4 THE PROPAGATION OF LIGHT

4.1 INTRODUCTION

We now consider a number of phenomena related to the propagation of light and its interaction with material media. In particular, we shall study the characteristics of lightwaves as they progress through various substances, crossing interfaces, and being reflected and refracted in the process. For the most part, we shall envision light as a classical electromagnetic wave whose velocity through any medium is dependent upon that material's electric and magnetic properties. It is an intriguing fact that many of the basic principles of optics are predicated on the wave aspects of light but are completely independent of the exact nature of the wave. As we shall see, this accounts for the longevity of *Huygens's principle*, which has served in turn to describe mechanical aether waves, electromagnetic waves, and now, after three hundred years, applies to quantum optics.

Suppose, for the moment, that a wave impinges on the interface separating two different media (e.g., a piece of glass in air). As we know from our everyday experiences, a portion of the incident flux density will be diverted back in the form of a *reflected wave*, while the remainder will be transmitted across the boundary as a *refracted wave*. On a submicroscopic scale we can envision an assemblage of atoms that scatter the incident radiant energy. The manner in which these emitted light wavelets superimpose and combine with each other will depend on the spatial distribution of the scattering atoms. As we know from the previous chapter, the scattering process is responsible for the *index of refraction*, as well as the resultant *reflected* and *refracted* waves. This atomistic description is quite satisfying conceptually, even though it is not a simple matter to treat analytically. It should, however, be kept in mind even when applying macroscopic techniques, as indeed we shall later on.

We now seek to determine the general principles governing or at least describing the propagation, reflection, and refraction of light. In principle it should be possible to trace the progress of radiant energy through any system by applying Maxwell's equations and the associated boundary conditions. In practice, however, this is often an impractical if not an impossible task (see Section 10.1). So we shall take a somewhat different route, stopping, when appropriate, to verify that our results are in accord with electromagnetic theory.

4.2 THE LAWS OF REFLECTION AND REFRACTION

4.2.1 Huygens's Principle

Recall that a wavefront is a surface over which an optical disturbance has a constant phase. As an illustration, Fig. 4.1 shows a small portion of a spherical wavefront Σ emanating from a monochromatic point source S in a homogeneous medium. Clearly, if the radius of the wavefront as shown is r, at some later time t it will simply be $(r + vt)$, where v is the phase velocity of the wave.

But suppose instead that the light passes through a nonuniform sheet of glass, as in Fig. 4.2, so that the wavefront itself is distorted. How can we determine its new form Σ'? Or for that matter, what will Σ' look like at some later time, if it is allowed to continue unobstructed?

A preliminary step toward the solution of this problem appeared in print in 1690 in the work entitled *Traité de la Lumière*, which had been written 12 years earlier by the Dutch physicist Christiaan Huygens. It was there that he enunciated what has since become known as **Huygens's principle**, that *every point on a primary wavefront serves as the source of spherical secondary wavelets, such that the primary wavefront at some later time is the envelope of these wavelets. Moreover, the wavelets advance with a speed and frequency equal to those of the primary wave at each point in space.* If the medium is homogeneous, the wavelets may be constructed with finite radii, whereas if it is inhomogeneous, the wavelets must have infinitesimal radii. Figure 4.3 should make this fairly clear; it shows a view of a wavefront Σ, as well as a number of spherical secondary wavelets, which, after a time t, have propagated out to a radius of vt. The envelope of all these wavelets is then asserted to correspond to the advanced primary wave Σ'. It is easy to visualize the process in terms of mechanical vibrations of an elastic medium. Indeed this is the way that Huygens envisioned it within the context of an all-pervading aether, as is evident from this comment by him:

> We have still to consider, in studying the spreading out of these waves, that each particle of matter in which a wave proceeds not only communicates its motion to the next particle to it, which is on the straight line drawn from the luminous point, but that it also necessarily gives a motion to all the others which touch it and which oppose its motion. The result is that around each particle there arises a wave of which this particle is a center.

We can make use of these ideas in two different ways. On one level, a mathematical representation of the wavelets will serve as the basis for a valuable analytical technique in treating diffraction theory. One can trace the progress of a primary wave past all sorts of apertures and obstacles by summing up the wavelet contributions

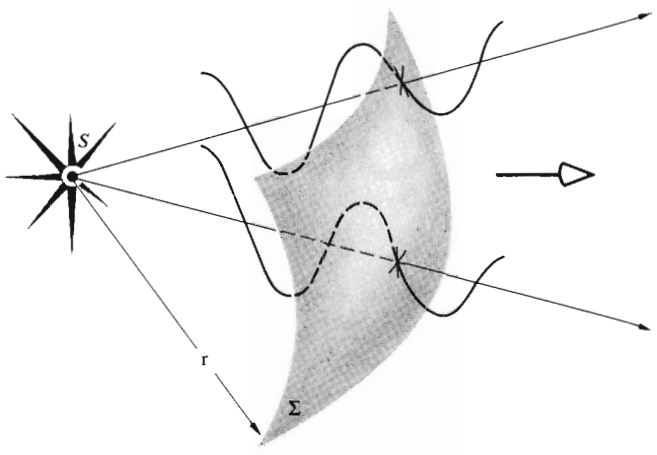

Figure 4.1 A segment of a spherical wave.

mathematically. On another level, Fig. 4.3 represents a graphical application of the essential ideas and as such is known as *Huygens's construction.*

Thus far we have merely stated Huygens's principle, without any justification or proof of its validity. As we shall see (Chapter 10), Fresnel successfully modified Huygens's principle somewhat in the 1800s. A little later on, Kirchhoff showed that the *Huygens–Fresnel principle* was a direct consequence of the differential wave equation (2.59), thereby putting it on a firm mathematical base. That there was a need for a reformulation

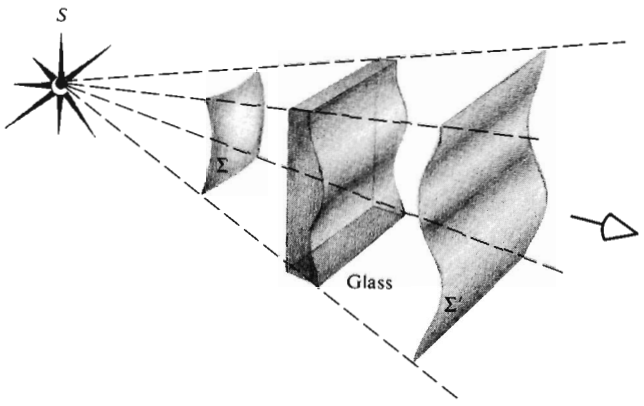

Figure 4.2 Distortion of a portion of a wavefront on passing through a material of nonuniform thickness.

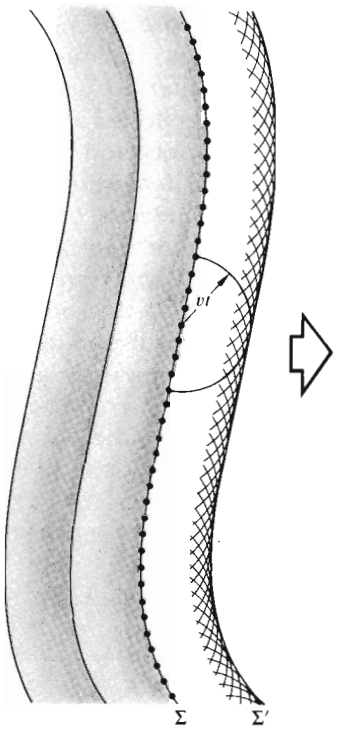

Figure 4.3 The propagation of a wavefront via Huygens's principle.

of the principle is evident from Fig. 4.3, where we deceptively only drew hemispherical wavelets.* Had we drawn them as spheres, there would have been a *backwave* moving toward the source—something that is not observed. Since this difficulty was taken care of theoretically by Fresnel and Kirchhoff, we need not be disturbed by it. In fact, we shall overlook it completely when applying Huygens's construction, which, in the end, is best thought of as a highly useful fiction.

Still, Huygens's principle fits in rather nicely with our earlier discussion of the atomic scattering of radiant energy. Each atom of a material substance that interacts with an incident primary wavefront can be regarded as a point source of scattered secondary wavelets. Things are not quite as clear when we apply the principle to the propagation of light through a vacuum. It is helpful, however, to keep in mind that at any point in empty space on the primary wavefront there exists both a time-varying **E**-field and a time-varying **B**-field. These in turn create new fields that move out from the point. In this sense each point on the wavefront is analogous to a physical scattering center.

* See E. Hecht, *Phys. Teach.* **18**, 149 (1980).

4.2.2 Snell's Law and the Law of Reflection

The fundamental laws of reflection and refraction can be derived in several different ways; the first approach to be used here is based on Huygens's principle. It should be said, however, that our intention at the moment is as much to elaborate on the use of the method as to arrive at the end results. Huygens's principle will provide a highly useful and fairly simple means of analyzing and visualizing some complex propagation problems, for example, those involving anisotropic media (p. 287) or diffraction (p. 392). Consequently, it is to our advantage to gain some practice in using the technique, even if it is not the most elegant procedure for deriving the desired laws.

