G. R. Fowles
Introduction to Modern Optics
Holt Rhinehart Winston, N.Y.
2nd ed. 1975

5.1 General Description of Diffraction

If an opaque object is placed between a point source of light and a white screen, it is found that the shadow that is cast by the object departs from the perfect sharpness predicted by geometrical optics. Close examination of the shadow edge reveals that some light goes over into the dark zone of the geometrical shadow and that dark fringes appear in the illuminated zone. This "smearing" of the shadow edge is closely related to another phenomenon, namely, the spreading of light after passing through a very small aperture, such as a pinhole or a narrow slit, as in Young's experiment. The collective name given to these departures from geometrical optics is *diffraction*.

The essential features of diffraction phenomena can be explained qualitatively by Huygens' principle. This principle in its original form states that the propagation of a light wave can be predicted by assuming that each point of the wave front acts as the source of a secondary wave that spreads out in all directions. The envelope of all the secondary waves is the new wave front.

We shall not attempt to treat diffraction by a direct application of Huygens' principle. We want a more quantitative approach. Our strategy will be to cast Huygens' principle into a precise mathematical form known as the *Fresnel-Kirchhoff formula*. This formula will then be applied to various specific cases of diffraction of light by obstacles and apertures.

5.2 Fundamental Theory

Let us recall Green's theorem[1] that states that if U and V are any two scalar-point functions that satisfy the usual conditions of continuity and integrability, then the following equality holds:

$$\iint (V \operatorname{grad}_n U - U \operatorname{grad}_n V)\, d\mathscr{A} = \iiint (V \nabla^2 U - U \nabla^2 V)\, d\mathscr{V} \qquad (5.1)$$

The left-hand integral extends over any closed surface $\mathscr{A}$, and the right-hand integral includes the volume $\mathscr{V}$ within that surface. By "grad$_n$" is meant the normal component of the gradient at the surface of integration.[1]

In particular, if both U and V are wave functions; that is, if they satisfy the regular wave equations

$$\nabla^2 U = \frac{1}{u^2}\frac{\partial^2 U}{\partial t^2} \qquad \nabla^2 V = \frac{1}{u^2}\frac{\partial^2 V}{\partial t^2}$$

and if they both have a harmonic time dependence of the form $e^{\pm i\omega t}$, then it is straightforward to show that the volume integral in Green's theorem is identically zero. The theorem then reduces to

$$\iint (V \operatorname{grad}_n U - U \operatorname{grad}_n V)\, d\mathscr{A} = 0 \qquad (5.2)$$

Now suppose that we take V to be the wave function

$$V = V_0 \frac{e^{i(kr \pm \omega t)}}{r} \qquad (5.3)$$

Figure 5.1. Surface of integration for proving the Kirchhoff integral theorem.

This particular function represents spherical waves converging to the point P ($r = 0$). We let the volume enclosed by the surface of integration include the point P. Since V becomes infinite at P, we must exclude that point from the integration. This is accomplished by the

[1] Green's theorem can be proved from the divergence theorem $\iint \operatorname{grad}_n F\, d\mathscr{A} = \iiint \nabla \cdot F\, d\mathscr{V}$ by setting $F = U \nabla V - V \nabla U$ and using the vector identity $\nabla \cdot (U \nabla V) = U \nabla^2 V + (\nabla U) \cdot (\nabla V)$.

standard method of subtracting an integral over a small sphere of radius ρ centered at P, as indicated in Figure 5.1. Over this small sphere, $r = \rho$ and $\text{grad}_n = -\partial/\partial r$. Hence we can write

$$\iint \left(\frac{e^{ikr}}{r}\,\text{grad}_n\, U - U\,\text{grad}_n\,\frac{e^{ikr}}{r}\right) d\mathscr{A} - \iint \left(\frac{e^{ikr}}{r}\frac{\partial U}{\partial r} - U\frac{\partial}{\partial r}\frac{e^{ikr}}{r}\right)_{r=\rho} \rho^2\, d\Omega = 0 \tag{5.4}$$

where $d\Omega$ is the element of solid angle on the sphere centered at P, and $\rho^2\, d\Omega$ is the corresponding element of area. The common factor $V_0 e^{i\omega t}$ has been canceled out.

We now let ρ shrink to zero. Then, in the limit as ρ approaches zero the integrand of the second integral approaches the value that U has at the point P, namely U_p. This is easily verified by performing the indicated operations. Consequently, the second integral itself, including the sign, approaches the value

$$\iint U_p\, d\Omega = 4\pi U_p \tag{5.5}$$

Equation (5.4) then becomes, on rearranging terms,

$$U_p = -\frac{1}{4\pi}\iint \left(U\,\text{grad}_n\,\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\,\text{grad}_n\, U\right) d\mathscr{A} \tag{5.6}$$

This equation is known as the *Kirchhoff integral theorem*. It relates the value of any scalar wave function at any point P *inside* an arbitrary closed surface to the value of the wave function *at* the surface.

