Chapter 3

PHYSICAL MODEL OF 3D IMAGE SYNTHESIS

3.1 Definition of color

Light is an electromagnetic wave, and its color is determined by the eye's perception of its spectral energy distribution. In other words, the color is determined by the frequency spectrum of the incoming light. Due to its internal structure, the eye is a very poor spectrometer since it actually samples and integrates the energy in three overlapping frequency ranges by three types of photopigments according to a widely accepted (but also argued) model of the eye. As a consequence of this, any color perception can be represented by a point in a three-dimensional space, making it possible to define color perception by three scalars (called **tristimulus** values) instead of complete functions.

A convenient way to define the axes of a coordinate system in the space of color sensations is to select three wavelengths where one type of photopigment is significantly more sensitive than the other two. This is the method devised by Grassmann, who also specified a criterion for separating the three representative wavelengths. He states in his famous laws that the representative wavelengths should be selected such that no one of them can be matched by the mixture of the other two in terms of color sensation. (This criterion is similar to the concept of linear independence.)

An appropriate collection of representative wavelengths is:

$$\lambda_{\text{red}} = 700 \text{ nm}, \quad \lambda_{\text{green}} = 561 \text{ nm}, \quad \lambda_{\text{blue}} = 436 \text{ nm}. \tag{3.1}$$

Now let us suppose that monochromatic light of wavelength λ is perceived by the eye. The equivalent portions of red, green and blue light, or (r, g, b) tristimulus values, can be generated by three **color matching functions** ($r(\lambda)$, $g(\lambda)$ and $b(\lambda)$) which are based on physiological measurements.

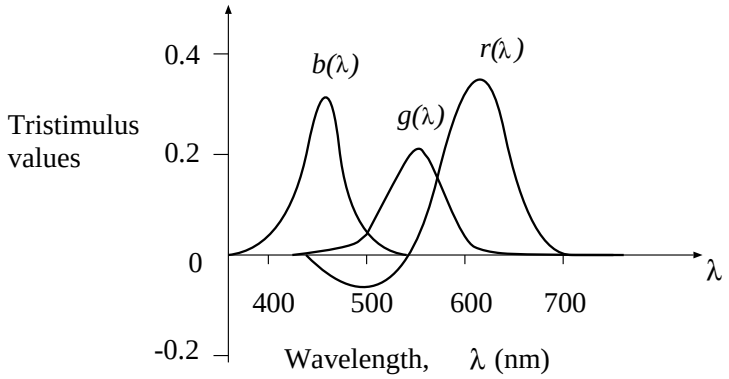

Figure 3.1: Color matching functions $r(\lambda)$, $g(\lambda)$ and $b(\lambda)$

If the perceived color is not monochromatic, but is described by an $L(\lambda)$ distribution, the tristimulus coordinates are computed using the assumption that the sensation is produced by an additive mixture of elemental monochromatic components:

$$r = \int_{\lambda} L(\lambda) \cdot r(\lambda)\, d\lambda, \qquad g = \int_{\lambda} L(\lambda) \cdot g(\lambda)\, d\lambda, \qquad b = \int_{\lambda} L(\lambda) \cdot b(\lambda)\, d\lambda. \tag{3.2}$$

Note the negative section of $r(\lambda)$ in figure 3.1. It means that not all the colors can be represented by positive (r, g, b) values, hence there are colors which cannot be produced, only approximated, on the computer screen.

This negative matching function can be avoided by careful selection of the axes in the color coordinate system, and in fact, in 1931 another standard, called the **CIE XYZ** system, was defined which has only positive weights [WS82].

For computer generated images, the color sensation of an observer watching a virtual world on the screen must be approximately equivalent to the color sensation obtained in the real world. If two energy distributions are associated with the same tristimulus coordinates, they produce the same color sensation, and are called **metamers**.

In computer monitors and on television screens red, green and blue phosphors can be stimulated to produce red, green and blue light. The objective, then, is to find the necessary stimulus to produce a metamer of the real energy distribution of the light. This stimulus can be controlled by the (R, G, B) values of the actual pixel. These values are usually positive numbers in the range of [0...255] if 8 bits are available to represent them.

Let the distribution of the energy emitted by red, green and blue phosphors be $P_R(\lambda, R)$, $P_G(\lambda, G)$ and $P_B(\lambda, B)$, respectively, for a given (R, G, B) pixel color. Since the energy distribution of a type of phosphor is concentrated around wavelength λ_{red}, λ_{green} or λ_{blue}, the tristimulus coordinates of the produced light will look like this:

$$r = \int_\lambda (P_R + P_G + P_B) \cdot r(\lambda)\, d\lambda \approx \int_\lambda P_R(\lambda, R) \cdot r(\lambda)\, d\lambda = p_R(R), \quad (3.3)$$

$$g = \int_\lambda (P_R + P_G + P_B) \cdot g(\lambda)\, d\lambda \approx \int_\lambda P_G(\lambda, G) \cdot g(\lambda)\, d\lambda = p_G(G), \quad (3.4)$$

$$b = \int_\lambda (P_R + P_G + P_B) \cdot b(\lambda)\, d\lambda \approx \int_\lambda P_B(\lambda, B) \cdot b(\lambda)\, d\lambda = p_B(B). \quad (3.5)$$

Expressing the necessary R, G, B values, we get:

$$R = p_R^{-1}(r), \qquad G = p_G^{-1}(g), \qquad B = p_B^{-1}(b). \quad (3.6)$$

Unfortunately p_R, p_G and p_B are not exactly linear functions of the calculated R, G and B values, due to the non-linearity known as γ-distortion of color monitors, but follow a $\text{const} \cdot N^\gamma$ function, where N is the respective R, G or B value. In most cases this non-linearity can be ignored, allowing $R = r$, $G = g$ and $B = b$. Special applications, however, require compensation for this effect, which can be achieved by rescaling the R, G, B values by appropriate lookup tables according to functions p_R^{-1}, p_G^{-1} and p_B^{-1}. This method is called **γ-correction**.

Now we can focus on the calculation of the $(r,\ g,\ b)$ values of the color perceived by the eye or camera through an (X,Y) point in the window.

According to the laws of optics, the virtual world can be regarded as a system that transforms a part of the light energy of the lightsources $(P_{\text{in}}(\lambda))$ into a light beam having energy distribution $P_{XY}(\lambda)$ and going to the camera through pixel (X,Y). Let us denote the transformation by functional L:

$$P_{XY}(\lambda) = L(P_{\text{in}}(\lambda)). \tag{3.7}$$

A tristimulus color coordinate, say r, can be determined by applying the appropriate matching function:

$$r_{XY} = \int_{\lambda} P_{XY}(\lambda) \cdot r(\lambda)\ d\lambda = \int_{\lambda} L(P_{\text{in}}(\lambda)) \cdot r(\lambda)\ d\lambda. \tag{3.8}$$

In order to evaluate this formula numerically, $L(P_{\text{in}}(\lambda))$ is calculated in discrete points $\lambda_1, \lambda_2, ..., \lambda_n$, and rectangular or trapezoidal integration rule is used:

$$r_{XY} \approx \sum_{i=1}^{n} L(P_{\text{in}}(\lambda_i)) \cdot r(\lambda_i) \cdot \Delta\lambda_i. \tag{3.9}$$

Similar equations can be derived for the other two tristimulus values, g and b. These equations mean that the calculation of the pixel colors requires the solution of the shading problem, or evaluating the L functional, for n different wavelengths independently, then the r, g and b values can be determined by summation of the results weighted by their respective matching functions. Examining the shape of matching functions, however, we can conclude that for many applications an even more drastic approximation is reasonable, where the matching function is replaced by a function of rectangular shape:

$$r(\lambda) \approx \hat{r}(\lambda) = \begin{cases} r_{\max} \text{ if } \lambda_{\text{red}} - \Delta\lambda_{\text{red}}/2 \le \lambda \le \lambda_{\text{red}} + \Delta\lambda_{\text{red}}/2 \\ \\ 0 \text{ otherwise} \end{cases} \tag{3.10}$$

Using this approximation, and assuming L to be linear in terms of the energy (as it really is) and $L(0) = 0$, we get:

$$r_{XY} \approx \int_{\lambda} L(P_{\text{in}}(\lambda)) \cdot \hat{r}(\lambda)\ d\lambda = L(\int_{\lambda} P_{\text{in}}(\lambda) \cdot \hat{r}(\lambda)\ d\lambda) = L(r_{\text{in}}), \tag{3.11}$$

where r_{in} is the first tristimulus coordinate of the energy distribution of the lightsources ($P_{\text{in}}(\lambda)$).

