A Ray Tracing Method for Illumination Calculation in Diffuse-Specular Scenes

Peter Shirley
Department of Computer Science
University of Illinois
1304 West Springfield Avenue
Urbana, Illinois 61801
USA

Abstract

Several ways of improving the realism of the results of traditional ray tracing are presented. The essential physical quantities of spectral radiant power and spectral radiance and their use in lighting calculations are discussed. Global illumination terms are derived by employing illumination ray tracing for calculation of quickly changing indirect lighting components, and radiosity ray tracing for slowly changing indirect lighting components. Direct lighting is calculated during the viewing phase allowing the use of bump maps. Finally, a method is introduced that reduces the total number of shadow rays to no more than the total number of viewing rays for a given picture.

Keywords: Bump Mapping, Illumination, Radiosity, Radiance, Ray Tracing, Realism, Stratified Sampling, Texture Mapping.

1 Introduction

The quest for accurate lighting models in computer graphics has taken two very different approaches in the 1980s. The first approach is based on ray tracing (point sampling) techniques made popular by Turner Whitted's landmark 1980 paper [Whi80]. The second approach, radiosity methods, is based on zonal techniques borrowed from the heat transfer literature [GTG84]. The results produced by both techniques suffer from discretization errors; ray tracing uses discrete sample points, and zonal methods partition the environment into finite patches.

Recent work has combined ray tracing and zonal techniques [WCG87, SP89]. This combination allows the zonal methods to be used where they perform best, and ray tracing to be used to handle mirrors, where the zonal methods break down. Both Rushmeier and Ward et al. have separated direct and indirect lighting to avoid high levels of discretization associated with pure zonal methods [Rus88, WRC88].

This paper extends the idea of the separate calculation of direct and indirect lighting to the calculation of hard and soft lighting, where hard lighting contains detail and soft lighting changes slowly. This is similar to the idea of patch and element substructuring of Radiosity [CGIB86], but also allows fine detail caused by hard indirect lighting such as the lighting that produces caustics. This is achieved with a three pass method that synthesizes conventional ray tracing, illumination ray tracing, and radiosity. The hard (detail producing) light paths are calculated using illumination ray tracing and an efficient method of view ray tracing, and the soft (slowly changing) indirect diffuse lighting is calculated with radiosity. Most of the discussion is limited to scenes containing only diffuse and specular surfaces, but extensions for imperfect specular surfaces are outlined.

2 Background

Commonly graphics programs calculate direct lighting analytically, and then add an empirically generated ambient term. If we do physically based global lighting calculations, we must be more careful with units because lighting components that are calculated separately must not be improperly scaled relative to each other. In this section some essential terms and formulae are presented that aid in maintaining correct proportionality in lighting calculations.

Since we are interested in the amount of light hitting a surface or film plane during a set time period, radiant flux (also called radiant power), the radiant energy...
per unit time, \( \Phi \), is often used. The quantity most often used in graphics is the \( \text{radiance} \), \( L \). Radiance shares many characteristics of the perceptual measure luminance; it gives an indication of surface brightness, dependent upon neither the size of the object being viewed, nor the distance to the viewer.

Diffuse surfaces can be characterized by a simple equation:

\[
L_\lambda = \frac{R_\lambda \Phi_\lambda}{\pi A}
\]

where \( L_\lambda \) is the \textit{spectral radiance} of the diffuse surface, \( R_\lambda \) is the reflectance of the surface at wavelength \( \lambda \), \( \Phi_\lambda \) is the \textit{spectral radiant power} incident on the surface from all directions, and \( A \) is the area of the surface.

From this definition, the direct lighting component due to a planar diffuse light source is:

\[
L_1 = \frac{R_1 L_1^1 \cos \theta_1 \cos \theta_2 A^1}{\pi d^2}
\]

where \( L_1^1 \) is the spectral radiance of the light source, \( \theta_1 \) is the angle the direction to the light source makes with the surface normal, \( \theta_2 \) is the angle the direction from the light source makes with the light source normal, \( A^1 \) is the area of the light, and \( d \) is the distance to the light.

The light that shades a surface can be divided into two types: \textit{hard}, and \textit{soft}. Soft lighting is the highly diffused light that is usually approximated with an ambient term. Hard light has quickly changing components that 'paint' detail onto surfaces, such as the hard edge of a shadow or caustic.

