# Radiosity and Hybrid Methods

LÁSZLÓ NEUMANN and ATTILA NEUMANN
Budapest

We examine various solutions to the global illumination problem, based on an exact mathematical analysis of the rendering equation. In addition to introducing efficient radiosity algorithms, we present a uniform approach to reformulate all of the basic radiosity equations used so far. Using hybrid methods we are able to analyze possible combinations of the view-dependent ray-tracing method and of the low-resolution radiosity-based method, and to offer new algorithms.

Categories and Subject Descriptors: I.3.3 [**Computer Graphics**]: Picture/Image Generation—*display algorithms*; I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism

General Term: Algorithms

Additional Key Words and Phrases: Complete two-pass method, conjugated gradient method, convergence criteria, coupling method, distributed ray tracing, double-patch method, non-diffuse ambient term, photosimulation, radiosity method, rendering equation, residual image, separable reflectance, Southwell algorithm

## 1. INTRODUCTION

The goal of this paper is to present efficient algorithms based on an exact mathematical analysis of the computer synthesis of *photorealistic images*. The *radiosity method*, which gives a complete but low-detail solution, is examined first, and then possible combinations with the view-dependent *distributed ray-tracing method* are analyzed. These *hybrid methods* combine the advantages of radiosity and ray tracing, and they are the most promising family of methods to solve the global illumination problem.

The radiosity method was introduced for diffuse systems based on analogous systems for heat transfer by Goral et al. [1984] and then was further developed by Cohen and Greenberg [1985], Cohen et al. [1986, 1988] and by Nishita and Nakamae [1985] to lead to practical solutions for complex systems. Extensions of the radiosity methods to include plane mirrors and refractive surfaces were discussed by Rushmeier and Torrance [1990]. Radiosity equations for environments including non-diffuse reflectors were con-

Authors' address: Maros u. 36, H-1122 Budapest, Hungary.

sidered by Immel et al. [1986] and by Neumann and Neumann [1989], while practical algorithms were proposed by Shao et al. [1988] and by Neumann and Neumann [1990]. A promising method based on *hierarchical radiosity* was introduced by Hanrahan and Salzman [1990].

The distributed ray-tracing method was introduced by Cook et al. [1984], and was further developed by Cook [1986], by Lee et al. [1985], and by Purgathofer [1987]. An effective variance reduction method is due to Kajiya [1986]. The first example of two-pass approaches combining the advantages of radiosity and ray tracing with diffuse and specular reflection appeared in a paper by Wallace et al. [1987], and the same approach was further developed by Sillion and Puech [1989].

The mathematical foundations of this work, from *radiometry*, were given in our previous paper [Neumann and Neumann 1989]. Section 2 of the present paper, mostly theoretical, explains important basic concepts (*albedo, efficiency*) and, as a result, presents the first proof of the convergence of the Neumann series for the *rendering equation in power norm*. In Section 3 the basic radiosity equations are considered and reformulated. The concept of *radiance*, or *radiosity* forms, as well as *power* forms are discussed, and different solutions are indicated. These two different forms have sometimes been confused in the literature. An important result of this paper is to introduce the *zero-order illumination component*, permitting the simultaneous treatment of *emissive* effects and *point light sources* in radiosity-based methods. All of the scenes in previous images from the literature are rather similar since only diffuse (Lambertian) sources were allowed. This restriction even applies to the non-diffuse systems used by Shao et al. [1988], as well as the two-pass methods. On the other hand, images generated by ray tracing generally have sharp shadow boundaries. The separate handling of the zero-order component without interreflection presented here allows the effective merging of the two types of effects. Another result is the identification of the *sorted shooting method* for diffuse environments with the *Southwell relaxation algorithm* [Krekó 1976]. This leads to a more comprehensive approach that permits the same formulation to handle the case of diffuse, separable, and general reflectance equations. We then introduce the *double-patch method* to obtain an efficient solution for non-diffuse environments. Because of its flexibility, this actually constitutes a family of methods. Section 3 then considers a particular new category of problems for *very bright environments*, where a version of the *conjugate gradient method* is showed to be an efficient solution. Section 4 considers the generalization of the *ambient term* for non-diffuse systems and its suitability to help integrate the various radiosity-based methods. Section 5 addresses the scope of hybrid methods, and we survey the possibilities of coupling ray-tracing and radiosity methods based on the rendering equation and on the fundamental concepts of *image of order K* and *residual image of order K*. After discussing the *direct residual image method* and the *coupling method*, we present a *complete two-pass method* with diffuse and specular decomposition that allows arbitrary illumination and that includes all of the possible diffuse/specular permutations of light transport.

## 2. THEORETICAL FOUNDATIONS

We will recapitulate concisely the formulations of light transport and the rendering equation for interreflection. The main result is the first exact analysis of the criteria for convergence. The concepts and symbols introduced in this section are used throughout the paper.

### 2.1 Bidirectional Response

The local reflecting behavior of surfaces can be described by their *bidirectional reflectance* function. When surfaces are partly translucent, one should also use a *bidirectional transmittance* function. Reflectance and/or transmittance are jointly described by means of the *bidirectional response*, whose definition, which is based on radiometry, was given by Neumann and Neumann [1989, 1990]. Let the unit vectors $\mathbf{V}$ and $\mathbf{L}$ point from a surface element toward the viewer and the light source. The bidirectional response $\rho$ is a scalar-valued function with two vector variables:

$$\rho(\mathbf{L},\mathbf{V}) \geq 0, \qquad \mathbf{L} \in F, \quad \mathbf{V} \in F, \tag{2.1}$$

where $F$ is the set of unit vectors whose origin is at the surface element. For opaque surfaces, the appropriate range for reflection is the half-space $H$, instead of $F$. The units of $\rho$ are $[sr^{-1}]$.

Bidirectional response has two important features in the case of linear, passive optical material. One is the essential symmetry derived from Helmholtz's *principle of reciprocity* (see Chandrasekhar 1960):

$$\rho(\mathbf{L},\mathbf{V}) = \rho(\mathbf{V},\mathbf{L}). \tag{2.2}$$

According to this principle, the value of $\rho$ remains unchanged upon exchanging the directions of the viewer and the light source, a fact put to use when determining $\rho$ experimentally. The other feature comes from energy conservation, which requires that the total energy reflected or transmitted is less than or equal to the incident energy. The difference, if any, is the energy dissipated or absorbed by the material. The proportion of nondissipated energy will be noted $a(\mathbf{L})$:

$$a(\mathbf{L}) = \int_{\mathbf{V}\in F} \rho(\mathbf{L},\mathbf{V})|\mathbf{NV}|\, d\omega_{\mathbf{v}}. \tag{2.3}$$

$\mathbf{N}$ is the unit vector normal to the surface element; hence, $|\mathbf{NV}| = \cos \Theta_{\mathbf{V}}$, that is, the cosine of the angle of incidence. The absolute value is needed since there is no stipulation on the direction of $\mathbf{N}$. $a(\mathbf{L})$ is the *albedo* for the direction $\mathbf{L}$. In its classic meaning, this term is used for diffuse materials where $a(\mathbf{L}) = a = \pi \cdot \rho$ and is constant irrespective of direction $\mathbf{L}$. For reflecting material, the albedo can also be called the *reflectivity*. With these notations one can state the conservation of energy as $a(\mathbf{L}) \leq 1$ for every $L \in F$.

The concept of *average albedo* can now be introduced. It is useful when it is required to express with one scalar the fraction of total power not absorbed by the surface, whether diffuse or non-diffuse. This fraction obviously de-

pends on the *spatial distribution* of the power of the incident light. Let us assume an illumination uniformly distributed in *radiance*[1] (intensity) in any direction of the (half-) space. Such an illumination allows us to define the average albedo $\bar{a}$ as a ratio between total input and output powers:

$$\bar{a} = \frac{1}{\pi} \int_{\mathbf{L} \in F} a(\mathbf{L}) \cdot |\mathbf{NL}| \, d\omega_L, \tag{2.4}$$

where $a(\mathbf{L})$ is the albedo function from eq. (2.3). We used reciprocity (eq. (2.2)) to deduce eq. (2.4).

### 2.2 Point and Distributed Light Sources

Assume a point source of radiant power $P$ at distance $r$ from a surface element $dA$. The output radiance in direction $\mathbf{V}$, assuming no absorbing medium, is given by

$$S^{out}(\mathbf{V}) = \frac{P}{4\pi r^2} \rho(\mathbf{L}, \mathbf{V}) |\mathbf{NL}|. \tag{2.5}$$

Surfaces within the scene may have emissivity of their own. For these *emitting* surfaces, we denote their radiance function as $S_E(\mathbf{V})$. The combined effect of the emission of the surface and of the reflected/transmitted light from the point light sources will be called the *zero-order* illumination component:

$$S_0^{out}(\mathbf{V}) = S_E(\mathbf{V}) + \sum_{i=1}^{L} \frac{P_i}{4\pi r_i^2} \rho(\mathbf{L}_i, \mathbf{V}) |\mathbf{NL}_i| h(\mathbf{L}_i), \tag{2.6}$$

where the factor $h(\mathbf{L}_i)$ with $0 \le h(\mathbf{L}_i) \le 1$ depends on the occlusion of the light source by other surfaces. It is 0 when the source is totally occluded, and 1 when the source is totally visible. The effect of a light-absorbing medium $d$ may be incorporated using $h(\mathbf{L}_i)$ as a multiplying factor. In the case of illumination by a "distributed" light source, the radiance of the source $S^{in}(\mathbf{L})$ has to be known for any direction $\mathbf{L} \in F$. In this case,

$$S^{out}(\mathbf{V}) = \int_{\mathbf{L} \in F} \rho(\mathbf{L}, \mathbf{V}) |\mathbf{NL}| S^{in}(\mathbf{L}) \, d\omega_L. \tag{2.7}$$

If the diffuse light source has a constant radiance, $S^{in}(\mathbf{L}) = 1$, for any $\mathbf{L} \in F$, and from Eqs. (2.7) and (2.2), the reflected radiance is given by

$$S^{out}(\mathbf{V}) = a(\mathbf{V}). \tag{2.8}$$

[1] *Spectral radiance* is a fundamental radiometric quantity in $[Wm^{-2}sr^{-1}]$ units of a given wavelength, to be derived from the radiant power of a differential surface element and a differential solid angle element. Its exact definition is given by Neumann and Neumann [1989]. According to the CIE recommendation, it is denoted $L_{e,\lambda}$, the notation that is sometimes used in computer graphics, but a more common terminology is *intensity*, denoted $I$. Throughout this paper spectral radiance is denoted $S$, as in Neumann and Neumann [1989, 1990].

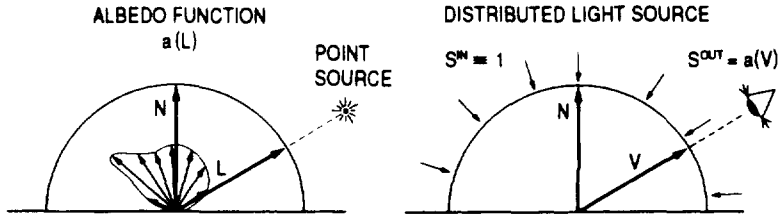


Figure 1

The term on the right-hand side is the albedo function (2.3) in the view direction **V**. Eqs. (2.3) and (2.8) are in some sense reciprocal of each other (see Figure 1).

## 2.3 The Efficiency

Assume a composite scene consisting of several surfaces. For every surface point $T'$ in the scene, we can define a set of directions $F(T')$ pointing from $T'$ to the points in the scene visible from $T'$. For the directions $\mathbf{V} \in F(T')$, the point closest to $T'$ in the direction **V** can be characterized by the light-absorption factor $0 \leq d(T', \mathbf{V}) \leq 1$. The *efficiency* is defined as

$$w(\mathbf{L}) = \int_{\mathbf{V} \in F(T')} \rho(\mathbf{L}, \mathbf{V}) |\mathbf{NV}| d(T', \mathbf{V})\, d\omega_{\mathbf{V}}. \tag{2.9}$$

As opposed to the albedo, the efficiency depends not only on the surface, but also on the other elements of the scene. The purpose of $w$ is to express the ratio of power effectively incident on the surfaces in the system to the input power from direction **L**. Efficiency will be of fundamental importance in the investigation of convergence.