In the application of Kirchhoff's theorem to diffraction, the wave function U is known as the "optical disturbance." Being a scalar quantity, it cannot accurately represent an electromagnetic field. However, in this so-called "scalar approximation" the square of the absolute value of U may be regarded as a measure of the irradiance at a given point.

The more rigorous theory of diffraction, which takes into account the vectorial nature of light, is beyond the scope of this book. Owing to the mathematical complexity of the rigorous theory, complete calculations have been carried out for only a relatively few simple cases [5].

The Fresnel-Kirchhoff Formula We now proceed to apply the Kirchhoff integral theorem to the general problem of diffraction of light. The diffraction is produced by an aperture of arbitrary shape in an otherwise opaque partition. This partition separates a light source from a receiving point (Figure 5.2).

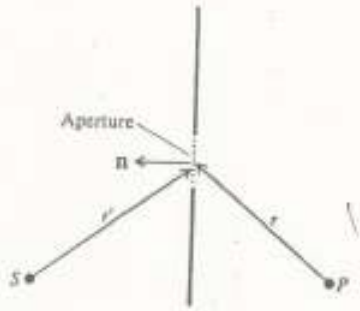


Figure 5.2. Geometry for the Fresnel-Kirchhoff formula.

Our task is to determine the optical disturbance reaching the receiving point P from the source S. In applying the Kirchhoff integral we choose a surface of integration such that it encloses the receiving point and includes, as a part of it, the aperture opening as indicated in the figure.

Two basic simplifying assumptions are introduced:

(1) The wave function U and its gradient contribute negligible amounts to the integral except at the aperture opening itself.
(2) The values of U and grad U at the aperture are the same as they would be in the absence of the partition.

Although the validity of these assumptions is open to considerable debate, the results are generally in good agreement with experimental observations.

If r' denotes the position of a point on the aperture relative to the source S, then the wave function at the aperture is given by the expression

$$U = U_0\,\frac{e^{i(kr'-\omega t)}}{r'} \tag{5.7}$$

which represents spherical monochromatic waves traveling outward from S. The Kirchhoff integral theorem then yields

$$U_p = \frac{U_0 e^{-i\omega t}}{4\pi}\iint \left(\frac{e^{ikr}}{r}\,\text{grad}_n\,\frac{e^{ikr'}}{r'} - \frac{e^{ikr'}}{r'}\,\text{grad}_n\,\frac{e^{ikr}}{r}\right) d\mathscr{A} \tag{5.8}$$

where the integration extends only over the aperture opening.

The operations indicated in the integrand are carried out as follows:

$$\text{grad}_n \frac{e^{ikr}}{r} = \cos(\mathbf{n},\mathbf{r}) \frac{\partial}{\partial r} \frac{e^{ikr}}{r} = \cos(\mathbf{n},\mathbf{r}) \left(\frac{ike^{ikr}}{r} - \frac{e^{ikr}}{r^2} \right) \quad (5.9)$$

$$\text{grad}_n \frac{e^{ikr'}}{r'} = \cos(\mathbf{n},\mathbf{r}') \frac{\partial}{\partial r'} \frac{e^{ikr'}}{r'} = \cos(\mathbf{n},\mathbf{r}') \left(\frac{ike^{ikr'}}{r'} - \frac{e^{ikr'}}{r'^2} \right) \quad (5.10)$$

where $(\mathbf{n},\mathbf{r}')$ and $(\mathbf{n},\mathbf{r}')$ denote the angles between the vectors and the normal to the surface of integration. Now, in Equations (5.9) and (5.10) the second terms in parentheses are negligibly small compared to the first terms in the normal situation where both r and r' are much larger than the wavelength of the radiation because $k = 2\pi/\lambda$. Consequently Equation (5.8) gives

$$U_p = -\frac{ikU_0 e^{-i\omega t}}{4\pi} \iint \frac{e^{ik(r+r')}}{rr'} [\cos(\mathbf{n},\mathbf{r}) - \cos(\mathbf{n},\mathbf{r}')] \, d\mathscr{A} \quad (5.11)$$

This equation is known as the *Fresnel Kirchhoff integral formula*. It is, in effect, a mathematical statement of Huygens' principle. This is most easily seen by applying the formula to a specific case, namely, that of a circular aperture with the source symmetrically located as shown in Figure 5.3. The surface of integration is taken to be a spher-

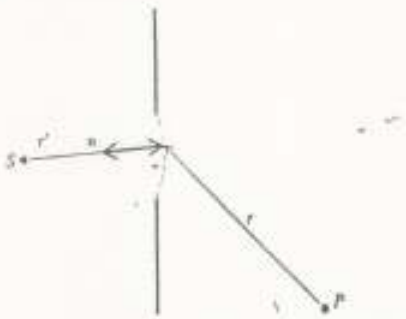


Figure 5.3. Diagram to show how Huygens' principle follows from the Kirchhoff integral formula.

