in the range $[0, 1]$. The initial energy is then the probability that the given photon enters the object which is given by the Fresnel equations.

Note that this 'virtual energy' has nothing in common with the real, physical energy of the photon (which depends on factors like the wavelength (i.e. the color) of the photon). It merely gives the probability that a photon is not reflected (or, as we will see later, absorbed) if it takes this way.

[1] Note that every time, there are an uncountable infi nite possible direction/distances over which to integrate.

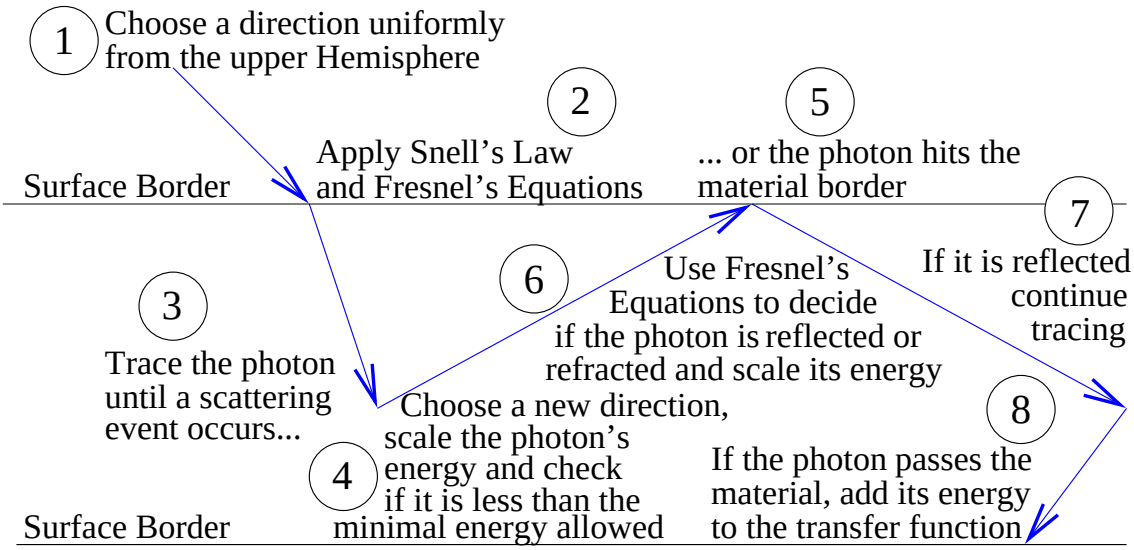


Figure 4.1: A possible photon path and the steps done in our Monte Carlo simulation. Details are given in the text.

Once the photon is inside the object we have to look if the photon first hits the border of the object, interacts with the material (i.e. is scattered) or is rejected.

If the photon hits the boundary of the object the Fresnel terms are computed and used to (randomly) decide if the photon leaves the object or if it is reflected. Here, simply scaling the energy as we have done when the photon entered the material would either result in no photon being reflected or in no photon being able to leave the object. Both effects are not desired since there either are possible photon paths not regarded or no photon (i.e. energy) will ever leave the object.

If the photon leaves the object, the 'virtual' energy remaining will be added to the BSS-RDF[2].

A wrong way to handle the Fresnel term would be to split the photon into two parts (one of which is reflected while the other is refracted) because then, the photons whose paths we are following would no longer be independent. Thus, we no longer have Monte Carlo integration but Quasi Monte Carlo integration, possibly introducing a undesired bias.

To check if the photon interacts with the material we can model the distance the photon traverses by $d := \frac{-\log(\xi)}{\sigma_t}$ if the material is uniform (where ξ is a canonical random variable and σ_t is the attenuation coefficient of the material, see Section 2.5.2). For non-uniform materials we have to choose d such that $\int_0^d \sigma_t(x)dx = -\log(\xi)$. The right value for d can

[2]Of course, the BSSRDF is stored discrete. The discretisation used in the simulation depends on the geometry. This topic will be covered in the next section in detail.

be easily approximated if the integral can be computed (it might be necessary to use Monte Carlo integration to do so). In our simulations, σ_t was either constant or a step function (since the material was stored as a voxel grid). In this case, all we had to do is to compute the integral per voxel, i.e. $\int_0^d \sigma_t(x)dx = \sum_{i=1} \int_{a_i}^{a_{i+1}} \sigma_t(x)dx$ where $0 = a_1 < a_2 < a_3 \ldots$ are the distances the photon must travel to enter the ith voxel. Since the material parameters are constant inside each voxel, each integral can be computed trivially. As soon as the sum is greater than $-\log(\xi)$, we know the d we are looking for is inside this voxel. The exact value can then be obtained by linear interpolation between the value of the sum before and after this cell.

In case of a photon material interaction, the energy is scaled with the albedo $\Lambda(x)$ of the material at the current position. Furthermore the direction of the photon changes according to the Henyey-Greenstein approximation (see Section 2.5.3) and the asymmetry parameter $g(x)$ of the material at the current position. Using Russian roulette would also be possible, but as soon as a photon is absorbed the work done to trace its way would be wasted. Thus, Russian roulette would require more time to yield comparable results.

We might also decide that the remaining energy of the photon is so small that it can as well be thrown away. This step is necessary because it is possible (although unlikely) that a photon will never leave the object again. This can especially happen in infinite objects such as the infinite half space used in the dipole model. In such a case, the Monte Carlo simulation would run forever (or a very long time) for a single photon whose influence is very small. Of course, throwing away a photon does alter the result of the computation. But, as will be shown later, the influence of these photons is rather small if the threshold under which the photon's energy must fall to be thrown away is chosen correctly.

4.1.1 Simulation Results and Post-Processing Tools

In this section, the possible outputs our Monte Carlo simulation can create are explained.

Photon Paths and Associated Modes

The first output mode we implemented is to save the traversal paths of all photons arriving close to a given point along with the energies the photons have between scatterings. This is useful to watch how photons traverse from one point to another. In order to be able to use this data the other way round (e.g. from the end point to the start point) it is useful to start the photons in a area identical to the area in which the photons must arrive[3]. This will usually be a small circle.

[3]Still, there is a difference because more photons will arrive in some regions (e.g. the regions that are nearer to the starting point) but this effect cannot easily be adapted for the starting area of the photons. However, this effect is reduced by shrinking the area in which the photons must arrive.

Afterwards, different average photon paths can be computed. The most primitive average photon path is obtained by averaging the photon paths seen as functions from $[0,1]$ to $\mathbb{R}^3$, where each path is parameterized uniformly with respect to its length. This is done approximatively by choosing equidistant values in $[0,1]$. For each value, the average function value of all paths yields a point on the average photon path[4]. The averaging step can also compute a weighted average of the function values by choosing the energy at the end of the path (i.e. the probability that the photon taking this way is not absorbed) as weights.

The photon paths can be used to generate a volume data grid that stores the energy passing through each cell. This volume data can be used to approximate an iso-surface of the photon density inside the object using standard tools like AMIRA.

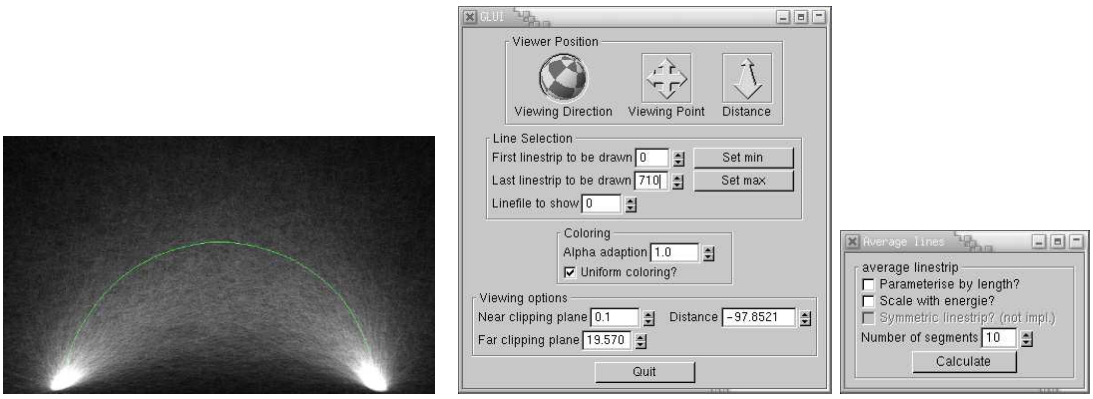
Another possibility is to reuse the photon paths for a new geometry: We wrote a program that loads the photon paths and checks which paths do not intersect with a new geometry. This way, a completely absorbing blocker can be added to an old geometry without redoing the Monte Carlo simulation.

The saved photon paths can also be displayed together with the geometry by a special viewer we wrote with GLUI [26]. A screenshot of this viewer is shown in Figure 4.2. This figure also illustrates why the region where most photons are is called the **photon banana**[4]. However, the photon paths are most of the time not useful because simple questions like the amount of energy that has arrived near the endpoint cannot be answered without further processing. Moreover, this question is often asked for every point on the object. To answer this question with post-processing of the photon paths would require to save about 2183 MB for a million photons (and we did simulations with up to one billion photons). Thus we also implemented other output modes that only summarizes how much energy has arrived where, without saving the photon paths. These are described in the next sections.

Two-dimensional Output Modes

The easiest output modes are two dimensional arrays containing the ratio of energy leaving the object in a certain area per square millimeter. Thus, values above 1 (i.e. a hundred percent) are possible if the area is smaller than one square millimeter. Using the right program, we can examine every point in such two dimensional arrays, which we will call **image** from now on. This is unfortunately not possible once the images are printed. That is, only relative changes can be seen in these images but the exact amount of energy leaving the object cannot be read from the images shown here. The images shown in this thesis are usually gamma-corrected and amplified. This way, small changes are easier to spot.

[4]Computing the average is in this case the same as computing a least square fitting.



(a) The display area

(b) The control elements

Figure 4.2: This figure shows a screenshot of the photon path viewer we wrote. The display area on the left shows the photon paths which are in this case the photon paths for an infinite half space. Thus, the photon banana can be seen there, as well as the average photon path the program guessed (green line). The right side of the figure shows the control elements of the program where various parameters can be changed.

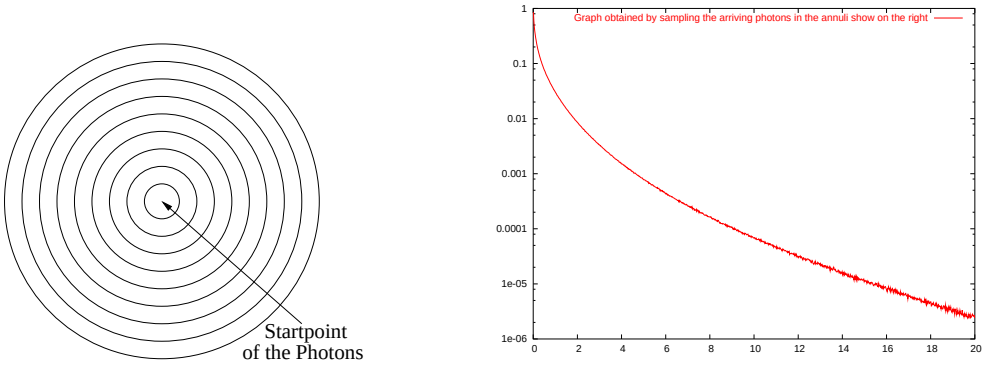
Theoretically, a lot of 2-manifolds can be mapped to such an image. The images shown in this thesis are either part of a plane or of a infinitely long, slit-up cylinder and thus in both cases planar. The cylinder is slit up along a straight line opposite to the photon starting point. That is, the photon start point is in the center of the image. Therefore these images are axis-symmetric on both the x and the y-axis if no noise occurs.