## 2.4 The Rendering Equation

Some notations will be introduced before we write down the rendering equation[2] to express the radiance incident on any point in the scene from any other unoccluded point. Let $T'$ be a point on a surface in the scene, and let $T$ and $T''$ be points visible from $T'$. The radiance output from $T'$ toward $T$ has two components: zero-order illumination according to eq. (2.6) and the radiance from the light reflected from all of the points $T''$ through $T'$ to $T$ according to eq. (2.7). We define the unit vectors $\mathbf{L} = (T'' - T')$ and $\mathbf{V} = (T - T')$; see Figure 2. If the bidirectional response at $T'$ is $\rho_{T'}$, then

$$S^{out}(\mathbf{V}) = S_0^{out}(\mathbf{V}) + \int_{\mathbf{L} \in F(T')} \rho_{T'}(\mathbf{L}, \mathbf{V}) |\mathbf{NL}| S^{in}(\mathbf{L})\, d\omega_L. \tag{2.10}$$

[2] The rendering equation was first given and analyzed in computer graphics by Kajiya [1986] using radiometric quantities. It has been given in a form closer to ours in a paper by Sillion and Puech [1989].

Next we consider the behavior of the whole system. For convenience we transform the "local" $S$ to[3]

$$\mathbf{s}(T, -\mathbf{V}) = S_T^{in}(-\mathbf{V}) = S_{T'}^{out}(\mathbf{V}) \cdot d(T', T). \tag{2.11}$$

$d$ is the absorption effect of the medium, and $\mathbf{s}$ is the radiance computed from the actual power transport between pairs of points. For each mutually visible point pair, both sides of eq. (2.10) are to be multiplied by $d(T' - T)$, yielding for the two-variable function $\mathbf{s}$ the linear functional equation

$$\mathbf{s} = \mathbf{s}^o + \mathscr{R}\mathbf{s}. \tag{2.12}$$

The linear operator $\mathscr{R}$ deriving from the integral term in eq. (2.10) is the *single reflection operator* (or *optical response operator*) for the whole system, which assigns the input radiance from the scene $\mathbf{s}$ by exactly one reflection to $\mathbf{s}$ itself, while $\mathbf{s}^o$ corresponds to the scene $S_0^{in}$. The total result after all of the interreflection, the radiance of the scene $\mathbf{s}$, is the limit of the series starting from an arbitrary $\mathbf{s}^o$ produced by successive applications of

$$\mathbf{s}_{k+1} = \mathbf{s}_o + \mathscr{R}\mathbf{s}_k, \tag{2.13}$$

converging if $\mathscr{R}$ is a contraction, that is, if there exists some $q < 1$:

$$\|\mathscr{R}\mathbf{s}\| \leq q \cdot \|\mathbf{s}\|. \tag{2.14}$$

Using the Neumann series expansion, the series (2.13) for $\mathbf{s}_o = \mathbf{s}^o$ converges to the solution of (2.12) in the norm above as

$$\mathbf{s} = \mathbf{s}^o + \mathscr{R}\mathbf{s}^o + \mathscr{R}^2\mathbf{s}^o + \cdots. \tag{2.15}$$

## 2.5 Convergence Criteria

We will now give an exact analysis, the first in the computer graphics literature, of the convergence of the operator eq. (2.12) and, hence, of the rendering equation.

Any visual perception relies on the detection of energy. Therefore, if we define the norm as the sum of the absolute values of the power assigned to $\mathbf{s}$, convergence to that norm will mean convergence as far as view is concerned, though it is not a pointwise convergence. Let us now find a sufficient condition for that type of convergence, that is, for $\mathscr{R}$ to be a contraction. Let us introduce a *restriction* of $\mathbf{s}$ to a small area and solid angle:

$$\mathbf{s}_{(A,\omega)}(T', \mathbf{L}) = \begin{cases} S^{in}(T', \mathbf{L}), & \text{if } T' \in A, \ \mathbf{L} \in \omega, \\ 0, & \text{otherwise.} \end{cases} \tag{2.16}$$

[3] In the presence of a light-absorbing medium ($d \neq 1$), it is meaningful to distinguish between input and output magnitudes. For simplicity's sake, in the rest of the paper it is not always specified which one is actually used, or whether we assumed $d \neq 1$ or not. In any case, the magnitudes can always be retrieved by properly multiplying by $d$ and can be easily converted to each other.

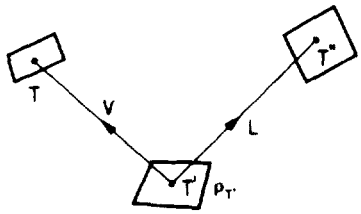


Figure 2

Using $d\mathbf{s}$ as the limit of $\mathbf{s}$ in the above restriction, the efficiency and radiance then become

$$w(T',\mathbf{L}) = \frac{\|\mathscr{R}d\mathbf{s}\|}{\|d\mathbf{s}\|}(T',\mathbf{L}), \qquad S^{in}(T',\mathbf{L}) = \frac{\|d\mathbf{s}\|}{d\omega dA|\mathbf{N}_{T'}\cdot\mathbf{L}|}(T',\mathbf{L}). \quad (2.17)$$

If $\mathscr{R}$ is a contraction, then condition $\|\mathscr{R}\mathbf{s}\|/\|\mathbf{s}\| \le q$ has also to hold for $\mathbf{s}_{(A,\omega)}$, and $w(T',\mathbf{L}) \le q$ has to hold for the efficiency for any $T'$ and $\mathbf{L}$. This is a sufficient condition for $\mathscr{R}$ to be a contraction, as seen from

$$\begin{aligned}
\|\mathscr{R}\mathbf{s}\| &= \int_{T'}\int_{\mathbf{L}} \frac{\|\mathscr{R}d\mathbf{s}\|}{d\omega dA}(T',\mathbf{L})d\omega_L\, dA_{T'} \\
&= \int_{T'}\int_{\mathbf{L}} \frac{\|\mathscr{R}d\mathbf{s}\|}{\|d\mathbf{s}\|}(T',\mathbf{L})\frac{\|d\mathbf{s}\|}{d\omega dA}(T',\mathbf{L})\, d\omega_L\, dA_{T'} \\
&= \int_{T'}\int_{\mathbf{L}} w(T',\mathbf{L})\frac{\|d\mathbf{s}\|}{d\omega dA}(T',\mathbf{L})d\omega_L\, dA_{T'} \le \int_{T'}\int_{\mathbf{L}} q\frac{\|d\mathbf{s}\|}{d\omega dA}(T',\mathbf{L})\, d\omega_L\, dA_{T'} \\
&= q\int_{T'}\int_{\mathbf{L}} \frac{\|d\mathbf{s}\|}{d\omega dA}(T',\mathbf{L})\, d\omega_L\, dA_{T'} = q\|\mathbf{s}\|. \qquad (2.18)
\end{aligned}$$

We can accordingly state the following:

THEOREM 1. *$\mathscr{R}$ is a contraction iff sup $w(T',\mathbf{L}) < 1$ is true.*

We can define the efficiency $w_K$ of order $K$ as the fraction of undissipated power remaining in the system after at most $K$ interreflections. Obviously, $w_K \ne w^K$. We can state the following (the detailed proof is omitted):

THEOREM 2. *$\mathscr{R}^K$ is a contraction iff $w_K(T',\mathbf{L}) < 1$ is true.*

If there is such a $K$, then the Neumann series in (2.15) is convergent, and its limit $\mathbf{s}$ is the real view. Thus, if after a given number of steps the total power assigned to $\mathbf{s}$ has decreased by more than a fraction $q$, then $\mathbf{s}$ will converge. Losses in energy may be due to surface absorption, absorption by a participating medium, and transfer of energy outside of the system. In particular, the observer's pupil always absorbs energy. Note that the proof of the theorems did not make use of the reciprocity condition (2.2).

## 3. RADIOSITY METHOD

The solution of the rendering equation by finite approximation is referred to in the computer graphics literature as the *radiosity method*. This section is

concerned with the systems of *radiosity equations* and the algorithms for their solution. The known results will be surveyed briefly using a uniform notation. The novel aspects will be the discussion in parallel of the radiance/radiosity and power forms, and the introduction of the zero-order illumination component including point sources. Then the *Southwell algorithm* will be described as an effective solution to the radiosity systems of equations, a method identical in the diffuse case to the *sorted shooting method*, a fact unnoticed until recently.[4] Then we will introduce the *double-patch method* to solve the system with nondiffuse reflectors. Finally, a variation of the *conjugate gradient method*, mainly for very bright environments, will be introduced, and the claim of its efficiency will be supported by test examples.

## 3.1 Reformulation of the Basic Radiosity Equations

3.1.1 *Fundamentals.* Let us establish the basic symbols. Let $A_i$, $i = 1, 2, \ldots, N$, be the area of a patch, considered to be homogeneous. Let $T_i$ be a distinguished point on patch $i$, for instance, the center if the patch is a rectangle. The solid angle under which $T_i$ views patch $j$ (only its visible part) is $\omega_{ij}$. If $A_j$ is completely visible, then

$$\omega_{ij} \approx \frac{A_j \cdot |\mathbf{N}_j \mathbf{V}_{ji}|}{r_{ij}^2} \quad [sr], \tag{3.1}$$

where $r_{ij}$ is the distance between $T_i$ and $T_j$, and where $\mathbf{V}_{ij}$ is the unit vector form $T_i$ to $T_j$. If patch $j$ is only partly visible from $T_i$, then $A_j$ in eq. (3.1) is the area of the visible part of the patch. We then define the *form factor* as[5]

$$F_{ij} = \omega_{ij} |\mathbf{N}_i \mathbf{V}_{ij}| \quad [sr]. \tag{3.2}$$

The absolute value in (3.2) (the cosine of the incident angle $\Theta_{ij}$) is needed because of transmissive materials where $\mathbf{N}_i$ could point either way. The equality stating the reciprocity of the form factors

$$A_i F_{ij} = A_j F_{ji} \tag{3.3}$$

is only approximate in our context, because of the discretization. There might be cases where $A_j$ is partly visible from $T_i$, but where $A_i$ is not seen from $T_j$. There might be large errors as well for patches perpendicular to large patches. For such problematic pairs $(i, j)$, eq. (3.3) yields new corrected form factors meeting the requirements of reciprocity by taking the arithmetic mean of the left and right sides.

[4] Since this paper was submitted in 1989, a technical report by Gortler et al. [1993] established the same connection.

[5] According to the usual notation in computer graphics, the form factor is as given in eq. (3.2), but divided by $\pi$. Our notation is advantageous by allowing discussion of the non-diffuse cases with a consistent notation, while the diffuse case remains essentially unaltered.

The radiant power $P_{ij}$ is emitted from patch $i$ to patch $j$. Obviously, in general, $P_{ij} \neq P_{ji}$. The spectral radiance derived from $P_{ij}$ is

$$S_{ij} = \frac{P_{ij}}{A_i F_{ij}} \quad [Wm^{-2}sr^{-1}]. \tag{3.4}$$

3.1.2 *Radiosity for Non-Diffuse Environments.* Let us now consider the most general case. Let patch $j$ be characterized by an arbitrary bidirectional response $\rho_j$ (reflectance or transmittance). The unit vectors $\mathbf{V}_{jk}$ and $\mathbf{V}_{ji}$ point from patch $j$ to patches $k$ and $i$, respectively.