ical cap bounded by the aperture opening. In this case r' is constant and $\cos(\mathbf{n},\mathbf{r}') = -1$. The Fresnel-Kirchhoff formula then reduces to

$$U_p = -\frac{ik}{4\pi} \iint \frac{U_{\mathscr{A}} e^{ik(r-\omega t)}}{r} [\cos(\mathbf{n},\mathbf{r}) + 1] \, d\mathscr{A} \quad (5.12)$$

where

$$U_{\mathscr{A}} = \frac{U_0 e^{ikr'}}{r'}$$

Equation (5.12) can be given the following simple interpretation: $U_{\mathscr{A}}$ is the complex amplitude of the incident primary wave at the aperture. From this primary wave each element $d\mathscr{A}$ of the aperture gives rise to a secondary spherical wave

$$\frac{U_{\mathscr{A}} e^{ik(r-\omega t)}}{r} d\mathscr{A}$$

The total optical disturbance at the receiving point P is obtained by summing the secondary waves from each element. However, in the summation it is necessary to take into account the factor $\cos(\mathbf{n},\mathbf{r}) - \cos(\mathbf{n},\mathbf{r}')$ known as the *obliquity factor*. In the case under discussion $\cos(\mathbf{n},\mathbf{r}') = -1$, so the obliquity factor is $\cos(\mathbf{n},\mathbf{r}) + 1$. In the forward direction $\cos(\mathbf{n},\mathbf{r}) = 1$, and the obliquity factor is then equal to 2, its maximum value. On the other hand, in the backward direction $\cos(\mathbf{n},\mathbf{r}) = -1$, so the obliquity factor is zero. This explains why there is no backward progressing wave created by the original wave front. Huygens' principle, as originally proposed, did not include the obliquity factor and thus could not account for the absence of a backward wave. The presence of the factor $-i$ means that the diffracted waves are shifted in phase by 90 degrees with respect to the primary incident wave. This feature was also lacking in the original form of Huygens' principle.

Complementary Apertures. Babinet's Principle Consider a diffracting aperture $\mathscr{A}$ that produces a certain optical disturbance U_p at a given observing point P. Suppose, now, that the aperture is divided into two portions $\mathscr{A}_1$ and $\mathscr{A}_2$ such that $\mathscr{A} = \mathscr{A}_1 + \mathscr{A}_2$. The two apertures $\mathscr{A}_1$ and $\mathscr{A}_2$ are then said to be *complementary*. An example is shown in Figure 5.4. From the form of the Fresnel-Kirchhoff formula, it is clear that

$$U_p = U_{1p} + U_{2p} \quad (5.13)$$

where U_{1p} is the optical disturbance at P produced by aperture $\mathscr{A}_1$ alone, and U_{2p} is the optical disturbance produced by aperture $\mathscr{A}_2$ alone. The equation above is one form of a theorem known as *Babinet's principle*.

Babinet's principle is useful in certain special cases. For instance, if $U_p = 0$, then $U_{1p} = -U_{2p}$. The complementary apertures in this case yield identical optical disturbances, except that they differ in phase by 180 degrees. The intensity at P, being equal to the absolute square of the optical disturbance, is therefore the same for the two apertures. Thus a collimated beam of light, such as that from a searchlight, will undergo diffraction scattering from a small spherical particle or a large number of particles such as fog. The condition $U_p = 0$ is then valid for those points P not in the direct beam, the

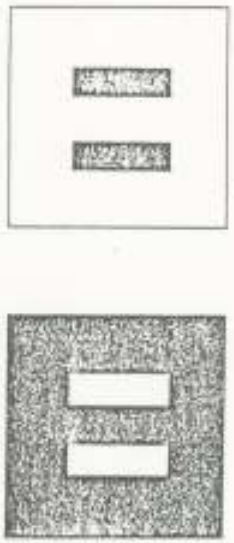


Figure 5.4. Complementary apertures.

aperture being defined by the size of the beam itself. Then, according to Babinet's principle, the same diffraction pattern will result if the beam is diffused by a screen containing a small circular hole or a large number of randomly placed holes the same size as the fog particles.

5.3 Fraunhofer and Fresnel Diffraction

In the detailed treatment of diffraction it is customary to distinguish between two general cases. These are known as *Fraunhofer diffraction* and *Fresnel diffraction*. Qualitatively speaking, Fraunhofer diffraction occurs when both the incident and diffracted waves are effectively plane. This will be the case when the distances from the source to the diffracting aperture and from the aperture to the receiving point are both large enough for the curvatures of the incident and diffracted waves to be neglected [Figure 5.5 (a)].

