CHAPTER 3

Foundations of Scalar Diffraction Theory

The phenomenon known as *diffraction* plays a role of utmost importance in the branches of physics and engineering that deal with wave propagation. In this chapter we consider some of the foundations of scalar diffraction theory. While the theory discussed here is sufficiently general to be applied in other fields, such as acoustic-wave and radio-wave propagation, the applications of primary concern will be in the realm of physical optics. To fully understand the properties of optical imaging and data processing systems, it is essential that diffraction and the limitations it imposes on system performance be appreciated. A variety of references to more comprehensive treatments of diffraction theory will be found in the material that follows.

3.1 HISTORICAL INTRODUCTION

Before beginning a discussion of diffraction, it is first necessary to mention another phenomenon with which diffraction should not be confused—namely *refraction*. Refraction can be defined as the bending of light rays that takes place when they pass through a region in which there is a gradient of the local velocity of propagation of the wave. The most common example occurs when a light wave encounters a sharp boundary between two regions having different refractive indices. The propagation velocity in the first medium, having refractive index n_1, is $v_1 = c/n_1$, c being the vacuum velocity of light. The velocity of propagation in the second medium is $v_2 = c/n_2$.

As shown in Fig. 3.1, the incident light rays are bent at the interface. The angles of incidence and refraction are related by *Snell's law*, which is the foundation of geometrical optics,

$$n_1 \sin\theta_1 = n_2 \sin\theta_2, \tag{3-1}$$

where in this example, $n_2 > n_1$ and therefore $\theta_2 < \theta_1$.

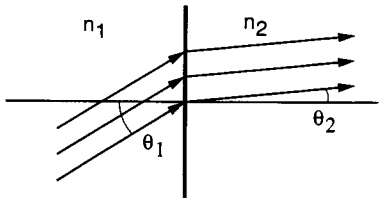

FIGURE 3.1
Snell's law at a sharp boundary.

Light rays are also bent upon *reflection*, which can occur at a metallic or dielectric interface. The fundamental relation governing this phenomenon is that the angle of reflection is always equal to the angle of incidence.

The term *diffraction* has been defined by Sommerfeld (Ref. [270]) as "any deviation of light rays from rectilinear paths which cannot be interpreted as reflection or refraction." Diffraction is caused by the confinement of the lateral extent of a wave, and is most appreciable when that confinement is to sizes comparable with a wavelength of the radiation being used. The diffraction phenomenon should also not be confused with the *penumbra effect*, for which the finite extent of a source causes the light transmitted by a small aperture to spread as it propagates away from that aperture (see Fig. 3.2). As can be seen in the figure, the penumbra effect does not involve any bending of the light rays.

There is a fascinating history associated with the discovery and explanation of diffraction effects. The first accurate report and description of such a phenomenon was made by Grimaldi and was published in the year 1665, shortly after his death. The measurements reported were made with an experimental apparatus similar to that shown in Fig. 3.3. An aperture in an opaque screen was illuminated by a light source, chosen small enough to introduce a negligible penumbra effect; the light intensity was observed across a plane some distance behind the screen. The corpuscular theory of light propagation, which was the accepted means of explaining optical phenomena at the time, predicted that the shadow behind the screen should be well defined, with sharp borders. Grimaldi's observations indicated, however, that the transition from light to shadow was gradual rather than abrupt. If the spectral purity of the light source had been better, he might have observed even more striking results, such as the presence of light and dark fringes extending far into the geometrical shadow of the screen. Such effects cannot be explained by a corpuscular theory of light, which requires rectilinear propagation of light rays in the absence of reflection and refraction.

The initial step in the evolution of a theory that would explain such effects was made by the first proponent of the wave theory of light, Christian Huygens, in the year

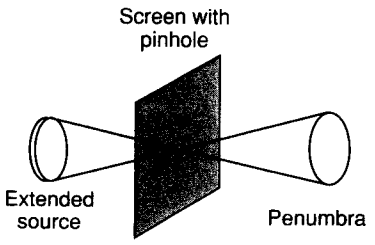

FIGURE 3.2
The penumbra effect.

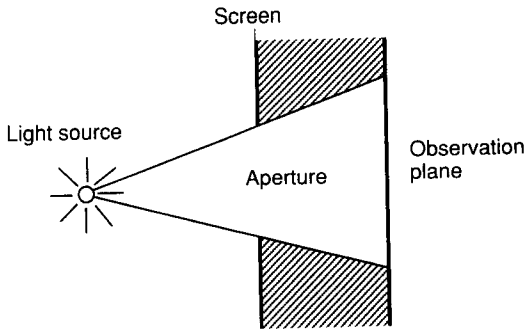

FIGURE 3.3
Arrangement used for observing diffraction of light.

1678. Huygens expressed the intuitive conviction that if each point on the wavefront of a disturbance were considered to be a new source of a "secondary" spherical disturbance, then the wavefront at a later instant could be found by constructing the "envelope" of the secondary wavelets, as illustrated in Fig. 3.4.