The quantity $\rho(k \rightarrow j \rightarrow i) = \rho_j(\mathbf{V}_{jk}, \mathbf{V}_{ji})$ is necessary to compute the contribution of patch $k$ $(k = 1, 2, \ldots, N;\ k \neq j)$ to the radiance $S_{ji}$. The finite approximation of (2.10), first given by Immel et al. [1986] with somewhat different symbols, is

$$S_{ji} = S^o_{ji} + \sum_{k=1}^{N} \rho(k \rightarrow j \rightarrow i) F_{jk} S_{kj}. \tag{3.5}$$

This equation can be written for any pair $(i, j)$ where $F_{ji} > 0$. The constant term $S^o_{ji}$ corresponds to the finite approximation of the zero-order illumination component from (2.6):

$$S^o_{ji} = S^E_{ji} + \frac{1}{4\pi} \sum_{k=1}^{L} \rho(k \rightarrow j \rightarrow i) |\mathbf{N}_j \mathbf{V}^*_{jk}| \cdot \frac{P_k}{r^2_{jk}} h^*_{kj}. \tag{3.6}$$

$S^E_{ji}$ is the value of the emissive component of the patch in the direction $\mathbf{V}_{ji}$. In addition, there is a term on the right-hand side for the illumination due to $L$ punctual light sources. The inclusion of punctual light sources is an obvious but important result here. The first reflection from the punctual light sources *is not included among patch-to-patch interreflection.* Accordingly, their effect appears in the right-hand-side *constant vector* of the system of equations, rather than in the matrix coefficients.

The superscript * denotes that the given magnitude refers to a point source corresponding to the subscript variable of the sum $\Sigma$. The factor $h^*_{ki}$ may assume any value from 0 to 1, depending on the proportion of the area of patch $i$ illuminated by the given light source (occlusions can be determined, e.g., with the *hemicube* technique; a point light source has to be surrounded by a full cube). In addition, the point source might emit anisotropically. This possibility is usually ignored in classic ray tracing, although it permits spectacular effects. Such an anisotropic light source may be, for example, a light from a slide projector or sunshine transmitted from a stained-glass window (slides, photos of stained-glass windows, or silhouettes of real objects can be digitized). Anisotropy in any direction can be described by a transmission factor $\tau \in [0, 1]$, which will be included as a multiplicative factor to $h$, together with the light-absorbing coefficient $d$ of the medium. For these $\tau$ and $d$ values, one also assumes mean values for patches. All of the effects averaged at the patch level are illustrated in Figure 3.

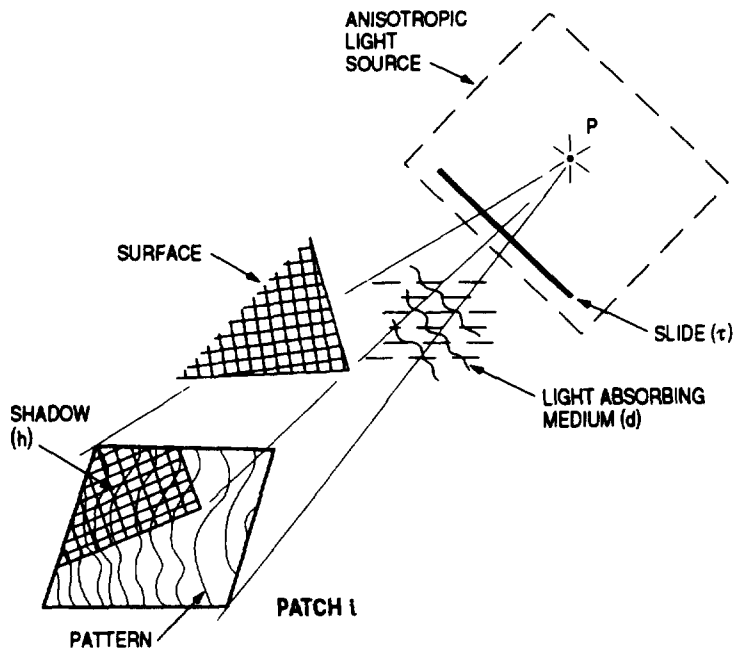


Figure 3

In Figure 3 the zero-order components from (3.6) are, of course, assumed at pixel resolution, handling the anisotropic effects suggested above by ray tracing. The computation of the interreflection effects at the pixel level is outlined in Section 5.

Equation (3.5) can be written not only for radiance, but also directly for *power* transport, using (3.3) and (3.4). This form of the equations has been introduced by Neumann and Neumann [1989]:

$$P_{ji} = F_{ji}\left( A_j S^o_{ji} + \sum_{k=1}^{N} \rho(k \to j \to i) P_{kj} \right). \tag{3.7}$$

A row of eq. (3.7) contains a single form factor $F_{ji}$ on the output side, since the form factors $F_{jk}$ are not explicit on the input side. The quantities $P_{kj}$ exist, however, only for pairs $(k, j)$ for which $F_{jk} > 0$. The difficulty is that (3.5) and (3.7) lead to *very large systems*, though with rather *sparse matrices*. For $N$ patches there are $O(N^2)$ unknowns. Their exact number depends on the occlusion conditions.

In the survey aspect of this paper, the basic equations such as these will always be considered in two forms in parallel: in the radiance/radiosity form and in the power form. The convergence characteristics and the solution methods differ for each form.

3.1.3 *Radiosity for Diffuse Environments.* Let us now discuss the best known case for these equations, the diffuse case. In terms of radiosity variables $[W \cdot m^{-2}]$, the equations become

$$B_i = E_i + \rho_i \cdot \sum_{j=1}^{N} F_{ij} B_j. \tag{3.8}$$

This equation was introduced by Goral et al. [1984] by analogy from radiative heat transfer. It has been further developed for occluded and complex systems by Nishita and Nakamae [1985], by Cohen and Greenberg [1985] and by Cohen et al. [1986]. As a matter of fact, the sorted shooting method [Cohen et al. [1988] relies on the power form of the diffuse equation, although (3.8) was explicitly used. The power form is

$$P_i = P_i^o + \rho_i \cdot \sum_{j=1}^{N} F_{ji} P_j, \tag{3.9}$$

where $P_i^o = A_i E_i$ and $P_i = A_i B_i$. For the convergence criteria [Neumann and Neumann 1989, 1990], we observe that eqs. (3.5), (3.7), (3.8), and (3.9) are of the form $\mathbf{x} = \mathbf{Ax} + \mathbf{b}$. In eqs. (3.5) and (3.8), the row norm of matrix $\mathbf{A}$ is less than 1, while for eqs. (3.7) and (3.9), its column norm is less than 1. The convergence of the classic, or Jacobi, iteration is guaranteed in either norm. A known sufficient condition for the convergence of the Gauss–Seidel iteration is met by the row norm. For the zero-order component, the constant term in (3.8) is:

$$E_i = B_i^E + \frac{1}{4\pi} a_i \cdot \sum_{k=1}^{L} |\mathbf{N}_i \mathbf{L}_{ik}^*| \cdot \frac{P_k^*}{r_{ik}^2} h_{ki}^*, \tag{3.10}$$

where $a_i$ is a dimensionless albedo (or reflectivity) constant, of value in the interval [0, 1]. $B_i^E$ is the emissivity of the patch, considered as a Lambertian diffuse emitter, which is totally independent of direction.

For diffuse environments there are $N$ unknowns and a relatively low number of nonzero matrix elements. But for very complex occlusion conditions, the matrix may be a sparse one.

3.1.4 *Radiosity for Separable Reflectance.* A mathematical generalization of diffuse environments is environments with a *separable* reflectance, introduced by Neumann and Neumann [1989]. This is the widest class of reflectance behavior where the distribution of the emitted radiance is independent of the distribution of the incident radiance (of course, its magnitude is not). The relevant system of equations includes the diffuse case as a special case and has $N$ unknowns as well. The bidirectional response $\rho$ is separable if it is of the form

$$\rho(\mathbf{L}, \mathbf{V}) = a(\mathbf{L}) \cdot \hat{a}(\mathbf{V}). \tag{3.11}$$

Due to reciprocity, the function $a$ has to be equal to the function $\hat{a}$ within a multiplicative constant. It is easy to see that $a$ is equal to the dimensionless albedo function if

$$\int_{\mathbf{V}\in F} \hat{a}(\mathbf{V})|\mathbf{NV}|d\omega_V = 1 \tag{3.12}$$

holds. If (3.12) does not hold, $\hat{a}[sr^{-1}]$ is the distribution density function of the reflected light. For diffuse materials, $a$ is the albedo, constant, and $\leq 1$, and $\hat{a} = constant = 1/\pi$. Using the notations $a_{ij} = a_i(\mathbf{V}_{ij})$ and $\hat{a}_{ji} = \hat{a}_j(\mathbf{V}_{ji})$, the system of equations becomes in the radiosity form

$$B_i = E_i + \sum_{j=1}^{N} F_{ij} a_{ij} \hat{a}_{ji} B_j \tag{3.13}$$

and in the power form

$$P_i = P_i^o + \sum_{j=1}^{N} F_{ji} a_{ij} \hat{a}_{ji} P_j \tag{3.14}$$

where, as before, $P_i^o = A_i E_i$ and $P_i = A_i B_i$. With these notations, the constant term of the equations becomes

$$E_i = B_i^E + \frac{1}{4\pi} \cdot \sum_{k=1}^{L} a_i(\mathbf{L}_{ik}^*) \cdot |\mathbf{N}_i \mathbf{L}_{ik}^*| \cdot \frac{P_k^*}{r_{ik}^2} h_{ki}^*. \tag{3.15}$$

Remember that (3.15) expresses the zero-order component, and so does the term $B_i^E$ within it, describing the emissivity of the patch according to a distribution given by $\hat{a}(\mathbf{V})$. Once the equations are solved, the radiosities found are also emitted according to this distribution, so that, in direction $\mathbf{V}$, patch $i$ is seen with a radiance $B_i \hat{a}_i(\mathbf{V})$. In the diffuse case, again, $\hat{a} = constant$, so that, instead of radiance, the radiosity can be used directly. Eq. (3.13) fails both conditions of row and column norms, while the column norm of the power form (3.14) is less than 1.

3.1.5 *A Generalization.* The enumeration of the fundamental radiosity equations will be concluded by an interesting generalization. In the diffuse and separable case, the spatial distribution of the emission from the patches could only be according to the distribution function $\hat{a}$. After reflection from a point source, the distribution has to be such, but in the case of the light directly *emitted* by the patch, it could follow an arbitrary distribution, as shown below. Let the zero-order illumination component have the radiance $S_{ji}^o$ in the direction $\mathbf{V}_{ji}$. The meaning of the radiosity and power variables also has to change. They are now variables without the zero-order component, which we call the *zero-order residual components.* With the new variables, the system becomes

$$B_i = \sum_{j=1}^{N} F_{ij} a_{ij} (\hat{a}_{ji} B_j + S_{ji}^o) \tag{3.16}$$

and

$$P_i = \sum_{j=1}^{N} F_{ji} a_{ij} (\hat{a}_{ji} P_j + S^o_{ji} A_j), \tag{3.17}$$

while the radiance observed by the viewer becomes

$$S_i(\mathbf{V}) = B_i \hat{a}_i(\mathbf{V}) + S^o_i(\mathbf{V}). \tag{3.18}$$

### 3.2 The Sorted Shooting Algorithm and the Southwell Algorithm

Now we will establish that the well-known Southwell relaxation algorithm (see, e.g., Krekó [1976]) is identical to the sorted shooting algorithm described by Cohen et al. [1988]. Again we use a uniform notation to discuss the shooting method for the already-known diffuse case and a similar method for the separable reflectance case.

The Southwell algorithm is derived from Gauss–Seidel elimination, but is a simpler method. In the Southwell algorithm, the row with the largest error in absolute value is selected, and the element from that row on the main diagonal is modified so that there is no error on that row. As well, the whole error vector is recomputed. The convergence criteria with this method are different from those in the Gauss–Seidel method. Convergence may be proved for either a positive definite matrix or for a strict column norm condition. None of these conditions is met by the matrix arising from the radiosity equations. However, for a zero initial vector, the convergence of the power form of the equations can be proved.