Figure 4.4 shows a monochromatic plane wave impinging normally down onto the smooth interface separating two homogeneous transparent media. When an incident wave comes into contact with the interface, it can be imagined as split into two: we observe one wave reflected upward and another transmitted downward. If we consider an incident wavefront Σ_i coincident with the interface splitting into Σ_r and Σ_t, both also congruent with the interface, we can utilize Huygens's construction (neglecting the back-waves). Every point on Σ_r serves as a source of secondary wavelets, which travel more or less upward into the incident medium at a speed v_i. At a time t later, the front will advance a distance $v_i t$ and appear as Σ_r'. Similarly, every point on the downward-moving front Σ_t will serve as a source for wavelets essentially heading down with a speed v_t. After a time t the transmitted front will appear a distance $v_t t$ below as Σ_t'.

The process is ongoing, repeating itself with the frequency of the incident wave.* The media are

* This assumes the use of light whose flux density is not so extraordinarily high that the fields are gigantic. With this assumption the medium will behave linearly, as is most often the case. In contrast, observable harmonics can be generated if the fields are made large enough (Section 14.4).

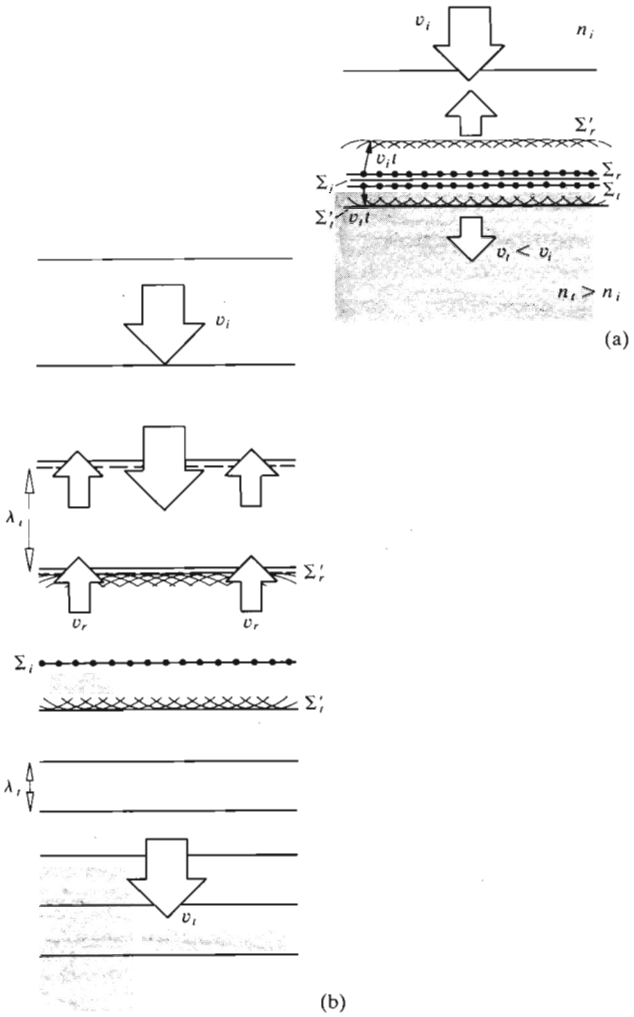

Figure 4.4 A monochromatic plane wave impinging down onto a homogeneous, isotropic medium of index n_t. Σ_i, Σ_r, and Σ_t should actually overlap.

assumed to respond linearly, so the reflected and transmitted waves have that same frequency (and period), as do all the secondary wavelets. Taking $n_t > n_i$, it follows that $c/v_t > c/v_i$, thus $v_t < v_i$, and the wavelengths (the distances between wavefronts drawn in consecutive intervals of τ) will be such that $\lambda_i > \lambda_t$ and $\lambda_i = \lambda_r$, as shown in Fig. 4.4(b). The incoming plane wave is perpendicular to the interface, and symmetry produces both reflected and transmitted plane waves that also travel out from the interface perpendicularly.

Now suppose the incident wave comes in at some other angle, as indicated in Fig. 4.5. Clearly, it sweeps across the interface again, essentially splitting into two waves: one reflected and one refracted. Let's follow the progress of a typical front in Fig. 4.6, envisioning the diagram as if it were a series of snapshots taken in successive intervals of time τ. Start when Σ_i makes contact with the interface at point a. At that point, both the reflected and transmitted wavefronts begin, so a, which lies on both fronts, can be taken as a source of both an upwardly emitted wavelet traveling at a speed v_i and a downwardly emitted wavelet traveling at a speed v_t. Now focus on another point, say, b on Σ_i.

After a time t_1 the plane Σ_i will have moved a distance in the incident medium of $v_i t_1$, so that b then corresponds to b'. Presumably, two wavelets will then propagate out from b' into the incident and transmitting media, contributing to the reflected, Σ_r', and transmitted, Σ_t', wavefronts. These wavelets are shown here after a time t_2, where $\tau = t_1 + t_2$. The rest of the diagram

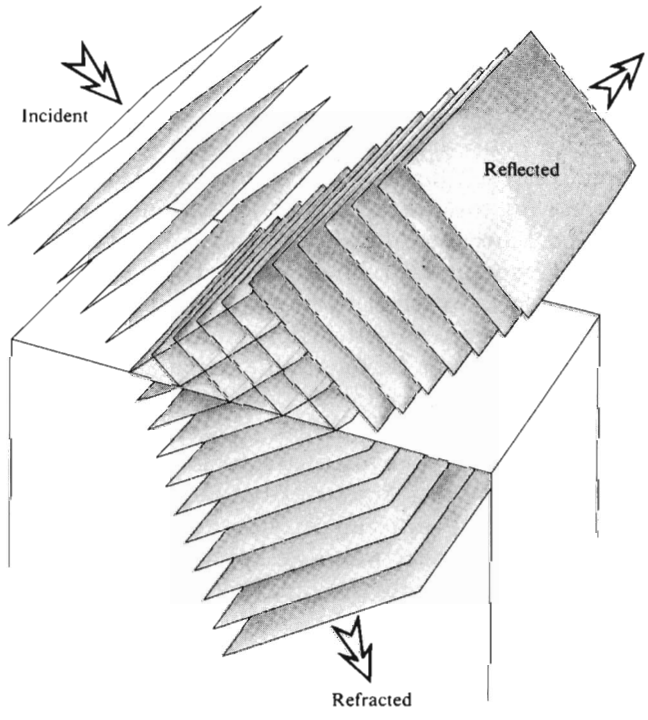

Figure 4.5 Reflection and transmission of plane waves.

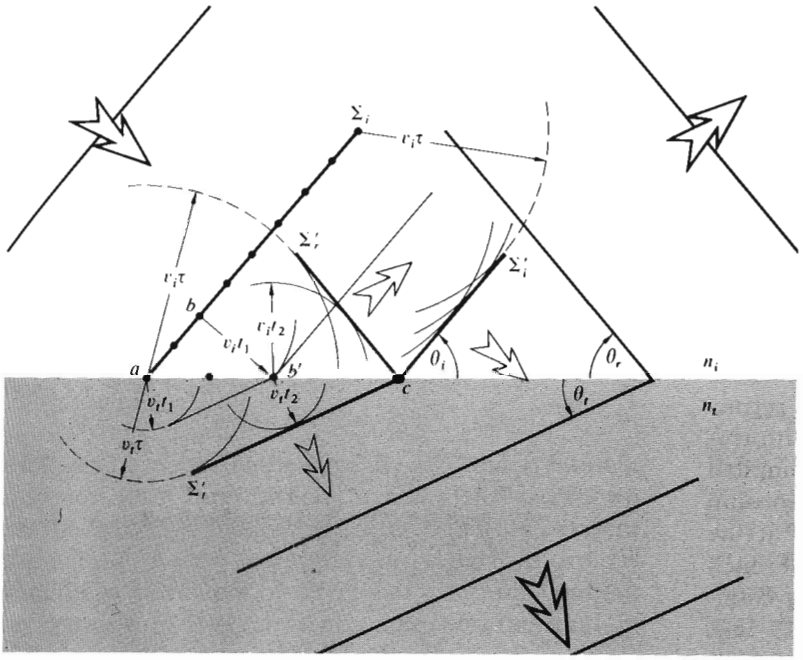

Figure 4.6 Reflection and transmission at an interface via Huygens's principle.

should be self-explanatory. Figure 4.7 is a somewhat simplified version in which θ_i, θ_r, and θ_t, as before, are the angles of *incidence*, *reflection*, and *transmission* (or *refraction*), respectively. Notice that

$$\frac{\sin\theta_i}{\overline{BD}} = \frac{\sin\theta_r}{\overline{AC}} = \frac{\sin\theta_t}{\overline{AE}} = \frac{1}{\overline{AD}}. \tag{4.1}$$

By comparison with Fig. 4.6, it should be evident that

$$\overline{BD} = v_i t, \qquad \overline{AC} = v_i t, \qquad \overline{AE} = v_t t,$$

so substituting into Eq. (4.1) and canceling t, we have

$$\frac{\sin\theta_i}{v_i} = \frac{\sin\theta_r}{v_i} = \frac{\sin\theta_t}{v_t}. \tag{4.2}$$

It follows from the first two terms that **the angle of incidence equals the angle of reflection**, that is,

$$\theta_i = \theta_r. \tag{4.3}$$

Known as the **law of reflection**, it first appeared in the book entitled *Catoptrics*, which was purported to have been written by Euclid.