This means that the tristimulus values of the pixel can be determined from the tristimulus values of the lightsources. Since there are three tristimulus coordinates (blue and green can be handled exactly the same way as red) the complete shading requires independent calculations for only three wavelengths. If more accurate color reproduction is needed, equation 3.9 should be applied to calculate r, g and b coordinates.

3.2 Light and surface interaction

Having separated the color into several (mainly three) representative frequencies, the problem to be solved is the calculation of the energy reaching the camera from a given direction, i.e. through a given pixel, taking into account the optical properties of the surfaces and the lightsources in the virtual world. Hereinafter, monochromatic light of a representative wavelength λ will be assumed, since the complete color calculation can be broken down to these representative wavelengths. The parameters of the equations usually depend on the wavelength, but for notational simplicity, we do not always include the λ variable in them.

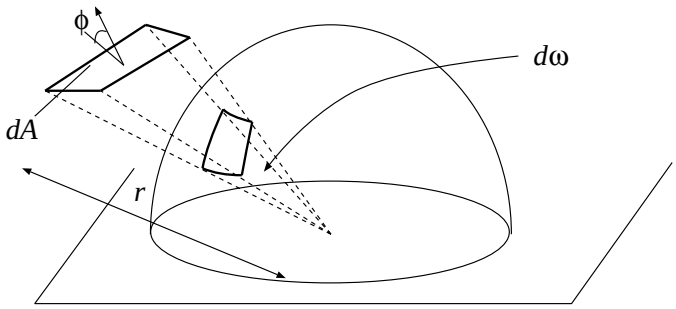

Figure 3.2: Definition of the solid angle

The directional property of the energy emission is described in a so-called **illumination hemisphere** which contains those solid angles to where the surface point can emit energy. By definition, a **solid angle** is a cone or a pyramid, with its size determined by its subtended area of a unit sphere

centered around the apex (figure 3.2). The solid angle, in which a differential dA surface can be seen from point $\vec{p}$, is obviously the projected area per the square of the distance of the surface. If the angle between the surface normal of dA and the directional vector from dA to $\vec{p}$ is ϕ, and the distance from dA to $\vec{p}$ is r, then this solid angle is:

$$d\omega = \frac{dA \cdot \cos\phi}{r^2}. \tag{3.12}$$

The intensity of the energy transfer is characterized by several metrics in computer graphics depending on whether or not the directional and positional properties are taken into account.

The light power or **flux** Φ is the energy radiated through a boundary per unit time over a given range of the spectrum (say $[\lambda, \lambda + d\lambda]$).

The **radiant intensity**, or **intensity** I for short, is the differential light flux leaving a surface element dA in a differential solid angle $d\omega$ per the projected area of the surface element and the size of the solid angle. If the angle of the surface normal and the direction of interest is ϕ, then the projected area is $dA \cdot \cos\phi$, hence the intensity is:

$$I = \frac{d\Phi(d\omega)}{dA \cdot d\omega \cdot \cos\phi}. \tag{3.13}$$

The total light flux radiated through the hemisphere centered over the surface element dA per the area of the surface element is called the **radiosity** B of surface element dA.

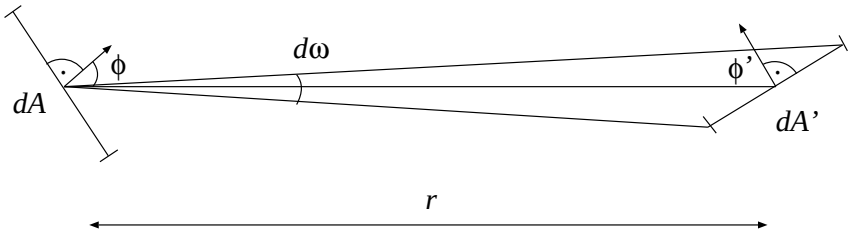

Figure 3.3: Energy transfer between two differential surface elements

Having introduced the most important metrics, we turn to their determination in the simplest case, where there are only two differential surface

elements in the 3D space, one (dA) emits light energy and the other (dA') absorbs it (figure 3.3). If dA' is visible from dA in solid angle $d\omega$ and the radiant intensity of the surface element dA is $I(d\omega)$ in this direction, then the flux leaving dA and reaching dA' is:

$$d\Phi = I(d\omega) \cdot dA \cdot d\omega \cdot \cos\phi. \tag{3.14}$$

according to the definition of the radiant intensity. Expressing the solid angle by the projected area of dA', we get:

$$d\Phi = I \cdot \frac{dA \cdot \cos\phi \cdot dA' \cdot \cos\phi'}{r^2}. \tag{3.15}$$

This formula is called the **fundamental law of photometry**.

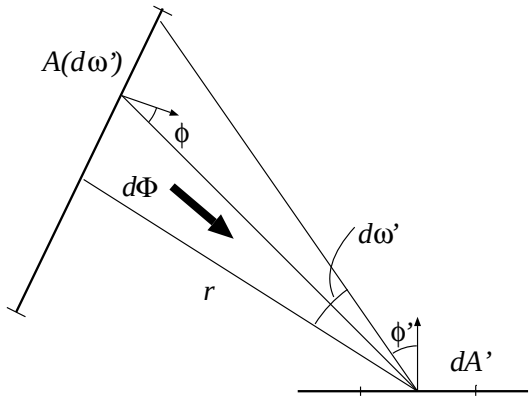

Figure 3.4: Radiation of non-differential surfaces

Real situations containing not differential, but finite surfaces can be discussed using as a basis this very simple case (figure 3.4). Suppose there is a finite radiating surface (A), and we are interested in its energy reaching a dA' element of another surface in the solid angle $d\omega'$. The area of the radiating surface visible in the solid angle $d\omega'$ is $A(d\omega') = r^2 \cdot d\omega'/\cos\phi$, so the flux radiating dA' from the given direction will be independent of the position and orientation of the radiating surface and will depend on its intensity only, since:

$$d\Phi = I \cdot \frac{A(d\omega') \cdot \cos\phi \cdot dA' \cdot \cos\phi'}{r^2} = I \cdot dA' \cdot \cos\phi' \cdot d\omega' = \text{const} \cdot I. \tag{3.16}$$

Similarly, if the flux going through a pixel to the camera has to be calculated (figure 3.5), the respective solid angle is:

$$d\omega_{\text{pix}} = \frac{dA_{\text{pix}} \cdot \cos \phi_{\text{pix}}}{r_{\text{pix}}^2}. \tag{3.17}$$

The area of the surface fragment visible through this pixel is:

$$A(d\omega_{\text{pix}}) = \frac{r^2 \cdot d\omega_{\text{pix}}}{\cos \phi}. \tag{3.18}$$