In other cases, however, there were much bigger areas that could be stacked to single points. Then, the result of the simulation has only one dimension, as described in the next chapter, i.e. only the decay of the energy with distance is of interest.

One-dimensional Output Modes

Another output mode that is only a bit more complex results in a one dimensional "**graph**" like the one in Figure 4.3(b). Here, the percentage of energy leaving at a certain distance (once again measured in percent per square millimeter) is plotted as y-coordinate with the corresponding distance as x-coordinate. To do so, all energy arriving within a certain region is gathered and divided through the area of the region.

The main advantage of these modes is that it is very easy to compare different graphs (since they can be plotted into the same figure). To do so, a small program that calls gnuplot to plot several graphs in one figure has been written. Moreover, it is rather easy



(a) A set of possible sample areas

(b) A corresponding graph to the sample areas on the left side

Figure 4.3: The left part of the figure shows a set of possible sample areas on a plane while a corresponding graph is given on the right side.

to read off approximate values from the graph even in this thesis which is not true for the two dimensional data, as explained above. Furthermore, since the data from several points are accumulated, the noise is reduced quickly.

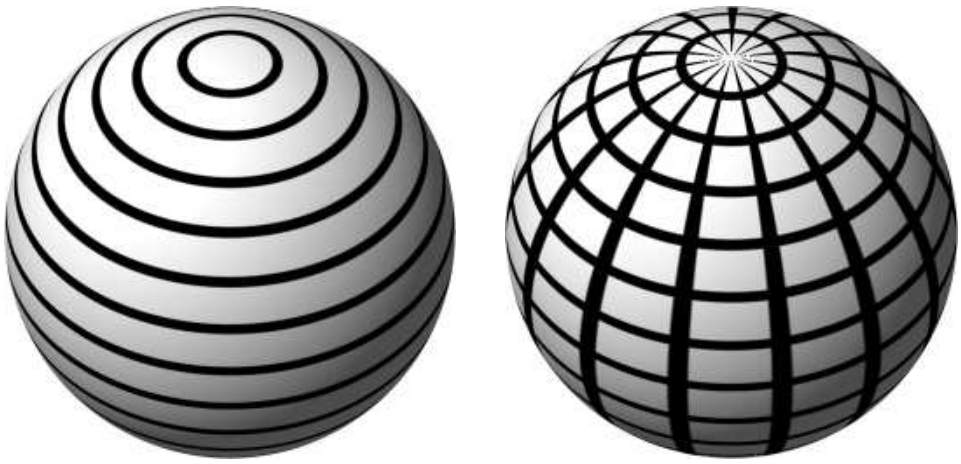
On the other side, gathering data in one dimensional modes only makes sense if the energy arriving at two different points with the same distance from the start point is equal. Thus the material and the object have to be symmetric[5].

There are three different kinds of such graphs in this thesis:

The easiest graph is just a line of a two-dimensional image, i.e. the measures energy along a straight line (usually either a horizontal or a vertical line) is plotted as a graph. Such a graph can easily be generated by one of our programs. A graph contains much less information than the image it was made from, but is useful to see small changes more easily. Moreover, it is quite simple to plot two different slices from one or two images into the same figure, allowing to see differences not visible in the image(s). If two different images touch or overlap at a certain point, it is also possible to draw a graph where the x-coordinate starts at 0 in one image and runs continuously over both images, as shown in Section 5.1. In this case, the distance is measured along this line. A tool that has been written for this purpose can perform this step automatically if desired.

If the area to be evaluated is planar, all photons arriving inside one annulus (or a single disc in the center) are summarized to one point in the graph, where the center of all annuli is the start point of the photons. The annuli and a possible resulting graph are shown

[5]The exact kind of symmetry required depends on the sampling method



(a) Sampling areas for a graph on the sphere

(b) Sampling areas for an image on the sphere

Figure 4.4: The first figure shows how a sphere can be divided into areas such that the energy leaving the sphere is approximately constant for all points in a given area. Again, a uniform material is required for this approximation to be valid. If this is not the case, a two-dimensional image can be generated by further dividing the sphere as shown in the right image.

in Figure 4.3. For each annulus (and for the inner circle), one point is plotted with the "average" distance of the annulus from the photon start point as x-coordinate and the average energy leaving the object (per unit area) inside the annulus as y-coordinate. The difference of the outer and inner radius of all annuli is always constant in this thesis. Obviously, reducing this distances results in more accurate graphs but higher noise.

It is also possible to gather energies for graphs on spheres, as demonstrated in Figure 4.4(a). Then, the energy arriving between two circles of latitude (or near the poles, respectively) is combined into one point of the graph. Once again, the granularity can be chosen arbitrarily. In this thesis, the graphs that came from spheres usually have circles of latitude with uniform distances from one pole.

4.2 Verification

This section will briefly cover the topic of correctness. That is, we try to verify that the simulation results are close to real observations and try to find possible errors.

First of all, it is easy to see that photon-object intersections[6] are handled correctly since a huge amount of photon paths have been saved and observed.

We described earlier (see Section 4.1, Page 42), that there is a minimal energy a photon can have before it is thrown away. In Figure 4.5, one can see several graphs. The only difference between them is the amount of energy the photons had when they left the material. Using this figure (and several others not shown here because they basically show a similar behavior) one can see how big the error by introducing the minimal energy is in which distance. For example, choosing a minimal energy of 1% of the initial energy results in a error of approximately five percent at ten millimeters from the photon start point. Reducing the minimal energy to 0.1% (that is the value we used in most of our computations) results in only 0.35% error at distance ten millimeters and about 14% error at a distance of twenty millimeters. Since we rarely gather data at more than ten millimeters from the initial start point, this is sufficient for our calculations.

As a final step, we compared a simulation to the analytic result of Jensen given in [16], i.e. the dipole model. In the simulation, single scatterings have been excluded since single scattering is not included in the dipole model, neither. One can see in Figure 4.6 that the shapes of the graphs are similar, but not identical. This is due to the fact that Fresnel's equations are not included in the dipole model. Figure 4.7 shows how the graphs look like if no refraction occurs. Moreover, the starting directions of the photons are sampled uniformly on a hemisphere before entering the material, but nothing comparable is known for the dipole approximation, making these two approaches hard to compare[7].

[6] The border of the objects was stored as triangles. We used a BSP tree and the algorithm described in [20], with slight modifi cations to check for photon-triangle intersections.

[7] I assume Jensen did sample the photon directions uniform after they are already in the material while we sampled them uniformly before the entered the material. However, this is only a guess.

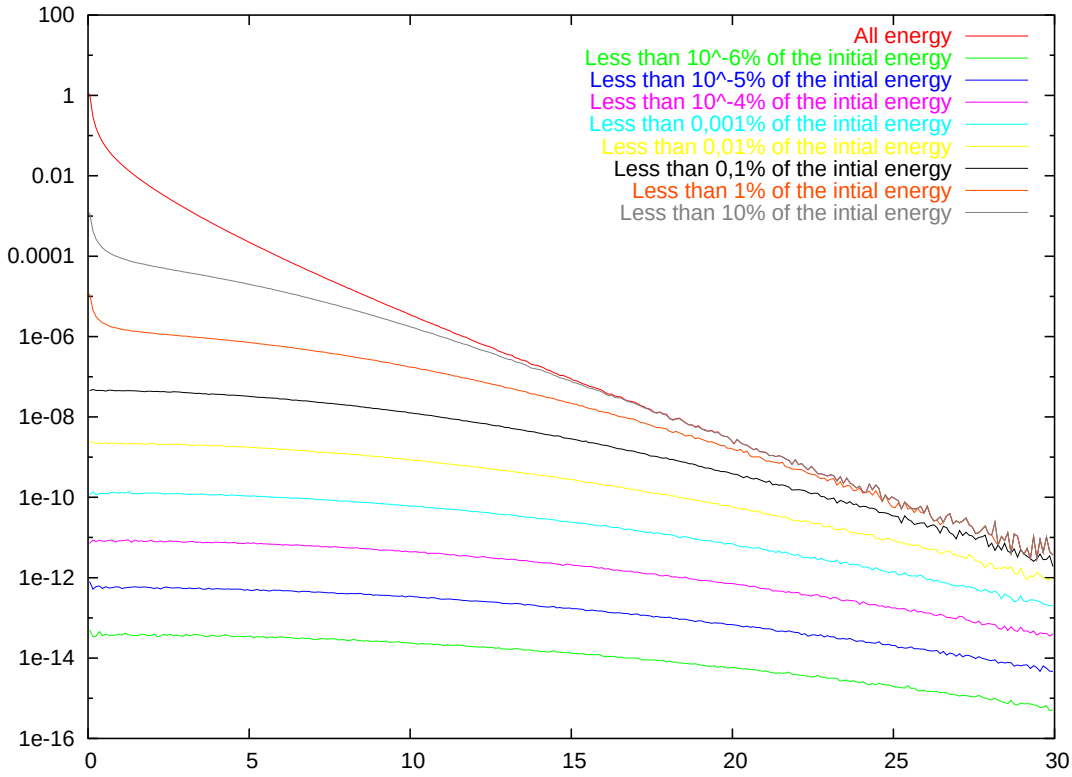


Figure 4.5: This figure shows the amount of energy that is lost due to the minimal energy requirement introduced for the infinite half space geometry and a shift-invariant material. The uppermost, red graph shows the energy arriving for a very low minimal energy (10^{-7}% of the initial energy.). The other graphs illustrate the amount of energy that is lost due to the minimal energy for a minimal energy of $10^{-i}, i \in \{1 \dots 8\}$. For example, by choosing a minimal energy of $10^{-5} = 10^{-3}$%, we throw away photons that would arrive at a distance of $5mm$ with a total energy of about $10^{-10} = 10^{-8}$% of the initial energy, per square millimeter. Compared to the energy arriving of about $0.0002 = 2$%, this effect can be ignored without a noticeable loss of precision.

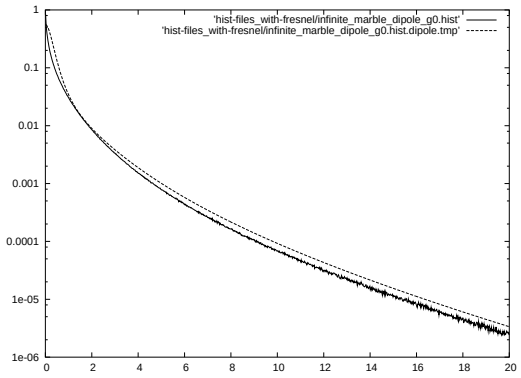


Figure 4.6: In this figure, the results from the dipole model (without the single scattering term) and from our simulation are compared. Here we chose a refractive material ($n = 1.5$) with material coefficients $\sigma_a = 0.0041, \sigma_s = 2.62$. These are the values given in [17] for the green wavelength in marble. We chose this values to compare these two approaches because they are the same as Jensen used to compare his model to his simulation. Here, the two graph do not really match because the Fresnel factors are not included in the dipole model.