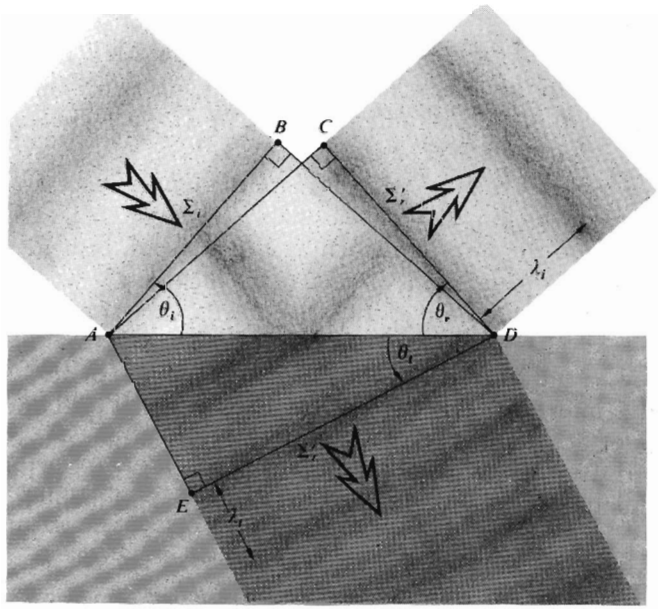

Figure 4.7 Reflected and transmitted wavefronts at a given instant.

The first and last terms of Eq. (4.2) yield

$$\frac{\sin \theta_i}{\sin \theta_t} = \frac{v_i}{v_t}, \tag{4.4}$$

or since $v_i/v_t = n_t/n_i$,

$$n_i \sin \theta_i = n_t \sin \theta_t. \tag{4.5}$$

This is the very important **law of refraction**, the physical consequences of which have been studied, at least on record, for over eighteen hundred years. On the basis of some fine observations, Claudius Ptolemy of Alexandria attempted unsuccessfully to divine the expression. Kepler nearly succeeded in deriving the law of refraction in his book *Supplements to Vitello* in 1604. Unfortunately he was misled by some erroneous data compiled earlier by Vitello (ca. 1270). The correct relationship seems to have been arrived at first by Snell* at the University of Leyden and then by the French mathematician Descartes.† In English-speaking countries Eq. (4.5) is generally referred to as **Snell's law**. Notice that it can be rewritten in the form

$$\frac{\sin \theta_i}{\sin \theta_t} = n_{ti}, \tag{4.6}$$

where $n_{ti} \equiv n_t/n_i$ is the *ratio of the absolute indices of refraction.* In other words, it is the *relative index of refraction of the two media.* It is evident in Fig. 4.6, where $n_{ti} > 1$ (i.e., $n_t > n_i$ and $v_i > v_t$), that $\lambda_i > \lambda_t$, whereas the opposite would be true if $n_{ti} < 1$.

One feature of the above treatment merits some further discussion. It was reasonably assumed that each point on the interface, such as c in Fig. 4.6, coincides with a particular point on each of the incident, reflected, and transmitted waves. In other words, there is a fixed phase relationship between each of the waves at points a, b, c, and so forth. As the incident front sweeps across the interface, every point on it in contact with the interface is also a point on both a corresponding reflected front and a corresponding transmitted front. This situation is known as *wavefront continuity*, and it will be

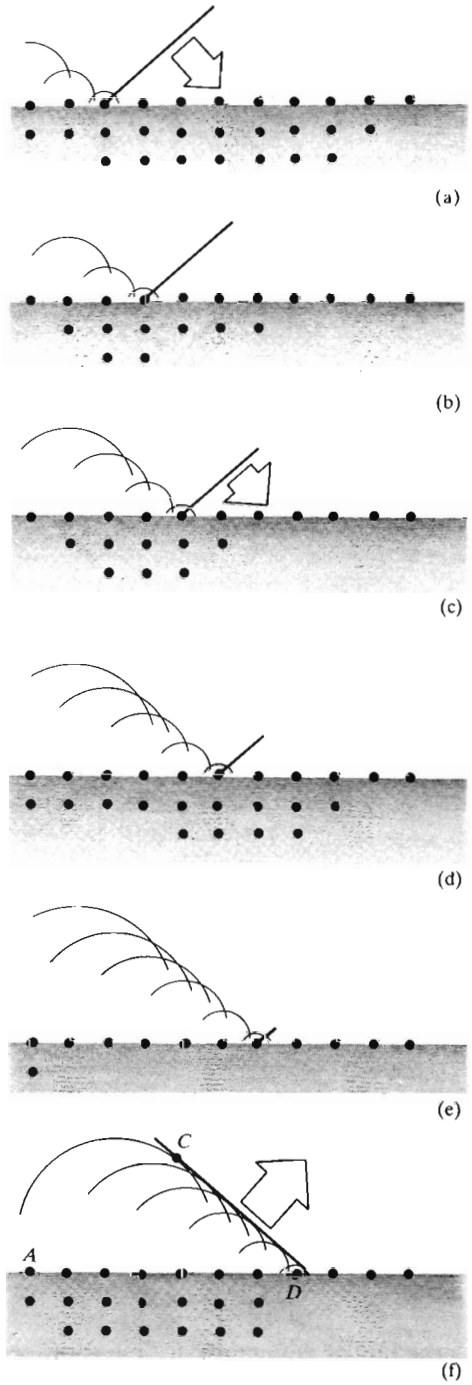

Figure 4.8 The reflection of a wave as the result of scattering.

* This is the common spelling, although Snel is probably more accurate.

† For a more detailed history, see Max Herzberger, "Optics from Euclid to Huygens," *Appl. Opt.* **5**, 1383 (1966).

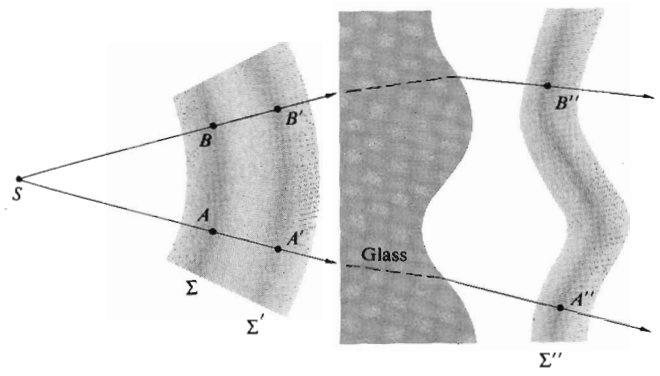

Figure 4.9 Wavefronts and rays.

justified in a more mathematically rigorous treatment in Section 4.3.1. Interestingly, Sommerfeld* has shown that the laws of reflection and refraction (independent of the kind of wave involved) can be derived directly from the requirement of wavefront continuity without any recourse to Huygens's principle, and the solution to Problem 4.9 demonstrates as much.

A far more physically appealing view of the whole process is depicted in Fig. 4.8. An electromagnetic disturbance whose wavelength (λ) is several thousand times larger than the spacing between the atoms ($d \approx 0.1$ nm) sweeps across an interface. Each atom is driven successively and scatters a wavelet. The tilt of the incident wave determines the phase delay between the scattering of each atom in turn (see Section 10.1.3 for the details). The front running from C to D is composed of wavelets that arrive in phase, superimpose, and interfere constructively. Since every point on the incident front (ranging from A to B in Fig. 4.7) has the same phase, if $\overline{AC} = \overline{BD}$, the distances traveled and therefore the phases of the wavelets arriving at C and D will be equal, as indeed they will be all across the front. From the geometry, this can happen only for a reflected wavefront propagating in the one direction such that $\theta_i = \theta_r$. This picture of scattered interfering wavelets is essentially an atomic version of the Huygens–Fresnel principle.