Let us start from a system of radiosity equations in the power form, that is, (3.9) for diffuse reflectance, (3.14) for a separable reflectance, or (3.7) for a general, non-separable case. This has the general form $\mathbf{x} = \mathbf{Ax} + \mathbf{b}$, which can be rearranged as $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{b}$, defining $\mathbf{C} = (\mathbf{I} - \mathbf{A})$, $\mathbf{Cx} = \mathbf{b}$. Let us define the error vector $\mathbf{e}$ as $\mathbf{e} = \mathbf{b} - \mathbf{Cx}$, and the initial approximation as $\mathbf{x}^o = \mathbf{0}$. Therefore, initially $\mathbf{e} = \mathbf{b}$, the constant vector of the system of equation, that is, the vector of the zero-order illumination components.

Let $\mathbf{e} = (e_1, e_2, \ldots, e_N)$. Among the elements of $\mathbf{e}$, the one with the maximum value $e_k$ will be chosen. For the starting vector given, all of the elements of $\mathbf{e}$ will remain non-negative during the iteration. The column $k$ of the matrix $\mathbf{C}$ is $(c_{1k}, c_{2k}, \ldots, c_{Nk})$. The sums of the elements of column $k$ and of the error vector are denoted $v_k$ and $E$, respectively. Now we can state the steps of the Southwell algorithm:

0. Repeat until $E < \epsilon$
1. Select $k$ such that $e_k$ is maximum
2. $x_k := x_k + e_k$ (alteration of a single element)
3. $e_i := e_i - e_k \cdot c_{ik}$ (for $i = 1, 2, \ldots, N$; compute the new error vector)
4. $E := E - v_k \cdot e_k$

The algorithm repeats until the absolute norm of the error vector (in this case, simply $E$) drops below a given value $\epsilon$. For the power form, the column norm of $\mathbf{C}$ is less than 1; hence, $v_k \leq V < 1$ $(K = 1, 2, \ldots, N)$, permitting us

to assert the fractional decrease in value of $E$ in step 4. Using $e_k \geq E/N$, the speed of convergence may be assessed as

$$E_{new} < E_{old} \cdot \left(1 - \frac{V}{N}\right). \tag{3.19}$$

The actual convergence is faster than that, especially in the first steps [Neumann and Neumann 1990]. The convergence of the Southwell algorithm is manifest for the power form of the equations, and the sorted shooting algorithm is identical to the Southwell algorithm. Let us look specifically at the diffuse case of eq. (3.14). The error vector is called the *unshot power* by Cohen et al. [1988]. The counterpart of the elements of the matrix $\mathbf{C}$ is $c_{ik} = -\rho_i F_{ki}$ for $k \neq i$, and $c_{ii} = 1$. The solution of vector $\mathbf{x}$ gives the power emitted by the patches as the result of interreflection. During the iteration the value of the power in $\mathbf{x}$ changes only in the row selected by the Southwell algorithm, so we do not have to rewrite all of the values of $\mathbf{x}$ at each step, but we have only to compute $\mathbf{x}$ for the image (in the form of $\mathbf{x} + \mathbf{e}$) at the last step after meeting the error criterion.

There is, of course, a possibility of image in the meantime. To do so, the solution of the *residual problem* has to be approximated. At some point in the iteration, the values of the solution vector and of the error vector be $\mathbf{x} = \mathbf{x}_{app}$ and $\mathbf{e} = \mathbf{e}_{app}$, respectively. The residual problem is of the form $\mathbf{C}\mathbf{x}_{res} = \mathbf{e}_{app}$. The problem is simply approximated by a single classic (Jacobi) iteration step, that is, $\mathbf{x}_{res} = \mathbf{e}_{app}$, and the image can be computed from the power values $\mathbf{x} = \mathbf{x}_{app} + \mathbf{x}_{res}$. Another solution is to use the ambient term discussed in Section 4.

The discussion above is valid for all of the shooting algorithms, whether with diffuse, separable, or general reflectance. Let us consider how the method works for extended systems in the diffuse and separable cases. The form factors are computed as by Cohen et al. [1988]; that is, the selected column of matrix $\mathbf{C}$ is always computed by solving a single hidden surface problem by the hemicube method. If only a small number of patches are emissive and the system contains no point light sources, then the constant vector $\mathbf{b}$ has few nonzero terms from which light is emitted into the system. At first, the zero components of the vector $\mathbf{b}$ increase slowly in the process, but generally are less than the initial power values of the emissive patches and are as a rule seldom selected. That means that this method is really effective for relatively dark, large systems and that, in fact, much fewer than $N$ hidden surface problems have to be solved. It also means that real interreflections are omitted by this method. This case corresponds in practice to a classic problem without interreflection with many light sources, needing $O(N)$ operations (providing that there are a $O(1)$ number of emissive patches). In cases where interreflection effects are accounted for, the main advantage of the method of Gauss–Seidel iteration is lost since the algorithm requires $O(N^2)$ computations; this is because the hidden surface problem has to be solved for most of the patches, which involves the whole matrix in the computation. An example of this is a simple light-colored interior illuminated

by a point source. Still, the Southwell algorithm is asymptotically faster than the Gauss–Seidel method.

Another advantage of the Southwell algorithm manifests itself in the case of general reflectance. This large problem (3.5) of a sparse matrix has been solved before by the Gauss–Seidel method [Immel et al. 1986] by intuitively selecting the order of operation and expecting to follow the propagation of light in the system. The Southwell algorithm specifies a rule for objective selection for solving the power form of the non-diffuse eq. (3.7). Although in practice, non-diffuse reflectance is effectively approximated by diffuse plus specular decomposition [Shao et al. 1988; Neumann and Neumann 1990], often, the high cost of the computation of the general case cannot be avoided, such as for very dull reflecting materials. In such cases, the Southwell algorithm introduced in Neumann and Neumann [1990] is the most efficient method known and can be complemented with the generalized ambient term described in Section 4.

Remember that step 4 of the algorithm may be omitted and that the error may be taken as $e_k$. The value of $E$, however, is better related to the convergence. The problem of finding a better selection criterion also arises. Selecting the variable for which the value $v_k \cdot e_k$ is maximum would obviously yield the maximum decrease in a given step; the algorithm so modified, however, seems to converge more slowly later in spite of this initial advantage. On the other hand, the column sums $v_k$ of matrix $\mathbf{C}$ will only be known if the hidden surface problem has been solved for all of the patches and if all of the form factors have been computed. The main advantage of the shooting method is that it is sufficient to solve the hidden surface problem only for the selected patches. If this advantage is to be kept, then there is no better selection rule than to select the maximum $e_k$ value. It is still possible to assign initially an average for the sum of the columns not yet selected, gradually replacing them with the sums of the columns already selected so that $v_k$ can be taken into account in the selection.

This method, however, can be further developed even when using maximum error selection and keeping the relaxation character of the method, where in one step a single coordinate is modified. Overrelaxation seems advisable where the maximum error coordinate drops to a negative value rather than to 0 and, therefore, where the other error components grow faster, anticipating a power increase due to interreflection. Step 2 of the modified algorithm will then be of the form $x_k := x_k + \tau \cdot e_k$, where the $\tau$ value will be chosen so as to effect a maximum decrease for a given error function $E(\tau)$. An error function such as $E = Constant \cdot \sum e_i^2 + (\sum e_i)^2$ is suggested for a numerical test.

### 3.3 The Double-Patch Method (DPM)

The solution of the radiosity equation for the general reflectance model ((3.5) and (3.7)) is rather laborious even with the Southwell algorithm. The complexity of $O(N^3)$ seems unavoidable for general, non-separable reflectance models. For such cases, an approximate solution of complexity $O(N^2)$, with exact zero- and first-order components, is given by the albedo-equivalent

separable model introduced by Neumann and Neumann [1990]. There is an exact and efficient radiosity solution in cases where the reflectance model can be decomposed into diffuse (or separable) and specular forms, so that the specular part is only relevant within a cone of a rather small solid angle.[6] The first example of this family of methods has been introduced by Shao et al. [1988], and their method exhibited convergence in practice in the cases they presented, but this convergence has not yet been proved. In this method, the solution is approximated by a set of equations with different matrices of the same size as the diffuse problem. Methods with fixed-size matrices, an essentially different approach, were introduced by Neumann and Neumann, relying on the power form equation system given in Neumann and Neumann [1990, sect. 3]. It has been solved by Jacobi iteration, and has been proved to converge, by the conjugate gradient method, as well as by the sorted gathering and shooting methods. These methods, as well as DPM, to be presented later, result in a pure radiosity solution relying on the decomposition between diffuse (or separable) and specular components. They are distinct from two-pass methods, which also use decomposition of the reflectance, but are characterized by the use of ray tracing to compute the image; the latter are the hybrid methods discussed in Section 5.6.

The DPM method is expected to reconcile the following two contradictory requirements: For the separable problem, an equation system of $N$ variables suffices, but every patch reflects toward any other non-occluded patch. In the case of scenes with purely specular surfaces, this equation system will have $O(N^2)$ variables, but a rather sparse matrix, since a power incident on a patch from a given direction will be reflected only toward a small number of patches within the effective specular cone. If one cannot take advantage of these two types of situations, then the original eqs. (3.5) and (3.7) will have to be used in spite of their high cost.

An efficient solution is still possible using the double-patch method, as suggested in Neumann and Neumann 1989. Consider the two reflection components of the same physical patch as belonging to two formally separate patches with the same incoming illumination. There is a restriction, namely, that the two patches cannot directly reflect light on each other. One member of the double patch can only have separable (or diffuse) reflectance, and the other can only have specular reflectance. The specular reflectance may be replaced by another arbitrary component, even of negative value, because in the method that counts is that the light is reflected only toward a small part of space.

The bidirectional response, either reflectance or transmittance, is of the form

$$\rho(\mathbf{L}, \mathbf{V}) = a(\mathbf{L})\hat{a}(\mathbf{V}) + \rho^s(\mathbf{L}, \mathbf{V}). \tag{3.20}$$

[6] This algorithm has a complexity of $C \cdot O(N^2)$, where $C$ is the average number of patches within the specular cones. It is difficult to obtain an exact theoretical complexity for $C$, but it is in the range $O(1)$ to $O(N)$.

The first term of the sum is a separable part of the form given in (3.11), while $\rho^s$ is the specular component. It has to verify the law of energy conservation:

$$a(\mathbf{L}) + \int \rho^s(\mathbf{L}, \mathbf{V})|\mathbf{NV}|\, d\omega_V \leq 1. \tag{3.21}$$

If the separable part is simply diffuse, the albedo $a(\mathbf{L}) = a$ is constant, and so is $\hat{a}(\mathbf{V}) = 1/\pi$. The simplest type of reflector for (3.20) is a diffuse material with a mirrorlike component (polished). The degree of polish should be somewhat moderate, so that the specular component is within a small solid angle, larger than for a pure mirror. A still more general model is a separable lacquer model [Neumann and Neumann 1989] with the same kind of polishing, actually making the lacquer model more similar to real varnish.

To ensure convergence, the power form equations are used. Let the total power of the separable term of the double patch $k$ be $P_k$, emitted according to the density function $\hat{a}$, as in eqs. (3.11) and (3.12). The patch $j$ receives the fraction $F_{kj}\hat{a}_{kj}$ of power $P_k$. Let $P^s_{kj}$ be the power going by specular reflection from the specular member of double patch $k$ to patch $j$. The system of equations written according to (3.7), (3.14), and (3.17) becomes

$$P_j = P_j^o + \sum_{k=1}^{N} a_{jk} \cdot \left(F_{kj}\hat{a}_{kj}P_k + P^S_{kj}\right), \tag{3.22a}$$

$$P^S_{ji} = P^{S,o}_{ji} + F_{ji} \cdot \sum_{k \in I(j,i)} \rho^S(k \to j \to i) \cdot \left(F_{kj}\hat{a}_{kj}P_k + P^S_{kj}\right), \tag{3.22b}$$

where $I(j, i)$ is the set of subscripts for the patches in important directions involving subscripts $k$ such that for input directions $k \to j$ the direction $j \to i$ is within the reflection cone (defined by $\mathscr{S}$ greater than a given $\epsilon$). The reciprocity of $\mathscr{S}$ implies the reciprocity of the cones; that is, $k \in I(i, j)$ iff $i \in I(j, k)$. In eqs. (3.22a) and (3.22b), the constant term may be zero. It seems advisable, however, to separate the term $P_j^o$ emitted according to the distribution function $a_j$ in the case of point sources (see eq. (3.15)).