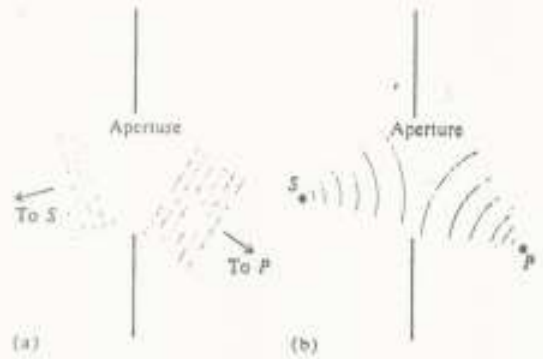


Figure 5.5. Diffraction by an aperture. (a) Fraunhofer case; (b) Fresnel case.

If either the source or the receiving point is close enough to the diffracting aperture so that the curvature of the wave front is significant, then one has Fresnel diffraction [Figure 5.5(b)]. There is, of course, no sharp line of distinction between the two cases. However, a quantitative criterion can be obtained as follows. Consider Figure 5.6, which shows the general geometry of the diffraction problem.

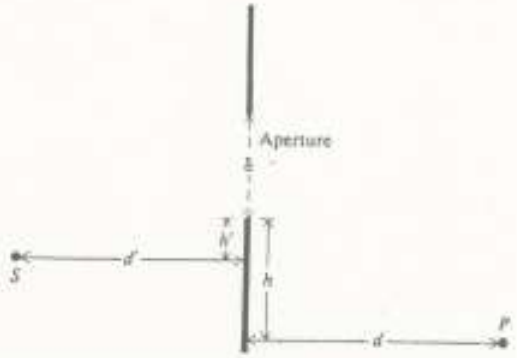


Figure 5.6. Geometry to show distinction between Fraunhofer diffraction and Fresnel diffraction.

The receiving point P is located a distance d from the plane of the diffracting aperture, and the source S is a distance d' from this plane. One edge of the aperture is located a distance h from the foot of the perpendicular drawn from P to the plane of the aperture. The corresponding distance for the source is h' as shown. The size of the aperture opening is δ. From the figure it is seen that the variation Δ of the quantity $r + r'$ from one edge of the aperture to the other is given by

$$\Delta = \sqrt{d'^2 + (h' + \delta)^2} + \sqrt{d^2 + (h + \delta)^2} - \sqrt{d'^2 + h'^2} - \sqrt{d^2 + h^2}$$

$$\approx \left(\frac{h'}{d'} + \frac{h}{d}\right)\delta + \frac{1}{2}\left(\frac{1}{d'} + \frac{1}{d}\right)\delta^2 + \cdots \qquad (5.14)$$

The quadratic term in the expansion above is essentially a measure of the curvature of the wave front. The wave is effectively plane over the aperture if this term is negligibly small compared to the wavelength of the light, that is, if

$$\frac{1}{2}\left(\frac{1}{d'} + \frac{1}{d}\right)\delta^2 \ll \lambda \qquad (5.15)$$

This is the criterion for Fraunhofer diffraction. If this condition does not obtain, the curvature of the wave front becomes important and the diffraction is of the Fresnel type. Similar considerations apply in the case of diffraction by an opaque object or obstacle. Then δ is the linear size of the object. (Note that Babinet's principle applies here.)

Examples of Fraunhofer and Fresnel diffraction by various types of apertures are treated in the sections that follow. Since the Fraunhofer case is, in general, mathematically simpler than the Fresnel case, Fraunhofer diffraction will be discussed first.

5.4 Fraunhofer Diffraction Patterns

The usual experimental arrangement for observing Fraunhofer diffraction is shown in Figure 5.7. Here the aperture is *coherently*

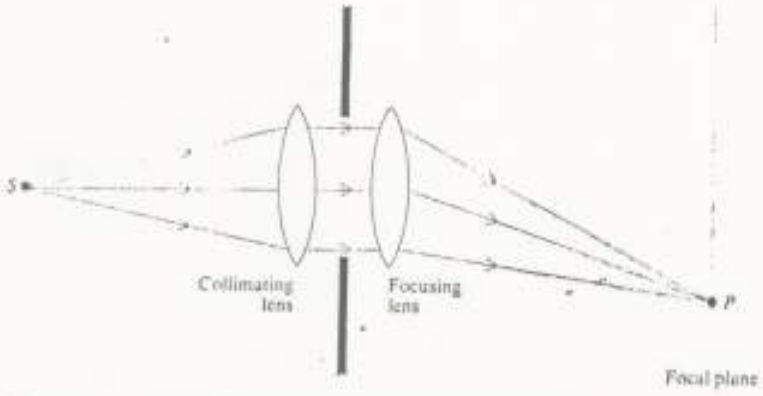


Figure 5.7. Arrangement for observing Fraunhofer diffraction.

illuminated by means of a point monochromatic source and a collimating lens. A second lens is placed behind the aperture as shown. The incident and diffracted wave fronts are therefore strictly plane, and the Fraunhofer case is rigorously valid. In applying the Fresnel-Kirchhoff formula [Equation (5.11)] to the calculation of the diffraction patterns, the following simplifying approximations are taken to be valid:

(1) The angular spread of the diffracted light is small enough for the obliquity factor $[\cos(\mathbf{n},\mathbf{r}) - \cos(\mathbf{n},\mathbf{r}')]$ not to vary appreciably over the aperture and to be taken outside the integral.