Although theoretically all the dipoles throughout the medium contribute to the reflected wave, the dominant effect is due to a surface layer only about $\frac{1}{2}\lambda$ thick, which is nonetheless typically several thousand atoms deep. Furthermore, the condition that only one beam is reflected is true provided that $\lambda \gg d$; it would not be the case with x-rays where $\lambda \approx d$, and there several scattered beams actually result; nor is it the case with a diffraction grating, where the separation between scatterers is again comparable to λ, and several reflected and transmitted beams are produced. A similar argument can be made for the scattering process giving rise to the transmitted wave and Snell's law, as Problem 4.11 establishes.

4.2.3 Light Rays

The concept of a light ray is one that will be of interest to us throughout our study of optics. *A* **ray** *is a line drawn in space corresponding to the direction of flow of radiant energy.* As such, it is a mathematical device rather than a physical entity. In practice one can produce very narrow *beams* or *pencils* of light (e.g., a laserbeam), and we might imagine a ray to be the unattainable limit on the narrowness of such a beam. Bear in mind that in an *isotropic medium* (i.e., one whose properties are the same in all directions) *rays are orthogonal trajectories of the wavefronts.* That is to say, *they are lines normal to the wavefronts at every point of intersection.* Evidently, *in such a medium a ray is parallel to the propagation vector* **k**. As you might suspect, this is not true in *anisotropic* substances, which we will consider later (see Section 8.4.1). *Within homogeneous isotropic materials, rays will be straight lines,* since by symmetry they cannot bend in any preferred direction, there being none. Moreover, because the speed of propagation is identical in all directions within a given medium, the spatial separation between two wavefronts, measured along rays, must be the same everywhere.* Points where a single ray intersects a set of wavefronts are called *corresponding points,* for example, A, A', and A'' in Fig. 4.9. *Evidently the separation in time between any two corresponding points on any two*

*A. Sommerfeld, *Optics*, p. 151. See also J. J. Sein, *Am. J. Phys.* **50**, 180 (1982).

* When the material is inhomogeneous or when there is more than one medium involved, it will be the *optical path length* (see Section 4.2.4) between the two wavefronts that is the same.

sequential wavefronts is identical. In other words, if wavefront Σ is transformed into Σ'' after a time t'', the distance between corresponding points on any and all rays will be traversed in that same time t''. This will be true even if the wavefronts pass from one homogeneous isotropic medium into another. This just means that each point on Σ can be imagined as following the path of a ray to arrive at Σ'' in the time t''.

If a group of rays is such that we can find a surface that is orthogonal to each and every one of them, they are said to form a *normal congruence.* For example, the rays emanating from a point source are perpendicular to a sphere centered at the source and consequently form a normal congruence.

We can now briefly consider an alternative to Huygens's principle that will also allow us to follow the progress of light through various isotropic media. The basis for this approach is the *theorem of Malus and Dupin* (introduced in 1808 by E. Malus and modified in 1816 by C. Dupin), according to which *a group of rays will preserve its normal congruence after any number of reflections and refractions* (as in Fig. 4.9). From our present vantage point of the wave theory, this is equivalent to the statement that rays remain orthogonal to wavefronts throughout all propagation processes in isotropic media. As shown in Problem 4.12, the theorem can be used to derive the law of reflection as well as Snell's law. It is often most convenient to carry out a ray trace through an optical system using the laws of reflection and refraction and then reconstruct the wavefronts. The latter can be accomplished in accord with the above considerations of equal transit times between corresponding points and the orthogonality of the rays and wavefronts.

Figure 4.10 depicts the parallel ray formation concomitant with a plane wave, where θ_i, θ_r, and θ_t, which have the exact same meanings as before, are now measured from the normal to the interface. The incident ray and the normal determine a plane known as the **plane of incidence**. Because of the symmetry of the situation, we must anticipate that both the reflected and transmitted rays will be undeflected from that plane. In other words, the respective unit propagation vectors $\hat{\mathbf{k}}_i$, $\hat{\mathbf{k}}_r$, and $\hat{\mathbf{k}}_t$ are coplanar.

In summary, then, the three basic laws of reflection and refraction are:

1. The incident, reflected, and refracted rays all lie in the plane of incidence.
2. $\theta_i = \theta_r$. *[4.3]*
3. $n_i \sin \theta_i = n_t \sin \theta_t$. *[4.5]*

These are illustrated rather nicely with a narrow light beam in the photographs of Fig. 4.11. Here, the incident medium is air ($n_i \approx 1.0$), and the transmitting medium is glass ($n_t \approx 1.5$). Consequently, $n_i < n_t$, and it follows

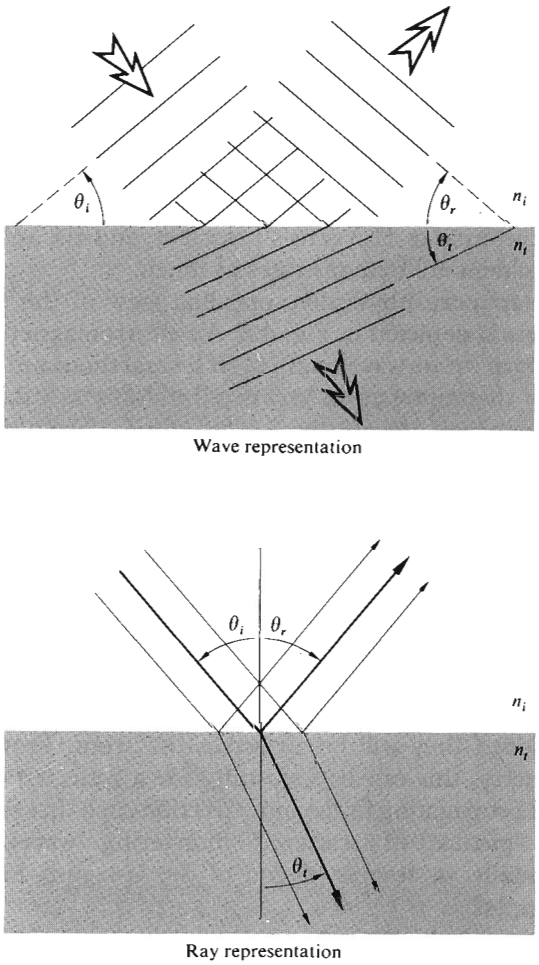

Figure 4.10 The wave and ray representations of an incident, reflected, and transmitted beam.

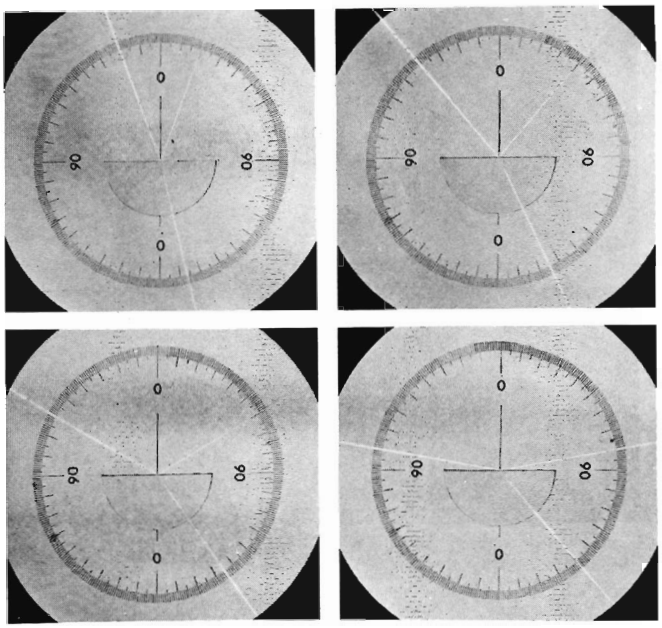

Figure 4.11 Refraction at various angles of incidence. (Photos courtesy *PSSC College Physics*, D. C. Heath & Co., 1968.)

from Snell's law that $\sin\theta_i > \sin\theta_t$. Since both angles, θ_i and θ_t, vary between 0° and 90°, a region over which the sine function is smoothly rising, it can be concluded that $\theta_i > \theta_t$. *Rays entering a higher-index medium from a lower one refract toward the normal* and vice versa. This much is evident in the figure. Notice that the bottom surface is cut circular so that the transmitted beam within the glass always lies along a radius and is therefore normal to the lower surface in every case. If a ray is normal to an interface, $\theta_i = 0 = \theta_t$, and it sails right through with no bending.

The incident beam in each portion of Fig. 4.11 is narrow and sharp, and the reflected beam is equally well defined. Accordingly, the process is known as **specular reflection** (from the word for a common mirror alloy in ancient times, speculum). In this case, as in Fig. 4.12(a), the reflecting surface is smooth, or more precisely, any irregularities in it are small compared with a wavelength.* In contrast, the **diffuse reflection** in Fig. 4.12(b) occurs when the surface is relatively rough. For example, "nonreflecting" glass used to cover pictures is actually glass whose surface is roughened so that it reflects diffusely. The law of reflection holds exactly over any region that is small enough to be considered smooth. These two forms of reflection are extremes; a whole range of intermediate behavior is possible. Thus, although the paper of this page was manufactured deliberately to be a fairly diffuse scatterer, the cover of the book reflects in a manner that is somewhere between diffuse and specular.