Thus, the energy defining the color of the pixel is:

$$d\Phi_{\text{pix}} = I \cdot \frac{A(d\omega_{\text{pix}}) \cdot \cos \phi \cdot dA_{\text{pix}} \cdot \cos \phi_{\text{pix}}}{r^2} = I \cdot dA_{\text{pix}} \cdot \cos \phi_{\text{pix}} \cdot d\omega_{\text{pix}} = \text{const} \cdot I. \tag{3.19}$$

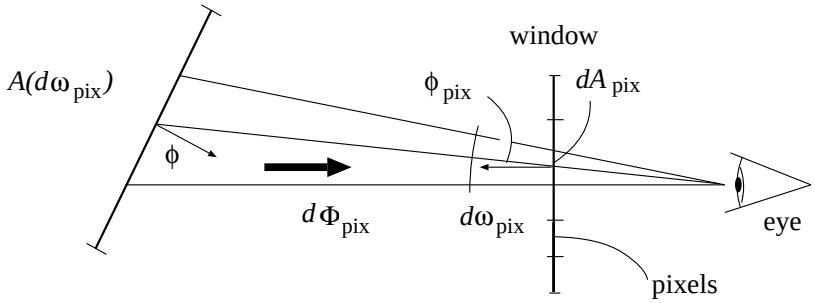

Figure 3.5: Energy transfer from a surface to the camera

Note that the intensity of a surface in a scene remains constant to an observer as he moves towards or away from the surface, since the inverse square law of the energy flux is compensated for by the square law of the solid angle subtended by the surface. Considering this property, the intensity is the best metric to work with in synthetic image generation, and we shall almost exclusively use it in this book.

In light-surface interaction the surface illuminated by an incident beam may reflect a portion of the incoming energy in various directions or it may absorb the rest. It has to be emphasized that a physically correct model must maintain energy equilibrium, that is, the reflected and the transmitted (or absorbed) energy must be equal to the incident energy.

Suppose the surface is illuminated by a beam having energy flux Φ from the differential solid angle $d\omega$. The surface splits this energy into reflected and transmitted components, which are also divided into **coherent** and **incoherent** parts.

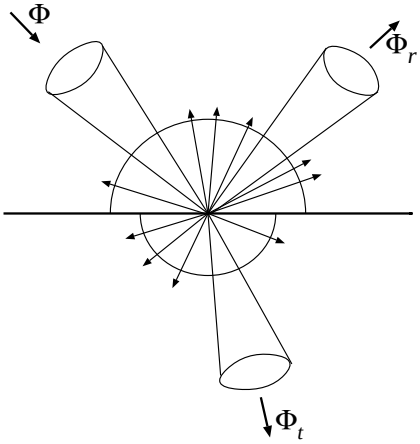

Figure 3.6: Transformation of the incident light by a surface

Optically perfect or smooth surfaces will reflect or transmit only coherent components governed by the laws of geometric optics, including the law of reflection and the Snellius–Descartes law of refraction. If the surface is optically perfect, the portions of reflection (Φ_r) and transmission (Φ_t) (figure 3.6) can be defined by the **Fresnel coefficients** F_r, F_t, namely:

$$\Phi_r = F_r \cdot \Phi, \qquad \Phi_t = F_t \cdot \Phi. \tag{3.20}$$

The energy equilibrium requires $F_t + F_r = 1$.

The incoherent components are caused by the surface irregularities reflecting or refracting the incident light in any direction. Since the exact nature of these irregularities is not known, the incoherent component is modeled by means of probability theory. Assume that a photon comes from the direction denoted by unit vector $\vec{L}$. The event that this photon will leave the surface in the reflection or in the refraction direction being in the solid angle $d\omega$ around unit vector $\vec{V}$ can be broken down into the following mutually exclusive events:

1. if $\vec{L}$ and $\vec{V}$ obey the reflection law of geometric optics, the probability of the photon leaving the surface exactly at $\vec{V}$ is denoted by k_r.

2. if $\vec{L}$ and $\vec{V}$ obey the **Snellius–Descartes law** of refraction — that is

$$\frac{\sin \phi_{\text{in}}}{\sin \phi_{\text{out}}} = \nu,$$

   where ϕ_{in} and ϕ_{out} are the incident and refraction angles respectively and ν is the refractive index of the material — then the probability of the photon leaving the surface exactly at $\vec{V}$ is denoted by k_t.

3. The probability of incoherent reflection and refraction onto the solid angle $d\omega$ at $\vec{V}$ is expressed by the **bi-directional reflection** and **refraction functions** $R(\vec{L},\vec{V})$ and $T(\vec{L},\vec{V})$ respectively:

$$R(\vec{L},\vec{V}) \cdot d\omega = \Pr\{\text{photon is reflected to } d\omega \text{ around } \vec{V} \mid \text{it comes from } \vec{L}\}, \tag{3.21}$$

$$T(\vec{L},\vec{V}) \cdot d\omega = \Pr\{\text{photon is refracted to } d\omega \text{ around } \vec{V} \mid \text{it comes from } \vec{L}\}. \tag{3.22}$$

Note that the total bi-directional probability distribution is a mixed, discrete-continuous distribution, since the probability that the light may obey the laws of geometric optics is non-zero. The energy equilibrium guarantees that the integration of the bi-directional probability density over the whole sphere is 1.

Now we are ready to consider the inverse problem of light-surface interaction. In fact, computer graphics is interested in the radiant intensity of surfaces from various directions due to the light energy reaching the surface from remaining part of the 3D space (figure 3.7).

The light flux (Φ^{out}) leaving the surface at the solid angle $d\omega$ around $\vec{V}$ consists of the following incident light components:

1. That portion of a light beam coming from incident direction corresponding to the $\vec{V}$ reflection direction, which is coherently reflected. If that beam has flux Φ_r, then the contribution to Φ^{out} is $k_r \cdot \Phi_r^{\text{in}}$.

2. That portion of a light beam coming from the incident direction corresponding to the $\vec{V}$ refraction direction, which is coherently refracted. If that beam has flux Φ_t^{in}, then the contribution to Φ^{out} is $k_t \cdot \Phi_t^{\text{in}}$.

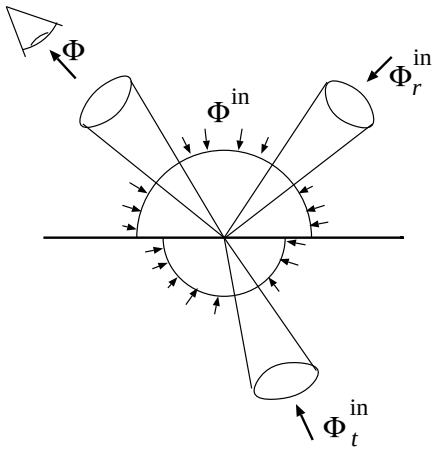

Figure 3.7: Perceived color of a surface due to incident light beams

3. The energy of light beams coming from any direction above the surface (or outside the object) and being reflected incoherently onto the given solid angle. This contribution is expressed as the integration of all the possible incoming directions $\vec{L}$ over the hemisphere above the surface:

$$\int^{2\pi} (R(\vec{L}, \vec{V})\ d\omega)\Phi^{\text{in}}(\vec{L}, d\omega_{\text{in}}). \tag{3.23}$$

4. The energy of light beams coming from any direction under the surface (or from inside the object) and being refracted incoherently onto the given solid angle. This contribution is expressed as the integration of all the possible incoming directions $\vec{L}$ over the hemisphere under the surface:

$$\int^{2\pi} (T(\vec{L}, \vec{V})\ d\omega)\Phi^{\text{in}}(\vec{L}, d\omega_{\text{in}}). \tag{3.24}$$