(2) The quantity $e^{ikr'}/r'$ is very nearly constant and can be taken outside the integral.

(3) The variation of the remaining factor e^{ikr}/r over the aperture comes principally from the exponential part, so the factor $1/r$ can be replaced by its mean value and taken outside the integral.

Consequently, the Fresnel-Kirchhoff formula reduces to the very simple equation

$$U_p = C \iint e^{ikr}\, d\mathscr{A} \tag{5.16}$$

where all constant factors have been lumped into one constant C. The formula above states that the distribution of the diffracted light is obtained simply by integrating the phase factor e^{ikr} over the aperture.

The Single Slit The case of diffraction by a single narrow slit is treated here as a one-dimensional problem. Let the slit be of length L

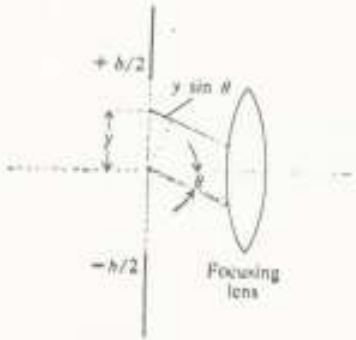


Figure 5.8. Definition of the variables for Fraunhofer diffraction by a single slit.

and of width b. The element of area is then $d\mathscr{A} = L\, dy$ as indicated in Figure 5.8. Furthermore, we can express r as

$$r = r_0 + y \sin\theta \tag{5.17}$$

where r_0 is the value of r for $y = 0$, and where θ is the angle shown. The diffraction formula (5.16) then yields

$$\begin{aligned} U &= Ce^{ikr_0} \int_{-b/2}^{+b/2} e^{iky\sin\theta} L\, dy \\ &= 2\,Ce^{ikr_0} L\, \frac{\sin(\tfrac{1}{2}kb\sin\theta)}{k\sin\theta} = C'\left(\frac{\sin\beta}{\beta}\right) \end{aligned} \tag{5.18}$$

where $\beta = \frac{1}{2}kb\sin\theta$, and $C' = e^{ikr_0}CbL$ is merely another constant.

Thus $C'(\sin\beta/\beta)$ is the total amplitude of the light diffracted in a given direction defined by β. This light is brought to a focus by the second lens, and the corresponding irradiance distribution in the focal plane is given by the expression

$$I = |U|^2 = I_0\left(\frac{\sin\beta}{\beta}\right)^2 \tag{5.19}$$

where $I_0 = |CLb|^2$, which is the irradiance for $\theta = 0$. The distribution is plotted in Figure 5.9. The maximum value occurs at $\theta = 0$, and zero

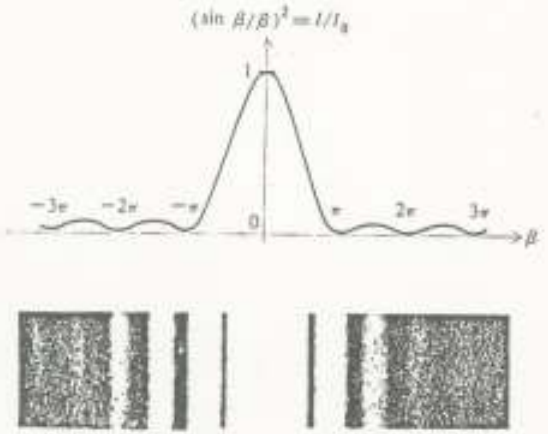


Figure 5.9. Fraunhofer diffraction pattern of a single slit.

values occur for $\beta = \pm\pi, \pm 2\pi, \ldots$, and so forth. Secondary maxima of rapidly diminishing value occur between these zero values. Thus the diffraction pattern at the focal plane consists of a central bright band. On either side there are alternating bright and dark bands. Table 5.1 gives the relative values of I of the first three sec-

Table 5.1. Relative Values of the Maxima of Diffraction Patterns of Rectangular and Circular Apertures

	Rectangular	*Circular*
Central Max	1	1
1st Max	0.0496	0.0174
2d Max	0.0168	0.0042
3rd Max	0.0083	0.0016

ondary maxima. The first minimum, $\beta = \pi$, corresponds to

$$\sin\theta = \frac{2\pi}{kb} = \frac{\lambda}{b} \tag{5.20}$$

Thus, for a given wavelength, the angular width of the diffraction pattern varies inversely with the slit width, and the amplitude of the central maximum is proportional to the area of the slit. For very narrow slits the pattern is dim but wide. It shrinks and becomes brighter as the slit is widened.