Let $\hat{\mathbf{u}}_n$ be a unit vector normal to the interface pointing in the direction from the incident to the transmitting medium (Fig. 4.13). As you will have the opportunity to prove in Problem 4.13, the first and third basic laws can be combined in the form of a *vector refraction equation:*

$$n_i(\hat{\mathbf{k}}_i \times \hat{\mathbf{u}}_n) = n_t(\hat{\mathbf{k}}_t \times \hat{\mathbf{u}}_n) \tag{4.7}$$

or, alternatively,

$$n_t\hat{\mathbf{k}}_t - n_i\hat{\mathbf{k}}_i = (n_t\cos\theta_t - n_i\cos\theta_i)\hat{\mathbf{u}}_n. \tag{4.8}$$

4.2.4 Fermat's Principle

The laws of reflection and refraction, and indeed the manner in which light propagates in general, can be viewed from an entirely different and intriguing perspective afforded us by **Fermat's principle**. The ideas that will unfold presently have had a tremendous influence on the development of physical thought in and beyond the study of classical optics. Apart from its implications in quantum optics (Section 13.6, p. 552), Fermat's principle provides us with an insightful and highly useful way of appreciating and anticipating the behavior of light.

Hero of Alexandria, who lived some time between 150 B.C. and 250 A.D., was the first to set forth what has since become known as a *variational principle.* In his formulation of the law of reflection, he asserted that *the path actually taken by light in going from some point S to a point P via a reflecting surface was the shortest possible one.* This can be seen rather easily in Fig. 4.14, which depicts a point source S emitting a number of rays that are

*If the surface ridges and valleys are small compared with λ, the scattered wavelets will still interfere constructively in only one direction ($\theta_i = \theta_r$).

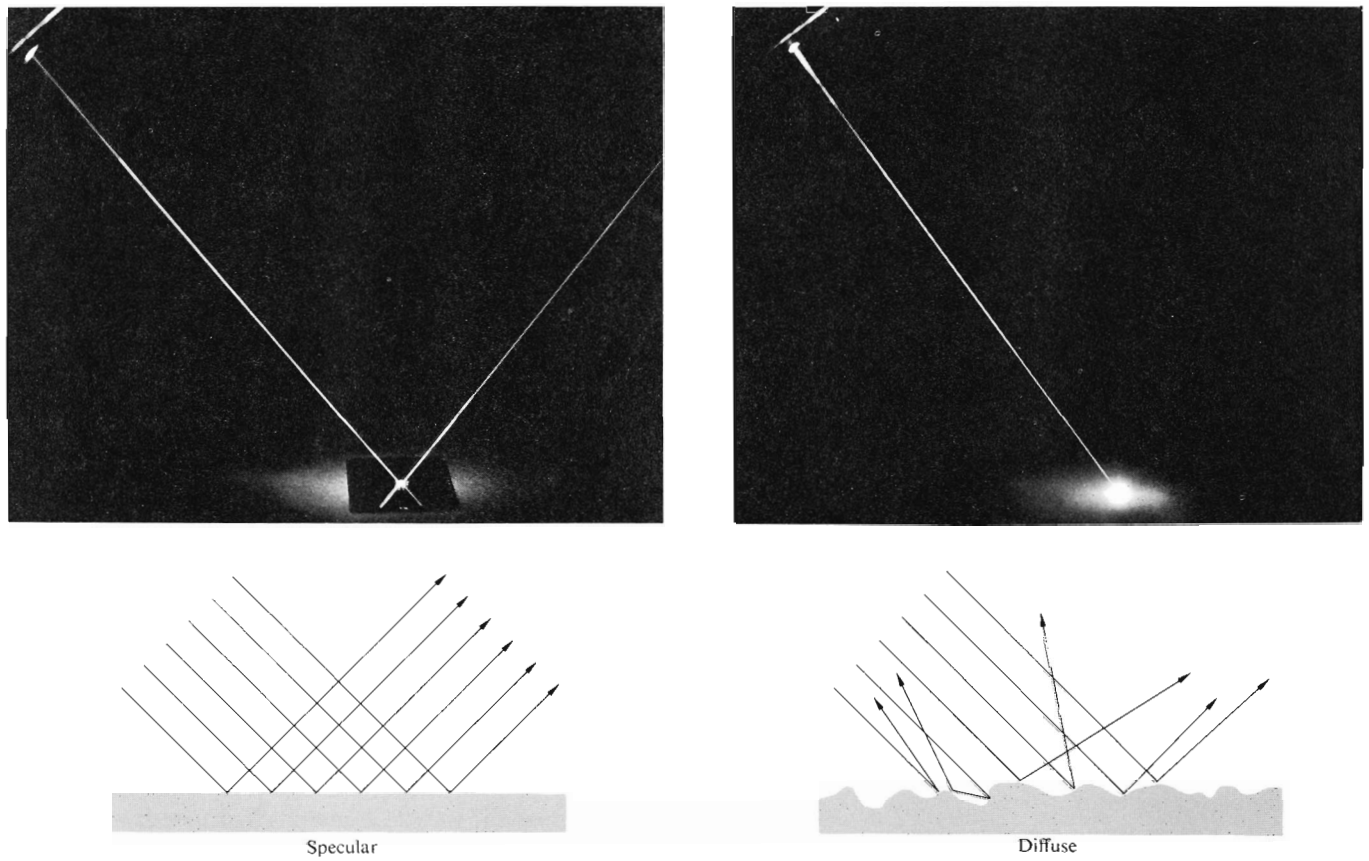

Figure 4.12 (a) Specular reflection. (b) Diffuse reflection. (Photos courtesy Donald Dunitz.)

then "reflected" toward P. Of course, only one of these paths will have any physical reality. If we simply draw the rays as if they emanated from S' (the image of S), none of the distances to P will have been altered (i.e., $SAP = S'AP$, $SBP = S'BP$, etc.). But obviously the straight-line path $S'BP$, which corresponds to $\theta_i = \theta_r$, is the shortest possible one. The same kind of reasoning (Problem 4.15) makes it evident that points S, B, and P must lie in what has previously been defined as the plane of incidence. For over fifteen hundred years Hero's curious observation stood alone, until in 1657 Fermat propounded his celebrated *principle of least time*, which encompassed both reflection and refraction. Obviously, a beam of light traversing an interface does not take a straight line or *minimum spatial path* between a point in the incident medium and one in the transmitting medium. Fermat consequently reformulated Hero's statement to read: *the actual path between two points taken by a beam of light is the one that is traversed in the least time.* As we shall see, even this form of the statement is somewhat incomplete and a bit erroneous at that. For the moment then, let us embrace it but not passionately.

As an example of the application of the principle to the case of refraction, refer to Fig. 4.15, where we minimize t, the transit time from S to P, with respect to the variable x. In other words, changing x shifts point O, thereby changing the ray from S to P. The smallest

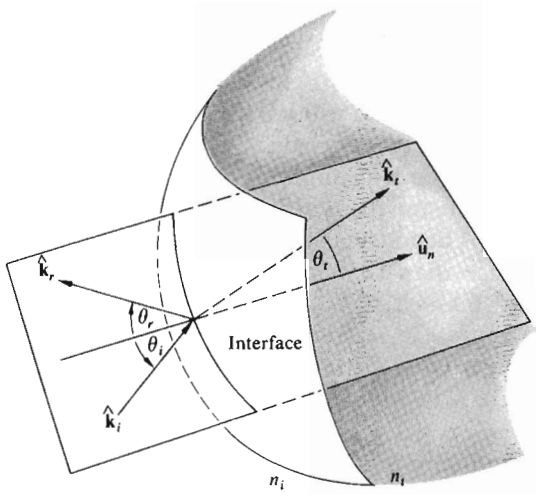

Figure 4.13 The ray geometry.

transit time will then presumably coincide with the actual path. Hence

$$t = \frac{\overline{SO}}{v_i} + \frac{\overline{OP}}{v_t}$$

or

$$t = \frac{(h^2 + x^2)^{1/2}}{v_i} + \frac{[b^2 + (a - x)^2]^{1/2}}{v_t}.$$

To minimize $t(x)$ with respect to variations in x, we set $dt/dx = 0$, that is,

$$\frac{dt}{dx} = \frac{x}{v_i(h^2 + x^2)^{1/2}} + \frac{-(a - x)}{v_t[b^2 + (a - x)^2]^{1/2}} = 0.$$

Using the diagram, we can rewrite the expression as

$$\frac{\sin \theta_i}{v_i} = \frac{\sin \theta_t}{v_t},$$

which is of course no less than Snell's law (Eq. 4.4). Thus if a beam of light is to advance from S to P in the least possible time, it must comply with the empirical law of refraction.