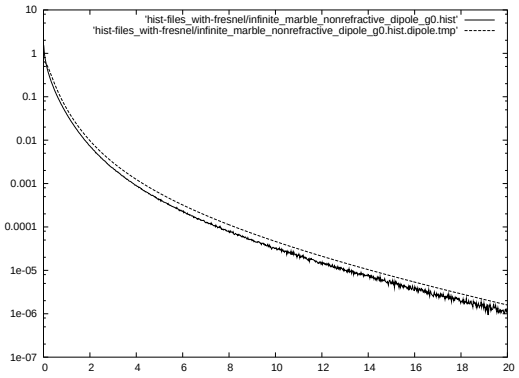


Figure 4.7: This figure shows the same situation as the figure on the left. This time, however, the refraction index has been set to 1 and thus, the Fresnel factors are all equal to one. As predicted, the graphs match a bit better this time but are still not equal due to the minimal possible energy introduced in the simulator. Thus, our results are always a bit smaller than the dipole approximation.

Chapter 5

Using the Dipole Model for Different Geometries

As discussed above, it is not clear whether the dipole approximation still yields acceptable results if one of the assumptions made is broken. Thus, we will compare the dipole model to simulations with different geometries in this chapter. The next chapter will deal with the problem of spatially varying materials. Here, as well as in the next chapter, all distances are given in millimeter, unless noted otherwise. Furthermore the asymmetry parameter g is set to 0, i.e. we assume that the scattering is isotropic.

The following geometries will be discussed: edges, cylinders, spheres and slabs. The details on how these geometries look like are discussed before the corresponding results in the following sections. Comparisons to the dipole model will be shown in Section 5.5.

For all simulations in this section we used the green wavelength of the material **skin2** ($\sigma_a = 0.07, \sigma_s = 1.59, n = 1.3$) from Jensen's paper [17] except in those for which an other material is given.

5.1 Edges

In the edge geometry, two half-planes share a common edge and between the two planes is a right angle. These two half-planes are the border of the material. In other words, the material is a infinite half space an infinite large part of which has been cut away. The cut is in the same direction than the (average) starting direction of the photons.

In our simulations, the results for this geometry is given as two images (see Section 4.1.1 on Page 43), one for each half-plane. The point where the photons enter the material is not on the edge but on one of the half-planes (compare Figure 5.1). Since the half-planes are

semi-infinite, the distance of this point from the edge is sufficient to uniquely characterize the point. Therefore, we may speak about an "edge with distance d", meaning an edge where the distance of the start point of the photons from the edge is equal to d. The half-plane on which the photons start will be called the **top plane**, while the other plane is called the **side plane**.

If the distance d increases, the result of the simulation will approach the result of the dipole model (with the limitations discussed above) since the geometry approaches the infinite half space.

To describe the whole simulation, the values of this distance d must be given, together with the refraction index of the material n, and the scattering and absorption coefficient σ_a, σ_s (or any other two of the four parameters $\sigma_a, \sigma_s, \sigma_t$ and Λ).

However, it is difficult so see anything specific on the images. Thus, only parts of the images will be further considered and shown as graphs (see Section 4.1.1, Page 44). They start at the photon start point and take a path on the cross section through the photon start point perpendicular to both planes.

The graph generated this way from an edge with distance $0.5mm$ is shown in Figure 5.2. This graph clearly points out that there are two points where the energy leaving the material abruptly rises. There are two different reasons for these increases.

The first increase can be noted at exactly the distance of the photon start point to the edge, which means that this boost happens because the path shown in Figure 5.1 changes the plane. The same jump can be clearly seen in Figure 5.3. This increase occurs because of the Fresnel equations. A photon hitting the side plane is more likely to be able to pass through the material border than a photon hitting the top plane (because a photon arriving near the edge must move towards the edge and is thus more likely to hit the side plane with the right angle to pass (i.e. nearly perpendicular) than the top plane).

The second jump, which can barely be seen for an edge with distance $2mm$ and not at all for an edge with distance $4mm$, will also occur in all other geometries in this chapter. It is due to photons that have not been scattered at all. In Figure 5.4, a part of the side plane corresponding to Figure 5.2 is shown. A bright spot at the place where the jump in the graph in Figure 5.2 is can clearly be seen. Figure 5.5 is an image for the same area. The only difference between the two images is that for Figure 5.5 only non-scattered photons have been considered. These figures clearly point out that the unscattered photons are responsible for this jump and therefore should be carefully considered in renderings.

5.2 Cylinders

The geometry used in this section is a infinitely long cylinder with a certain radius r. Once again, as r increases, this geometry approaches the infinite half space. The start point of

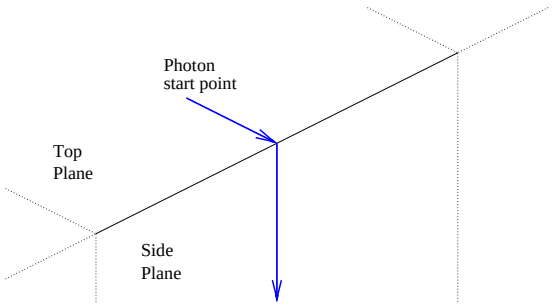


Figure 5.1: In this figure, the path along which the graphs (made from the two 2D-images obtained for a edge geometry) run is shown.

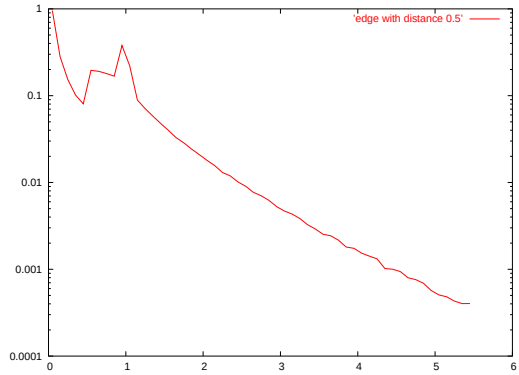
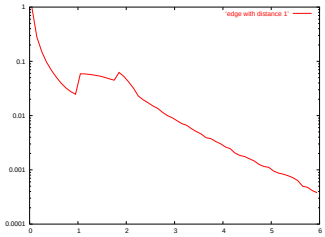
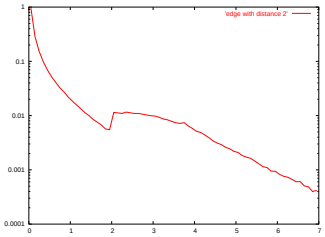


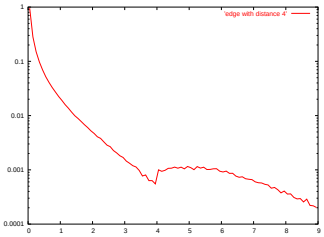
Figure 5.2: Here, a graph for an edge geometry with distance 0.5*mm* is shown. The two sudden increases in the amount of energy, which might be surprising at first, are explained in the text.



(a) distance 1*mm*



(b) distance 2*mm*

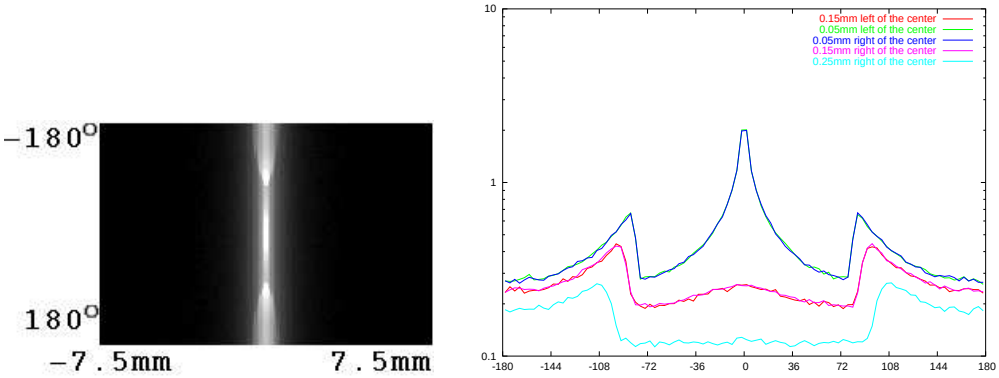


(c) distance 4*mm*

Figure 5.3: Here, more graphs for edge geometries with different distances (1, 2 and 4 millimeters) are given. They illustrate that both increases in the amount of energy noted in Figure 5.2 are reduced with increasing distance from the edge.

Figure 5.4: Here, a part of the image done with an edge geometry with distance 0.5*mm* is shown. The edge itself is along the upper border of the image. A bright spot is clearly visible near the center of the image.

Figure 5.5: This Figure shows the same situation as Figure 5.4. However, only photons that have not been scattered have been considered this time. Obviously the bright spot in Figure 5.4 does come from photons that have not been scattered.



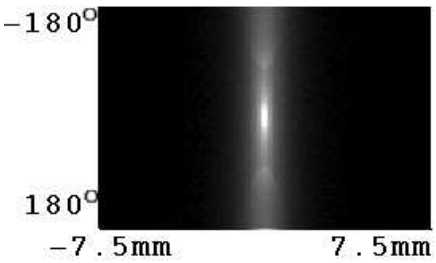
(a) Image for a cylinder with radius $0.5mm$

(b) A few graphs for a cylinder with radius $0.5mm$

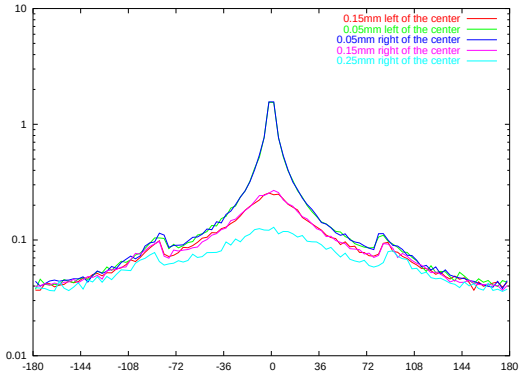
Figure 5.6: The left part of this figure shows an image for a slid-up cylinder with radius $0.5mm$, while the right part shows several graphs created from the left image. The graphs are parameterized by angle and show some central vertical lines through the image. To illustrate the noise from the simulation, two pairs of graphs (blue-green and red-magenta) that should be identical are shown. In both sub-figures, the non-scattered photons can clearly be seen.

the photons is of no importance since all points of a cylinder are indistinguishable. Thus, the value of r is sufficient to uniquely describe the geometry. Together with the scattering and the absorption coefficient(s) σ_a, σ_s and the refraction index n, this is sufficient to describe the complete simulation. The graphs showing vertical lines from images resulting from this geometry are parameterized by degrees.

The two-dimensional images from simulations with cylinder geometries are shown in such a way that the highest and lowest horizontal line of the image must be glued together to get the cylinder (compare Figure 5.6). As can be seen in that figure, the non-scattered photons can clearly be seen for small radii, but their influence is quickly reduced (see Figure 5.7) until these photons can not be noted at all (compare Figure 5.8). In these figures, only graphs along vertical lines are shown. In Figure 5.9 horizontal lines through the photon start point for several different radii are shown. As already mentioned, both the vertical and the horizontal lines will approach the graph for the infinite half space and thus the dipole model.

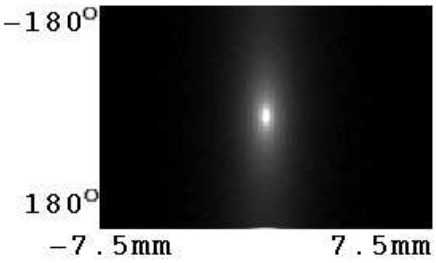


(a) Image for a cylinder with radius 1*mm*

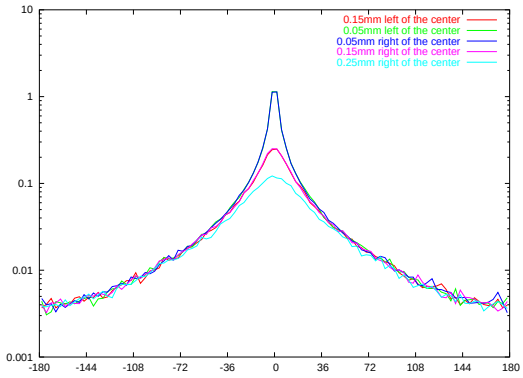


(b) A few graphs for a cylinder with radius 1*mm*

Figure 5.7: This figure shows the same situation as Figure 5.6 with the exception that the radius of the cylinder is 1*mm*.