The Rectangular Aperture The case of diffraction by a single aperture of rectangular shape is treated in the same way as the single slit, except that one must now integrate in two dimensions, say x and y as shown in Figure 5.10. It is left as a problem to show that the ir-

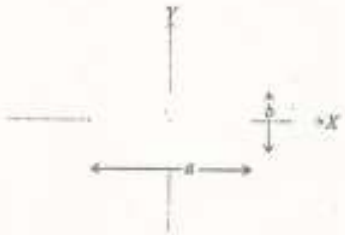


Figure 5.10. Rectangular aperture.

radiance distribution is given by the product of two single-slit distribution functions. (See Section 5.6.) The result is

$$I = I_0\left(\frac{\sin\alpha}{\alpha}\right)^2\left(\frac{\sin\beta}{\beta}\right)^2 \tag{5.21}$$

where $\alpha = \frac{1}{2}ka\sin\phi$, $\beta = \frac{1}{2}kb\sin\theta$. The dimensions of the aperture are a and b and the angles ϕ and θ define the direction of the diffracted ray. The resulting diffraction pattern (Figure 5.11) has lines of zero irradiance defined by $\alpha = \pm\pi, \pm 2\pi, \ldots$ and $\beta = \pm\pi, \pm 2\pi \ldots$. As with the slit, the scale of the diffraction pattern bears an inverse relationship to the scale of the aperture.

The Circular Aperture To calculate the diffraction pattern of a circular aperture, we choose y as the variable of integration, as in the case of the single slit. If R is the radius of the aperture, then the element of area is taken to be a strip of width dy and length $2\sqrt{R^2 - y^2}$ (Figure 5.12).

The amplitude distribution of the diffraction pattern is then given by

$$U = Ce^{ikr_0}\int_{-R}^{+R} e^{iky\sin\theta}\,2\sqrt{R^2 - y^2}\,dy \tag{5.22}$$

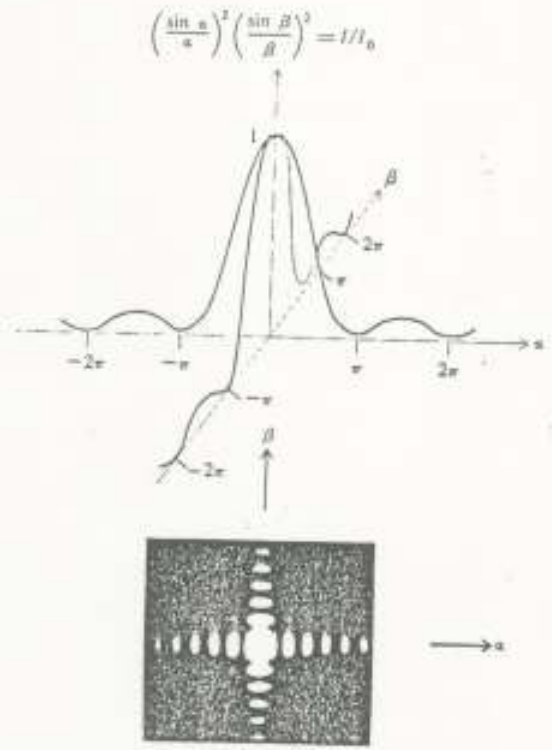


Figure 5.11. Fraunhofer diffraction pattern of a rectangular aperture.

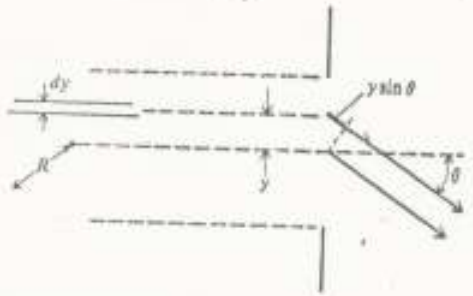


Figure 5.12. Circular aperture.

We introduce the quantities u and ρ defined by $u = y/R$ and $\rho = kR \sin\theta$. The integral in Equation 5.22 then becomes

$$\int_{-1}^{+1} e^{i\rho u}\sqrt{1-u^2}\,du \tag{5.23}$$

This is a standard integral. Its value is $\pi J_1(\rho)/\rho$ where J_1 is the Bessel function of the first kind, order one [27]. The ratio $J_1(\rho)/\rho \to \frac{1}{2}$ as $\rho \to 0$. The irradiance distribution is therefore given by

$$I = I_0\left[\frac{2J_1(\rho)}{\rho}\right]^2 \tag{5.24}$$

where $I_0 = (C\pi R^2)^2$, which is the intensity for $\theta = 0$.