5. If the surface itself emits energy, that is, if it is a lightsource, then the emission also contributes to the output flux:

$$\Phi^{e}(\vec{V}). \tag{3.25}$$

Adding the possible contributions we get:

$$\Phi^{\text{out}} = \Phi^{e} + k_r \cdot \Phi_r^{\text{in}} + k_t \cdot \Phi_t^{\text{in}} +$$

$$\int^{2\pi} (R(\vec{L},\vec{V})\ d\omega)\Phi^{\text{in}}(\vec{L}, d\omega_{\text{in}}) + \int^{2\pi} (T(\vec{L},\vec{V})\ d\omega)\Phi^{\text{in}}(\vec{L}, d\omega_{\text{in}}). \qquad (3.26)$$

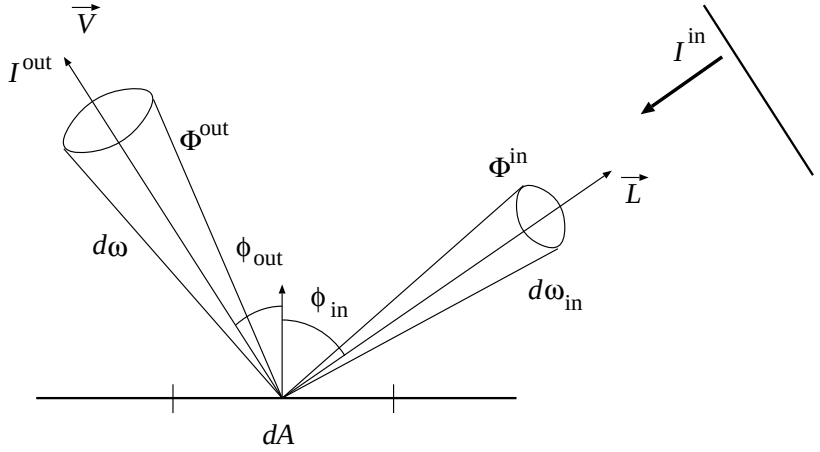

Figure 3.8: Interdependence of intensity of surfaces and the energy flux

Recall that the radiant intensity is the best metric to deal with, so this equation is converted to contain the intensities of surfaces involved. Using the notations of figure 3.8 and relying on equation 3.16, the flux of the incident light beam can be expressed by the radiant intensity of the other surface (I^{in}) and the parameters of the actual surface thus:

$$\Phi^{\text{in}}(\vec{L}, d\omega_{\text{in}}) = I^{\text{in}} \cdot dA \cdot \cos\phi_{\text{in}} \cdot d\omega_{\text{in}}. \qquad (3.27)$$

Applying this equation for Φ_r^{in} and Φ_t^{in} the intensities of surfaces in the reflection direction (I_r^{in}) and in the refraction direction (I_t^{in}) can be expressed.

The definition of the radiant intensity (equation 3.13) expresses the intensity of the actual surface:

$$\Phi^{\text{out}}(\vec{V}, d\omega) = I^{\text{out}} \cdot dA \cdot \cos\phi_{\text{out}} \cdot d\omega. \qquad (3.28)$$

Substituting these terms into equation 3.26 and dividing both sides by $dA \cdot d\omega \cdot \cos\phi_{\text{out}}$ we get:

$$I^{\text{out}} = I_e + k_r \cdot I_r \cdot \frac{\cos\phi_r \cdot d\omega_r}{\cos\phi_{\text{out}} \cdot d\omega} + k_t \cdot I_t \cdot \frac{\cos\phi_t \cdot d\omega_t}{\cos\phi_{\text{out}} \cdot d\omega} +$$

$$\int^{2\pi} I^{\text{in}}(\vec{L}) \cdot \cos\phi_{\text{in}} \cdot \frac{R(\vec{L},\vec{V})}{\cos\phi_{\text{out}}}\, d\omega_{\text{in}} + \int^{2\pi} I^{\text{in}}(\vec{L}) \cdot \cos\phi_{\text{in}} \cdot \frac{T(\vec{L},\vec{V})}{\cos\phi_{\text{out}}}\, d\omega_{\text{in}}. \quad (3.29)$$

According to the reflection law, $\phi_{\text{out}} = \phi_r$ and $d\omega = d\omega_r$. If the refraction coefficient ν is about 1, then $\cos\phi_{\text{out}} \cdot d\omega \approx \cos\phi_t \cdot d\omega_t$ holds.

Using these equations and introducing $R^*(\vec{L},\vec{V}) = R(\vec{L},\vec{V})/\cos\phi_{\text{out}}$ and $T^*(\vec{L},\vec{V}) = T(\vec{L},\vec{V})/\cos\phi_{\text{out}}$, we can generate the following fundamental formula, called the **shading, rendering** or **illumination equation**:

$$I^{\text{out}} = I_e + k_r I_r + k_t I_t +$$

$$\int^{2\pi} I^{\text{in}}(\vec{L}) \cdot \cos\phi_{\text{in}} \cdot R^*(\vec{L},\vec{V})\, d\omega_{\text{in}} + \int^{2\pi} I^{\text{in}}(\vec{L}) \cdot \cos\phi_{\text{in}} \cdot T^*(\vec{L},\vec{V})\, d\omega_{\text{in}}. \quad (3.30)$$

Formulae of this type are called **Hall equations**. In fact, every color calculation problem consists of several Hall equations, one for each representative frequency. Surface parameters $(I_e, k_r, k_t, R^*(\vec{L},\vec{V}), T^*(\vec{L},\vec{V}))$ obviously vary in the different equations.

3.3 Lambert's model of incoherent reflection

The incoherent components are modeled by bi-directional densities in the Hall equation, but they are difficult to derive for real materials. Thus, we describe these bi-directional densities by some simple functions containing a few free parameters instead. These free parameters can be used to tune the surface properties to provide an appearance similar to that of real objects.

First of all, consider diffuse — optically very rough — surfaces reflecting a portion of the incoming light with radiant intensity uniformly distributed in all directions. The constant radiant intensity (I_d) of the diffuse surface lit by a collimated beam from the angle ϕ_{in} can be calculated thus:

$$I_d = \int^{2\pi} I^{\text{in}}(\vec{L}) \cdot \cos\phi_{\text{in}} \cdot R^*(\vec{L},\vec{V})\, d\omega_{\text{in}}. \quad (3.31)$$

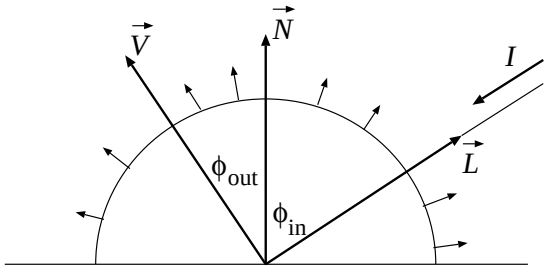

Figure 3.9: Diffuse reflection

The collimated beam is expressed as a directional delta function, $I^{\text{in}} \cdot \delta(\vec{L})$, simplifying the integral as:

$$I_d = I^{\text{in}} \cdot \cos \phi_{\text{in}} \cdot R^*(\vec{L}, \vec{V}). \tag{3.32}$$

Since I_d does not depend on $\vec{V}$ or ϕ_{out}, the last term is constant and is called the **diffuse reflection coefficient** k_d:

$$k_d = R^*(\vec{L}, \vec{V}) = \frac{R(\vec{L}, \vec{V})}{\cos \phi_{\text{out}}}. \tag{3.33}$$

The radiant intensity of a diffuse surface is:

$$I_d(\lambda) = I^{\text{in}}(\lambda) \cdot \cos \phi_{\text{in}} \cdot k_d(\lambda). \tag{3.34}$$

This is **Lambert's law** of diffuse reflection. The term $\cos \phi_{\text{in}}$ can be calculated as the dot product of unit vectors $\vec{N}$ and $\vec{L}$. Should $\vec{N} \cdot \vec{L}$ be negative, the light is incident to the back of the surface, meaning it is blocked by the object. This can be formulated by the following rule:

$$I_d(\lambda) = I^{\text{in}}(\lambda) \cdot k_d(\lambda) \cdot \max\{(\vec{N} \cdot \vec{L}), 0\}. \tag{3.35}$$

This rule makes the orientation of the surface normals essential, since they always have to point outward from the object.