(a) Image for a cylinder with radius 2*mm*



(b) A few graphs for a cylinder with radius 2*mm*

Figure 5.8: In this figure the situation described in Figure 5.6 is shown once again, this time with a cylinder of radius 2*mm*. The non-scattered photons are not noticeable any more.

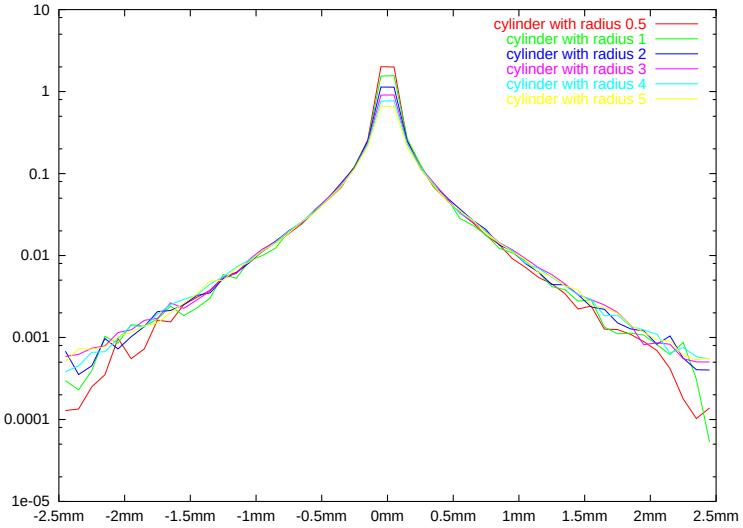


Figure 5.9: Instead of vertical lines, which have been shown in the Figures 5.6 to 5.8, horizontal lines through the photon start point for cylinders with different radii are shown here. The distance from the photon start point is given on the x-axis.

5.3 Spheres

Here, the geometry of the material through which the photons pass is a sphere with radius r. The start point of the photons is still not important (since all points of a sphere are indistinguishable anyways) and the geometry again approaches the infinite half space for $r \to \infty$. That's why, once again, knowing r, σ_a, σ_s and the refraction index n is sufficient to completely describe a simulation.

Again, the non-scattered photons have a great influence in the objects appearance, as seen in Figure 5.10. Moreover, this figure also shows the importance of the single scattering term. In Figure 5.11, graphs for spheres with different radii are plotted together with a graph for an infinite half space with the same material properties. As expected, the graphs of the sphere geometries approach the graph of the infinite half space. However, this happens very slowly. Note again that for the spheres, the distance (i.e. the x-values of points on the graphs) is measured in absolute distance here and is not the distance along the surface of the sphere. Using the distance on the surface of the sphere would results in better results as shown later.

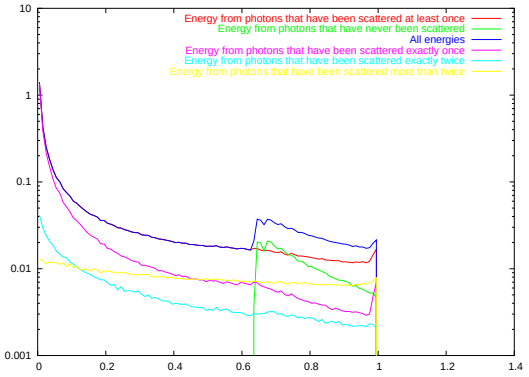


Figure 5.10: In this figure, the energies of the photons arriving for a sphere geometry with radius $0.5mm$ have been split up depending on the number of scatterings done. Once again it is clearly visible that photons that have not been scattered are responsible for the sudden increase in the energy at about $0.63mm$ in this figure.

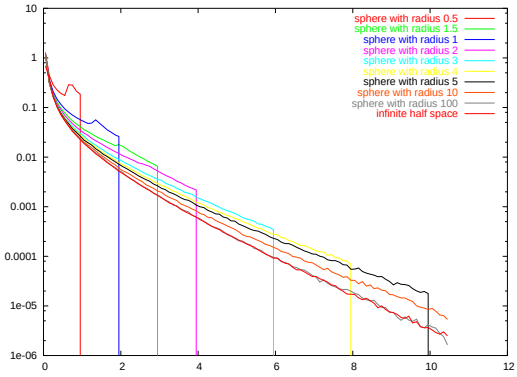


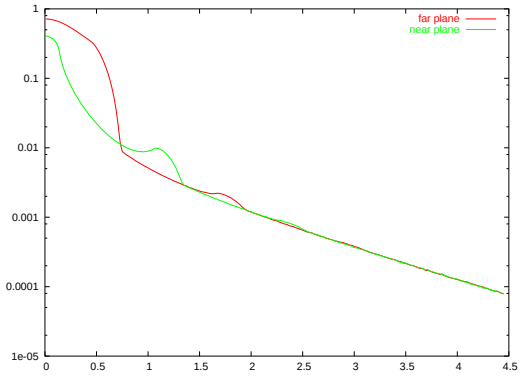
Figure 5.11: Here, the graphs from sphere geometries with different radii are shown. As expected, the graphs converge against the graph of the infinite half space. This leads to the assumption that, for the upper half spheres, the dipole model should be an acceptable approximation, especially for spheres with large radii.

5.4 Slabs

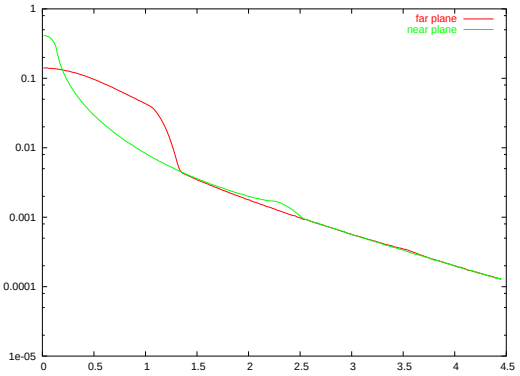
The last kind of geometry that we will present is a slab, i.e. the border of the material consists of two parallel infinite planes. If we speak about a slab with distance d, we mean that these planes have the distance d. As with all other geometries in this section, the limit geometry for $d \to \infty$ is the infinite half space.

The plane of which the photons start will be called the **near plane**, while the plane parallel to the near plane is called the **far plane**. The graphs for both planes with slab geometries and the skin1 material ($\sigma_a = 0.17, \sigma_s = 0.88$) with distances $0.5mm$ and $1mm$ can be seen in Figure 5.12. In the first figure (i.e. 5.12(a)) the non scattered photons can once again be seen (red line from approximately $0mm$ to $0.75mm$ in the first figure). The other visible jumps in the graphs (green graph in the area $0.95mm - 1.35mm$, red graph in $1.65mm - 1.95mm$ etc.) are also due to unscattered photons. This time, however, the photons have been reflected once ($0.95mm - 1.35mm$), twice($1.65mm - 1.95mm$) or more often.

These bumps are, however, especially important in slabs with a small distance. Figure 5.14 shows graphs for slabs with bigger distances. While the features can also (partly) be seen in these graphs, they are significantly smoothed there. The same is true for slab geometries with the *skin*2 material. This material has a higher scattering coefficient σ_s which results in far less photons that have not been scattered (see Figure 5.13). This figure



(a) Slab with distance 0.5*mm*



(b) Slab with distance 1*mm*

Figure 5.12: Both the graphs for the near and the far plane of a slab geometry (and the skin1 material) are shown in this figures. In the left figure the distance is 0.5*mm* while it is 1*mm* in the right one. Apart from the reflected non-scattered photons that are visible, the interesting thing noticeable in this figures is that the graphs can be considered as identical after about 2*mm*. Without the non-scattered photon, this would even happen earlier.

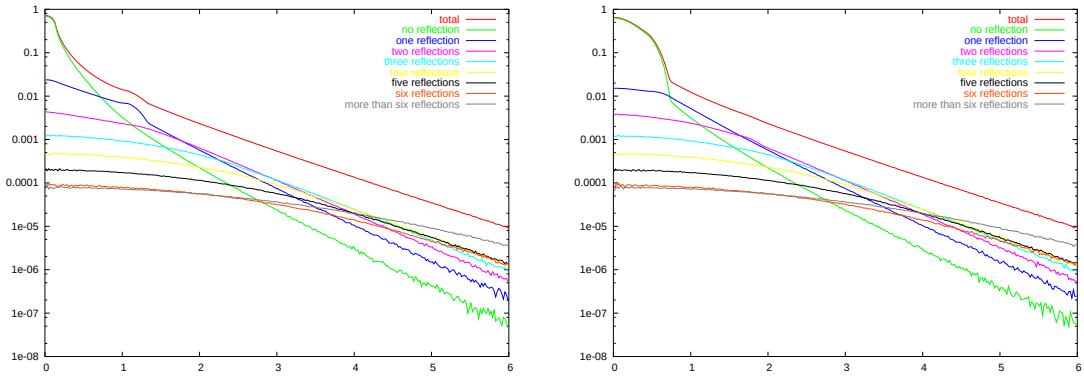
shows graphs for a similar situation (skin2 material), split up according to the number of reflections done. As one should expect, the influence of photons that are never reflected is the smaller the further away from the photon start point one is.

Figure 5.15 shows several near planes. It is not surprising that graphs for a geometry with a higher distance are closer to the graph of the infinite half space. However you should note how quickly this happens.

The next figure (5.16) shows the far planes. These will obviously converge against the zero function. This happens very quickly at distance 0*mm* (especially because less non-scattered photons arrive with increasing distance) but takes longer at greater distances.

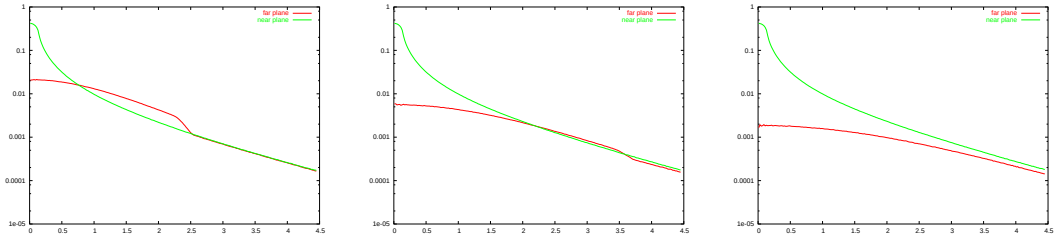
5.5 Comparing different Geometries to the Dipole Model

In this section we will compare the geometries discussed in the first four sections of this chapter to the dipole model by Jensen and to the simulation results from the infinite half space. As discussed in Section 4.2, the dipole model and the simulation results from the infinite half space are similar but not identical. This is basically due to the fact that the directions of the incident and outgoing light play an important role in Jensen's model (especially in the single scattering term). However, in our simulations the initial



(a) Near plane (b) Far plane

Figure 5.13: In this figure graphs for the near (left figure) and far (right figure) plane of a slab geometry with distance 0.5*mm* (with the skin2 material) are shown. The total energy (red graphs) is split up according to the number of reflections.