Soft lighting occurs whenever light is bounced off a diffuse reflector. Any fine geometric features of the light will be lost when diffuse reflection occurs. A mirror does not eliminate detail in the light it reflects, it merely attenuates it. A curved mirror or clear object can even increase detail in the light. This detail will not be evident to the viewer until the light hits a diffuse surface and 'paint' its pattern on the surface.

This distinction between hard and soft lighting is taken advantage of in the next section to make the rendering algorithm faster.

### 3 Algorithm

Because of the fine structure of hard light, it must be calculated accurately. For light to produce detail on a surface, its path from the light source cannot have included a diffuse reflector. When a diffuse reflector is encountered detail is lost and one enters the realm of soft lighting. Conventional ray tracing will produce the most important details: shadows on diffuse surfaces, and reflections of shadows. The detail producing light paths not accounted for by conventional ray tracing are those that hit some number of specular surfaces before a diffuse surface. An example of this effect is the fine patterns seen on a surface that is painted by light that has just passed through a bottle or reflected off a wristwatch. This shortcoming can be dealt with by tracing illumination rays from the light source and storing ray hits on diffuse objects in illumination tables.

The soft light missing from the ray tracing calculations can be estimated by radiosity. Since this radiosity solution is used for only soft lighting, a very coarse environment partitioning can be used. Usually, radiosity solutions account for all components of the lighting, so care must be taken to remove the detailed components from the radiosity calculation as described in Section 3.2.

The three phases of the algorithm are described in order of their execution. First the illumination ray step and some optimizations is presented in Section 3.1. Second, the radiosity pass is described in Section 3.2. Finally, the viewing stage, along with an optimization for shadow calculation, are presented in Section 3.3.

#### 3.1 Illumination Ray Tracing

The first pass of the algorithm calculates lighting that hits diffuse surfaces after hitting at least one specular surface. This type of lighting is calculated using Arvo's method of simulating light paths by sending illumination rays from the light source.

In Arvo's method, each light source emits some large number of rays. Each ray represents a fraction of the outgoing energy that will be transported to other surfaces. Arvo accomplishes this by defining a standard texture map on each diffuse surface. When an energy bundle hits a surface, the energy is divided between the four texture nodes that surround it. To convert the energy maps to radiance maps the area represented...
Figure 1: Two dimensional illumination ray tracing with a specular sphere and a diffuse plane. Top: First pass feeler rays hit the specular sphere in two of the zones. Bottom: Dense energy carrying rays are sent in the two zones where the sphere was hit, and in the two zones adjacent to the hit. Energy rays that hit the ground plane directly are ignored, while those hitting the sphere first are tabulated in the ground plane's illumination map.

by each texture node must be estimated. When this phase is completed, the lighting will be stored as radiance maps on the diffuse reflectors. Direct lighting is omitted from the maps (by not recording direct hits by energy rays) and calculated in the conventional method during the viewing phase. This calculation is view-independent.

Because direct lighting is calculated in the viewing phase, illumination rays need only be sent toward specular surfaces (which produce indirect lighting not handled by radiosity). To find the specular surfaces where illumination rays should be sent, a set of 'feeler' rays is sent in all directions from the source, and then a dense set of illumination rays is sent in directions where the feeler rays encounter specular surfaces. Since the edges of specular objects should not be ignored by illumination rays, illumination rays are generated in zones where feeler rays hit specular objects, as well in adjacent zones (Figure 1).

Stratified sampling [Coo86] is used to choose both feeler and illumination ray directions. The light sources are assumed to be diffuse and the hemisphere of directions is partitioned into zones in a diffuse distribution. This partitioning is achieved by mapping a partitioned unit square to the hemisphere [WRC88]. The top of Figure 2 shows an initial partition into 36 zones, and a further subdivision of three of these zones. This illustrates the initial partitioning for feeler rays, and a finer partitioning for the actual illumination rays. The bottom of Figure 2 shows how this idea can be used to choose points on a disk [3].

The illumination rays can generate fine patterns such as those inside the metal ring shown in Figure 3.

If there are imperfect mirrors (such as brushed steel) in the scene, illumination rays should be reflected probabilistically. This is similar to the reflection of viewing rays in distributed ray tracing [CPC84], and produces noise in the illumination maps which can only be reduced by sending more illumination rays. Figure 4 shows four plates of variable smoothness and the reflection they cast on a diffuse surface. The same number of rays are shot at each plate. The level of noise on the diffuse surfaces increases with the spread of the reflection functions of the plates. To reduce the noise, a greater number of illumination rays could be sent, or the resolution of the texture map on the diffuse surface could be decreased.