Suppose that we have a stratified material composed of m layers, each having a different index of refraction, as in Fig. 4.16. The transit time from S to P will then be

$$t = \frac{s_1}{v_1} + \frac{s_2}{v_2} + \cdots + \frac{s_m}{v_m}$$

or

$$t = \sum_{i=1}^{m} s_i/v_i,$$

where s_i and v_i are the path length and speed, respectively, associated with the ith contribution. Thus

$$t = \frac{1}{c} \sum_{i=1}^{m} n_i s_i, \tag{4.9}$$

in which the summation is known as the **optical path length** (OPL) traversed by the ray, in contrast to the spatial path length $\sum_{i=1}^{m} s_i$. Clearly, for an inhomogeneous medium where n is a function of position, the summation must be changed to an integral:

$$(\text{OPL}) = \int_S^P n(s)\, ds. \tag{4.10}$$

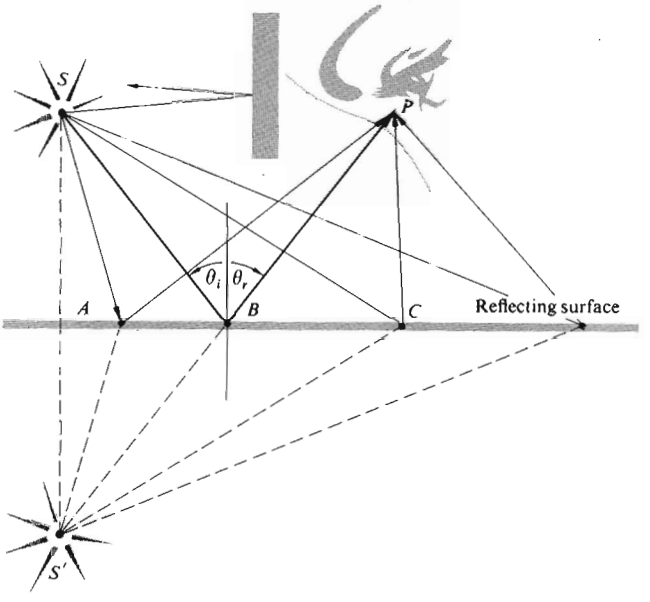

Figure 4.14 Minimum path from the source S to the observer's eye at P.

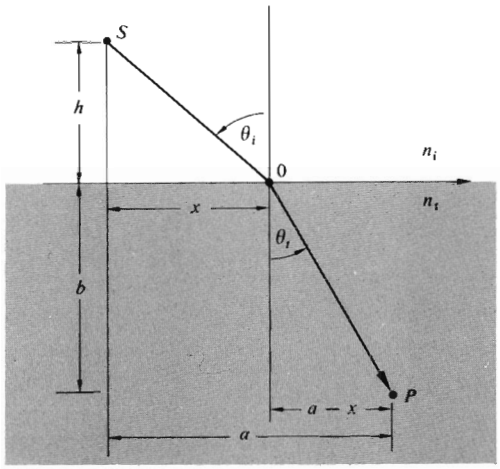

Figure 4.15 Fermat's principle applied to refraction.

Inasmuch as $t = (\text{OPL})/c$, we can restate Fermat's principle: *light, in going from points S to P, traverses the route having the smallest optical path length.* Accordingly, when light rays from the Sun pass through the inhomogeneous atmosphere of the Earth, as shown in Fig. 4.17(a), they bend so as to traverse the lower, denser regions as abruptly as possible, thus minimizing the OPL. Ergo, one can still see the Sun after it has actually passed below the horizon. In the same way, a road viewed at a glancing angle, as in Fig. 4.17(b), will appear to reflect the environs as if it were covered with a sheet of water. The air near the roadway will be warmer and less dense than that farther above it. Rays will bend upward, taking the shortest optical path, and in so doing they will appear to be reflected from a mirrored surface. The effect is particularly easy to see on long modern highways. The only requirement is that you look at the road at near glancing incidence, because the rays bend very gradually.

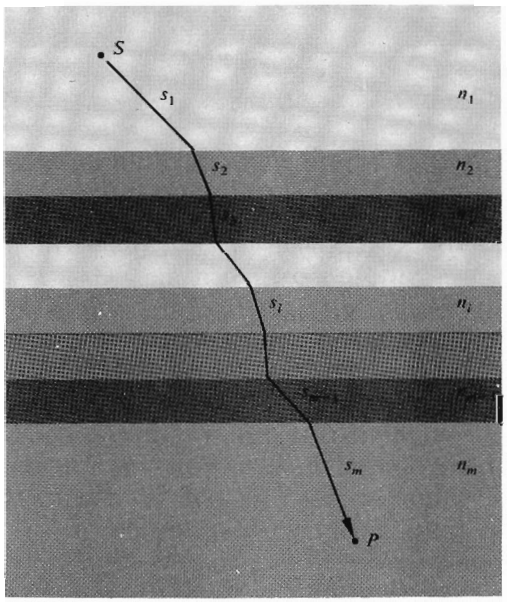

Figure 4.16 A ray propagating through a layered material.

The original statement of Fermat's *principle of least time* has some serious failings and is, as we shall see, in need of alteration. To that end, recall that if we have a function, say $f(x)$, we can determine the specific value of the variable x that causes $f(x)$ to have a *stationary* value by setting $df/dx = 0$ and solving for x. By a stationary value we mean one for which the slope of $f(x)$ versus x is zero or equivalently where the function has a maximum ⌒, minimum ⌣, or a point of inflection with a horizontal tangent ⌒⌣.

Fermat's principle in its modern form reads: *a light ray in going from point S to point P must traverse an optical path length that is stationary with respect to variations of that path.* In other words, the OPL for the true trajectory will equal, to a first approximation, the OPL of paths immediately adjacent to it.* Thus there will be many curves neighboring the actual one, which would take nearly the same time for the light to traverse. This latter point makes it possible to begin to understand how light manages to be so clever in its meanderings. Suppose that we have a beam of light advancing through a homogeneous isotropic medium so that a ray passes from points S to P. Atoms within the material are driven by the incident disturbance, and they reradiate in all directions. Generally, wavelets originating in the immediate vicinity of a stationary path will arrive at P by routes that differ only slightly and will therefore

* The first derivative of the OPL vanishes in its Taylor series expansion, since the path is stationary.

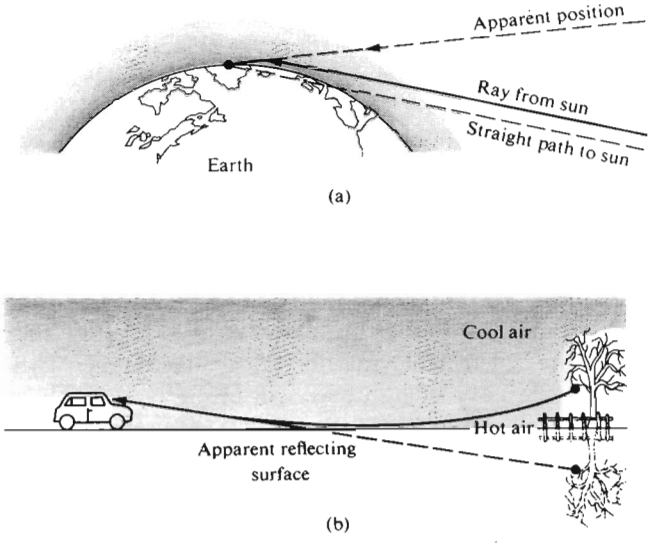

Figure 4.17 The bending of rays through inhomogeneous media.

arrive nearly in phase and reinforce each other (see Section 7.1). Wavelets taking other paths will arrive at P out of phase and will therefore tend to cancel each other. That being the case, energy will effectively propagate along that ray from S to P that satisfies Fermat's principle.