Progress on further understanding diffraction was impeded throughout the entire 18th century by the fact that Isaac Newton, a scientist with an enormous reputation for his many contributions to physics in general and to optics in particular, favored the corpuscular theory of light as early as 1704. His followers supported this view adamantly. It was not until 1804 that further significant progress occurred. In that year, Thomas Young, an English physician, strengthened the wave theory of light by introducing the critical concept of *interference*. The idea was a radical one at the time, for it stated that under proper conditions, light could be added to light and produce darkness.

The ideas of Huygens and Young were brought together in 1818 in the famous memoir of Augustin Jean Fresnel. By making some rather arbitrary assumptions about the amplitudes and phases of Huygens' secondary sources, and by allowing the various wavelets to mutually interfere, Fresnel was able to calculate the distribution of light in diffraction patterns with excellent accuracy.

At Fresnel's presentation of his paper to a prize committee of the French Academy of Sciences, his theory was strongly disputed by the great French mathematician S. Poisson, a member of the committee. He demonstrated the absurdity of the theory

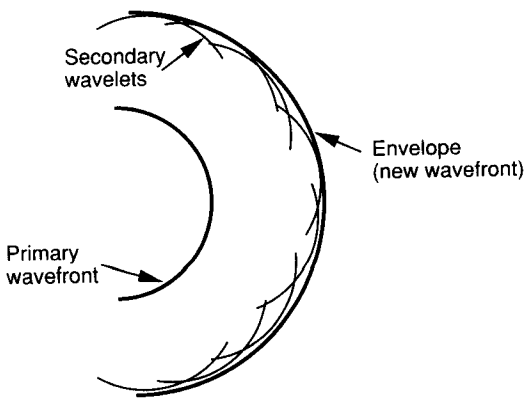

FIGURE 3.4
Huygens' envelope construction.

by showing that it predicted the existence of a bright spot at the center of the shadow of an opaque disk. F. Arago, who chaired the prize committee, performed such an experiment and found the predicted spot. Fresnel won the prize, and since then the effect has been known as "Poisson's spot".

In 1860 Maxwell identified light as an electromagnetic wave, a step of enormous importance. But it was not until 1882 that the ideas of Huygens and Fresnel were put on a firmer mathematical foundation by Gustav Kirchhoff, who succeeded in showing that the amplitudes and phases ascribed to the secondary sources by Fresnel were indeed logical consequences of the wave nature of light. Kirchhoff based his mathematical formulation upon two assumptions about the boundary values of the light incident on the surface of an obstacle placed in the way of propagation of light. These assumptions were later proved to be inconsistent with each other, by Poincaré in 1892 and by Sommerfeld in 1894.[1] As a consequence of these criticisms, Kirchhoff's formulation of the so-called *Huygens-Fresnel* principle must be regarded as a first approximation, although under most conditions it yields results that agree amazingly well with experiment. Kottler [174] attempted to resolve the contradictions by reinterpreting Kirchhoff's boundary value problem as a *saltus* problem, where *saltus* is a Latin word signifying a discontinuity or jump. The Kirchhoff theory was also modified by Sommerfeld, who eliminated one of the aforementioned assumptions concerning the light amplitude at the boundary by making use of the theory of Green's functions. This so-called *Rayleigh-Sommerfeld diffraction theory* will be treated in Section 3.5.

It should be emphasized from the start that the Kirchhoff and Rayleigh-Sommerfeld theories share certain major simplifications and approximations. Most important, light is treated as a *scalar* phenomenon, neglecting the fundamentally vectorial nature of the electromagnetic fields. Such an approach neglects the fact that, at boundaries, the various components of the electric and magnetic fields are coupled through Maxwell's equations and cannot be treated independently. Fortunately, experiments in the microwave region of the spectrum [262] have shown that the scalar theory yields very accurate results if two conditions are met: (1) the diffracting aperture must be large compared with a wavelength, and (2) the diffracting fields must not be observed too close to the aperture. These conditions will be well satisfied in the problems treated here. For a more complete discussion of the applicability of scalar theory in instrumental optics the reader may consult Ref. [28] (Section 8.4). Nonetheless, there do exist important problems for which the required conditions are *not* satisfied, for example in the theory of diffraction from high-resolution gratings and from extremely small pits on optical recording media. Such problems are excluded from consideration here, since the vectorial nature of the fields *must* be taken into account if reasonably accurate results are to be obtained. Vectorial generalizations of diffraction theory do exist, the first satisfactory treatment being due to Kottler [172].

The first truly rigorous solution of a diffraction problem was given in 1896 by Sommerfeld [268], who treated the two-dimensional case of a plane wave incident on an infinitesimally thin, perfectly conducting half plane. Kottler [173] later compared Sommerfeld's solution with the corresponding results of Kirchhoff's scalar treatment.