For an area light source, rays should be sent in from all points on the light source. If this is not done, the lighting map will not have soft shadows or 'soft' caustics. To reduce noise in the illumination map, rays can be sent exclusively from the center of the light source. This is similar to the way radiosity hemicubes are fixed to a certain point. If direct lighting is calculated in the viewing phase we can still produce soft shadows, but these soft shadows will be mixed with sharp caustics. For soft caustics, either the light must be subdivided or the illumination rays must originate at random points on the source. Either method will require more rays to avoid aliasing or noise.

3.2 Radiosity for Ray Tracing

The second step of the algorithm uses radiosity to calculate the soft lighting. This lighting involves at least two diffuse reflectors in all light paths. The modified radiosity algorithm used in this paper differs from the standard radiosity method because it omits direct lighting.

Radiosity calculations can be done by ray tracing in a very simple manner. Rays can be sent between all pairs

[3] The partitioning of the disk is required for lens effects of distributed ray tracing. The partitioning of the disk on the right is preferred because the zones have better locality. This is the central idea of designing partitioning schemes.
of surfaces [BFP83, NN85, WEH89], or can be sent in a cosine distribution from each surface toward all unobscured surfaces [MBG86, Mal88, SP89]. The second technique has the advantage of allowing indirect transport of specular surfaces, so it is the method of choice for scenes containing mirrors. Recently, progressive refinement techniques have been used to implicitly solve the linear equations of radiance [CCWG88, SP89]. This approach avoids the \( O(n^3) \) storage requirement of explicit techniques.

The progressive refinement method can be thought of as straightforward transport simulation. Though we want to find the radiances of all surfaces in an environment, it is easier to formulate the problem in terms of power, and convert to radiance once the solution is obtained. Suppose we are given the emitted power, \( \Phi_i \), and reflectance, \( R_i \), of each surface, \( i \), in the environment, and we wish to find the total power, \( \Phi_i \), coming from each surface. If the rays are sent in random directions rather than at specific patches, and there are \( n_r \) diffuse surfaces, this yields:

\[
\text{for } i = 1 \text{ to } n_r, \quad \Phi_{i \text{ source}} = \Phi_i = \Phi_i \\
\text{while (not converged)} \quad \text{choose surface } s_{\text{source}} \text{ to emit unsent power} \\
\text{choose } n_r \text{ rays } r_{j} \text{ to send} \\
\text{for } j = 1 \text{ to } n_r, \quad \text{find surface } s_{\text{hit}} \text{ hit by } r_{j} \\
\Phi_{\text{hit}} = \Phi_{\text{hit}} + (R_{\text{hit}} \Phi_{\text{source}})/n_r \\
\Phi_{\text{unsent}} = \Phi_{\text{unsent}} + (R_{\text{hit}} \Phi_{\text{source}})/n_r \\
\Phi_{\text{source}} = 0
\]

If mirrors are added to the scene, light can bounce directly off a mirror back to the source patch, so the algorithm becomes:

\[
\text{for } i = 1 \text{ to } n_r, \quad \Phi_{i \text{ source}} = \Phi_i = \Phi_i \\
\text{while (not converged)} \quad \text{choose surface } s_{\text{source}} \text{ to emit unsent power} \\
\text{choose } n_r \text{ rays } r_{j} \text{ to send} \\
\Phi = \Phi_{\text{source}} \\
\Phi_{\text{source}} = 0 \\
\text{for } j = 1 \text{ to } n_r, \quad k = 1.0 \\
\text{while (not done)} \quad \text{find surface } s_{\text{hit}} \text{ hit by } r_{j} \\
\text{if } s_{\text{hit}} \text{ is diffuse} \quad \Phi_{\text{hit}} = \Phi_{\text{hit}} + k R_{\text{hit}} \Phi / n_r \\
\Phi_{\text{unsent}} = \Phi_{\text{unsent}} + k R_{\text{hit}} \Phi / n_r \\
\text{done} = \text{True} \\
\text{else} \quad k = k R_{\text{hit}} \\
\text{reflect } r_{j} \text{ off surface } s_{\text{hit}}
\]

The \( \Phi_i \) can then be converted to radiances using Equation 1. Accuracy can be improved by making the number of rays \( n_r \), proportional to the energy being represented by the particular set of rays. The directions are chosen in a stratified pattern as is done with the illumination feeler rays in Section 3.1.

Since this solution is to be combined with the standard viewing and illumination ray tracing solutions, the direct and direct-specular transport components must be removed. This is done by shooting from each light source in turn and then setting \( \Phi_i \) to zero. Finally, the patch radiances must be transported to patch vertices as described in [CG85].