The system of equations for the DPM is of the form $\mathbf{x} = \mathbf{Ax} + \mathbf{b}$. If eq. (3.21) holds, then the column norm of matrix $\mathbf{A}$ is less than 1. For extremely bright systems, this value may be greater than 1 because of errors of finite approximation, but this is not a concern in practice. The DPM equations can be solved by Jacobi iteration. It is possible to speed up the solution of eq. (3.9) if $I(j, i)$ is much less than $N$. For instance, if an environment has 2,000 patches and, on the average, 20 patches affect the important directions, then the computation for eq. (3.22b) is accelerated 100 times (in principle, there may be as many as 4 million variables).

The DPM equations can also be solved by the Southwell algorithm. This method is especially appropriate for the double-patch approach, since depending on whether the separable or the specular parts dominate, the frequency of selecting eq. (3.22a) or eq. (3.22b) may be rather different.

The flexibility of the DPM method makes it suitable for other situations as well. For example, in a case with a very bright separate term the following procedure may be applied: One first solves separately the separable part of

eq. (3.22a) with fixed specular variables by the conjugate gradient method, and then one updates the specular variables of eq. (3.22b) by the Jacobi or Southwell method. These two steps are then repeated until a prescribed error criterion is met. For the final step, the formula to compute the radiance for the image becomes

$$S_{kj} = \frac{P_k}{A_k} \cdot \hat{a}_{kj} + \frac{P^S_{kj}}{A_k F_{kj}}. \tag{3.23}$$

### 3.4 The Conjugate Gradient Method

A version of the gradient method for solving the linear equation system by the radiosity method was described in Neumann and Neumann 1989. The paper also suggested applying the two-parameter method for the least error terms. These methods are convenient for moderately bright environments.

Up to now, however, there was no fast converging method applicable both to systems with a mean albedo close to 1 and to systems with a low albedo. To correct this we give below a version of the conjugate gradient method that can be applied in these cases.

3.4.1 *Notations and Algorithms.* The system of equations for radiosity again is of the form $\mathbf{Cx} = \mathbf{b}$. Rather than to apply the usual symmetric positive definite form $\mathbf{CC}^*$ [Marcsuk 1976], let us define an error function $F(\mathbf{x})$ as the sum of the square of errors in each row:

$$2 \cdot F(\mathbf{x}) = \sum_{i=1}^{N} (b_i - \mathbf{c}_i\mathbf{x})^2 = \sum_{i=1}^{N} e_i(\mathbf{x})^2. \tag{3.24}$$

The minimum point of $F(\mathbf{x})$ is the solution of eq. $C\mathbf{x} = b$ if the minimum value is zero, which always holds in our case. Let us minimize $F(\mathbf{x})$ by means of the *Stiefel–Hestenes* version of the generalized conjugate gradient method [Kòsa 1979]. The basic step of that method is to determine the gradient vector $\mathbf{g}(\mathbf{x}) = \nabla F(\mathbf{x})$. The coordinate $j$ of the gradient vector ($j = 1, 2, \ldots, N$) is given by

$$g_j = -\sum_{i=1}^{N} c_{ij} \cdot e_i(\mathbf{x}). \tag{3.25}$$

It means that to compute $\mathbf{g}$ knowing the error vector $\mathbf{e}$ requires $N^2$ multiplications. Let the initial vector be the zero vector; that is, $x_i^o = 0$ and $e_i^o = b_i$. Furthermore, let $\mathbf{p}^o = \mathbf{g}^o = \mathbf{g}(\mathbf{x}^o)$. Initially setting $k = 0$, we have the following algorithm:

1. $\tau_k = \dfrac{\sum_{i=1}^{N} e_i^k \cdot \mathbf{c}_i\mathbf{p}^k}{\sum_{i=1}^{N} (\mathbf{c}_i\mathbf{p}^k)^2}$

2. $\mathbf{x}^{k+1} = \mathbf{x}^k + \tau_k \cdot \mathbf{p}^k$

Table I. Quadratic Mean Values of Row-wise Errors by the Conjugated Gradient Method for Three Test Problems

| Iteration step number | I | II | III |
|---|---|---|---|
| 0 | 56.20166 | 56.20166 | 56.20166 |
| 1 | 9.76466 | 35.21932 | 48.12593 |
| 2 | 0.12021 | 3.25783 | 47.65600 |
| 3 | 0.00005 | 0.00560 | 0.00003 |
| 4 | 0.00000 | 0.00000 | 0.00000 |

3. $\mathbf{e}_i^{k+1} = \mathbf{e}_i^k - \tau_k \mathbf{c}_i \mathbf{p}^k$ $(i = 1, 2, \ldots, N)$ if $\max(\mathbf{e}_i^{k+1}) < \epsilon$, then output the image and stop

4. $g_j^{k+1} = -\sum_{i=1}^{N} c_{ij} e_i^{k+1}$ $(j = 1, 2, \ldots, N)$

5. $\beta_k = \dfrac{\mathbf{g}^{k+1}(\mathbf{g}^{k+1} - \mathbf{g}^k)}{\mathbf{g}_k^2}$ (the parameter of the Stiefel-Hestenes method)

6. $\mathbf{p}^{k+1} = \mathbf{g}^{k+1} + \beta_k \cdot \mathbf{p}^k$; $k := k + 1$; **GOTO** 1.

3.4.2 *Numerical Observations.* After tests with the diffuse case, we can make some observations on the numerical computations. Simple unoccluded environments with random reflectivity distribution have been generated for $N = 1000$. A dark (I), a medium-bright (II), and an extremely bright (III) system have each been investigated (see Table I). The constant vector was the same in all three cases with the components, which are also the initial values of the error vector, being uniformly distributed random numbers in the interval [0, 100]. The values of albedo for cases I, II, and III are, respectively, in the intervals [0, 0.3], [0, 1], and [0.9998, 1], with a uniform random distribution. That makes the respective average light reflection to be 15 percent, 50 percent, and 99.99 percent (!). The computed values are, in general, typical of unoccluded or slightly occluded environments. The latter case is understood as meaning nearly *zero mean optical distance* [Neumann and Neumann 1989] between pairs of patches, which also means a minimum number of reflections in the light path from one patch to the other. One step of the conjugate gradient method involves twice the number of operations needed in Jacobi iteration. It, therefore, becomes equivalent to the classic method for systems with average brightness needing five or six Jacobi iteration steps, but replacing them with three steps of the conjugate gradient method. The conjugate gradient method is doubtless the most efficient method for bright and extremely bright systems, where it is slow or almost impossible to compute the image by other methods. For this system the conjugate gradient method shows an interesting behavior. The mean error hardly varies in the first two steps, while it abruptly drops in the third step. Combining the solution with an ambient term, a single step may suffice for dark systems (equivalent to the classic gradient method), while for medium-bright systems, two steps are enough. As the complexity of the occlusion conditions increases, of course, the number of steps also increases. As a very coarse rule, it can be stated that the increase is proportional to the mean optical distance.

Table II. $\rho = c \cdot \mathbf{NH}^n$

| $n$ | $c_{max}$ | $\bar{a}$ (mean albedo) |
|---|---|---|
| 0.00 | $\frac{1}{\pi} = 0.318309$ | 1.000 (diffuse) |
| 0.25 | 0.326031 | 0.964 |
| 0.50 | 0.333823 | 0.932 |
| 1.00 | 0.349615 | 0.879 |
| 2.00 | 0.381971 | 0.800 |
| 4.00 | 0.449378 | 0.706 |
| 8.00 | 0.592204 | 0.620 |
| 16.00 | 0.895028 | 0.562 |
| 32.00 | 1.521918 | 0.531 |
| 64.00 | 2.790185 | 0.516 |
| 128.00 | 5.334177 | 0.508 |
| 256.00 | 10.425892 | 0.505 |
| 512.00 | 20.611186 | 0.504 |

## 4. AMBIENT TERM

The ambient-term method gives a coarse approximation for global illumination without solving the radiosity equation. This approximation, however, is only suitable for illustrative purposes. Its main advantage is to generate an image easily since it uses little a priori information, but of course, the result is only approximate. There are several ways to extend the ambient-term method by combining it with the shooting-type solution of diffuse systems (introduced by Cohen et al. [1988]).

### 4.1 The Mean Albedo

Consider a system with general, nondiffuse reflectance. The average reflectivity of patch $i$ may be described by the mean albedo value $\bar{a}$, as in (2.4). The fraction of energy reradiated into the half-space depends, of course, on the spatial distribution of the illumination. Without any a priori information, in conformity with the Bayes principle used in statistics, we assume a uniform distribution, as in eq. (2.4). As an illustration, consider Table II, which was computed by numerical integration and which used the mean albedo values from the Phong [1975] model, that is, a bidirectional reflectance of the form $\rho = c \cdot \mathbf{NH}^n$, with various $n$. The albedo function is admissible; that is, $a(\mathbf{L}) \leq 1$ when condition $c \leq c_{max}$ is met for parameter $c$. If $c = c_{max}$, then $a(\mathbf{L}) = 1$ is met for a perpendicularly incident light. The mean albedo in Table II refers to $c = c_{max}$. The system as a whole has a mean albedo of $a_{ave}$, computed from the mean albedo of patches weighted by their areas:

$$a_{ave} = \frac{\sum_{i=1}^{N} A_i \bar{a}_i}{\sum_{i=1}^{N} A_i}. \tag{4.1}$$

### 4.2 Computation of the Zero-Order Component

Let $P^o$ be the overall power due to all of the zero-order components of the system. $P^o$ is the sum of all of the terms ($P^{o,E}$), the power emitted by the emissive patches (the extended light sources), and the terms ($P^{o,P}$), the power reflected by the patches directly illuminated by the punctual light sources. Now consider closed environments without any light-absorbing medium. Open environments can be made closed by surrounding them with black surfaces. In this case it is advisable to find or approximate the minimum enclosing surface that does not change occlusions in the original system (otherwise, the mean efficiencies have to be used instead of the mean albedo, which requires in practice that the values of the form factors be known). The emissive patch $i$ irradiates a power given by

$$P_i^{o,E} = A_i \cdot \int_{V \in H_i} S_i^o(\mathbf{V}) |\mathbf{N}_i \mathbf{V}| \, d\omega_V . \tag{4.2}$$

The value given by (4.2) can be determined numerically. The effect of point light sources is determined as follows: Let each point light source $k$ ($k = 1, 2, 3, \ldots, N$) be surrounded by a cube whose faces are covered by a square mesh, and let it be used to compute occlusion by forming a depth-buffer on them, as in the *hemi-cube algorithm*. Take an elementary square of the mesh of a face, and call $\Delta\omega$ the solid angle it supports from light source $k$. If the closest visible surface is patch $i$, with an albedo function $a_i$, then the fraction of power reaching patch $i$ and re-radiated in the half-space above is

$$\Delta P = P_k \frac{\Delta\omega}{4\pi} a_i(\mathbf{V}_{ik}^*). \tag{4.3}$$

The power elements $\Delta P$ have to be summed for all of the mesh elements and all of the point lights. To compute (4.3) more efficiently, it is advisible to tabulate the albedo function for the isotropic reflectance model as a function of the angle of incidence.