A graph of the intensity function is shown in Figure 5.13. The diffraction pattern is circularly symmetric and consists of a bright cen-

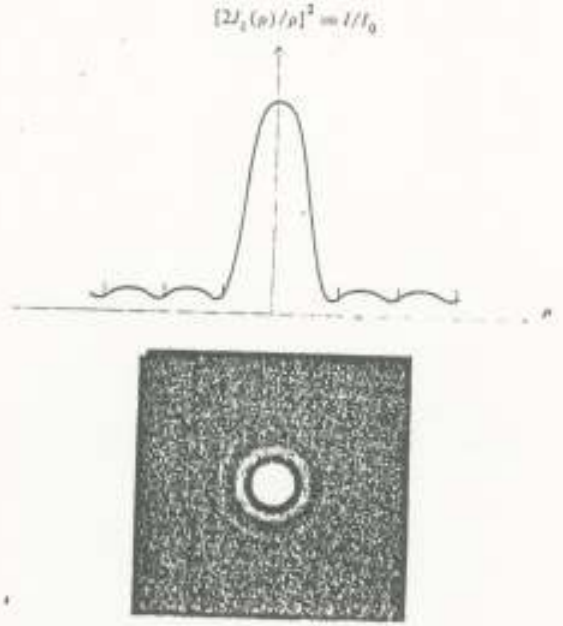


Figure 5.13. Fraunhofer diffraction pattern of a circular aperture.

tral disk surrounded by concentric circular bands of rapidly diminishing intensity. The bright central area is known as the *Airy disk*. It extends to the first dark ring whose size is given by the first zero of the Bessel function, namely, $\rho = 3.832$. The angular radius of the first dark ring is thus given by

$$\sin\theta = \frac{3.832}{kR} = \frac{1.22\lambda}{D} \approx \theta \tag{5.25}$$

which is valid for small values of θ. Here $D = 2R$ is the diameter of the aperture.

The angular size of the Airy disk is thus slightly larger than the corresponding value λ/b for the bright central band of the diffraction pattern of the rectangular aperture or slit. In Table 5.1 are listed the values of the first few maxima of the diffraction patterns of rectangular and circular apertures.

Optical Resolution The image of a distant point source formed at the focal plane of an optical-telescope lens or a camera lens is actually a Fraunhofer diffraction pattern for which the aperture is the lens opening. Thus the image of a composite source is a superposition of many Airy disks. The resolution of detail in the image therefore depends on the size of the individual Airy disks. If D is the diameter of the lens opening, then the angular radius of an Airy disk is approximately $1.22\,\lambda/D$. This is also the approximate minimum angular separation between two equal point sources such that they can be just barely resolved, because at this angular separation the central maximum of the image of one source falls on the first minimum of the other (Figure 5.14). This condition for optical resolution is known as

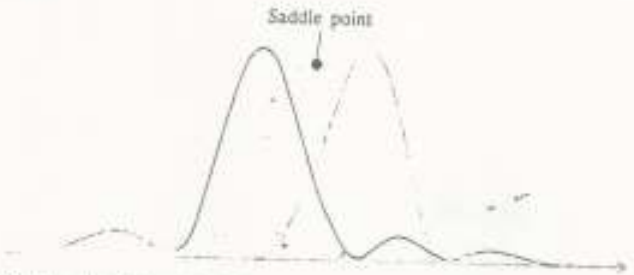


Figure 5.14. Rayleigh criterion.

the *Rayleigh criterion*. It is more convenient to use, in this case, than the Taylor criterion mentioned earlier.

In the case of the rectangular aperture, the minimum angular separation according to the Rayleigh criterion is just λ/b, where b is the width of the aperture. The intensity at the saddle point in this case is $8/\pi^2 = 0.81$ times the maximum intensity. The proof of this statement is left as an exercise.

The Double Slit Let us consider a diffracting aperture consisting of two parallel slits, each of width b and separated by a distance h (Figure 5.15). As with the single slit, we treat this case as a one-

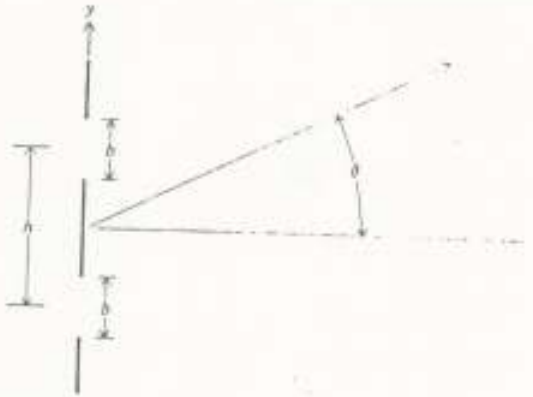


Figure 5.15. Double-slit aperture.