(a) Slab with distance 2*mm* (b) Slab with distance 3*mm* (c) Slab with distance 4*mm*

Figure 5.14: In this figure, the graphs for slab geometries (with the skin1 material) with distances 2*mm*, 3*mm* and 4*mm* are shown. As expected the influence of non-scattered photons is reduced, and the graphs are (nearly) identical much later.

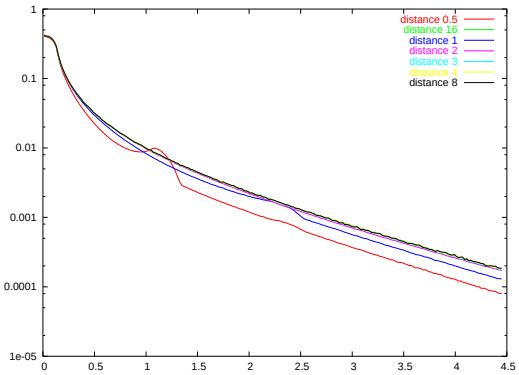


Figure 5.15: Only the graphs from the near planes (with the skin1 material) of different slab geometries are shown here. The graphs converge against the graph from the infinite half space quite quickly.

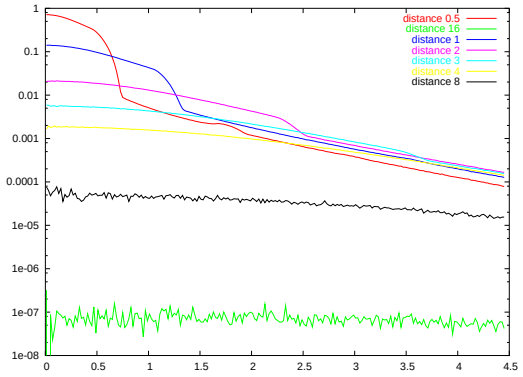


Figure 5.16: In contrast to the figure on the left, only the graphs on the far planes are shown here. The material is the same as in the figure on the left. These graphs are much more different than the graphs on the left as explained in the text.

direction was sampled uniformly and the direction of the light leaving the material was not considered at all (after applying Fresnel's Equations).

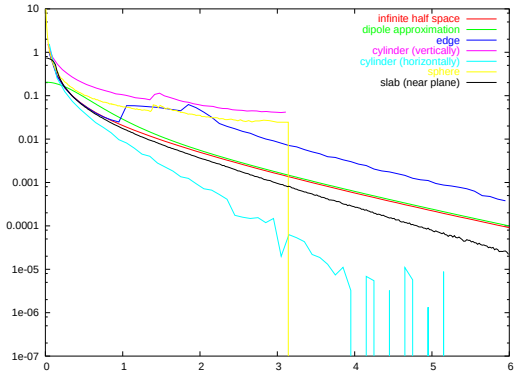
The Figures 5.17 and 5.18 show several graphs that have one thing in common: All graphs have been generated using the *skin*2 material. In Figure 5.19, the graphs for two different materials (skin2 and skin1) are compared.

Apart from the result of the infinite half space (red graphs) and the dipole model (green graphs) the graphs generated for geometries discussed earlier in this chapter are shown. That is, graphs for an edge geometry (blue), two graphs each from a cylinder geometry (magenta - vertically (i.e. along a circle), cyan - horizontally (i.e. along a straight line)), yellow graphs for spheres and graphs from the near plane of slab geometries (black) are plotted in these figures. In contrast to some graphs in the last sections, the x-coordinate is always the distance along the surface of the object[1]. An alternative would have been to use the Euclidean distance in $\mathbb{R}^3$. For all graphs where this is different to the distance on the surface, however, the results would be even further away from the dipole model.

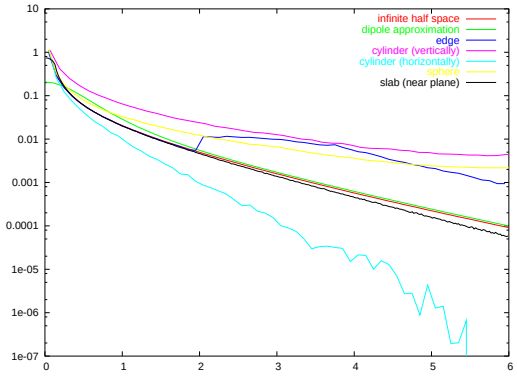
The graphs for the spheres and the vertical lines from the cylinder stop at $r \cdot \pi$ (where r is the radius of the sphere or cylinder respectively). This happens because this point is already the point most distant from the photon start point.

Although the horizontal line through the cylinders come from the simulation with the most photons, these lines are still very noisy due to the low amount of photons that arrive

[1] This was not always the case for the cylinder and sphere geometries discussed earlier.



(a) distance 1*mm*



(b) distance 2*mm*

Figure 5.17: The left figure show several graphs (for the infinite half space, edge, cylinder (twice) and sphere geometries) for geometries with radius 1*mm* (or, for the edge, with distance 1*mm*), together with the dipole approximation. In the right figure, the radius (or distance, respectively) is 2*mm*. For all graphs, the skin2 material has been used.

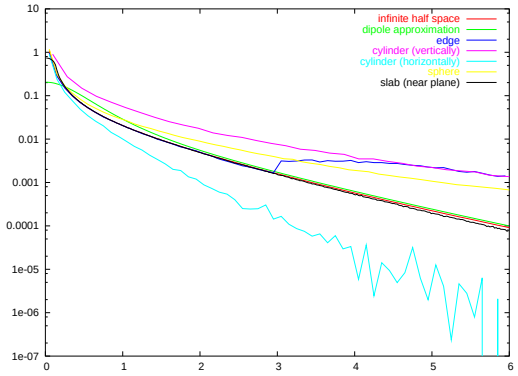
near this line. The situation when using the slab geometry is much better because we could gather photons in large areas.

5.5.1 Discussion

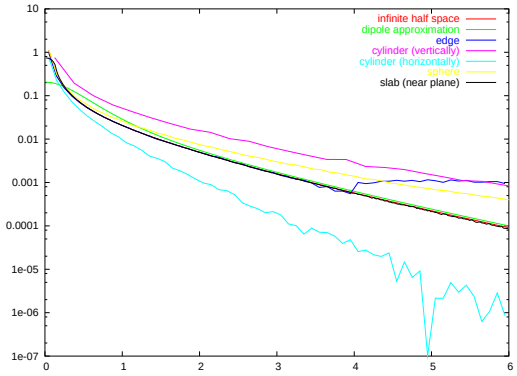
There are some things worth noticing while comparing these graphs:

- Non-scattered photons can have a great influence on the appearance of objects in which this phenomenon occurs, especially for short distances and if the attenuation coefficient is small.
- In the first hundred μm, the differences between the different geometries are small and can be neglected since single scattering is the most important effect.
- The slab geometry is close to the dipole model and the graph from the infinite half space, especially with larger distances.
- Cutting away parts of the material usually results in less energy arriving.
- The larger the angle β between the normal at the photon start point and the normal at the endpoint, the more energy arrives at this point[2].

[2]Of course both normals must either point towards or away from the object.



(a) distance 3*mm*



(b) distance 4*mm*

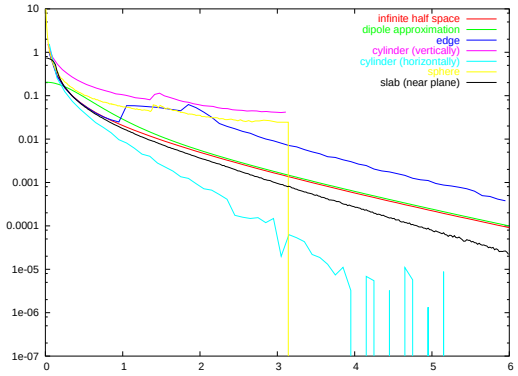
Figure 5.18: Here, the graphs for the same geometries and materials as in Figure 5.17 are shown, this time with a distance/radius of 3*mm* and 4*mm*.

While the first two points are rather obvious, the other might require more information.

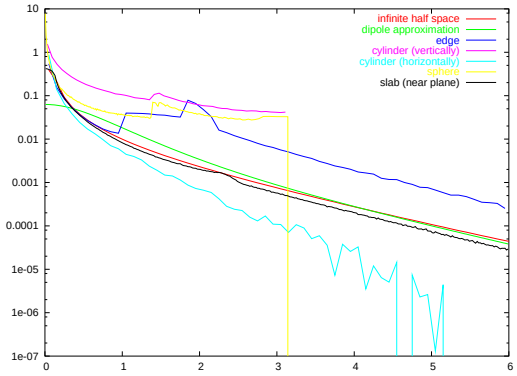
It might be surprising that the arriving energy in the slab geometry is close to the energy arriving in the infinite half space. This is due to the fact that photons arriving at the surface far away from the photon start point usually penetrate the object rather deeply (as seen in Figure 4.2). Since the slab geometry is not very deep, the photons must be reflected to be able to traverse that far (reflections are common in this geometry, as seen in Figure 5.13).

If a part of the material is cut away, the part of the photons that did pass through this part will arrive at a different point, often on the part cut away. Thus, less energy arrives at the "old" border of the material. This effect can be seen for the slabs and the "horizontal" lines of the cylinders the graphs of which are all lower than the one from the infinite half space.

The last point is especially interesting because it can be used to find out which graphs are higher than the dipole (or the infinite half space). The border of the material in the dipole model is a plane. Thus, the angle β is 0° which is the minimal possible value. For the sphere and ("vertical") cylinder geometries, β slowly changes from $\beta = 0^\circ$ to $\beta = 180^\circ$ during which the energy arriving constantly increases. The effect is even better visible for the edge geometries in which β abruptly changes at the edge. This sudden change of β also results in a significant increase in the amount of energy arriving. The reason why this angle is important for the amount of arriving energy is that it is more likely that the photons fly in the direction they started with than changing this direction. They therefore arrive more often (with the right angle such that they are not reflected) at areas with a large β.



(a) skin2



(b) skin1

Figure 5.19: The left part of this figure is exactly the same as Figure 5.17(a) while the right part shows the graphs for the same geometries with the skin1 material.

Summarizing the results so far, we can give the following hints:

- Include light that has been scattered once, especially for points near the points where light hits the object.
- Include light that has not been scattered, especially for points opposite and close to the places where light arrives.
- Use the distance along the surface of the object and not the distance in $\mathbb{R}^3$.
- All geometries except the slab geometry result in graphs that are quite different to the dipole model.
- Including the angle β might be a good idea. Further research is necessary to decide how this should be done, though.

Chapter 6

Spatially Varying Materials

In this chapter, another pre-requirement of the dipole model is broken: The material through which the photons are traveling is no longer uniform. That means that the scattering and absorption coefficient depend on the position of the photon. A typical case when this happens is when two different materials share a common border or in materials like marble whose appearances changes with position.

We start by showing some simulations where two different materials hit each other along a common border, and continue our study with a complex, generated marble material with veins going into different directions.

6.1 Discrete materials

The first spatially varying material we will examine consists of two "blocks" both of which have a uniform material. The first block has the shape of a slab (compare Section 5.4, Page 57) while the second is an infinite half space as used in the dipole model. We did simulations with several different thicknesses d of the first block. The resulting graphs[1] (Figures 6.1 and 6.2) are plotted together with the results from the simulations where $d = 0$ and $d = \infty$, i.e. with the two limit cases where the pre-requirements for the dipole model are given. As scattering and absorption coefficients, we used the green wavelength of the materials **skin1** ($\sigma_a = 0.17, \sigma_s = 0.88$) and **skin2** ($\sigma_a = 0.07, \sigma_s = 1.59$) from Jensen's paper [17]. Both materials had a refraction index of $n = 1.3$. Graphs for the same materials, this time with a refraction index of 1, are show in the Figures 6.3 and 6.4.