### 4.3 Approximation of Interreflection

The effect of interreflections is expressed by an equation with a single unknown:

$$P = P^o + a_{ave} \cdot P, \tag{4.4}$$

where $P$ is the total power within the entire scene emitted by interreflection. The solution of (4.4) can be expressed as

$$P = P^o + \overline{P}^o = P^o + \frac{a_{ave}}{1 - a_{ave}} \cdot P^o. \tag{4.5}$$

There is no need for additional information about the scene, and the excess of power $P$ over $P^o$, noted $\overline{P}^o$, is distributed among patches according to albedo

and area. The power $P_i$ irradiated by patch $i$ is given by

$$P_i = P_i^o + \bar{P}_i^o = P_i^o + \frac{A_i \cdot \bar{a}_i}{\sum_{i=1}^{N} A_i \cdot \bar{a}_i} \cdot \bar{P}^o. \tag{4.6}$$

### 4.4 Image Computation for Nondiffuse Environments

The albedo function allows us to determine the radiances $S_i(\mathbf{V})$ at a point of patch $i$ seen by a viewer from direction $\mathbf{V}$. Again, it is assumed that the surface receives a uniform illumination as a result of the interreflection. Now, using eq. (2.8), (3.11), and (3.12), we have

$$S_i(\mathbf{V}) = S_i^o(\mathbf{V}) + \frac{\bar{P}_i^o}{A_i} \cdot \hat{a}_i(\mathbf{V}). \tag{4.7}$$

In this case the ambient term for the known diffuse case is included. It should be pointed out that the ambient term can be determined without computing the form factors. This method can be seen as the simplest form of the hybrid method that is discussed in Section 5. The first term (4.7), the zero-order component according to (2.6), can be computed at pixel resolution by the classic ray-tracing method, while the effect of interreflections is represented by the second term of the sum in (4.6), the ambient term, as a rough approximation. This method is rather convenient for scenes illuminated by a few point sources, and ray-tracing programs (many are available) are simply complemented by the ambient term. In the nondiffuse case, the radiance or color of the ambient term varies for each pixel within the same surface according to the view vector $\mathbf{V}$.

The drawback of this method is that it does not include either soft penumbra from interreflection or any typical local interreflection effect, depending on the occlusion conditions. It takes into account, however, global interreflection effects such as a global color shift. Namely, any dominant color manifests itself in the mean albedo of the system. This is also enhanced by the resultant interreflection effect, especially for very bright environments. As an illustration consider a system with a mean albedo (.7, .8, .9) corresponding to the color components RGB, illuminated by a point source. The scene is then bright and of rather unsaturated color with a B/R ratio of $9/7 = 1.222$. With an approximation of the ambient term, the color resulting from interreflection becomes (3.33, 5, 10); that is, the ratio B/R goes to 3! Therefore, the resulting color will be more saturated in the most intense color component. Bright environments with cold white surfaces will look bluish, while a warm white will acquire an overall brownish color because of the ambient term.

### 4.5 Other Generalizations

An important extension for the ambient term is as follows: Rather than using a shooting-type method alone, every iterative procedure can be complemented by ambient terms. Consider a radiosity equation system with arbitrary power

variables to be written in the usual form $\mathbf{Cx} = \mathbf{b}$. For an arbitrary approximate solution $\mathbf{x}_{app}$, row $i$ is affected by an error $e_i = b_i - \mathbf{c}_i\mathbf{x}_{app}$ ($i = 1, 2, \ldots, N$). The solution can be directly obtained using a Jacobi, Southwell, or gradient-type solution. Separating $\mathbf{x}_{app}$, the missing component $\mathbf{x}_{res}$ of the solution is sought as a solution of $\mathbf{Cx}_{res} = \mathbf{e}$, where $\mathbf{e} = (e_1, e_2, \ldots, e_N)$. Vector $\mathbf{e}$ in the equation may be considered as the zero-order illumination component of the modified problem. For this modified problem, the ambient term can be determined as described earlier. Any procedure suitable for a convergent system of equations can be complemented by an ambient term updated in a stepwise fashion. Obviously, when $\mathbf{x}_{app} \rightarrow \mathbf{x}$, that is, when $|\mathbf{e}| \rightarrow 0$, the ambient term goes to zero and gradually fades as we approach the real solution. Note that $e_i$ may be negative, for instance, with the gradient-type methods. These zero-order components are then power sinks, rather than sources.

A qualitatively new form for the ambient term, better suited to the radiosity problem, can be written even for very complex environments by grouping the surfaces. A linear system of equations may be written for the power transport between groups. The total power within a group is distributed between the surfaces by analogy with the classic ambient-term method. The equations above, in particular, (4.4), realize the case where the surfaces have been put in a single group. The simplest grouping is defined by the histogram of $A_i a_i$ or orientation of $\mathbf{N}_i$, etc. A generalized ambient-term method, taking the occlusion conditions into consideration, and, therefore, half way between the ambient term and the radiosity method, will be reported in a subsequent paper by the same authors.

## 5. HYBRID METHODS

### 5.1 Comparison between Radiosity and Ray Tracing

Ray-tracing or radiosity methods may involve errors usually avoidable by a proper combination of the two methods. That is the motivation for hybrid methods. The ray-tracing method is suitable to compute an image from a given viewpoint at high resolution, but when taking increasing levels of interreflection into account, the variance of the radiance in the image increases abruptly. More precisely, the cost of computation for a given error tolerance increases almost exponentially and cannot be dealt with by variance reduction methods. The increase of the variance is relatively slower for highly specular reflectance. This is exactly the property two-pass methods make use of.

In the radiosity method, the rendering equation is approximated by the radiosity equations. The radiosity solution is view-independent; that is, it produces a complete solution, but unfortunately, only at patch level and, therefore, at low resolution. Partial images of order $j$ (containing exactly $j$ patch-to-patch interreflections) decrease in accuracy with increasing $j$ values. This is especially true for highly specular reflectance in a radiosity system for an environment consisting of mirrors. The mirror images of emissive patches will be larger and with a flatter light distribution, involving

more and more patches. The system of equations yields initially an approximate matrix, and the successive powers of this matrix are increasingly rough approximations of the corresponding operators for the infinite problem. This *degeneration* is a phenomenon characteristic of the radiosity method. It is less serious for diffuse materials, but very relevant for very bright and specular environments. The radiosity method is nevertheless widely applicable, and its main advantage is that the solution of its equation system is a complete approximate solution that includes all of the power transported by interreflection processes.

The hybrid method we will present combines the advantages of the two basic procedures described above. The advantage of the ray-tracing procedure is that it works at pixel resolution and that the variance is still acceptable for first- or second-order interreflection. On the other hand, the radiosity method offers a total solution, and the higher-order reflection effects are obtained at a bearable cost, even if biased by degeneration.

Specifically, for two-pass methods we use the decomposition into diffuse and specular reflectance. The variance of the specular part can be managed by ray tracing. The diffuse part, a single diffuse reflection of the complete solution, is computed by the radiosity method, with a degeneration less than that for the complete solution. Note that, for bright environments containing many highly reflective specular surfaces, even the hybrid methods give an approximation with large errors.

### 5.2 Definitions

In addition to the concepts and notations introduced in Section 2, we now introduce some new ones.

The convergent conditions of the Neumann series have been discussed in Section 2.5. If these are met, the image can be computed from eq. (2.15):

$$\mathbf{s} = \sum_{j=0}^{\infty} \mathscr{R}^j \mathbf{s}_0. \tag{5.1}$$

Although $\mathbf{s}$ is the vector of radiances within the complete scene and the image is the vector of radiances from a single viewpoint, for the sake of simplicity we will call $\mathbf{s}$ an image. $\mathbf{s}$ as in (5.1) is called the *complete image*, and component $\mathscr{R}^j \mathbf{s}_o$, a *partial image of order j*. It consists of rays involved in exactly $j$ interreflections between surfaces. The zero-order illumination component $\mathbf{s}_o$ is identical to the partial zero-order image. It includes emissive surfaces and the effect of point light sources. Their energies get into the system by reflection or refraction, but there is no interreflection between surfaces and, hence, no power term involving $\mathscr{R}$. The sum of the partial images of order 0, 1, 2, ..., $K$ is the image of order $K$:

$$\mathbf{s}_K = \sum_{j=0}^{K} \mathscr{R}^j \mathbf{s}_o. \tag{5.2}$$

The difference between the complete image and the image of order $K$ is the *residual image of order K*:

$$\bar{\mathbf{s}}_K = \mathbf{s} - \mathbf{s}_K = \sum_{j=K+1}^{\infty} \mathscr{P}^j \mathbf{s}_0. \tag{5.3}$$

In finite approximations, Jacobi iterations are the counterpart of the Neumann series. It is the method producing $K$-order images from zero-order illuminations as initial approximations.

## 5.3 Direct Residual Image Method

5.3.1 *Fundamental Case.* This section deals with the simplest hybrid method appropriate for diffuse or separable environments illuminated by point light sources. The efficiency of this method resides in the ability to compute the image given the radiosity solution while obtaining the sharp shadow edges from point sources at pixel accuracy without further adaptive refinement. The zero-order image $\mathbf{s}_0$ including the direct effects of point sources can be computed by ray tracing at pixel resolution. It is natural to complement this sharp image at high resolution by the zero-order residual image component $\bar{\mathbf{s}}_0$, including all of the interreflection effects. By definition, the zero-order residual image $\bar{\mathbf{s}}_0 = \mathbf{s} - \mathbf{s}_0$ is the overall radiosity solution minus the constant vector of the system. Its image at pixel resolution can be computed by interpolation, as in Gouraud shading. For a diffuse environment, obviously radiosity may be used instead of radiance. The method is especially advantageous for images with a large zero-order component and many shadow boundaries. This method is also suitable for textures at pixel resolution. This is obvious for the zero-order component. For the residual image in the diffuse case, for a given patch at a given pixel the interpolation yields the radiosity $B_{ave}$, without the texture. Since, in the radiosity approach $\rho_{ave}$ for the patch is the area weighted average of $\rho$ within the patches, in the computation of the residual image the radiosity for a given pixel is given by

$$B_{pixel} = B_{ave} \cdot \frac{\rho_{pixel}}{\rho_{patch}}. \tag{5.4}$$

Remember that, instead of the complete radiosity solution $\mathbf{s}$, its approximations by the first- or second-order image and their ambient terms can be used. In some cases, the ambient term as a rough approximation of interreflection is enough, providing that the zero-order image is dominant (see Section 4.3).

In the general case, the direct residual image method relies on the decomposition $\mathbf{s}_K + \bar{\mathbf{s}}_K$. For a low $K$ value (in practice, $K < 2$), the component $\mathbf{s}_K$ can be computed at pixel resolution by distributed ray tracing, while the residual image $\bar{\mathbf{s}}_K$ is computed by interpolation. In the case $K > 0$, the residual image can be computed by solving the radiosity problem by Jacobi iteration up to a $K$-order image, which is then stored. The iteration may then be solved by another method, and finally, the $K$-order image would be obtained from this solution.

The disadvantage of the direct residual image method is that the residual image can be dim and that the fine penumbra transitions are often incorrectly rendered at pixel resolutions. This can be helped by adaptive refinement of the residual image, but this is costly. The necessity to use a local threshold value for the brightness gradient if refinement is used forces us to take into account the brightness value of the image component determined by ray tracing in order to avoid overrefining. Further improvements can be obtained by the method presented below, where the residual image is not directly used; instead, its first- or higher-order reflection obtained by distributed ray tracing is used.

5.3.2 *Direct Residual Image Method with Extended Radiosity.* An important generalization is based on the extended radiosity method of Rushmeier and Torrance [1990]. It permits the inclusion of some mirrors (and/or refractions) in the radiosity system, by means of an *extended interpretation* of form factors. The drawback of this method is in the exponential increase of the number of virtual viewpoints or virtual worlds as a function of the number of mirrors, with a proportional increase in the cost of computation. In the case of extended radiosity, the zero-order ray-traced image may be obtained by the classic tree-graph ray-tracing method of Bouville et al. [1985], which also includes the effects of emissive surfaces and point sources across mirrors and refractions. The difference between the complete solution of the extended radiosity problem and the extended zero-order component is the zero-order residual image with extended interpretation. The extended interpretation may also be applied to the hybrid method below.