### 3.3 Viewing and Direct Lighting

Direct lighting is calculated in the final, viewing, stage. The viewing rays operate as in a standard distributed ray tracer [Co086], except that the radiosity and illumination map values are substituted for the conventional ambient term. The only change made in this algorithm is in the treatment of shadow rays, described in the remainder of this section.

One problem with Whitted-style ray tracing is that every viewing ray that hits a diffuse surface generates a shadow ray for every light in the environment. This is computationally expensive if the number of light sources is large. The shadow ray intersection test is unusual in that it does not matter which object is hit by the ray. This observation led to the use of the light-buffer, which optimized shadow testing for point light sources [HG86]. Another way to optimize shadow testing is to reduce the total number of shadow rays as suggested by Kajiya [Kaj86]. This second method is employed in this paper because it preserves the ability to produce soft shadows, and keeps the algorithm simple.

In a realistic ray tracer many viewing rays are sent through each pixel, so a less accurate direct lighting estimate for each viewing ray is acceptable. The number of shadow rays determines the accuracy of the direct lighting component. As a first approximation, there should be about as many shadow rays per pixel as there are viewing rays per pixel. The number of viewing rays for a pixel must be sufficient to provide an adequate statistical estimate of whether the pixel contains an edge between a bright and a dark region. The shadow rays for a particular pixel must provide an adequate statistical estimate for a region partially in shadow and partially not in shadow. Since the shadowed regions will be the dark values for the image, and the unshadowed regions will be the bright values for the image, the shadow

---

The light sources themselves are the brightest parts of the image. This can safely be ignored because the brightness
rays are averaging values similar to those of the viewing rays. This means each viewing ray that eventually reaches a diffuse surface should generate approximately one shadow ray.

In traditional distributed ray tracing, shadow rays are sent to random points on each light source\cite{CPC84}. Instead one ray can be sent to a random point chosen from all the light source surfaces. First a target selection probability is assigned to each source. This probability is proportional to the energy coming from the light source to the intersection point as calculated by Equation 2. This ensures that rays are likely to be sent to the nearest and brightest light sources. Once the candidate source is chosen, a ray is sent to a random point on the candidate light. If the ray hits an obstruction no direct lighting component is added. Otherwise it is assumed to be illuminated by all sources. For example, suppose there are ten lights of equal brightness. If there are twenty viewing rays through the pixel that views the point, then about two rays will be sent to each light. If the point can ‘see’ four of the lights, about eight rays will not be blocked and each of these will return the lighting for all ten lights. The other twelve shadow rays will be blocked and return zero radiance. The average of the samples will be four times the ten light contributions divided by the twenty samples, or four light contributions, as expected.

The random values used to generate the shadow rays are stratified using uncorrelated jittering as described by Cook\cite{Co086}.

4 Results

The algorithm of Section 3 was implemented in C++ on an Encore Multimax. The primitives used are polygonal meshes. Each mesh made of a diffuse material has an associated illumination map with user set resolution. When a viewing ray hits a polygon, the direct lighting is calculated, a value is retrieved from the illumination map of the parent mesh, and soft radiosity lighting is interpolated from the vertices of the polygon. The direct, illumination map, and radiosity components are then summed and multiplied by the surface reflectance to give the total radiance value seen along the ray. Because the illumination maps are associated with the meshes, and radiosity values are stored at polygon vertices, the resolution of the illumination maps is independent of the level of mesh polygonalization.

If the mesh polygonalization or the illumination map of the light sources is usually clamped to a value near the brightest non-light because of the limited dynamic range of display devices.

<table>
<thead>
<tr>
<th>fig.</th>
<th>total rays</th>
<th>view rays</th>
<th>shad. rays</th>
<th>illum. rays</th>
<th>radios. rays</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>26,482,766</td>
<td>61%</td>
<td>39%</td>
<td>0.2%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>15,924,534</td>
<td>55%</td>
<td>45%</td>
<td>0.3%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>52,842,224</td>
<td>74%</td>
<td>25%</td>
<td>0.9% 0.2%</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Ray counts for figures

resolution is too coarse, Mach banding appears on diffuse surfaces. Higher resolutions are needed when hard lighting is important. This is more true for illumination maps than for mesh polygonalization because the radiosity values approximate the slowly changing soft lighting components. Very fine illumination maps are usually needed only when light is focused, such as in the caustic in the ring of Figure 3.