[1] Using a two-dimensional output mode was not necessary.

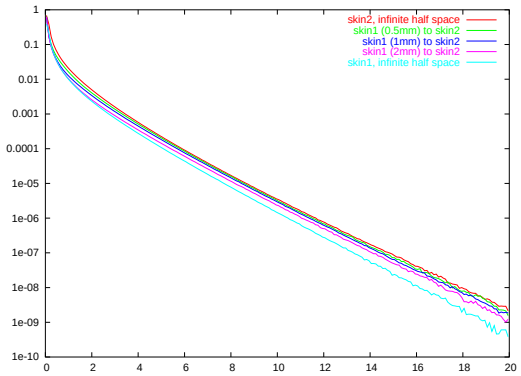


Figure 6.1: This graphs are the result of simulations where the first block has the "skin1" material and the second block the material "skin2". The graphs show the behavior one might first expect because the graphs for the spatially varying material is always between the two limit cases.

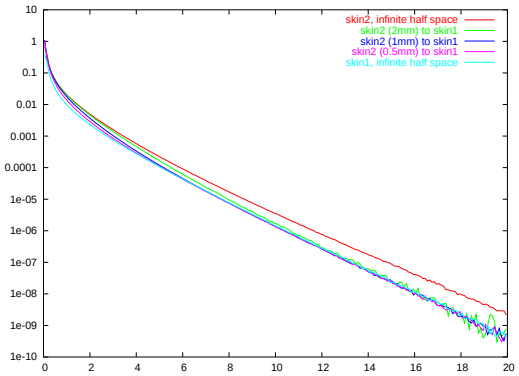


Figure 6.2: Basically the same situation as in the last figure is shown here. However, the first material is "skin2" while the second block has the "skin1" material this time. It is not clear if the graphs with the shift-variant materials stay between the graphs of the limit cases or if they pass the cyan (skin1) graph.

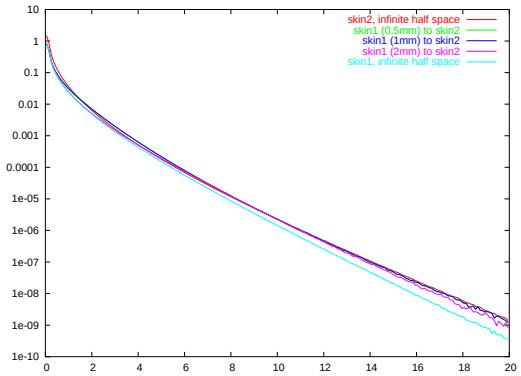


Figure 6.3: The geometry and the material parameter are the same as in Figure 6.1 except that the refraction index is 1 this time. Here it is clearly visible that the graphs of the inhomogeneous media do cross the graph for the infinite half space with the skin2 material.

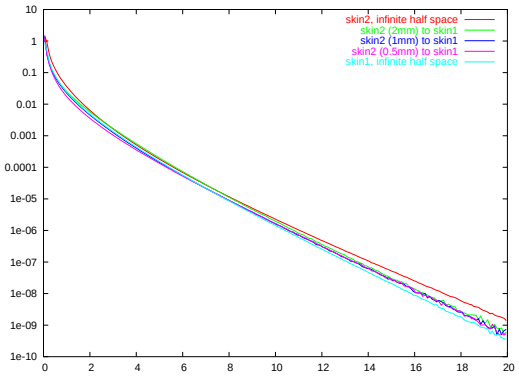


Figure 6.4: Again, the Fresnel Equations are ignored by setting the refraction index to 1. And once again, the graphs obtained from the spatially varying materials do not stay between the graphs of the limit cases.

The first thing one should note in these figures is that the graph of the mixed material does not always lie between the graphs of the limit case. On second though, this is not as surprising as one might think:

It has been shown that, the further away a photon arrives from the photon start point, the further it was usually inside the material ([4], compare also Figure 4.2(a) where the photon banana is shown.). However, the attenuation coefficient of skin1 is smaller than the attenuation coefficient of skin2. This basically means that it is easier for a photon to pass a few millimeters through the skin1 material than to do the same in the skin2 material. Thus more photons do get deeper inside the material (and thus arrive further form the photon start point) if they are first in a skin1 material. Similar, it is easier to return to the border if the last step goes through a material with a low attenuation coefficient.

Therefore the graphs of the "mixed" materials in Figure 6.3 have some higher values than the graph obtained from a skin2 material.

On the other hand these graphs are also higher than the graphs from a pure skin1 material. One reason for this is that more photons are scattered back towards the border in the skin2 material while more photons get lost in the depths of the skin1 material. Another is the higher absorption coefficient of the skin1 material. Similar reasons lead to the results in Figure 6.4. The Fresnel Equations, however, seem to wash out these effects.

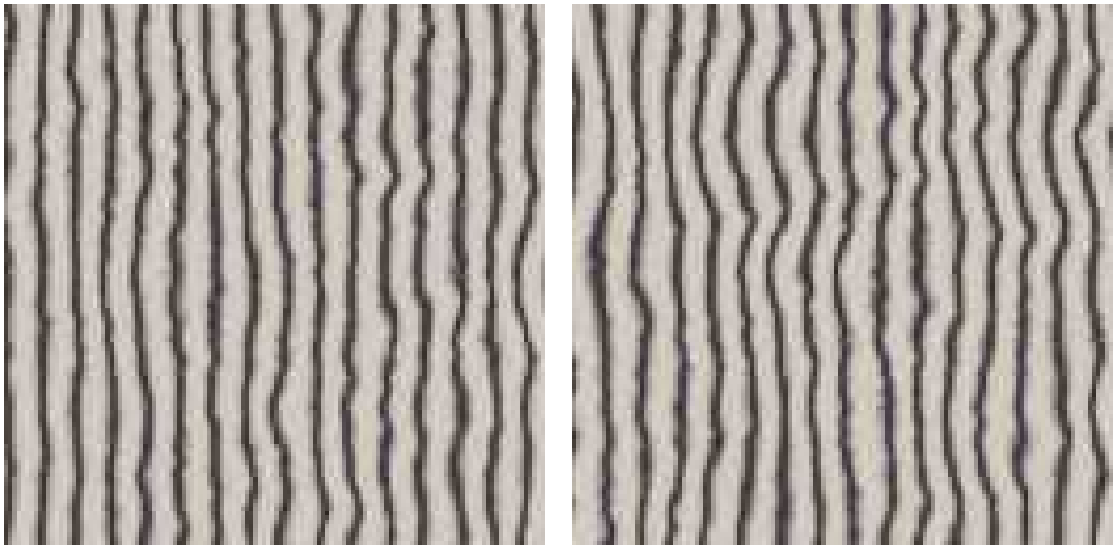
6.2 Continuous materials

Another case we researched was the case of continuous materials in the infinite half space which basically means that the material constantly changes its properties. More accurately, the materials we analyzed consisted of very small blocks inside which the material is constant. Thus the material was not really continuous. A simulation with a material that is really continuous would strongly lengthen the simulations as it would require to solve an integral (using Monte Carlo Integration) every time the photon changes its direction to decide how long the photon traverses before the next scattering event occurs. Therefore, we considered the benefit from using a really continuous material as too small.

Instead of storing the whole infinite half space, we stored only a finite cuboid in the simulations in this section. However, the size of the block was big enough that it virtually never happened that a photon hit one of the other borders of the block[2]. Thus the simulation results will be the same as when using an infinite half space of the same non-homogeneous material. [3].

[2] With one of the "other borders", we mean all borders except the one that is also part of the surface of the infinite half space.

[3] The minimal energy and the noise from the calculations will be much bigger since such a photon is very unlike to return to the surface and usually already has a low energy.



(a) A slice at the upper surface

(b) A slice trough the center of the marble

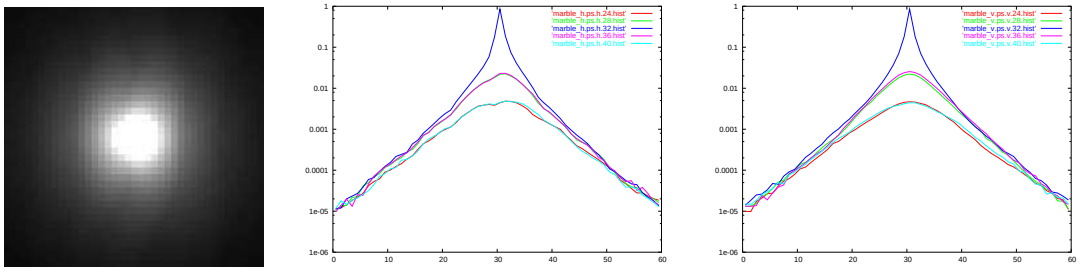
Figure 6.5: These two figures show two different slices of the generated marble block we used for the simulations described in this and the following section.

Figure 6.5 shows two slices through the cuboid we used at different heights. It is an automated generated marble function used by Lensch et al. in [18], although the dark veins were thinner there. We increased the thickness of the veins to increase their influence[4]. In this marble data, the absorption and scattering coefficients are interpolated between two extreme values. A sine, together with a Perlin noise [23], [24] is used to get the interpolation parameter. The two extreme values are $\sigma_a = 0.0041, \sigma_s = 2.62$ (which is the green wavelength of the marble from Jensen's paper [17] again) and $\sigma_a = 0.364, \sigma_s = 2.09$ as dark part. The refraction index was constant at $n = 1.5$.

If the veins are parallel to the surface (i.e. parallel to the infinite plane that is the border of the infinite half space) the effect is very small. This could be expected after the last section because the effect was even rather small for two discrete blocks, a situation where changes should be easier visible.

The situation is different when the veins are perpendicular to the material border: The position of some veins can be seen in the output of our Monte Carlo simulation, as seen in Figure 6.6. If such relatively small errors are not important, the dipole approximation can be used. Then the marble block has the appearance of a homogeneous material, of course. However, one should note that the error is small because the material was illuminated only

[4]The influence of the veins is still rather small.



(a) An image created with a shift-variant marble

(b) Horizontal lines through the left fi gure

(c) Vertical lines through the left fi gure

Figure 6.6: On the left one can see an image created from an “infinite half space” geometry (see text) with a generated marble material. The dark veins run vertically. This can be seen in the figure (which is gamma-corrected and amplified) if it is observed carefully. However it can be seen much clearer in the graphs in the middle. There, horizontal lines through the image on the left are shown. Comparing these to the vertical lines shown on the right, it becomes clear that the small bumps in the graphs in the middle are not noise or other simulation artifacts but are due to the veins which have a higher absorption coefficient.

at one point because this leads to a large range of brightness. This is not the case if the block is illuminated evenly. Better results might be achieved by other approaches like the one discussed in the next chapter.

Chapter 7

Fitting a Simple Model to Simulation Data

Finally, we will try to fit an easy model to match the appearance of a marble block, using 16641 simulations with at least four million photon each, i.e. with more than 66,4 billion photons total. The simulation time was about four months on a 3GHz Intel Xeon CPU with sufficient memory to store all data without swapping. The results of this simulations occupied about 780 MB of disk space after packing. In this chapter we will first describe the simulations done and explain which model will be fitted to the simulation results.