## 5.4 Coupling Methods

These methods rely on performing some ray reflections by distributed ray tracing, and then *coupling* the ray to the complete radiosity image. The radiosity image results from all of the rays sharing any number of interreflections. The ray coupled to this network automatically includes all higher-order image components. We then have

$$\mathscr{R}^K \mathbf{s} = \mathscr{R}^K \cdot \sum_{j=0}^{\infty} \mathscr{R}^j \mathbf{s}_0 = \sum_{j=K}^{\infty} \mathscr{R}^j \mathbf{s}_0 = \bar{\mathbf{s}}_{K-1}. \tag{5.5}$$

Using eq. (5.5), the complete image is

$$\mathbf{s} = \mathbf{s}_{K-1} + \bar{\mathbf{s}}_{K-1} = \mathbf{s}_o + \mathscr{R}\mathbf{s}_o + \cdots + \mathscr{R}^{K-1}\mathbf{s}_o + \mathscr{R}^K \mathbf{s}. \tag{5.6}$$

Equation (5.6) summarizes the coupling method. The distributed image of order $(K-1)$ is determined by distributed ray tracing. The details of this method have been presented, for instance, by Kajiya [1986]. The first step is to compute the zero-order image $\mathbf{s}_o$. For each pixel, a ray starts from the viewpoint to the closest surface in the scene. Here, it assumes the value of the zero-order illumination component at that point according to eq. (2.6). From this first intersection point, the ray branches out in several directions as in distributed ray tracing, again producing intersection points, and so on itera-

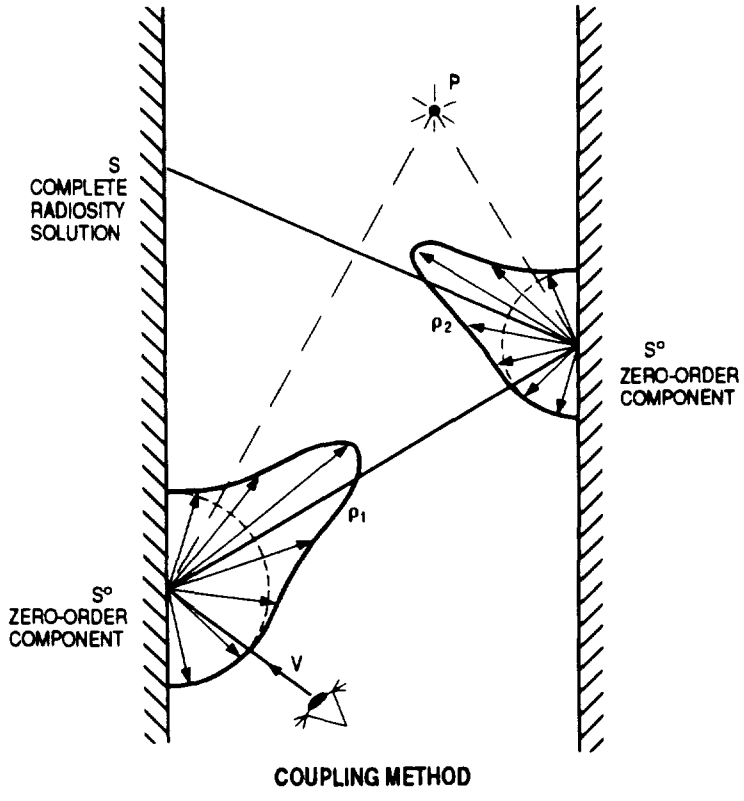


Figure 4

tively. The algorithm involves repetition of two types of fundamental operations: multiplication by the actual bidirectional reflectance values (belonging to the pair of directions before and after reflections) at the actual reflection point of the rays, and multiplications by the zero-order component at the actual intersection point. For more details about this algorithm and adaptive refinement to reduce the variance, see Kajiya 1986. Once the $(K - 1)$-order image is thus produced, the next step is similar to that of computing the $K$-order image, but in this step the rays at the intersection points are multiplied by the complete solution $\mathbf{s}$ known from the radiosity solution. The latter is the total interreflection radiance, as opposed to $\mathbf{s}_0$ as used before. This is the step where coupling to the radiosity solution, the gist of the method, occurs. The $\mathbf{s}$ is known at patch resolution, so its value at a given point has to be interpolated. It should be stressed that in the last step the coupling is to the complete solution $\mathbf{s}$ rather than to some residual image of order $K$, as might be assumed for reasons of complementarity. The method's efficiency can be compared with importance sampling techniques and with the selection of a different appropriate $K$ values for each ray.

Let $K = 1$ in eq. (5.6). Now the image is in the form $\mathbf{s}_0 + \mathscr{R}\mathbf{s}$, where $\mathbf{s}_0$ is a classic zero-order ray-traced image at pixel resolution. The term $\mathscr{R}\mathbf{s}$ may be

computed by adaptive refinement at pixel resolution. This can be done by the method of Cohen et al. [1986] for diffuse environments. If $\mathbf{s}_o$ contains only the diffuse emissive component, then in the case where $K = 1$, the coupling method is the same as that of Cohen et al. [1986]. For non-diffuse environments, the integral $\mathscr{R}\mathbf{s}$ is not strictly computed by patchwise summation, since in the important directions one patch may require several samples. In this case, for better accuracy the method introduced in Section 5.3 may be applied, with the residual term $\bar{\mathbf{s}}_o$ at the given point computed from patchwise values by interpolation. The term $\mathbf{s}_o$ including the effects of texture and anisotropic light sources may be handled separately.

Let us compare the direct residual image method and the case $K = 1$ of the coupling method. Especially for non-diffuse environments, there is a marked difference between the direct image equation and the coupling equation. In the latter the patches blurred by interpolation are not seen directly, but only through the reflections by $\mathscr{R}$. This is an enormous difference mainly because reflectance and occlusion conditions may vary abruptly at each pixel. The advantage of coupling may be illustrated by the fact that when perceived across a very blurred reflection an object can be replaced by its rough finite element approximation without visible difference. That is why we can use approximations to $\mathbf{s}$ instead of the complete solution.

Let us have a look at further possibilities of generalization of the coupling method, even though they are still only of theoretical significance.

### 5.5 Coupling to a Residual Image

The coupling method can be further generalized by coupling the ray from ray tracing to the radiosity solution at an intermediary reflection/refraction point, rather than at its end point. This effectively couples the ray to a residual image, and according to eqs. (5.4) and (5.6), this method of *residual image coupling* approximates the real image as

$$\mathbf{s} = \mathbf{s}_o + \mathscr{R}\mathbf{s}_o + \cdots + \mathscr{R}^L\mathbf{s}_o + \mathscr{R}^L\bar{\mathbf{s}}_K + \mathscr{R}^{L+1}\mathbf{s}_o + \cdots + \mathscr{R}^{L+K}\mathbf{s}_o, \qquad (5.7)$$

where the term $\mathscr{R}^L\bar{\mathbf{s}}_K$ is obtained by coupling. $L$ and $K$ are arbitrary nonnegative integers. $L = 0$ corresponds to the direct residual method described in Section 5.3. The case $K = 0$ is also of interest and constitutes an exact handling of the zero-order component $\mathbf{s}_o$ at the ray end point. This case can be applied to eqs. (5.6) and (5.9), as seen below.

### 5.6 Complete Two-Pass Method

In Section 5.1 we have compared the advantages and disadvantages of ray tracing and radiosity. The two-pass methods take advantage of diffuse reflection to limit the radiosity result to a small degeneration, while the purely specular (but not totally mirrorlike) reflectance results in an acceptable variance even after several steps of distributed ray tracing. In this approach, the smaller the cone of *important directions* determined by the specular reflection (see Section 3.3), the less is the variance.

Let us again decompose the bidirectional reflectances of some patches in the scene into diffuse and specular, $\rho = \rho^D + \rho^S$. Other decompositions are possible, as long as they create one separable term of flat response and one term above some threshold $\epsilon$ only over a small solid angle. Transmittance can also be included with both terms of the decomposition. The reflection operator $\mathcal{R}$ characteristic of the system has a decomposition of the form $\mathcal{R} = \mathcal{D} + \mathcal{S}$.

The Neumann series for the solution becomes

$$\mathbf{s} = \mathbf{s}_0 + (\mathcal{D} + \mathcal{S})\mathbf{s}_0 + (\mathcal{D} + \mathcal{S})^2\mathbf{s}_0 + \ldots, \tag{5.8}$$

including every possible sequence of diffuse or specular transfer exactly once. Successive applications of the equality $\mathbf{s} = \mathbf{s}_0 + \mathcal{D}\mathbf{s} + \mathcal{S}\mathbf{s}$ show that for $K \geq 0$ we have the relation

$$\mathbf{s} = (\mathcal{S}^0 + \mathcal{S}^1 + \cdots + \mathcal{S}^K) \cdot (\mathbf{s}_0 + \mathcal{D}\mathbf{s}) + \mathcal{S}^{K+1}\mathbf{s}. \tag{5.9}$$

The first term on the right-hand side contains the chains starting with at most $K$ times $\mathcal{S}$, and the second term, those starting with at least $K + 1$ times $\mathcal{S}$. Eq. (5.9) suggests effective ways to apply ray tracing and radiosity. The value of $K$ is limited in practice by the increase in variance during ray tracing, so let us consider the case $K = 1$ as an illustration. Eq. (5.9) becomes

$$\mathbf{s} = \mathbf{s}_0 + \mathcal{D}\mathbf{s} + \mathcal{S}(\mathbf{s}_0 + \mathcal{D}\mathbf{s}) + \mathcal{S}^2\mathbf{s}. \tag{5.10}$$

From eq. (5.10) we conclude that the procedure comprises the following steps:

(1) A complete radiosity solution for the general non-diffuse problem, computing the spectral radiances $\mathbf{s}$ for pairs of patches.

(2) The determination of $\mathcal{D}\mathbf{s}$ from $\mathbf{s}$ known at patch resolution. This involves a single diffuse interreflection, to be computed at patch resolution by matrix multiplications (the form factors have to be known).

(3) The computation of image $\mathbf{s}_0$ at pixel resolution. This is the classic ray-traced image of the zero-order component, including direct reflection from point sources in addition to the direction-dependent emission.

(4) The computation of the image for the directly visible component $\mathcal{D}\mathbf{s}$ at pixel resolution, from the value at patch resolution, either by Gouraud shading or by adaptive refinement based on the brightness gradient.

(5) The computation of the first term with operator $\mathcal{S}$. This is a distributed ray-tracing step using the specular components $\rho^S$. It can be done, for example, using a depth-buffer within the reflection cone (cf. Wallace et al. [1987]). We compute the first intersection point with the scene of a ray centered at each pixel. At this point the ray branches out, and at the end of these secondary rays, when they intersect another surface, the primary rays within the cone are multiplied by the auxiliary image value $(\mathbf{s}_0 + \mathcal{D}\mathbf{s})$. The first term $\mathbf{s}_0$ of the auxiliary image includes emission of patches and tracing to unoccluded light sources; the second term is computed by interpolation at the given point from the stored $\mathcal{D}\mathbf{s}$ values for patches. At

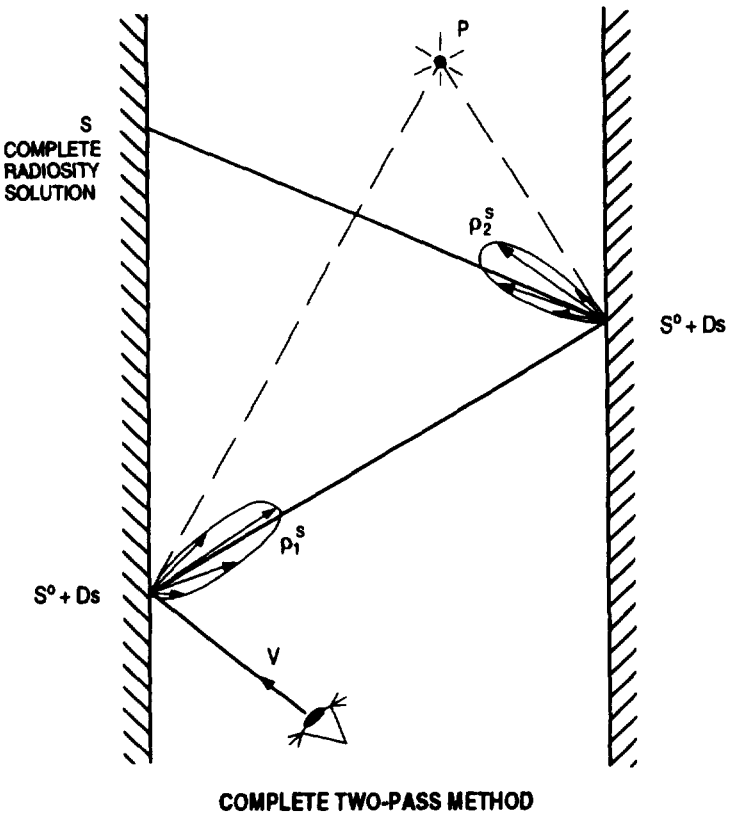


Figure 5

this level, interpolation is sufficient, and there is no need for adaptive refinement since $\mathscr{D}\mathbf{s}$ is obtained indirectly through specular reflection.