---

[1] For a more detailed discussion of these inconsistencies, see Section 3.5.

Needless to say, an historic introduction to a subject so widely mentioned in the literature can hardly be considered complete. The reader is therefore referred to more comprehensive treatments of diffraction theory, for example Refs. [13], [29], and [145].

3.2 FROM A VECTOR TO A SCALAR THEORY

The most fundamental beginning for our analysis is Maxwell's equations. In MKS units and in the absence of free charge, the equations are given by

$$\begin{aligned} \nabla \times \vec{\mathcal{E}} &= -\mu \frac{\partial \vec{\mathcal{H}}}{\partial t} \\ \nabla \times \vec{\mathcal{H}} &= \epsilon \frac{\partial \vec{\mathcal{E}}}{\partial t} \\ \nabla \cdot \epsilon \vec{\mathcal{E}} &= 0 \\ \nabla \cdot \mu \vec{\mathcal{H}} &= 0. \end{aligned} \tag{3-2}$$

Here $\vec{\mathcal{E}}$ is the electric field, with rectilinear components $(\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z)$, and $\vec{\mathcal{H}}$ is the magnetic field, with components $(\mathcal{H}_X, \mathcal{H}_Y, \mathcal{H}_Z)$. μ and ϵ are the permeability and permittivity, respectively, of the medium in which the wave is propagating. $\vec{\mathcal{E}}$ and $\vec{\mathcal{H}}$ are functions of both position P and time t. The symbols $\times$ and $\cdot$ represent a vector cross product and a vector dot product, respectively, while $\nabla = \frac{\partial}{\partial x}\hat{i} + \frac{\partial}{\partial y}\hat{j} + \frac{\partial}{\partial z}\hat{k}$, where $\hat{i}$, $\hat{j}$ and $\hat{k}$ are unit vectors in the x, y, and z directions, respectively.

We assume that the wave is propagating in a dielectric medium. It is important to further specify some properties of that medium. The medium is *linear* if it satisfies the linearity properties discussed in Chapter 2. The medium is *isotropic* if its properties are independent of the direction of polarization of the wave (i.e. the directions of the $\vec{\mathcal{E}}$ and $\vec{\mathcal{H}}$ vectors). The medium is *homogeneous* if the permittivity is constant throughout the region of propagation. The medium is *nondispersive* if the permittivity is independent of wavelength over the wavelength region occupied by the propagating wave. Finally, all media of interest in this book are *nonmagnetic*, which means that the magnetic permeability is always equal to μ_0, the vacuum permeability.

Applying the $\nabla\times$ operation to the left and right sides of the first equation for $\vec{\mathcal{E}}$, we make use of the vector identity

$$\nabla \times (\nabla \times \vec{\mathcal{E}}) = \nabla(\nabla \cdot \vec{\mathcal{E}}) - \nabla^2 \vec{\mathcal{E}}. \tag{3-3}$$

If the propagation medium is linear, isotropic, homogeneous (constant ϵ), and nondispersive, substitution of the two Maxwell's equations for $\vec{\mathcal{E}}$ in Eq. (3-3) yields

$$\nabla^2 \vec{\mathcal{E}} - \frac{n^2}{c^2} \frac{\partial^2 \vec{\mathcal{E}}}{\partial t^2} = 0 \tag{3-4}$$

where n is the *refractive index* of the medium, defined by

$$n = \left(\frac{\epsilon}{\epsilon_0} \right)^{1/2}, \tag{3-5}$$

ϵ_0 is the vacuum permittivity, and c is the velocity of propagation in vacuum, given by

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}. \tag{3-6}$$

The magnetic field satisfies an identical equation,

$$\nabla^2 \vec{\mathcal{H}} - \frac{n^2}{c^2} \frac{\partial^2 \vec{\mathcal{H}}}{\partial t^2} = 0.$$

Since the vector wave equation is obeyed by both $\vec{\mathcal{E}}$ and $\vec{\mathcal{H}}$, an identical scalar wave equation is obeyed by all components of those vectors. Thus, for example, $\mathcal{E}_X$ obeys the equation

$$\nabla^2 \mathcal{E}_X - \frac{n^2}{c^2} \frac{\partial^2 \mathcal{E}_X}{\partial t^2} = 0,$$

and similarly for $\mathcal{E}_Y, \mathcal{E}_Z, \mathcal{H}_X, \mathcal{H}_Y$, and $\mathcal{H}_Z$. Therefore it is possible to summarize the behavior of all components of $\vec{\mathcal{E}}$ and $\vec{\mathcal{H}}$ through a single scalar wave equation,

$$\nabla^2 u(P, t) - \frac{n^2}{c^2} \frac{\partial^2 u(P, t)}{\partial t^2} = 0, \tag{3-7}$$

where $u(P, t)$ represents any of the scalar field components, and we have explicitly introduced the dependence of u on both position P in space and time t.