7.1 The Simulations and the Model

We made the following simulations: The object for which the model was fitted is a square on the border of the infinite half space. The material was the same marble as shown in Figure 6.5 and used in Section 6.2, which we will once more assume to occupy an infinite half space[1]. The veins where perpendicular to the surface of the infinite half space. Simulations has been done for the following 129×129 light source positions:

$$L := \left\{ \left(\left(\frac{i}{4} - 16 \right) mm, \left(\frac{j}{4} - 16 \right) mm \right) | i, j \in \{0, 1, 2 \ldots 128\} \right\}$$

That is, the photon start points were spread on a grid in the rectangle with corner points $(-16mm, -16mm)$ and $(16mm, 16mm)$. The measures of the marble block were $64mm \times 64mm \times 32mm$ where $32mm$ is the depth of the marble block as seen from the surface of the infinite half space.

[1] See Section 6.2 for the limitations on this approach.

The model we wanted to fit was:

$$\begin{aligned} & L \times L \to \mathbb{R}, \\ & (m,n,m',n') \mapsto \alpha'(m',n') \left(c_1(m,n) e^{d_1(m,n)x} + c_2(m,n) e^{d_2(m,n)x} \right) \\ & \text{with } x = \sqrt{(m-m')^2 + (n-n')^2} \end{aligned}$$

That is, for every pair of start (m,n) and end point (m',n') we try to model the energy traveling from (m,n) to (m',n') using the sum of two scaled exponential functions (parameterized by the distance between start and end point), which depend only on the start point, and a multiplicative factor that depends only on the end point. Thus we fit 83.205 parameters to 276.922.881 results. The idea behind this model is that it is close to the dipole model but has an additional parameter $\alpha'(p)$ to include the local properties near the point p. That is, the parameters c_1, c_2, d_1, d_2 model global features while α' is responsible for local effects.

This model might lead to feasible results since the dipole model itself basically consists of the sum of two exponential functions. However, exact results can not be expected because these functions are only parameterized by the distance, which is not sufficient for exact results: More energy traverses in the same direction as the veins.

The fitting is done with the following algorithm:

1. Choose two intervals $n := [min1, max1]$ and $f := [min2, max2]$ with $min1 < min2$, $max1 < max2$.
2. Initialize all α'.
3. Estimate all first pairs of parameters (c_1, d_1) using only data from points with a distance inside the interval n.
4. Estimate all second pairs of parameters (c_2, d_2) using only data from points with a distance inside the interval f.
5. Estimate a trade-off factor and multiply c_1 and c_2 with it.
6. Estimate all α'.
7. Continue at step 3 until the fit is considered sufficient.

The estimations in the steps $3-6$ are all done using least square approximations.

In step 1, the intervals n (for near) and f (for far) are chosen. This is done by the user of the fitting program. Usually $min1 = 0$ is chosen and $min2$ is smaller than $max1$. These

intervals have a rather big influence on the results of the fitting. More details about this parameters are given in the next section.

Step 2 is merely an initialization necessary because the α' are needed in step 3 and 4. We estimated the α' by checking the amount of energy that returned to the photon start point. The α' of the smallest energy was set to 0.5, the α' of the largest to 1.5. For all other α' linear interpolation was used. However, setting all α' to a constant value also works.

In step 3 and 4 we do the following for each photon start point $p = (m,n)$: Store all points whose distance to the photon start point is inside the given interval in a set A. The parameters c_1, d_1 (or c_2, d_2 respectively) are then chosen in such a way that the sum of the squared errors along the y-axis is minimal in logarithmic space[2]. That is, the function

$$\begin{aligned}&\sum_{i\in A}\left(\left(\log\left(\alpha'(i)\cdot c(p)e^{d(p)|p-i|}\right)\right)-\log\left(s(p,i)\right)\right)^2\\ =&\sum_{i\in A}\left(\left(\log(\alpha'(i))+\log(c(p))+d(p)|p-i|\right)-\log\left(s(p,i)\right)\right)^2\end{aligned}$$

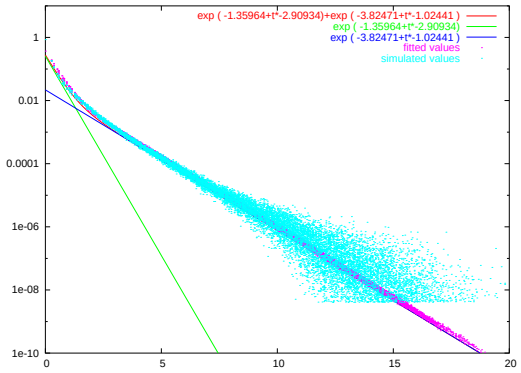
is minimized where $s(p,i)$ is the energy that traverses from p to i according to the simulations, $|p-i|$ is the Euclidean distance between p and i and $\alpha'(i)$ is the α' that belongs to the point i. This least square fitting requires solving a 2×2 linear system. We used determinants to do so and checked if the solution is correct afterwards. Up to now, this was always the case. Implementing a numerically stable way to solve the system was therefore not necessary.

Step 5 is necessary because we are fitting the pairs (c_1,d_1) and (c_2,d_2) independently and therefore, the sum of these two exponential functions will be greater than the simulated data in the means (because the sum of these two positive functions is always greater than one of these functions, and only one of the functions is used in the fit). We collect all photon start points whose distance is less than or equal to $max2$ in the set A and minimize the $t(p)$ in the function:

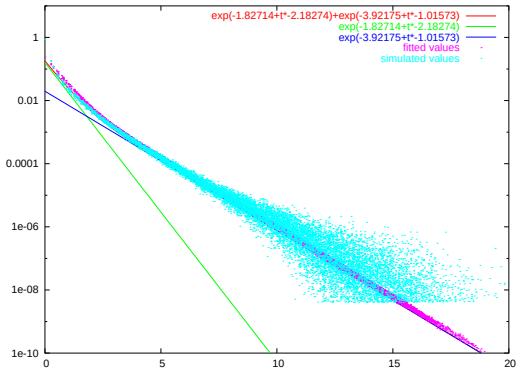
$$\sum_{i\in A}\left(\left(\log(t(p))+\log(\alpha'(i))+\log\left(c_1(p)e^{d_1(p)|p-i|}+c_2(p)e^{d_2(p)|p-i|}\right)\right)-\log\left(s(p,i)\right)\right)^2$$

This factor is not included in the model because it can be included into c_1 and c_2, i.e. after computing t, we multiply c_1 and c_2 with t and can ignore t afterwards (because it is always one outside this step).

[2]These steps are done in logarithmic space because it is hard to fit the parameters otherwise. In logarithmic space, however, the part of the model we try to fit is linear for which computing the least square fit is rather easy.



(a) near: 0mm - 1mm, far: 0.5mm - 10mm



(b) near: 0mm - 1.5mm, far: 1mm - 10mm

Figure 7.1: These two figures show the result of the fit of the central point of the marble block for different near and far intervals. The red points are the results from fitting while the blue points come from a (still too noisy) simulation. Furthermore, both exponential functions and their sum are shown in both images.

In step 6 we do basically the same as in step 5, but with a few modifications. This time the following is done for each endpoint i: Collect all photon start points whose distance is less than or equal to $max2$ in a set A and minimize the $\alpha'(p)$ in:

$$\sum_{i \in A} \left(\left(\log(\alpha'(p)) + \log\left(c_1(p) e^{d_1(p)|p-i|} + c_2(p) e^{d_2(p)|p-i|} \right) \right) - \log(s(i,p)) \right)^2$$

Note that this time, p is the point at which the photons arrived, not the point at which the light started as in the steps above. This step is done in the logarithmic space, too, because the points with a large energy would have a very large influence otherwise.

7.2 Results

This section will present some results of the fitting steps described in the last section. Figure 7.1 shows the fits for two points of such a fit with $n = [0mm, 1mm]$, $f = [0.5mm, 10mm]$ and $n = [0mm, 1.5mm]$, $f = [1mm, 10mm]$, respectively. The model which we fitted against the simulation results basically consists of the scaled sum of two lines in logarithmic space. These two lines are also shown in Figure 7.1 (a green line from the near interval and a blue line from the far interval), together with their sum (red graph). The near interval should be chosen in such a way that the slope of the green line approximately matches the "slope of the simulated data" in the near interval.

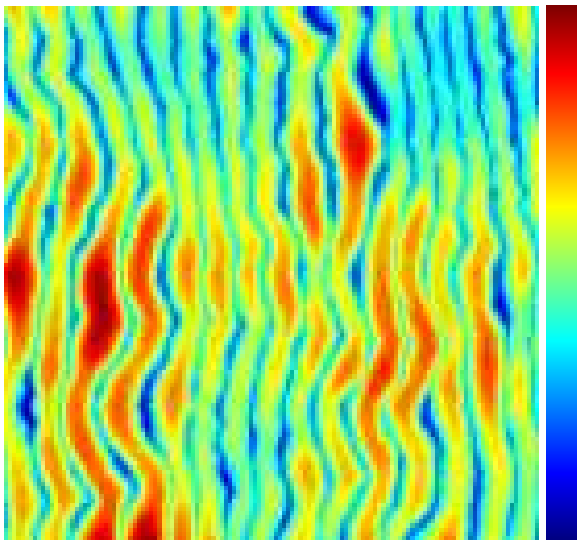


Figure 7.2: The left figure shows a false-color plot of all the α' from the same fit as in Figure 7.1(a). The right figure shows the color gradient used. In this figure, the minimal value (i.e. the dark blue points) corresponds to an α' of 0.840189 while the maximal value (dark red points) is at $\alpha' = 1.60504$.

The results of the fitting are not bad, except of the very first point that is barely visible in the figure. The energy arriving at this point (according to the simulation) is always underestimated in the fit. This should be a deficiency in the model and not in the fitting step. With increasing distance, however, the simulated data gets very noisy making a good fit practically impossible. Another deficit of the model is that the light traveling in the bright part of the material parallel to the veins is less reduced than the light passing several veins, which cannot currently be modeled by our model.

Furthermore, it can be seen that the points from the simulation are never smaller than a certain value. This is due to the fact that the minimal energy possibly arriving at a certain point (in the simulation) is given by the minimal energy possible in the simulation divided by the number of photons traced, multiplied by the area of the sampling region. This limit of the Monte Carlo Simulation has been reached in the images. This is one of the reasons why the far interval should have an upper bound.

All α' fitted for $n = [0mm, 1mm]$, $f = [0.5mm, 10mm]$ are shown in Figure 7.2. Even in this image, the veins are clearly visible. The same is true for the figures in which the parameters from the near interval c_1, d_1 (Figure 7.3) and the parameter c_2 from the far interval (Figure 7.4) are shown. However, this is not the case for the d_2 parameters which rather seems to model a more global behavior of the material.

Moreover, the parameter d_2 is nearly constant; the ratio between the maximal and the minimal value is approximately 1.09, but the same ratio for the parameter d_1 is 1.29. The

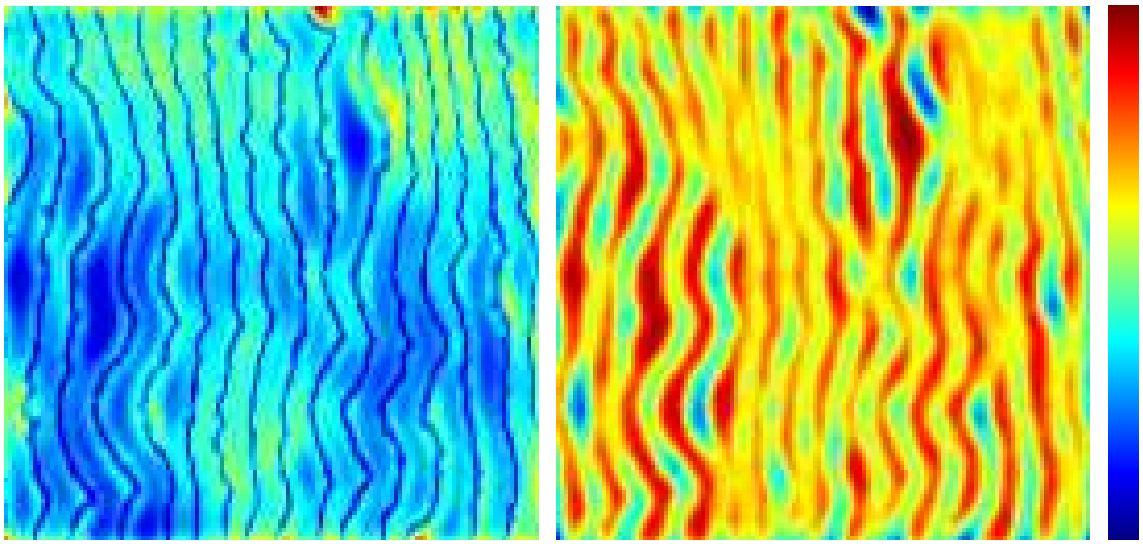


Figure 7.3: In this figures, the coefficients c_1 (ranging from 0.200354 to 0.367901, left image) and d_1 (ranging from -3.42545 to -2.64922, right image) are shown. As mentioned these parameters come (mainly) from the near interval.