It is interesting to examine the properties of the diffuse coefficient k_d. Suppose the diffuse surface reflects a fraction r of the incoming energy, while the rest is absorbed.

The following interdependence holds between k_d and r:

$$r = \frac{\Phi^{\text{out}}}{\Phi^{\text{in}}} = \frac{dA \cdot \int^{2\pi} I_d \cdot \cos\phi_{\text{out}}\, d\omega}{dA \cdot \int^{2\pi} I^{\text{in}} \cdot \delta(\vec{L}_{\text{in}}) \cdot \cos\phi_{\text{in}}\, d\omega_{\text{in}}} =$$

$$\frac{\int^{2\pi} I^{\text{in}} \cdot \cos\phi_{\text{in}} \cdot k_d \cdot \cos\phi_{\text{out}}\, d\omega}{I^{\text{in}} \cdot \cos\phi_{\text{in}}} = k_d \cdot \int^{2\pi} \cos\phi_{\text{out}}\, d\omega = k_d \cdot \pi. \qquad (3.36)$$

Note that diffuse surfaces do not distribute the light flux evenly in different directions, but follow a $\cos\phi_{\text{out}}$ function, which is eventually compensated for by its inverse in the projected area of the expression of the radiant intensity. According to equation 3.36, the k_d coefficient cannot exceed $1/\pi$ for physically correct models. In practical computations however, it is usually nearly 1, since in the applied models, as we shall see, so many phenomena are ignored that overemphasizing the computationally tractable features becomes acceptable.

Since diffuse surfaces cannot generate mirror images, they present their "**own color**" if they are lit by white light. Thus, the spectral dependence of the diffuse coefficient k_d, or the relation of k_d^{red}, k_d^{green} and k_d^{blue} in the simplified case, is primarily responsible for the surface's "own color" even in the case of surfaces which also provide non-diffuse reflections.

3.4 Phong's model of incoherent reflection

A more complex approximation of the incoherent reflection has been proposed by Phong [Pho75]. The model is important in that it also covers shiny surfaces. Shiny surfaces do not radiate the incident light by uniform intensity, but tend to distribute most of their reflected energy around the direction defined by the reflection law of geometric optics.

It would seem convenient to break down the reflected light and the bidirectional reflection into two terms; a) the diffuse term that satisfies Lambert's law and b) the specular term that is responsible for the glossy reflection concentrated around the mirror direction:

$$R(\vec{L},\vec{V}) = R_d(\vec{L},\vec{V}) + R_s(\vec{L},\vec{V}), \qquad (3.37)$$

$$I^{\text{out}} = I_d + I_s = I^{\text{in}} \cdot k_d \cdot \cos\phi_{\text{in}} + I^{\text{in}} \cdot \cos\phi_{\text{in}} \cdot \frac{R_s(\vec{L},\vec{V})}{\cos\phi_{\text{out}}}. \qquad (3.38)$$

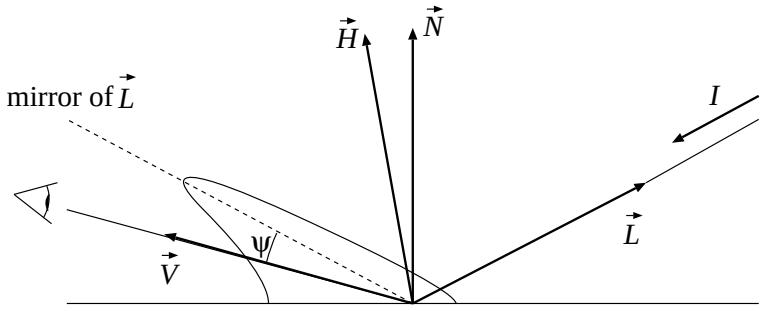

Figure 3.10: Specular reflection

Since $R_s(\vec{L},\vec{V})$ is relevant only when $\vec{V}$ is close to the mirror direction of $\vec{L}$:

$$\cos\phi_{\text{in}} \cdot \frac{R_s(\vec{L},\vec{V})}{\cos\phi_{\text{out}}} \approx R_s(\vec{L},\vec{V}). \tag{3.39}$$

To describe the intensity peak mathematically, a bi-directional function had to be proposed, which is relatively smooth, easy to control and simple to compute. Phong used the $k_s \cdot \cos^n\psi$ function for this purpose, where ψ is the angle between the direction of interest and the mirror direction, n is a constant describing how shiny the surface is, and k_s is the **specular coefficient** representing the fraction of the specular reflection in the total reflected light.

Comparing this model to real world measurements we can conclude that the specular coefficient k_s does not depend on the object's "own color" (in the highlights we can see the color of the lightsource rather than the color of the object), but that it does depend on the angle between the mirror direction and the surface normal, as we shall see in the next section.

The simplified illumination formula is then:

$$I^{\text{out}}(\lambda) = I^{\text{in}}(\lambda) \cdot k_d(\lambda) \cdot \cos\phi_{\text{in}} + I^{\text{in}}(\lambda) \cdot k_s(\lambda,\phi_{\text{in}}) \cdot \cos^n\psi. \tag{3.40}$$

Let the **halfway unit vector** of $\vec{L}$ and $\vec{V}$ be $\vec{H} = (\vec{L}+\vec{V})/|\vec{L}+\vec{V}|$. The term $\cos\psi$ can be calculated from the dot product of unit vectors $\vec{V}$ and $\vec{H}$, since according to the law of reflection:

$$\psi = 2 \cdot \text{angle}(\vec{N},\vec{H}). \tag{3.41}$$

By trigonometric considerations:

$$\cos\psi = \cos(2 \cdot \text{angle}(\vec{N}, \vec{H})) = 2 \cdot \cos^2(\text{angle}(\vec{N}, \vec{H})) - 1 = 2 \cdot (\vec{N} \cdot \vec{H})^2 - 1 \tag{3.42}$$

Should the result turn out to be a negative number, the observer and the lightsource are obviously on different sides of the surface, and thus the specular term is zero. If the surface is lit not only by a single collimated beam, the right side of this expression has to be integrated over the hemisphere, or if several collimated beams target the surface, their contribution should simply be added up. It is important to note that, unlike Lambert's law, this model has no physical interpretation, but it follows nature in an empirical way only.

3.5 Probabilistic modeling of specular reflection

Specular reflection can be more rigorously analyzed by modeling the surface irregularities by probability distributions, as has been proposed by Torrance, Sparrow, Cook and Blinn. In their model, the surface is assumed to consist of randomly oriented perfect mirrors, so-called **microfacets**. As in the previous section, the reflected light is broken down into diffuse and specular components. The diffuse component is believed to be generated by multiple reflections on the microfacets and also by emission of the absorbed light by the material of the surface. The diffuse component is well described by Lambert's law. The specular component, on the other hand, is produced by the direct reflections of the microfacets. The bi-directional reflection function is also broken down accordingly, and we will discuss the derivation of the specular bi-directional reflection function $R_s(\vec{L}, \vec{V})$:

$$R(\vec{L}, \vec{V}) = R_d(\vec{L}, \vec{V}) + R_s(\vec{L}, \vec{V}) = k_d \cdot \cos\phi_{\text{out}} + R_s(\vec{L}, \vec{V}). \tag{3.43}$$

Returning to the original definition, the bi-directional reflection function is, in fact, an additive component of a probability density function, which is true for R_s as well.

$$R_s(\vec{L}, \vec{V}) \cdot d\omega =$$

$$\Pr\{\text{photon is reflected directly to } d\omega \text{ around } \vec{V} \mid \text{coming from } \vec{L}\}. \tag{3.44}$$

Concerning this type of reflection from direction $\vec{L}$ to $d\omega$ around direction $\vec{V}$, only those facets can contribute whose normal is in $d\omega_H$ around the halfway unit vector $\vec{H}$. If reflection is to happen, the facet must obviously be facing in the right direction. It should not be hidden by other facets, nor should its reflection run into other facets, and it should not absorb the photon for the possible contribution.