From above we conclude that in a dielectric medium that is linear, isotropic, homogeneous, and nondispersive, all components of the electric and magnetic field behave identically and their behavior is fully described by a single scalar wave equation. How, then, is the scalar theory only an approximation, rather than exact? The answer becomes clear if we consider situations other than propagation in the uniform dielectric medium hypothesized.

For example, if the medium is inhomogeneous with a permittivity $\epsilon(P)$ that depends on position P (but not on time t), it is a simple matter to show (see Prob. 3-1) that the wave equation satisfied by $\vec{\mathcal{E}}$ becomes

$$\nabla^2 \vec{\mathcal{E}} + 2\nabla(\vec{\mathcal{E}} \cdot \nabla \ln n) - \frac{n^2}{c^2} \frac{\partial^2 \vec{\mathcal{E}}}{\partial t^2} = 0, \tag{3-8}$$

where n and c are again given by Eqs. (3-5) and (3-6). The new term that has been added to the wave equation will be nonzero for a refractive index that changes over space. More importantly, that term introduces a *coupling* between the various components of the electric field, with the result that $\mathcal{E}_X, \mathcal{E}_Y$, and $\mathcal{E}_Z$ may no longer satisfy the *same* wave equation. This type of coupling is important, for example, when light propagates through a "thick" dielectric diffraction grating.

A similar effect takes place when boundary conditions are imposed on a wave that propagates in a homogeneous medium. At the boundaries, coupling is introduced between $\vec{\mathcal{E}}$ and $\vec{\mathcal{H}}$, as well as between their various scalar components. As a consequence, even when the propagation medium is homogeneous, the use of a scalar theory entails some degree of error. That error will be small provided the boundary conditions have effect over an area that is a small part of the area through which a wave may be passing.

In the case of diffraction of light by an aperture, the $\vec{\mathcal{E}}$ and $\vec{\mathcal{H}}$ fields are modified only at the edges of the aperture where light interacts with the material of which the edges are composed, and the effects extend over only a few wavelengths into the aperture itself. Thus if the aperture has an area that is large compared with a wavelength, the coupling effects of the boundary conditions on the $\vec{\mathcal{E}}$ and $\vec{\mathcal{H}}$ fields will be small. As will be seen, this is equivalent to the requirement that the diffraction angles caused by the aperture are small.

With these discussions as background, we turn away from the vector theory of diffraction to the simpler scalar theory. We close with one final observation. Circuit theory is based on the approximation that circuit elements (resistors, capacitors, and inductors) are *small* compared to the wavelength of the fields that appear within them, and for this reason can be treated as lumped elements with simple properties. We need not use Maxwell's equations to analyze such elements under these conditions. In a similar vein, the scalar theory of diffraction introduces substantial simplifications compared with a full vectorial theory. The scalar theory is accurate provided that the diffracting structures are *large* compared with the wavelength of light. Thus the approximation implicit in the scalar theory should be no more disturbing than the approximation used in lumped circuit theory. In both cases it is possible to find situations in which the approximation breaks down, but as long as the simpler theories are used only in cases for which they are expected to be valid, the losses of accuracy will be small and the gain of simplicity will be large.

3.3 SOME MATHEMATICAL PRELIMINARIES

Before embarking on a treatment of diffraction itself, we first consider a number of mathematical preliminaries that form the basis of the later diffraction-theory derivations. These initial discussions will also serve to introduce some of the notation used throughout the book.

3.3.1 The Helmholtz Equation

In accord with the previous introduction of the scalar theory, let the light disturbance at position P and time t be represented by the scalar function $u(P, t)$. Attention is now restricted to the case of a purely monochromatic wave, with the generalization to polychromatic waves being deferred to Section 3.8.

For a monochromatic wave, the scalar field may be written explicitly as

$$u(P, t) = A(P) \cos[2\pi\nu t + \phi(P)] \tag{3-9}$$

where $A(P)$ and $\phi(P)$ are the amplitude and phase, respectively, of the wave at position P, while ν is the optical frequency. A more compact form of (3-9) is found by using complex notation, writing

$$u(P, t) = Re\{U(P) \exp(-j2\pi\nu t)\}, \tag{3-10}$$

where $Re\{\}$ signifies "real part of", and $U(P)$ is a complex function of position (sometimes called a *phasor*),

$$U(P) = A(P) \exp[-j\phi(P)]. \tag{3-11}$$

If the real disturbance $u(P, t)$ is to represent an optical wave, it must satisfy the scalar wave equation

$$\nabla^2 u - \frac{n^2}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \tag{3-12}$$

at each source-free point. As before, ∇^2 is the Laplacian operator, n represents the refractive index of the dielectric medium within which light is propagating, and c represents the vacuum velocity of light. The complex function $U(P)$ serves as an adequate description of the disturbance, since the time dependence is known a priori. If (3-10) is substituted in (3-12), it follows that U must obey the time-independent equation

$$(\nabla^2 + k^2)U = 0. \tag{3-13}$$

Here k is termed the *wave number* and is given by

$$k = 2\pi n\frac{\nu}{c} = \frac{2\pi}{\lambda},$$

and λ is the wavelength in the dielectric medium ($\lambda = c/n\nu$). The relation (3-13) is known as the *Helmholtz equation*; we may assume in the future that the complex amplitude of any monochromatic optical disturbance propagating in vacuum ($n = 1$) or in a homogeneous dielectric medium ($n > 1$) must obey such a relation.