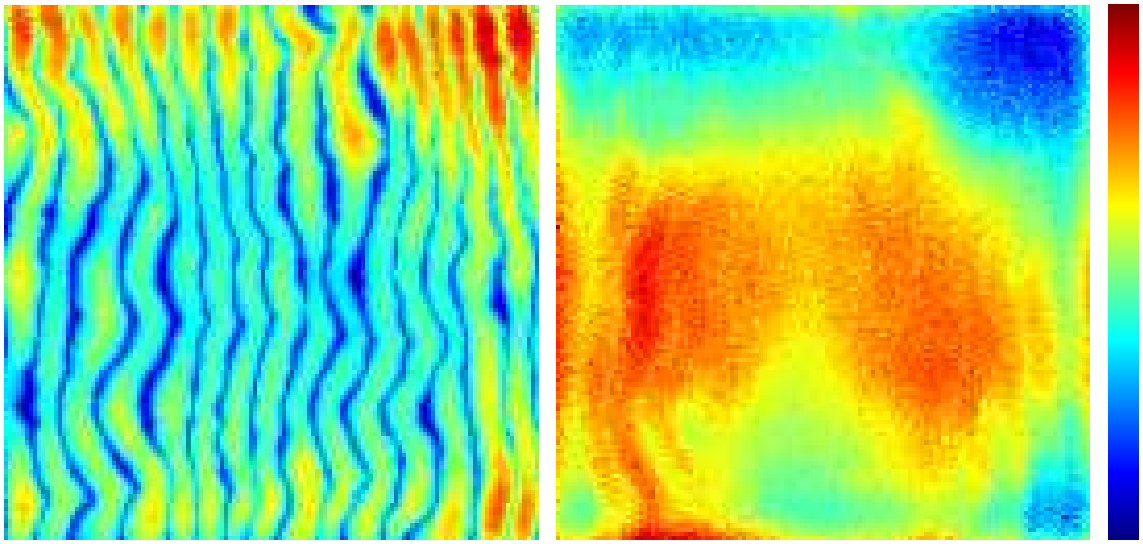


Figure 7.4: This figure shows the coefficients c_2 (left) and d_2 (right) from the same fit as in the last figures. The ranges of the images are from 0.0150702 to 0.0317442 (c_2) and from -1.09233 to -0.999242 (d_2), respectively.

other parameters have a ratio near 2.

Chapter 8

Discussion, Conclusion and Future work

We summarize the results we obtained, point out deficiencies of our model and the simulation and give an outlook on future work.

8.1 Evaluation of the Dipole Approximation

We discussed the differences between the dipole approximation and the simulation results from four different types of geometries, pointing out for which geometry the greatest differences to the dipole approximation occur. However, it is not clear yet if this differences are big enough to be noted by a human observer. Still, we proposed a few hints how to improve the dipole approximation in certain situations.

Furthermore we have clearly shown that the single scattering term that is neglected in many current rendering systems is of great importance. Even more important is that unscattered photons can create visible bright spots in certain geometries. These bright spots should be far more important for thin objects than the missing single scattering term because the single scattering term only amplifies the brightest region while the unscattered photons create a new feature. Still, the dipole model is not capable to create these new highlights in its current form. Adding a "no scattering term" to the dipole approximation is easy (for homogeneous materials) using Beer's law and should improve the appearance of translucent objects. For non-homogeneous materials the no scattering term will require to solve an integral. Once again, however, we will have to include our simulation results into a rendering system to verify the assumption that these spots are important.

As already mentioned, it might also be reasonable to include a factor in the dipole approximation that is based on the angle β (which is the angle between the normal at the

photon start point and the normal at the endpoint). The simulations we did clearly show that abrupt changes in β results in sudden changes in the amount of energy leaving the material. The nature of this factor, especially its dependency to the refraction index n, might be the topic of future work.

Another goal of future work might be to estimate the graphs of more complex geometries based on the results obtained so far and compare them to real results.

We illustrated that the energies arriving in discrete shift-variant materials are not always in between the energies for the corresponding shift-invariant materials. However, the graphs of discrete shift-variant materials seems to be smooth. Thus, these differences should be of minor importance. However, only very few simulations have been done, yet. More such simulations might enable us to derive a model that can predict the behavior of discrete inhomogeneous materials.

The quality of the dipole approximation if the material is neither shift-invariant nor has the form of an infinite half space is another point we did not consider yet. It might also be interesting to consider objects in which the refraction ratio changes inside the object.

8.2 Fitting of the Simple Model

Moreover, we proposed a simple model for a continuous inhomogeneous material and fitted the model to a generated marble block. Although the results from the simulation were quite noisy, the results of the fit are promising. More different materials must be considered to verify this approach, though, as well as rendering the marble with this model as validation.

It might be possible to further improve the model by exchanging the parameter α', which currently depends on the endpoint by a parameter that is dependent on the endpoint and the direction from the photon start point to the endpoint. For materials which have veins (such as the marble block we used) this should make the difference from the model to the simulation results smaller. However, other materials might not benefit from this change at all.

The last chapter also suggests that setting the parameter d_2 to a constant value might not noticeably reduce the quality of the model. However, it would reduce the number of parameters by 20%.

8.3 Conclusion

The dipole model is a frequently used approximation to the real behavior of translucent objects. We investigated its accuracy compared to Monte Carlo simulations for a variety

of setups and showed its strengths and weaknesses. Considering the visual importance of subsurface scattering, we believe that future work in this area is promising and will help to achieve improved results with respect to both rendering costs and visual realism.

Bibliography

[1] Max Born and Emil Wolf. *Principles of Optics*. Cambridge University Press, 1999.

[2] Julie Dorsey, Alan Edelman, Henrik Wann Jensen, Justin Legakis, and Hans Køhling Pedersen. Modeling and rendering of weathered stone. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques / SIGGRAPH*, pages 225–234. ACM Press/Addison-Wesley Publishing Co., 1999.

[3] Philip Dutré, Philippe Bekaert, and Kavita Bala. *Advanced Global Illumination*.

[4] Shechao Feng, Fanan Zeng, and Britton Chance. Monte Carlo Simulation of Photon Migration Path Distribution in Multiple Scattering Media. *SPIE*, 1888:78–89, 1993.

[5] Michael Goesele. *New Acquisition Techniques for Real Objects and Light Sources in Computer Graphics*. PhD thesis, Max-Planck-Institut für Informatik, 2004.

[6] Michael Goesele, Hendrik P. A. Lensch, Jochen Lang, Christian Fuchs, and Hans-Peter Seidel. DISCO: Acquisition of Translucent Objects. *ACM Trans. Graph. / SIGGRAPH*, 23(3):835–844, 2004.

[7] Paul S. Heckbert. *Derivation of Refraction Formulas*. Academic Press, 1989. Also available at www.cs.cmu.edu/∼ph/refr.ps.gz.

[8] Paul S. Heckbert and Pat Hanrahan. Beam tracing polygonal objects. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques / SIGGRAPH*, pages 119–127. ACM Press, 1984.

[9] Louis G. Henyey and Jesse L. Greenstein. Diffuse Radiation in the Galaxy. *Astrophys. Journal*, 93:70–83, 1941.

[10] http://en.wikipedia.org/wiki/cantor_function. Wikipedia, the free encyclopedia, last visited September 09, 2004.

[11] http://en.wikipedia.org/wiki/feynman. Wikipedia, the free encyclopedia, last visited September 09, 2004.

[12] http://en.wikipedia.org/wiki/polarization. Wikipedia, the free encyclopedia, last visited September 09, 2004.

[13] http://en.wikipedia.org/wiki/wave-particle_duality. Wikipedia, the free encyclopedia, last visited September 09, 2004.

[14] Akira Ishimaru. *Wave Propagation and Scattering in Random Media*, volume 1 and 2. Academic Press, New York, 1978.

[15] Hendrik Wann Jensen. *Realistic Image synthesis using photon mapping*. A.K. Peters, 2001.

[16] Hendrik Wann Jensen et al. State of the Art in Monte Carlo Ray Tracing for Realistic Image Synthesis. In *Proceedings SIGGRAPH*. ACM Press / SIGGRAPH Course Notes, 2001.

[17] Henrik Wann Jensen, Stephen R. Marschner, Marc Levoy, and Pat Hanrahan. A practical model for subsurface light transport. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques / SIGGRAPH*, pages 511–518. ACM Press, 2001.

[18] Hendrik P. A. Lensch, Michael Goesele, Philippe Bekaert, Jan Kautz, Marcus A. Magnor, Jochen Lang, and Hans-Peter Seidel. Interactive rendering of translucent objects. In *Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*, page 214. IEEE Computer Society, 2002.

[19] Don P. Mitchell. Spectrally optimal sampling for distribution ray tracing. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques / SIGGRAPH*, pages 157–164. ACM Press, 1991.

[20] Tomas Möller and Ben Trumbore. Fast, Minimum Storage Ray-Triangle Intersection. *Journal of graphics tools*, 2(1):21–28, 1997.

[21] F. E. Nicodemus, J. C. Richmond, J. J. Hsia, I. W. Ginsberg, and T. Limberis. *Geometric Considerations and Nomenclature for Reflectance*. Monograph 161. National Bureau of Standards (US), oct 1977.

[22] Optimizing fresnel reflection code. http://tigrazone.narod.ru/raytrace1.htm.

[23] Ken Perlin. A Unified Texture/Reflectance Model, Course on Advanced Image Synthesis. volume 18. ACM Press / SIGGRAPH Course Notes.

[24] Ken Perlin. Improving noise. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 681–682. ACM Press / SIGGRAPH, 2002.

[25] Matt Pharr and Pat Hanrahan. Monte carlo evaluation of non-linear scattering equations for subsurface reflection. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques / SIGGRAPH*, pages 75–84. ACM Press/Addison-Wesley Publishing Co., 2000.

[26] Paul Rademacher. *GLUI - A GLUT-Based User Interface Library*, 2.0 edition, June 1999.

[27] Jos Stam. An illumination model for a skin layer bounded by rough surfaces. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pages 39–52. Springer-Verlag, 2001.

[28] Eric W. Weisstein. Definition of phase function in eric weisstein's world of physics, wolfram research. http://scienceworld.wolfram.com/physics/PhaseFunction.html, last visited September 09, 2004.

[29] Turner Whitted. An improved illumination model for shaded display. *Commun. ACM*, 23(6):343–349, 1980.

List of Figures

Index