To show that the OPL for a ray need not always be a minimum, examine Fig. 4.18, which depicts a segment of a hollow three-dimensional ellipsoidal mirror. If the source S and the observer P are at the foci of the ellipsoid, then by definition the length SQP will be constant, regardless of where on the perimeter Q happens to be. It is also a geometrical property of the ellipse that $\theta_i = \theta_r$ for any location of Q. All optical paths from S to P via a reflection are therefore precisely equal—none is a minimum, and the OPL is clearly stationary with respect to variations. Rays leaving S and striking the mirror will arrive at the focus P. From another viewpoint we can say that radiant energy emitted by S will be scattered by electrons in the mirrored surface such that the wavelets will substantially reinforce each other only at P, where they have traveled the same distance and have the same phase. In any case, if a plane mirror was tangent to the ellipse at Q, the exact same

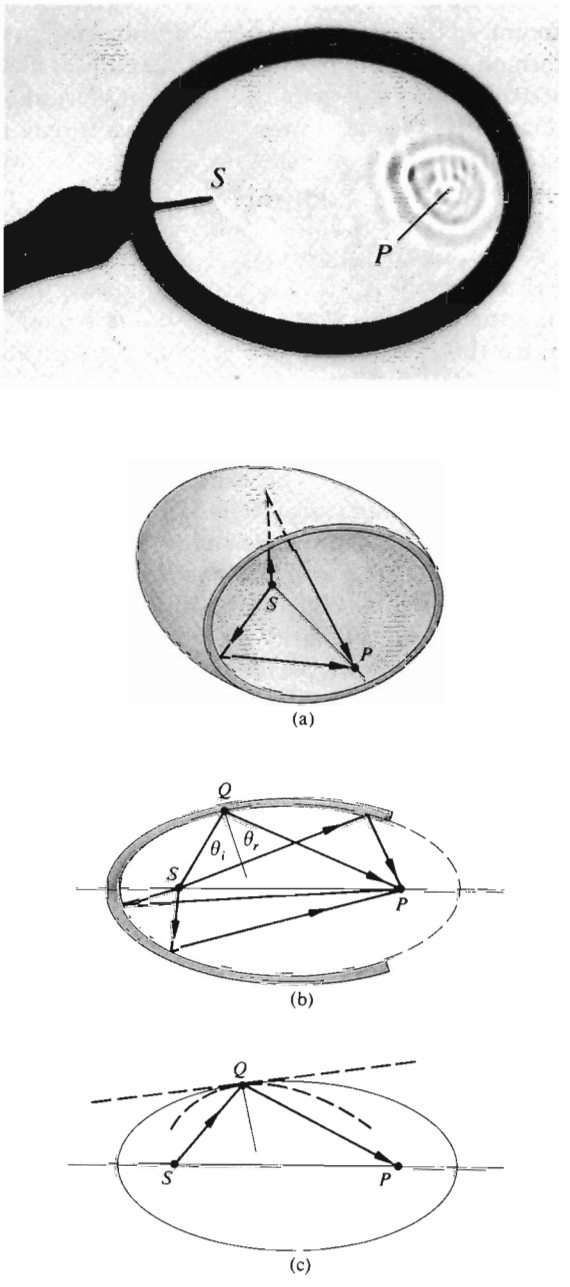

Figure 4.18 Reflection off an ellipsoidal surface. Observe the reflection of waves using a frying pan filled with water. Even though these are usually circular it is well worth playing with. (Photo courtesy *PSSC College Physics*, D. C. Heath & Co., 1968.)

path *SQP* traversed by a ray would then be a relative minimum. At the other extreme, if the mirrored surface conformed to a curve lying within the ellipse, like the dashed one shown, that same ray along *SQP* would now negotiate a relative maximum OPL. This is true even though other unused paths (where $\theta_i \neq \theta_r$) would actually be shorter (i.e., apart from inadmissible curved paths). Thus in all cases the rays travel a stationary OPL in accord with the reformulated Fermat's principle. Note that since the principle speaks only about the path and not the direction along it, a ray going from P to S will trace the same route as one from S to P. This is the very useful *principle of reversibility.*

Fermat's achievement stimulated a great deal of effort to supersede Newton's laws of mechanics with a similar variational formulation. The work of many men, notably Pierre de Maupertuis (1698–1759) and Leonhard Euler, finally led to the mechanics of Joseph Louis Lagrange (1736–1813) and hence to the *principle of least action*, formulated by William Rowan Hamilton (1805–1865). The striking similarity between the principles of Fermat and Hamilton played an important part in Schrödinger's development of quantum mechanics. In 1942 Richard Phillips Feynman (b. 1918) showed that quantum mechanics can be fashioned in an alternative way using a variational approach. The continuing evolution of variational principles brings us back to optics via the modern formalism of quantum optics (see Chapter 13).

Fermat's principle is not so much a computational device as it is a concise way of thinking about the propagation of light. It is a statement about the grand scheme of things without any concern for the contributing mechanisms, and as such it will yield insights under a myriad of circumstances.

4.3 THE ELECTROMAGNETIC APPROACH

Thus far we have been able to deduce the laws of reflection and refraction using three different approaches: *Huygens's principle*, the *theorem of Malus and Dupin*, and *Fermat's principle*. Each yields a distinctive and valuable point of view. Yet another and even more powerful approach is provided by the electromagnetic theory of light. Unlike the previous techniques, which say nothing about the incident, reflected, and transmitted radiant flux densities (i.e., I_i, I_r, I_t, respectively), the electromagnetic theory treats these within the framework of a far more complete description.

The body of information that forms the subject of optics has accrued over many centuries. As our knowledge of the physical universe becomes more extensive, the concomitant theoretical descriptions must become ever more encompassing. This, quite generally, brings with it an increased complexity. And so, rather than using the formidable mathematical machinery of the quantum theory of light, we will often avail ourselves of the simpler insights of simpler times (e.g., Huygens's and Fermat's principles). Thus even though we are now going to develop another and more extensive description of reflection and refraction, we will not put aside those earlier methods. In fact, throughout this study we shall use the simplest technique that can yield sufficiently accurate results for our particular purposes.

4.3.1 Waves at an Interface

Suppose that the incident monochromatic lightwave is planar, so that it has the form

$$\mathbf{E}_i = \mathbf{E}_{0i} \exp [i(\mathbf{k}_i \cdot \mathbf{r} - \omega_i t)] \tag{4.11}$$

or, more simply,

$$\mathbf{E}_i = \mathbf{E}_{0i} \cos (\mathbf{k}_i \cdot \mathbf{r} - \omega_i t). \tag{4.12}$$

Assume that $\mathbf{E}_{0i}$ is constant in time, that is, the wave is linearly or plane polarized. We'll find in Chapter 8 that any form of light can be represented by two orthogonal linearly polarized waves, so that this doesn't actually represent a restriction. Note that just as the origin in time, $t = 0$, is arbitrary, so too is the origin O in space, where $\mathbf{r} = 0$. Thus, making no assumptions about their directions, frequencies, wavelengths, phases, or amplitudes, we can write the reflected and transmitted waves as

$$\mathbf{E}_r = \mathbf{E}_{0r} \cos (\mathbf{k}_r \cdot \mathbf{r} - \omega_r t + \varepsilon_r) \tag{4.13}$$

and

$$\mathbf{E}_t = \mathbf{E}_{0t} \cos (\mathbf{k}_t \cdot \mathbf{r} - \omega_t t + \varepsilon_t). \tag{4.14}$$

Here ε_r and ε_t are *phase constants* relative to $\mathbf{E}_i$ and are introduced because the position of the origin is not unique. Figure 4.19 depicts the waves in the vicinity of the planar interface between two homogeneous lossless dielectric media of indices n_i and n_t.

The laws of electromagnetic theory (Section 3.1) lead to certain requirements that must be met by the fields, and these are referred to as the boundary conditions. Specifically, one of these is that the component of the electric field intensity $\mathbf{E}$ that is tangent to the interface must be continuous across it (the same is true for $\mathbf{H}$). In other words, the total tangential component of $\mathbf{E}$ on one side of the surface must equal that on the other (Problem 4.22). Thus since $\hat{\mathbf{u}}_n$ is the unit vector normal to the interface, regardless of the direction of the electric field within the wavefront, the cross-product of it with $\hat{\mathbf{u}}_n$ will be perpendicular to $\hat{\mathbf{u}}_n$ and therefore tangent to the interface. Hence

$$\hat{\mathbf{u}}_n \times \mathbf{E}_i + \hat{\mathbf{u}}_n \times \mathbf{E}_r = \hat{\mathbf{u}}_n \times \mathbf{E}_t \tag{4.15}$$

or

$$\begin{aligned}\hat{\mathbf{u}}_n \times \mathbf{E}_{0i} \cos(\mathbf{k}_i \cdot \mathbf{r} - \omega_i t) \\ + \hat{\mathbf{u}}_n \times \mathbf{E}_{0r} \cos(\mathbf{k}_r \cdot \mathbf{r} - \omega_r t + \varepsilon_r) \\ = \hat{\mathbf{u}}_n \times \mathbf{E}_{0t} \cos(\mathbf{k}_t \cdot \mathbf{r} - \omega_t t + \varepsilon_t).\end{aligned} \tag{4.16}$$

This relationship must obtain at any instant in time and at any point on the interface ($y = b$). Consequently, $\mathbf{E}_i$, $\mathbf{E}_r$, and $\mathbf{E}_t$ must have precisely the same functional dependence on the variables t and r, which means that

$$\begin{aligned}(\mathbf{k}_i \cdot \mathbf{r} - \omega_i t)|_{y=b} &= (\mathbf{k}_r \cdot \mathbf{r} - \omega_r t + \varepsilon_r)|_{y=b} \\ &= (\mathbf{k}_t \cdot \mathbf{r} - \omega_t t + \varepsilon_t)|_{y=b}.\end{aligned} \tag{4.17}$$