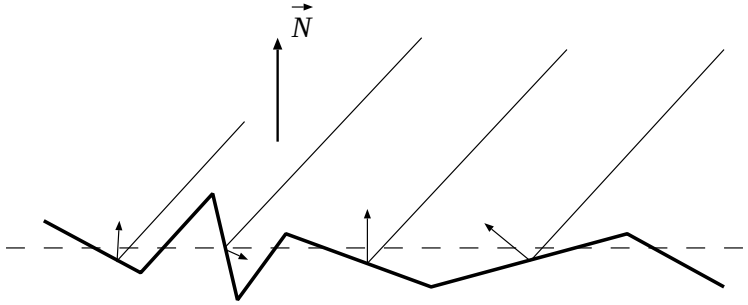

Figure 3.11: Microfacet model of the reflecting surface

Considering these facts, the event that "a photon is reflected directly to $d\omega$ around $\vec{V}$" can be expressed as the logical AND connection of the following stochastically independent events:

1. **Orientation**: In the path of the photon there is a microfacet having its normal in $d\omega_H$ around $\vec{H}$.

2. **No shadowing or masking**: The given microfacet is not hidden by other microfacets from the photon coming from the lightsource, and the reflected photon does not run into another microfacet.

3. **Reflection**: The photon is not absorbed by the perfect mirror.

Since these events are believed to be stochastically independent, their probability can be calculated independently, and the probability of the composed event will be their product.

Concerning the probability of the microfacet normal being in $d\omega_H$, we can suppose that all facets have equal area f. Let the probability density of the number of facets per unit area surface, per solid angle of facet normal

be $P(\vec{H})$. Blinn [Bli77] proposed **Gaussian distribution** for $P(\vec{H})$, since it seemed reasonable due to the central value theorem of probability theory:

$$P(\vec{H}) = \text{const} \cdot e^{-(\alpha/m)^2}. \tag{3.45}$$

where α is the angle of the microfacet with respect to the normal of the mean surface, that is the angle between $\vec{N}$ and $\vec{H}$, and m is the root mean square of the slope, i.e. a measure of the roughness.

Later Torrance and Sparrow showed that the results of the early work of Beckmann [BS63] and Davies [Dav54], who discussed the scattering of electromagnetic waves theoretically, can also be used here and thus Torrance proposed the **Beckmann distribution** function instead of the Gaussian:

$$P(\vec{H}) = \frac{1}{m^2 \cos^4 \alpha} \cdot e^{-\left(\frac{\tan^2 \alpha}{m^2}\right)}. \tag{3.46}$$

If a photon arrives from direction $\vec{L}$ to a surface element dA, the visible area of the surface element will be $dA \cdot (\vec{N} \cdot \vec{L})$, while the total visible area of the microfacets having their normal in the direction around $\vec{H}$ will be

$$f \cdot P(\vec{H}) \cdot d\omega_H \cdot dA \cdot (\vec{H} \cdot \vec{L}).$$

The probability of finding an appropriate microfacet aligned with the photon can be worked out as follows:

$$\Pr\{\text{orientation}\} = \frac{f \cdot P(\vec{H}) \cdot d\omega_H \cdot dA \cdot (\vec{H} \cdot \vec{L})}{dA \cdot (\vec{N} \cdot \vec{L})} = \frac{f \cdot P(\vec{H}) \cdot d\omega_H \cdot (\vec{H} \cdot \vec{L})}{(\vec{N} \cdot \vec{L})}. \tag{3.47}$$

The visibility of the microfacets from direction $\vec{V}$ means that the reflected photon does not run into another microfacet. The collision is often referred to as masking. Looking at figure 3.12, we can easily recognize that the probability of masking is l_1/l_2, where l_2 is the one-dimensional length of the microfacet, and l_1 describes the boundary case from where the beam is masked. The angles of the triangle formed by the bottom of the microfacet wedge and the beam in the boundary case can be expressed by the angles $\alpha = \text{angle}(\vec{N}, \vec{H})$ and $\beta = \text{angle}(\vec{V}, \vec{H}) = \text{angle}(\vec{L}, \vec{H})$ by geometric considerations and by using the law of reflection. Applying the sine law for this triangle, and some trigonometric formulae:

$$\Pr\{\text{not masking}\} = 1 - \frac{l_1}{l_2} = 1 - \frac{\sin(\beta + 2\alpha - \pi/2)}{\sin(\pi/2 - \beta)} = 2 \cdot \frac{\cos\alpha \cdot \cos(\beta + \alpha)}{\cos\beta}. \tag{3.48}$$

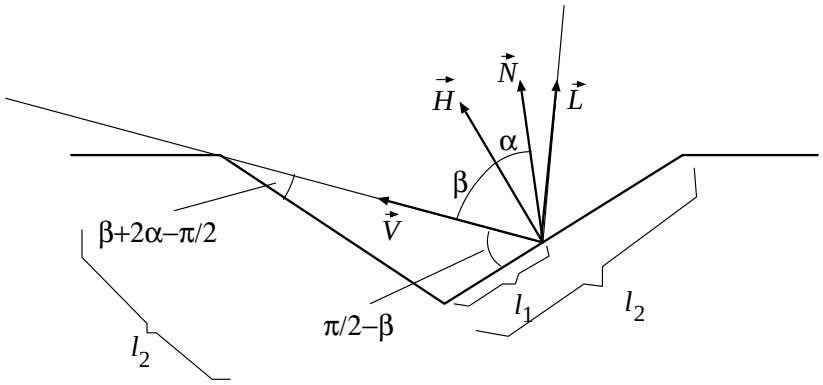

Figure 3.12: Geometry of masking

According to the definitions of the angles $\cos\alpha = \vec{N}\cdot\vec{H}$, $\cos(\beta+\alpha) = \vec{N}\cdot\vec{V}$ and $\cos\beta = \vec{V}\cdot\vec{H}$.