3.3.2 Green's Theorem

Calculation of the complex disturbance U at an observation point in space can be accomplished with the help of the mathematical relation known as *Green's theorem.* This theorem, which can be found in most texts on advanced calculus, can be stated as follows:

> Let $U(P)$ and $G(P)$ be any two complex-valued functions of position, and let S be a closed surface surrounding a volume V. If U, G, and their first and second partial derivatives are single-valued and continuous within and on S, then we have
>
> $$\iiint_V (U\nabla^2 G - G\nabla^2 U)\,dv = \iint_S \left(U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right) ds \tag{3-14}$$
>
> where $\frac{\partial}{\partial n}$ signifies a partial derivative in the *outward* normal direction at each point on S.

This theorem is in many respects the prime foundation of scalar diffraction theory. However, only a prudent choice of an auxiliary function G and a closed surface S will allow its direct application to the diffraction problem. We turn now to the former of these problems, considering Kirchhoff's choice of an auxiliary function and the consequent integral theorem that follows.

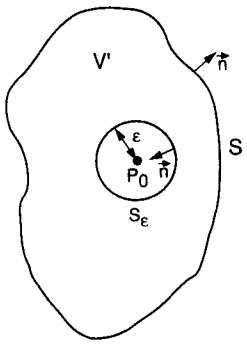

FIGURE 3.5
Surface of integration.

3.3.3 The Integral Theorem of Helmholtz and Kirchhoff

The Kirchhoff formulation of the diffraction problem is based on a certain integral theorem which expresses the solution of the homogeneous wave equation at an arbitrary point in terms of the values of the solution and its first derivative on an arbitrary closed surface surrounding that point. This theorem had been derived previously in acoustics by H. von Helmholtz.

Let the point of observation be denoted P_0, and let S denote an arbitrary closed surface surrounding P_0, as indicated in Fig. 3.5. The problem is to express the optical disturbance at P_0 in terms of its values on the surface S. To solve this problem, we follow Kirchhoff in applying Green's theorem and in choosing as an auxiliary function a unit-amplitude spherical wave expanding about the point P_0 (the so-called *free space* Green's function). Thus the value of Kirchhoff's G at an arbitrary point P_1 is given by[2]

$$G(P_1) = \frac{\exp(jkr_{01})}{r_{01}}, \tag{3-15}$$

where we adopt the notation that r_{01} is the length of the vector $\vec{r}_{01}$ pointing from P_0 to P_1.

Before proceeding further, a short diversion regarding Green's functions may be in order. Suppose that we wish to solve an inhomogeneous linear differential equation of the form

$$a_2(x)\frac{d^2U}{dx^2} + a_1(x)\frac{dU}{dx} + a_0(x)U = V(x) \tag{3-16}$$

where $V(x)$ is a driving function and $U(x)$ satisfies a known set of boundary conditions. We have chosen a one-dimensional variable x but the theory is easily generalized to a multidimensional $\vec{x}$. It can be shown (see Chapter 1 of [223] and [16]) that if $G(x)$ is the solution to the same differential equation (3-16) when $V(x)$ is replaced by the

[2]The reader may wish to verify that, for our choice of clockwise rotation of phasors, the description of an expanding wave should have a + sign in the exponential.

impulsive driving function $\delta(x - x')$ and with the same boundary conditions applying, then the general solution $U(x)$ can be expressed in terms of the specific solution $G(x)$ through a convolution integral

$$U(x) = \int G(x - x')\, V(x')\, dx'. \tag{3-17}$$

The function $G(x)$ is known as the *Green's function* of the problem, and is clearly a form of impulse response. Various solutions to the scalar diffraction problem to be discussed in the following sections correspond to results obtained under different assumptions about the Green's function of the problem. The function G appearing in Green's theorem may be regarded either as simply an auxiliary function which we cleverly choose to solve our problem, or it may eventually be related to the Green's function of the problem. Further consideration of the theory of Green's functions is beyond the scope of this treatment.

Returning now to our central discussion, to be legitimately used in Green's theorem, the function G (as well as its first and second partial derivatives) must be continuous within the enclosed volume V. Therefore to exclude the discontinuity at P_0, a small spherical surface S_ϵ, of radius ϵ, is inserted about the point P_0. Green's theorem is then applied, the volume of integration V' being that volume lying between S and S_ϵ, and the surface of integration being the composite surface

$$S' = S + S_\epsilon$$

as indicated in Fig. 3.5. Note that the "outward" normal to the composite surface points outward in the conventional sense on S, but inward (towards P_0) on S_ϵ.