With this as the case, the cosines in Eq. (4.16) cancel, leaving an expression independent of t and r, as indeed it must be. Inasmuch as this has to be true for all values of time, the coefficients of t must be equal, to wit

$$\omega_i = \omega_r = \omega_t. \tag{4.18}$$

Recall that the electrons within the media are undergoing (linear) forced vibrations at the frequency of the incident wave. Clearly, whatever light is scattered has that same frequency. Furthermore,

$$\begin{aligned}(\mathbf{k}_i \cdot \mathbf{r})|_{y=b} &= (\mathbf{k}_r \cdot \mathbf{r} + \varepsilon_r)|_{y=b} \\ &= (\mathbf{k}_t \cdot \mathbf{r} + \varepsilon_t)|_{y=b},\end{aligned} \tag{4.19}$$

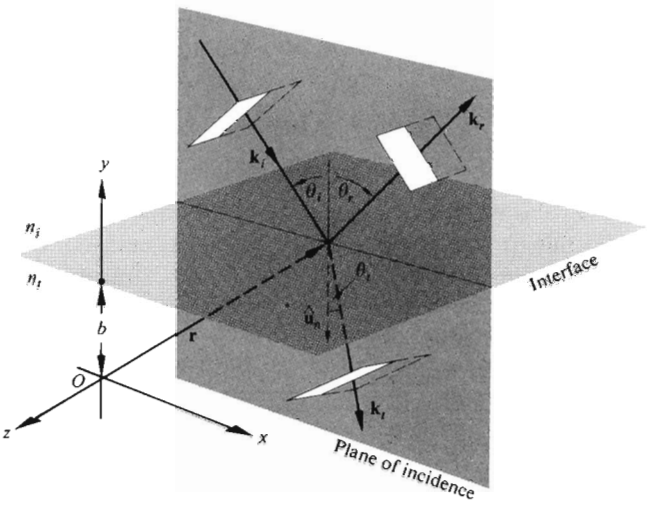

Figure 4.19 Plane waves incident on the boundary between two homogeneous, isotropic, lossless dielectric media.

wherein $\mathbf{r}$ terminates on the interface. The values of ε_r and ε_t correspond to a given position of O, and thus they allow the relation to be valid regardless of that location. (For example, the origin might be chosen such that $\mathbf{r}$ was perpendicular to $\mathbf{k}_i$ but not to $\mathbf{k}_r$ or $\mathbf{k}_t$.) From the first two terms we obtain

$$[(\mathbf{k}_i - \mathbf{k}_r) \cdot \mathbf{r}]_{y=b} = \varepsilon_r. \tag{4.20}$$

Recalling Eq. (2.42), this expression simply says that the endpoint of $\mathbf{r}$ sweeps out a plane (which is of course the interface) perpendicular to the vector $(\mathbf{k}_i - \mathbf{k}_r)$. To phrase it slightly differently, $(\mathbf{k}_i - \mathbf{k}_r)$ is parallel to $\hat{\mathbf{u}}_n$. Notice, however, that since the incident and reflected waves are in the same medium, $k_i = k_r$. From the fact that $(\mathbf{k}_i - \mathbf{k}_r)$ has no component in the plane of the interface, that is, $\hat{\mathbf{u}}_n \times (\mathbf{k}_i - \mathbf{k}_r) = 0$, we conclude that

$$k_i \sin\theta_i = k_r \sin\theta_r;$$

hence we have the law of reflection, that is,

$$\theta_i = \theta_r.$$

Furthermore, since $(\mathbf{k}_i - \mathbf{k}_r)$ is parallel to $\hat{\mathbf{u}}_n$ all three vectors, $\mathbf{k}_i$, $\mathbf{k}_r$, and $\hat{\mathbf{u}}_n$, are in the same plane, the plane of incidence. Again, from Eq. (4.19) we obtain

$$[(\mathbf{k}_i - \mathbf{k}_t) \cdot \mathbf{r}]_{y=b} = \varepsilon_t, \tag{4.21}$$

and therefore $(\mathbf{k}_i - \mathbf{k}_t)$ is also normal to the interface.

Thus $\mathbf{k}_i$, $\mathbf{k}_r$, $\mathbf{k}_t$, and $\hat{\mathbf{u}}_n$ are all coplanar. As before, the tangential components of $\mathbf{k}_i$ and $\mathbf{k}_t$ must be equal, and consequently

$$k_i \sin \theta_i = k_t \sin \theta_t. \tag{4.22}$$

But because $\omega_i = \omega_t$, we can multiply both sides by c/ω_i to get

$$n_i \sin \theta_i = n_t \sin \theta_t,$$

which is Snell's law. Finally, if we had chosen the origin O to be in the interface, it is evident from Eqs. (4.20) and (4.21) that ε_r and ε_t would both have been zero. That arrangement, although not as instructive, is certainly simpler, and we'll use it from here on.

4.3.2 Derivation of the Fresnel Equations

We have just found the relationship that exists among the phases of $\mathbf{E}_i(\mathbf{r}, t)$, $\mathbf{E}_r(\mathbf{r}, t)$, and $\mathbf{E}_t(\mathbf{r}, t)$ at the boundary. There is still an interdependence shared by the amplitudes $\mathbf{E}_{0i}$, $\mathbf{E}_{0r}$, and $\mathbf{E}_{0t}$, which can now be evaluated. To that end, suppose that a plane monochromatic wave is incident on the planar surface separating two isotropic media. Whatever the polarization of the wave, we shall resolve its **E**- and **B**-fields into components parallel and perpendicular to the plane of incidence and treat these constituents separately.

Case 1: E perpendicular to the plane of incidence. We now assume that **E** is perpendicular to the plane of incidence and that **B** is parallel to it (Fig. 4.20). Recall that $E = vB$, so that

$$\hat{\mathbf{k}} \times \mathbf{E} = v\mathbf{B} \tag{4.23}$$

and, of course,

$$\hat{\mathbf{k}} \cdot \mathbf{E} = 0 \tag{4.24}$$

(i.e., **E**, **B**, and the unit propagation vector $\hat{\mathbf{k}}$ form a right-handed system). Again making use of the continuity of the tangential components of the **E**-field, we have at the boundary at any time and any point

$$\mathbf{E}_{0i} + \mathbf{E}_{0r} = \mathbf{E}_{0t}, \tag{4.25}$$

where the cosines cancel. Realize that the field vectors as shown really ought to be envisioned at $y = 0$ (i.e., at the surface), from which they have been displaced for the sake of clarity. Note too that although $\mathbf{E}_r$ and $\mathbf{E}_t$ must be normal to the plane of incidence by symmetry, *we are guessing that they point outward* at the interface when $\mathbf{E}_i$ does. The directions of the **B**-fields then follow from Eq. (4.23).

We will need to invoke another of the boundary conditions in order to get one more equation. The presence of material substances that become electrically polarized by the wave has a definite effect on the field configuration. Thus, although the tangential component of **E** is continuous across the boundary, its normal component is not. Instead the normal component of the product $\epsilon\mathbf{E}$ is the same on either side of the

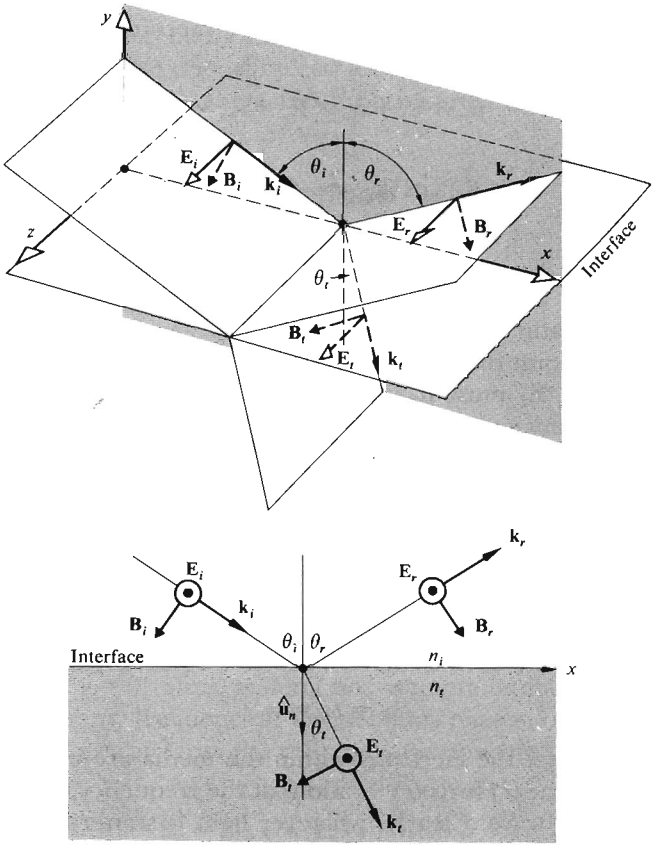

Figure 4.20 An incoming wave whose **E**-field is normal to the plane of incidence.