Since most of the work done in ray tracing code is expended in ray intersection tests, ray counts are more informative than machine/implementation dependent timings. Ray totals and the proportion of work done for viewing, shadowing, illumination ray tracing, and radiosity for figures 5, 6, and 7 are shown in Table 1. These three figures have 1024 by 768 resolution.

Figure 5 shows a room with nine lights. It was traced with sixteen viewing rays per pixel. Illumination ray tracing was not used for this figure. Because the radiosity calculations were done on a coarse grid (32 patches for each wall and each table panel), very little computation was spent on this stage (see Table 1). Fewer shadow rays than viewing rays were required for this picture because there is no direct lighting on the ceiling, so none of the viewing rays hitting the ceiling generated shadow rays.

Figure 6 shows a room with one light. A bump map has been added to the walls to simulate the depressions in the mortar between the blocks\cite{Bli78}. This figure illustrates both the strengths and weaknesses of bump mapping; the back wall looks fairly good, while the left wall is close enough to reveal the painted-on quality of the bump map. The sharp shadows coming from the chair and table legs, usually the bane of radiosity, are achieved easily because the direct lighting is calculated in the viewing stage.

Figure 7 shows a room with two lights and a bottle and a glass. Illumination rays are responsible for the caustics on the table. The banding in the shadow of the bottle results from the inadequate polygonal representation of the bottle; shadows of clear objects reveal imperfections not evident in the appearances of the objects themselves. The large number of viewing rays required for this picture results from the deep viewing ray trees for primary rays that hit the glass objects.
Figure 8 shows six simple rooms. On the top left is a ray traced picture with no ambient lighting. On the top middle is the same picture with an added ambient term (Whitted-style algorithm). On the top right the ambient term is replaced by a radiosity calculation. On the bottom left illumination ray tracing has added a caustic under the ball. On the bottom middle the light source has been shrunk causing sharper shadows. Since the shadows are generated in the viewing stage, the higher frequencies do not lengthen the radiosity calculations. On the bottom right the glass ball has been replaced by a second diffuse block.

5 Conclusion

The ray tracing algorithm presented in this paper is useful for producing pictures of diffuse/specular environments. The detail producing lighting components are calculated using viewing and illumination rays. The lower frequency diffuse lighting is calculated by a radiosity method. This separation follows the spirit of sub-structuring in conventional radiosity[CGIB86]. Direct lighting calculations allow the use of bump maps and are minimized by limiting the total number of shadow rays.

The conversion of a standard ray tracer to follow this algorithm is fairly easy. The ray tracing and texture map tools needed for the illumination ray tracer already exist in most ray tracing codes. The radiosity calculations are straightforward except for the interpolation to patch vertices. These calculations can be done on polygonal or higher order patches. Since the radiosity calculations model slowly changing lighting components, a very coarse discretisation can be used.

The most important conclusion of this paper is that all ray tracers that could benefit from global illumination effects should include them. This is demonstrated by the small relative performance cost of global illumination calculations as shown in Table 1. Since the execution time of the global illumination calculations is independent of image resolution, and the viewing time is proportional to image resolution, global illumination calculations are particularly inexpensive for large images.

One problem with the algorithm is that the illumination mapping requires a coordinate system to be applied to all diffuse surfaces. The radiosity calculations require decomposition of the environment into patches. The resolution of the illumination maps and size of radiosity patches has an effect on image run time and quality. These parameters are not selected automatically, so some experience is needed to use the algorithm.

Possible extensions to the algorithm include the addition of an isotropically scattering medium[RT87]. The medium could be broken into elements and treated in the radiosity step. Dispersion could also be added as done by Thomas[Tho86].

6 Acknowledgements

Thanks to Bill Kubitz (advisor for this project), Holly Rushmeier, Jean Buckley, Allan Tuchman, Greg Rogers, Kelvin Sung, and Bill Brown for helpful comments on the paper. Thanks also to the reviewers for their frank and detailed suggestions.

References


Figure 2: Mapping from partitioned square to hemisphere and circle.

Figure 3: Caustic inside ring simulated with illumination rays.
Figure 4: Reflected light from metal plates of different roughness.

Figure 5: Room with nine lights, where only one shadow ray is sent for each viewing ray. Sixteen viewing rays are shot for each pixel.

Figure 6: Room with bump mapped walls. Separate calculation of direct lighting allows directional shading on the bump maps.

Figure 7: Wine glass and bottle with hard indirect lighting in the shadows of the bottle and glass calculated using illumination ray tracing.

Figure 8: Importance of lighting components.

Graphics Interface '90