In the general case, step (5) has to be repeated $K$ times.

(6) The tree of distributed rays is continued within the narrow cones determined by $\rho^S$. When the rays intersect surfaces, they are multiplied by the complete solution **s**, obtained by interpolation of the complete non-diffuse radiosity solution. This computes the term $\mathscr{S}^2\mathbf{s}$ in eq. (5.10).

We can make the following remarks about this algorithm: There are several efficient methods available for solving the non-diffuse radiosity problem in step (1). The methods utilizing diffuse and specular decomposition, in addition to that by Shao et al. [1988], are the effective gathering and shooting method introduced by Neumann and Neumann or the double-patch method described in Section 3.3.

The ray tracing in steps (5) and (6) consists of generating uniformly distributed directions inside the cone and then multiplying by $\rho^S$ at the point of ray intersection.

Step (6) involves coupling to $\mathbf{s}$. This can be refined by the decomposition $\mathbf{s}_0 + \bar{\mathbf{s}}_0$ discussed earlier where $\mathbf{s}_0$ is exactly known, so that it is sufficient to use a less exact interpolated approximation for the term $\bar{\mathbf{s}}_0$.

Let us compare the above procedure to the two types of two-patch methods known from the literature. The two-pass method has been introduced by Wallace et al. [1987]. Their diffuse and specular transfer chains miss many terms. In particular, between each pair of diffuse pixels a single specular transfer was allowed that in practice restricts it to mirrorlike reflection. Curiously, in spite of this restriction, this method resulted in rather spectacular images. As generalized by Sillion and Puech [1989], the general two-pass method theoretically involves all of the possible diffuse and specular terms. The only serious restriction, occurring also in Wallace et al. 1987, is that emission can only be diffuse, that is, Lambertian. As well, out of all of the possible purely specular chains between diffuse–diffuse transfer this method takes only few into account. This is achieved by computing $O(N^2)$ additional form factors by ray tracing. Even then this short chain can only be efficiently determined for the mirrorlike specular component or refractive transmittance. Of course, in cases where the ratio of specular to diffuse components is low, or if all of the system is of moderate brightness, the images made by this method are hardly distinguishable from the exact complete solution, since anything omitted is negligible.

In the method presented, the full radiosity solution or its default has been used, taking all possible diffuse–specular into consideration. The $\mathcal{S}^{K-1}\mathbf{s}$ terms in (5.9) take into account the higher-order reflection terms, for instance, for an object made entirely of shiny metal. Its decisive advantage is with non-diffuse emission with the inclusion of the zero-order component presented in Section 2. This then unites the application of point light sources preferred in pure ray tracing and of emissive surfaces preferred in classic radiosity, and keeps the definite advantages of the two-pass approach.

## 6. CONCLUSIONS

After considering various solutions for radiosity equations and hybrid methods, the question now is which to apply and when? In the case of diffuse and separable reflectance, as well as for the case where decomposition into diffuse and specular is not appropriate, the Southwell algorithm is the most effective for dark- or medium-bright systems. For very bright systems where slowly decaying higher-order interreflections are significant, the conjugate gradient method, in the form introduced in this paper, is the most effective. In radiosity, most of the difficulties are due to the general bidirectional reflectance. If it can be decomposed into diffuse and specular components (with a small solid angle), then several solution methods are available. Two methods seem to be the most effective: (1) the double-patch method, solved by the Southwell algorithm; or (2) the sorted gathering and shooting method introduced earlier [Neumann and Neumann 1990]. For very bright environments, the solution is either the double-patch method or the power form equation introduced in Neumann and Neumann [1990], solved by the conjugate gradi-

ent method. It means that for every case there is a solution more efficient than the Gauss–Seidel method. Jacobi iteration is useful if the $K$-order image is to be used by itself, for instance, in the hybrid methods. All of the radiosity methods may be complemented by the generalized ambient term.

We have discussed extensively hybrid methods, which compute the image with high-resolution ray tracing, starting from the low-resolution radiosity solution. The simplest such method is that of direct residual images, yielding a fast solution for diffuse scenes with point light sources. Another new method presented is the coupling method, a general form of ray tracing coupled with a radiosity solution. Last, the complete two-pass method using diffuse and specular decomposition has been presented, allowing anisotropic emissive surfaces and spotlights, as opposed to earlier two-pass approximations. In particular, it includes all of the possible diffuse–specular permutations.

Beside giving the advantages of these methods, this paper also points to the limits of their applicability. For ray tracing, it is primarily the increase of variance; for radiosity, it is the phenomenon of degeneration.

The many methods discussed provide efficient new algorithms coping with a wide range of practical problems. We are still faced with the problem of undertaking comparative numerical testing of the large number of available radiosity and hybrid methods in standard environments. The test results should then be included in a decision tree or in an expert system to guide one toward an optimal solution method for actual problems.

## ACKNOWLEDGMENTS

The authors acknowledge the helpful comments of the anonymous reviewers, and the considerable efforts of Alain Fournier and Bob Lewis in improving the presentation of the results, which made publication of this paper possible.

## REFERENCES

BOUVILLE, C., BRUSQ, R., DUBOIS, J. L., AND MARCHAL, I. 1985. Generating high quality pictures by ray-tracing. *Comput. Graph. Forum 4*, 87–99.

PHONG, B.-T. 1975. Illumination for computer-generated pictures. *Commun. ACM 18*, 6, (June), 311–317.

CHANDRASEKHAR, S. 1960. *Radiative Transfer*. Dover Publication, New York.

COHEN, M. F., AND GREENBERG, D. P. 1985. The hemi-cube: A radiosity solution of complex environments. In Proceedings of SIGGRAPH 85. *Comput. Graph. 19*, 3 (July), 31–41.

COHEN, M. F., GREENBERG, D. P., IMMEL, D. S., AND BROCK, P. J. 1986. An efficient radiosity approach for realistic image synthesis. *IEEE Comput. Graph. Appl. 6*, 3 (March), 26–35.

COHEN, M. F., SHENCHANG, E. C., WALLACE, J. R., AND GREENBERG, D. P. 1988. A progressive refinement approach to fast radiosity image generation. In Proceedings of SIGGRAPH 88. *Comput. Graph. 22*, 4 (Aug.), 75–84.

COOK, R. L. 1986. Stochastic sampling in computer graphics. *ACM Trans. Graph. 5*, 1, 51–72.

COOK, R. L., PORTER T., AND CARPENTER L. 1984. Distributed ray tracing. In Proceedings of SIGGRAPH 84. *Comput. Graph. 18*, 3 (July), 137–145.

GORAL, C. M., TORRANCE, K. E., GREENBERG, D. P., AND BATTAILE B. 1984. Modelling the interaction of light between diffuse surfaces. In Proceedings of SIGGRAPH 84. *Comput. Graph. 18*, 3 (July), 213–222.

GORTLER, S. J., COHEN, M., AND SLUSSALLEK, P. 1993. Radiosity and relaxation methods: Progressive refinement is Southwell relaxation. Tech. Rep. CS-TR-408-93, Dept. of Computer Science, Princeton Univ., Princeton, N.J., Feb.

HANRAHAN, P., AND SALZMAN, D. 1990. A rapid hierarchical radiosity algorithm for unoccluded environments. In *Proceedings of Eurographics Workshop on Photosimulation, Realism and Physics in Computer Graphics* (Rennes, France, June), 151–170.

IMMEL, D. S., COHEN, M. F., AND GREENBERG, D. P. 1986. A radiosity method for non-diffuse environments. In Proceedings of SIGGRAPH 86. *Comput. Graph. 20*, 4, 133–142.

KAJIYA, J. T. 1986. The rendering equation. In Proceedings of SIGGRAPH 86. *Comput. Graph. 20*, 4 (Aug.), 143–150.

KOSA, A. 1979. *Optimumszàmitàsi modelle K (Optimization Models.)* Müszaki Köngvkiadò, Budapest, Hungary, 113–116.

KREKÓ, B. 1976. Lineáris algebra (Linear algebra). Jogi és Közgazdasági Könyvkiadó, Budapest, Hungary, 497–501.

LEE, M. E., REDNER, R. A., AND USELTON, S. 1985. Statistically optimized sampling for distributed ray tracing. In Proceedings of SIGGRAPH 85. *Comput. Graph. 19*, 3 (July), 61–67.

MARCSUK, G. I. 1976. A gépi matematika numerikus módszerei (Numerical methods for computerized mathematics). Müszaki Könyvkiad. Budapest, Hungary.

SHAO, M.-Z., PENG, O.-S., AND LIANG, Y.-D. 1988. A new radiosity approach by procedural refinements for realistic image synthesis. In Proceedings of SIGGRAPH 88. *Comput. Graph. 22*, 4 (Aug.), 93–101.

NEUMANN, L., AND NEUMANN, A. 1989. Photosimulation: Interreflection with arbitrary reflectance models and illumination. *Comput. Graph. Forum 8*, 1, 21–34.

NEUMANN L., AND NEUMANN A. 1990. Efficient radiosity methods for non-separable reflectance models. In *Proceedings, Eurographics Workshop on Photosimulation, Realism and Physics in Computer Graphics* (Rennes, France, June), 85–97.

NISHITA, T., AND NAKAMAE, E. 1985. Continuous tone representation of three-dimensional objects taking account of shadows and interreflection. In Proceedings of SIGGRAPH 85. *Comput. Graph. 19*, 3 (July), 22–30.

NISHITA, T., AND NAKAMAE, E. 1986. Continuous tone representation of three-dimensional objects illuminated by sky light. In Proceedings of SIGGRAPH 86. *Comput. Graph. 20*, 3 (Aug.), 125–132.

PURGATHOFER, W. 1987. A statistical method for adaptive stochastic sampling. *Comput. Graph. 11*, 2, 157–162.

RUSHMEIER, H. E. 1986. Extending the radiosity method to transmitting and specularly reflecting surfaces. Masters Thesis, Cornell Univ., Ithaca, N.Y.

RUSHMEIER, H. E., AND TORRANCE K. E. 1990. Extending the radiosity method to include specularly reflecting and translucent materials. *ACM Trans. Graph. 9*, 1 (Jan.), 1–27.

SILLION, F., AND PUECH, C. 1989. A general two-pass method integrating specular and diffuse reflection. In Proceedings of SIGGRAPH 89. *Comput. Graph. 23*, 4 (July), 335–344.

WALLACE, J. R., COHEN, M. F., AND GREENBERG, D. P. 1987. A two-pass solution to the rendering equation: A synthesis of ray tracing and radiosity methods. In Proceedings of SIGGRAPH 87. *Comput. Graph. 21*, 4 (July), 311–320.

WALLACE, J. R., ELMQUIST, K. A., AND HAINES, E. A. 1989. A ray tracing algorithm for progressive radiosity. Proceedings of SIGGRAPH 89. *Comput. Graph. 23*, 4 (July), 315–324.

YOUNG, D. M. 1971. *Iterative Solution of Large Linear Systems*. Academic Press, New York.

Received August 1989; revised March 1991, October 1994, and June 1995; accepted June 1995

Editor: Alain Fournier