If the angle of incident light and the facet normal do not allow the triangle to be formed, the probability of no masking taking place is obviously 1. This situation can be recognized by evaluating the formula without any previous considerations and checking whether the result is greater than 1, then limiting the result to 1. The final result is:

$$\Pr\{\text{not masking}\} = \min\{2\cdot\frac{(\vec{N}\cdot\vec{H})\cdot(\vec{N}\cdot\vec{V})}{(\vec{V}\cdot\vec{H})}, 1\}. \qquad (3.49)$$

The probability of shadowing can be derived in exactly the same way, only $\vec{L}$ should be substituted for $\vec{V}$:

$$\Pr\{\text{not shadowing}\} = \min\{2\cdot\frac{(\vec{N}\cdot\vec{H})\cdot(\vec{N}\cdot\vec{L})}{(\vec{L}\cdot\vec{H})}, 1\}. \qquad (3.50)$$

The probability of neither shadowing nor masking taking place can be approximated by the minimum of the two probabilities:

$$\Pr\{\text{no shadow and mask}\} \approx$$

$$\min\{2\cdot\frac{(\vec{N}\cdot\vec{H})\cdot(\vec{N}\cdot\vec{V})}{(\vec{V}\cdot\vec{H})}, 2\cdot\frac{(\vec{N}\cdot\vec{H})\cdot(\vec{N}\cdot\vec{L})}{(\vec{L}\cdot\vec{H})}, 1\} = G(\vec{N},\vec{L},\vec{V}). \qquad (3.51)$$

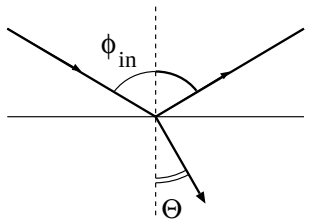

Figure 3.13: Reflection and refraction of a surface

Even perfect mirrors absorb some portion of the incident light, as is described by the **Fresnel equations** of physical optics, expressing the reflection (F) of a perfectly smooth mirror in terms of the refractive index of the material, ν, the extinction coefficient, κ which describes the conductivity of the material (for nonmetals $\kappa = 0$), and the angle of the incidence of the light beam, ϕ_{in}. Using the notations of figure 3.13, where ϕ_{in} is the incident angle and θ is the angle of refraction, the Fresnel equation expressing the ratio of the energy of the reflected beam and the energy of the incident beam for the directions parallel and perpendicular to the electric field is:

$$F_{\perp}(\lambda, \phi_{\text{in}}) = |\frac{\cos\theta - (\nu + \kappa\jmath)\cdot\cos\phi_{\text{in}}}{\cos\theta + (\nu + \kappa\jmath)\cdot\cos\phi_{\text{in}}}|^2, \tag{3.52}$$

$$F_{\parallel}(\lambda, \phi_{\text{in}}) = |\frac{\cos\phi_{\text{in}} - (\nu + \kappa\jmath)\cdot\cos\theta)}{\cos\phi_{\text{in}} + (\nu + \kappa\jmath)\cdot\cos\theta}|^2, \tag{3.53}$$

where $\jmath = \sqrt{-1}$. These equations can be derived from Maxwell's fundamental formulae describing the basic laws of electric waves. If the light is unpolarized, that is, the parallel ($\vec{E}_{\parallel}$) and the perpendicular ($E_{\perp}$) electric fields have the same amplitude, the total reflectivity is:

$$F(\lambda, \phi_{\text{in}}) = \frac{|F_{\parallel}^{1/2}\cdot\vec{E}_{\parallel} + F_{\perp}^{1/2}\cdot\vec{E}_{\perp}|^2}{|\vec{E}_{\parallel} + \vec{E}_{\perp}|^2} = \frac{F_{\parallel} + F_{\perp}}{2}. \tag{3.54}$$

Note that F is wavelength dependent, since n and κ are functions of the wavelength.

Parameters ν and κ are often not available, so they should be estimated from measurements, or from the value of the normal reflection if the extinction is small. At normal incidence ($\phi_{\text{in}} = 0$), the reflection is:

$$F_0(\lambda) = |\frac{1-(\nu+\kappa j)}{1+(\nu+\kappa j)}|^2 = \frac{(\nu-1)^2+\kappa^2}{(\nu+1)^2+\kappa^2} \approx [\frac{\nu-1}{\nu+1}]^2. \tag{3.55}$$

Solving for ν gives the following equation:

$$\nu(\lambda) = \frac{1+\sqrt{F_0(\lambda)}}{1-\sqrt{F_0(\lambda)}}. \tag{3.56}$$

F_0 can easily be measured, thus this simple formula is used to compute the values of the index of refraction ν. Values of $\nu(\lambda)$ can then be substituted into the Fresnel equations (3.52 and 3.53) to obtain reflection parameter F for other angles of incidence.

Since F is the fraction of the reflected energy, it also describes the probability of a photon being reflected, giving:

$$\Pr\{\text{reflection}\} = F(\lambda, \vec{N}\cdot\vec{L})$$

where variable ϕ_{in} has been replaced by $\vec{N}\cdot\vec{L}$.

Now we can summarize the results by multiplying the probabilities of the independent events to express $R_s(\vec{L},\vec{V})$:

$$R_s(\vec{L},\vec{V}) = \frac{1}{d\omega}\Pr\{\text{orientation}\}\cdot\Pr\{\text{no mask and shadow}\}\cdot\Pr\{\text{reflection}\} =$$

$$\frac{d\omega_H}{d\omega}\frac{f\cdot P(\vec{H})\cdot(\vec{H}\cdot\vec{L})}{(\vec{N}\cdot\vec{L})}\cdot G(\vec{N},\vec{L},\vec{V})\cdot F(\lambda,\vec{N}\cdot\vec{L}). \tag{3.57}$$

The last problem left is the determination of $d\omega_H/d\omega$ [JGMHe88]. Defining a spherical coordinate system (ϕ,θ), with the north pole in the direction of $\vec{L}$ (figure 3.14), the solid angles are expressed by the product of vertical and horizontal arcs:

$$d\omega = dV_{\text{hor}}\cdot dV_{\text{vert}}, \qquad d\omega_H = dH_{\text{hor}}\cdot dH_{\text{vert}}. \tag{3.58}$$

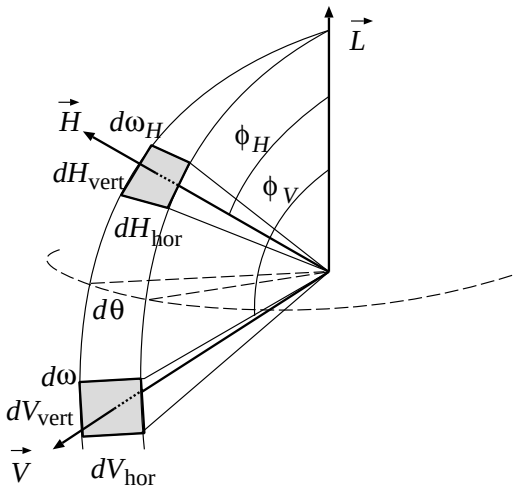

Figure 3.14: Calculation of $d\omega_H/d\omega$

By using geometric considerations and applying the law of reflection, we get:

$$dV_{\text{hor}} = d\theta \cdot \sin\phi_V, \qquad dH_{\text{hor}} = d\theta \cdot \sin\phi_H, \qquad dV_{\text{vert}} = 2dH_{\text{vert}}. \tag{3.59}$$

This in turn yields:

$$\frac{d\omega_H}{d\omega} = \frac{\sin\phi_H}{2\sin\phi_V} = \frac{\sin\phi_H}{2\sin 2\phi_H} = \frac{1}{4\cos\phi_H} = \frac{1}{4(\vec{L}\cdot\vec{H})}. \tag{3.60}$$

since $\phi_V = 2 \cdot \phi_H$.

The final form of the specular term is:

$$R_s(\vec{L},\vec{V}) = \frac{f \cdot P(\vec{H})}{4(\vec{N}\cdot\vec{L})} \cdot G(\vec{N},\vec{L},\vec{V}) \cdot F(\lambda, \vec{N}\cdot\vec{L}). \tag{3.61}$$

3.6 Abstract lightsource models

Up to now we have dealt with lightsources as ordinary surfaces with positive emission I_e. Simplified illumination models, however, often make a distinction between "normal" surfaces, and some abstract objects called **"lightsources"**. These abstract lightsources cannot be seen on the image directly, they are only responsible for feeding energy into the system and thus making the normal surfaces visible for the camera.

The most common types of such lightsources are the following:

1. **Ambient light** is assumed to be constant in all directions and to be present everywhere in 3D space. Its role in restoring the energy equilibrium is highlighted in the next section.

2. **Directional lightsources** are supposed to be at infinity. Thus the light beams coming from a directional lightsource are parallel and their energy does not depend on the position of the surface. (The sun behaves like a directional lightsource.)