Within the volume V', the disturbance G, being simply an expanding spherical wave, satisfies the Helmholtz equation

$$(\nabla^2 + k^2)G = 0. \tag{3-18}$$

Substituting the two Helmholtz equations (3-13) and (3-18) in the left-hand side of Green's theorem, we find

$$\iiint_{V'} (U\nabla^2 G - G\nabla^2 U)\, dv = -\iiint_{V'} (UGk^2 - GUk^2)\, dv \equiv 0.$$

Thus the theorem reduces to

$$\iint_{S'} \left(U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right) ds = 0$$

or

$$-\iint_{S_\epsilon} \left(U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right) ds = \iint_{S} \left(U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right) ds. \tag{3-19}$$

Note that, for a general point P_1 on S', we have

$$G(P_1) = \frac{\exp(jkr_{01})}{r_{01}}$$

and

$$\frac{\partial G(P_1)}{\partial n} = \cos(\vec{n}, \vec{r}_{01})\left(jk - \frac{1}{r_{01}}\right)\frac{\exp(jkr_{01})}{r_{01}} \tag{3-20}$$

where $\cos(\vec{n}, \vec{r}_{01})$ represents the cosine of the angle between the outward normal $\vec{n}$ and the vector $\vec{r}_{01}$ joining P_0 to P_1. For the particular case of P_1 on S_ϵ, $\cos(\vec{n}, \vec{r}_{01}) = -1$, and these equations become

$$G(P_1) = \frac{e^{jk\epsilon}}{\epsilon} \qquad \text{and} \qquad \frac{\partial G(P_1)}{\partial n} = \frac{e^{jk\epsilon}}{\epsilon}\left(\frac{1}{\epsilon} - jk\right).$$

Letting ϵ become arbitrarily small, the continuity of U (and its derivatives) at P_0 allows us to write

$$\lim_{\epsilon \to 0} \iint_{S_\epsilon} \left(U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right) ds$$

$$= \lim_{\epsilon \to 0} 4\pi\epsilon^2 \left[U(P_0)\frac{\exp(jk\epsilon)}{\epsilon}\left(\frac{1}{\epsilon} - jk\right) - \frac{\partial U(P_0)}{\partial n}\frac{\exp(jk\epsilon)}{\epsilon}\right] = 4\pi U(P_0).$$

Substitution of this result in (3-19) (taking account of the negative sign) yields

$$U(P_0) = \frac{1}{4\pi}\iint_S \left\{\frac{\partial U}{\partial n}\left[\frac{\exp(jkr_{01})}{r_{01}}\right] - U\frac{\partial}{\partial n}\left[\frac{\exp(jkr_{01})}{r_{01}}\right]\right\} ds. \tag{3-21}$$

This result is known as the *integral theorem of Helmholtz and Kirchhoff*; it plays an important role in the development of the scalar theory of diffraction, for it allows the field at any point P_0 to be expressed in terms of the "boundary values" of the wave on any closed surface surrounding that point. As we shall now see, such a relation is instrumental in the further development of scalar diffraction equations.

3.4 THE KIRCHHOFF FORMULATION OF DIFFRACTION BY A PLANAR SCREEN

Consider now the problem of diffraction of light by an aperture in an infinite opaque screen. As illustrated in Fig. 3.6, a wave disturbance is assumed to impinge on the screen and the aperture from the left, and the field at the point P_0 behind the aperture is to be calculated. Again the field is assumed to be monochromatic.

3.4.1 Application of the Integral Theorem

To find the field at the point P_0, we apply the integral theorem of Helmholtz and Kirchhoff, being careful to choose a surface of integration that will allow the calculation to be performed successfully. Following Kirchhoff, the closed surface S is chosen to consist of two parts, as shown in Fig. 3.6. Let a plane surface, S_1, lying directly behind the diffracting screen, be joined and closed by a large spherical cap, S_2, of radius R and

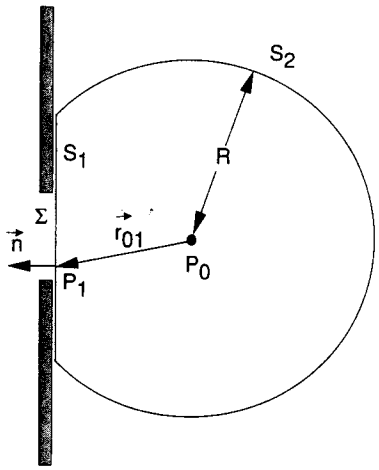

FIGURE 3.6
Kirchhoff formulation of diffraction by a plane screen.