3. **Positional or point lightsources** are located at a given point in the 3D space and are concentrated on a single point. The intensity of the light at distance d is $I_l(d) = I_0 \cdot f(d)$. If it really were a point-like source, $f(d) = 1/d^2$ should hold, but to avoid numerical instability for small distances, we use $f(d) = 1/(a \cdot d + b)$ instead, or to emphasize atmospheric effects, such as fog $f(d) = 1/(a \cdot d^m + b)$ might also be useful (m, a and b are constants).

4. **Flood lightsources** are basically positional ligthsources with radiant intensity varying with the direction of interest. They have a main radiant direction, and as the angle (ξ) of the main and actual directions increases the intensity decreases significantly. As for the Phong model, the function $\cos^n \xi$ seems appropriate:

$$I_l(d, \xi) = I_0 \cdot f(d) \cdot \cos^n \xi. \tag{3.62}$$

These abstract lightsources have some attractive properties which ease color computations. Concerning a point on a surface, an abstract lightsource may only generate a collimated beam onto the surface point, with

the exception of ambient light. This means that the integral of the rendering equation can be simplified to a summation with respect to different lightsources, if the indirect light reflected from other surfaces is ignored. The direction of the collimated beam can also be easily derived, since for a directional lightsource it is a constant parameter of the lightsource itself, and for positional and flood lightsources it is the vector pointing from the point-like lightsource to the surface point. The reflection of ambient light, however, can be expressed in closed form, since only a constant function has to be integrated over the hemisphere.

3.7 Steps for image synthesis

The final objective of graphics algorithms is the calculation of pixel colors, or their respective (R, G, B) values. According to our model of the camera (or eye), this color is defined by the energy flux leaving the surface of the visible object and passing through the pixel area of the window towards the camera. As has been proven in the previous section, this flux is proportional to the intensity of the surface in the direction of the camera and the projected area of the pixel, and is independent of the distance of the surface if it is finite (equation 3.19).

Intensity $I(\lambda)$ has to be evaluated for that surface which is visible through the given pixel, that is the nearest of the surfaces located along the line from the camera towards the center of the pixel. The determination of this surface is called the **hidden surface problem** or **visibility calculation** (chapter 6). The computation required by the visibility calculation is highly dependent on the coordinate system used to specify the surfaces, the camera and the window. That makes it worth transforming the virtual world to a coordinate system fixed to the camera, where this calculation can be more efficient. This step is called **viewing transformation** (chapter 5). Even in a carefully selected coordinate system, the visibility calculation can be time-consuming if there are many surfaces, so it is often carried out after a preprocessing step, called **clipping** (section 5.5) which eliminates those surface parts which cannot be projected onto the window.

Having solved the visibility problem, the surface visible in the given pixel is known, and the radiant intensity may be calculated on the representative wavelengths by the following **shading** equation (the terms of diffuse and

specular reflections have already been substituted into equation 3.30):

$$I^{\text{out}} = I_e + k_r \cdot I_r + k_t \cdot I_t + \int^{2\pi} k_d \cdot I^{\text{in}} \cdot \cos\phi_{\text{in}} \, d\omega_{\text{in}} + \int^{2\pi} k_s \cdot I^{\text{in}} \cdot \cos^n \psi \, d\omega_{\text{in}}. \quad (3.63)$$

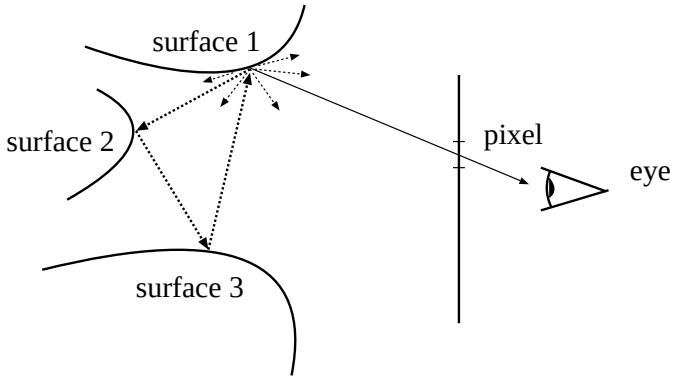

Figure 3.15: Multiple reflections of light

Due to multiple reflections of light beams, the calculation of the intensity of the light leaving a point on a surface in a given direction requires the intensities of other surfaces visible from this point, which of course generates new visibility and **shading** problems to solve (figure 3.15). It must be emphasized that these other surfaces are not necessarily inside the clipping region. To calculate those intensities, other surfaces should be evaluated, and our original point on the given surface might contribute to those intensities. As a consequence of that, the formula has complicated coupling between its left and right sides, making the evaluation difficult.

There are three general and widely accepted approaches to solve this integral equation in 3D continuous space.

1. **The analytical solution**

   Analytical methods rely on numerical techniques to solve the integral equation in its original or simplified form. One of the most popular numerical techniques is the finite element method. Its first step is the subdivision of the continuous surfaces into elemental surface patches, making it possible to approximate their intensity distribution

by constant values independent of the position on the surface. Taking advantage of this homogeneous property of the elemental patches, a stepwise constant function should be integrated in our formula, which can be substituted by a weighted sum of the unknown patch parameters. This step transforms the integral equation into a linear equation which can be solved by straightforward methods. It must be admitted that this solution is not at all simple because of the size of the resulting linear system, but at least we can turn to the rich toolset of the numerical methods of linear algebra. The application of the analytical approach to solve the integral equation of the shading formula in 3D space leads to the family of **analytical shading models**, or as it is usually called, the **radiosity method**.

2. **Constraining the possible coupling**

   Another alternative is to eliminate from the rendering equation those energy contributions which cause the difficulties, and thus give ourselves a simpler problem to solve. For example, if coherent coupling of limited depth, say n, were allowed, and we were to ignore the incoherent component coming from non-abstract lightsources, then the number of surface points which would need to be evaluated to calculate a pixel color can be kept under control. Since the illumination formula contains two terms regarding the coherent components (reflective and refracting lights), the maximum number of surfaces involved in the color calculation of a pixel is two to the power of the given depth, i.e. 2^n. The implementation of this approach is called **recursive ray tracing**.

3. **Ignoring the coupling**

   An even more drastic approach is to simplify or disregard completely all the terms causing problems, and to take into account only the incoherent reflection of the light coming from the abstract lightsources, and the coherent transmission supposing the index of refraction to be equal to 1. Mirrors and refracting objects cannot be described in this model. Since the method is particularly efficient when used with incremental hidden surface methods and can also be implemented using the incremental concept, it is called the **incremental shading method**.

The three different approaches represent three levels of the compromise between image generation speed and quality. By ignoring more and more terms in the illumination formula, its calculation can be speeded up, but the result inevitably becomes more and more artificial. The shading methods based on radiosity and ray tracing techniques or on the combination of the two form the family of **photorealistic image generation**. Simple, incremental shading algorithms, on the other hand, are suitable for very fast hardware implementation, making it possible to generate real-time animated sequences.

The simplification of the illumination formula has been achieved by ignoring some of its difficult-to-calculate terms. Doing this, however, violates the energy equilibrium, and causes portions of objects to come out extremely dark, sometimes unexpectedly so. These artifacts can be reduced by reintroducing the ignored terms in simplified form, called **ambient light**. The ambient light represents the ignored energy contribution in such a way as to satisfy the energy equilibrium. Since this ignored part is not calculated, nothing can be said of its positional and directional variation, hence it is supposed to be constant in all directions and everywhere in the 3D space. From this aspect, the role of ambient light also shows the quality of the shading algorithm. The more important a role it has, the poorer quality picture it will generate.