centered at the observation point P_0. The total closed surface S is simply the sum of S_1 and S_2. Thus, applying (3-21),

$$U(P_0) = \frac{1}{4\pi} \iint_{S_1+S_2} \left(G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n} \right) ds$$

where, as before,

$$G = \frac{\exp(jkr_{01})}{r_{01}}.$$

As R increases, S_2 approaches a large hemispherical shell. It is tempting to reason that, since both U and G will fall off as $1/R$, the integrand will ultimately vanish, yielding a contribution of zero from the surface integral over S_2. However, the area of integration increases as R^2, so this argument is incomplete. It is also tempting to assume that, since the disturbances are propagating with finite velocity c/n, R will ultimately be so large that the waves have not yet reached S_2, and the integrand will be zero on that surface. But this argument is incompatible with our assumption of monochromatic disturbances, which must (by definition) have existed for all time. Evidently a more careful investigation is required before the contribution from S_2 can be disposed of.

Examining this problem in more detail, we see that, on S_2,

$$G = \frac{\exp(jkR)}{R}$$

and, from (3-20),

$$\frac{\partial G}{\partial n} = \left(jk - \frac{1}{R} \right) \frac{\exp(jkR)}{R} \approx jkG$$

where the last approximation is valid for large R. The integral in question can thus be reduced to

$$\iint_{S_2} \left[G \frac{\partial U}{\partial n} - U(jkG) \right] ds = \int_{\Omega} G \left(\frac{\partial U}{\partial n} - jkU \right) R^2 \, d\omega,$$

where Ω is the solid angle subtended by S_2 at P_0. Now the quantity $|RG|$ is uniformly bounded on S_2. Therefore the entire integral over S_2 will vanish as R becomes arbitrarily large, provided the disturbance has the property

$$\lim_{R\to\infty} R\left(\frac{\partial U}{\partial n} - jkU\right) = 0 \tag{3-22}$$

uniformly in angle. This requirement is known as the *Sommerfeld radiation condition* [269] and is satisfied if the disturbance U vanishes at least as fast as a diverging spherical wave (see Prob. 3-2). It guarantees that we are dealing only with *outgoing* waves on S_2, rather than incoming waves, for which the integral over S_2 might not vanish as $R \to \infty$. Since only outgoing waves will fall on S_2 in our problem, the integral over S_2 will yield a contribution of precisely zero.

3.4.2 The Kirchhoff Boundary Conditions

Having disposed of the integration over the surface S_2, it is now possible to express the disturbance at P_0 in terms of the disturbance and its normal derivative over the infinite plane S_1 immediately behind the screen, that is,

$$U(P_0) = \frac{1}{4\pi}\iint_{S_1}\left(\frac{\partial U}{\partial n}G - U\frac{\partial G}{\partial n}\right)ds. \tag{3-23}$$

The screen is opaque, except for the open aperture which will be denoted Σ. It therefore seems intuitively reasonable that the major contribution to the integral (3-23) arises from the points of S_1 located within the aperture Σ, where we would expect the integrand to be largest. Kirchhoff accordingly adopted the following assumptions [162]:

1. Across the surface Σ, the field distribution U and its derivative $\partial U/\partial n$ are exactly the same as they would be in the absence of the screen.
2. Over the portion of S_1 that lies in the geometrical shadow of the screen, the field distribution U and its derivative $\partial U/\partial n$ are identically zero.

These conditions are commonly known as the *Kirchhoff boundary conditions*. The first allows us to specify the disturbance incident on the aperture by neglecting the presence of the screen. The second allows us to neglect all of the surface of integration except that portion lying directly within the aperture itself. Thus (3-23) is reduced to

$$U(P_0) = \frac{1}{4\pi}\iint_{\Sigma}\left(\frac{\partial U}{\partial n}G - U\frac{\partial G}{\partial n}\right)ds. \tag{3-24}$$

While the Kirchhoff boundary conditions simplify the results considerably, it is important to realize that neither can be exactly true. The presence of the screen will inevitably perturb the fields on Σ to some degree, for along the rim of the aperture certain boundary conditions must be met that would not be required in the absence of the screen. In addition, the shadow behind the screen is never perfect, for fields will inevitably extend behind the screen for a distance of several wavelengths. However, if the dimensions of the aperture are large compared with a wavelength, these fringing

effects can be safely neglected,[3] and the two boundary conditions can be used to yield results that agree very well with experiment.

3.4.3 The Fresnel-Kirchhoff Diffraction Formula

A further simplification of the expression for $U(P_0)$ is obtained by noting that the distance r_{01} from the aperture to the observation point is usually many optical wavelengths, and therefore, since $k \gg 1/r_{01}$, Eq. (3-20) becomes

$$\frac{\partial G(P_1)}{\partial n} = \cos(\vec{n}, \vec{r}_{01})\left(jk - \frac{1}{r_{01}}\right)\frac{\exp(jkr_{01})}{r_{01}}$$

$$\approx jk\cos(\vec{n}, \vec{r}_{01})\frac{\exp(jkr_{01})}{r_{01}}. \tag{3-25}$$

Substituting this approximation and the expression (3-15) for G in Eq. (3-24), we find

$$U(P_0) = \frac{1}{4\pi}\iint_{\Sigma}\frac{\exp(jkr_{01})}{r_{01}}\left[\frac{\partial U}{\partial n} - jkU\cos(\vec{n}, \vec{r}_{01})\right]ds. \tag{3-26}$$

Now suppose that the aperture is illuminated by a single spherical wave,

$$U(P_1) = \frac{A\exp(jkr_{21})}{r_{21}}$$

arising from a point source at P_2, a distance r_{21} from P_1 (see Fig. 3.7). If r_{21} is many optical wavelengths, then (3-26) can be directly reduced (see Prob. 3-3) to

$$U(P_0) = \frac{A}{j\lambda}\iint_{\Sigma}\frac{\exp[jk(r_{21} + r_{01})]}{r_{21}r_{01}}\left[\frac{\cos(\vec{n}, \vec{r}_{01}) - \cos(\vec{n}, \vec{r}_{21})}{2}\right]ds. \tag{3-27}$$

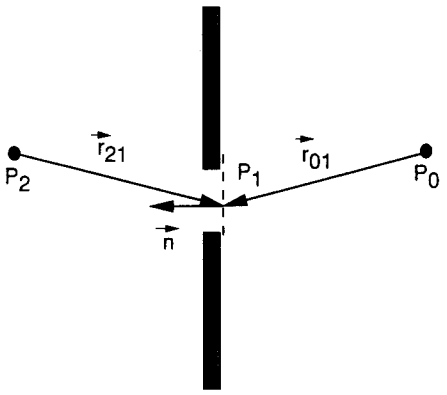

FIGURE 3.7
Point-source illumination of a plane screen.

[3] As we shall see, objections to the use of the Kirchhoff boundary conditions arise, not because of the fringing effects, but rather because of certain internal inconsistencies.

This result, which holds only for an illumination consisting of a single point source, is known as the *Fresnel-Kirchhoff diffraction formula.*

Note that Eq. (3-27) is symmetrical with respect to the illumination point source at P_2 and the observation point at P_0. Thus a point source at P_0 will produce at P_2 the same effect that a point source of equal intensity placed at P_2 will produce at P_0. This result is referred to as the *reciprocity theorem of Helmholtz.*

Finally, we point out an interesting interpretation of the diffraction formula (3-27), to which we will return later for a more detailed discussion. Let that equation be rewritten as follows:

$$U(P_0) = \iint_{\Sigma} U'(P_1)\frac{\exp(jkr_{01})}{r_{01}}\,ds \tag{3-28}$$

where

$$U'(P_1) = \frac{1}{j\lambda}\left[\frac{A\exp(jkr_{21})}{r_{21}}\right]\left[\frac{\cos(\vec{n},\vec{r}_{01}) - \cos(\vec{n},\vec{r}_{21})}{2}\right]. \tag{3-29}$$

Now (3-28) may be interpreted as implying that the field at P_0 arises from an infinity of fictitious "secondary" point sources located within the aperture itself. The secondary sources have certain amplitudes and phases, described by $U'(P_1)$, that are related to the illuminating wavefront and the angles of illumination and observation. Assumptions resembling these were made by Fresnel rather arbitrarily in his combination of Huygens' envelope construction and Young's principle of interference. Fresnel *assumed* these properties to hold in order to obtain accurate results. Kirchhoff showed that such properties are a natural consequence of the wave nature of light.

Note that the above derivation has been restricted to the case of an aperture illumination consisting of a single expanding spherical wave. However, as we shall now see, such a limitation can be removed by the Rayleigh-Sommerfeld theory.

3.5 THE RAYLEIGH-SOMMERFELD FORMULATION OF DIFFRACTION

The Kirchhoff theory has been found experimentally to yield remarkably accurate results and is widely used in practice. However, there are certain internal inconsistencies in the theory which motivated a search for a more satisfactory mathematical development. The difficulties of the Kirchhoff theory stem from the fact that boundary conditions must be imposed on *both* the field strength and its normal derivative. In particular, it is a well-known theorem of potential theory that if a two-dimensional potential function and its normal derivative vanish *together* along any finite curve segment, then that potential function *must vanish over the entire plane.* Similarly, if a solution of the three-dimensional wave equation vanishes on any finite surface element, it must vanish in all space. Thus the two Kirchhoff boundary conditions together imply that the field is zero everywhere behind the aperture, a result which contradicts the known physical situation. A further indication of these inconsistencies is the fact that the Fresnel-Kirchhoff diffraction formula can be shown to fail to reproduce the assumed boundary conditions as the observation point approaches the screen or aperture. In view of these