dimensional problem. The relevant diffraction integral is evaluated as follows:

$$\begin{aligned}\int_{\mathcal{A}} e^{iky\sin\theta}\,dy &= \int_0^b e^{iky\sin\theta}\,dy + \int_h^{h+b} e^{iky\sin\theta}\,dy \\ &= \frac{1}{ik\sin\theta}\left(e^{ikb\sin\theta} - 1 + e^{ik(h+b)\sin\theta} - e^{ikh\sin\theta}\right) \\ &= \left(\frac{e^{ikb\sin\theta}-1}{ik\sin\theta}\right)\left(1 + e^{ikh\sin\theta}\right) \qquad (5.26)\\ &= 2b\,e^{i\beta}\,e^{i\gamma}\,\frac{\sin\beta}{\beta}\cos\gamma\end{aligned}$$

where $\beta = \frac{1}{2}kb\sin\theta$ and $\gamma = \frac{1}{2}kh\sin\theta$. The corresponding irradiance distribution function is

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2\cos^2\gamma \qquad (5.27)$$

The factor $(\sin\beta/\beta)^2$ is the previously found distribution function for a single slit. Here this factor constitutes an envelope for the interference fringes given by the term $\cos^2\gamma$. A plot is shown in Figure 5.16. Bright fringes occur for $\gamma = 0, \pm\pi, \pm 2\pi$, and so forth. The angular separation between fringes is given by $\Delta\gamma = \pi$, or, approximately, in terms of the angle θ

$$\Delta\theta \approx \frac{2\pi}{kh} = \frac{\lambda}{h} \qquad (5.28)$$

It is interesting to note that this is equivalent to the result of the analysis of Young's experiment [Equation (3.9)].

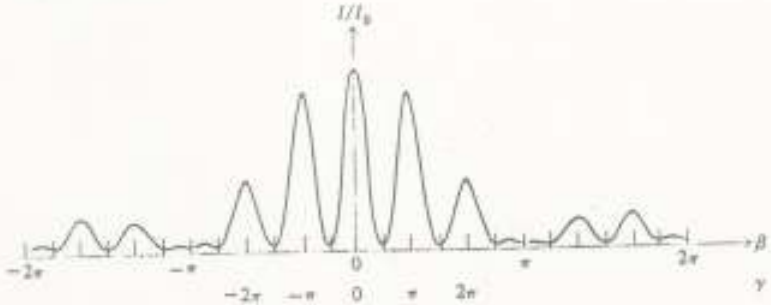


Figure 5.16. Fraunhofer diffraction pattern of a double-slit aperture.

Multiple Slits. Diffraction Gratings Let the aperture consist of a grating, that is, a large number N of identical parallel slits of width b and separation h (Figure 5.17). The evaluation of the diffractional integral is carried out in a manner similar to that of the double slit:

$$\begin{aligned}\int_{\mathscr{A}} e^{ik y \sin\theta}\,dy &= \int_0^b + \int_h^{h+b} + \int_{2h}^{2h+b} + \cdots + \int_{(N-1)h}^{(N-1)h+b} e^{ik y\sin\theta}\,dy \\ &= \frac{e^{ikb\sin\theta}-1}{ik\sin\theta}\left[1 + e^{ikh\sin\theta} + \cdots + e^{ik(N-1)h\sin\theta}\right] \\ &= \frac{e^{ikb\sin\theta}-1}{ik\sin\theta}\cdot\frac{1-e^{ikNh\sin\theta}}{1-e^{ikh\sin\theta}} \\ &= be^{i\beta}e^{i(N-1)\gamma}\left(\frac{\sin\beta}{\beta}\right)\left(\frac{\sin N\gamma}{\sin\gamma}\right)\end{aligned} \tag{5.29}$$

Figure 5.17. Multiple-slit aperture or diffraction grating.

where $\beta = \frac{1}{2}kb\sin\theta$ and $\gamma = \frac{1}{2}kh\sin\theta$. This yields the following intensity distribution function:

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2\left(\frac{\sin N\gamma}{N\sin\gamma}\right)^2 \tag{5.30}$$

The factor N has been inserted in order to normalize the expression. This makes $I = I_0$ when $\theta = 0$.

Again the single-slit factor $(\sin\beta/\beta)^2$ appears as the envelope of the diffraction pattern. Principal maxima occur within the envelope at $\gamma = n\pi$, $n = 0, 1, 2, \ldots$, that is,

$$n\lambda = h\sin\theta \tag{5.31}$$

which is the grating formula giving the relation between wavelength and angle of diffraction. The integer n is called the *order of diffraction*.

Secondary maxima occur near $\gamma = 3\pi/2N$, $5\pi/2N$, and so forth, and zeros occur at $\gamma = \pi/N$, $2\pi/N$, $3\pi/N$, ... A graph is shown in Figure 5.18(a). If the slits are very narrow, then the factor $\sin\beta/\beta \approx 1$. The first few primary maxima, then, all have approximately the same value, namely, I_0.

Resolving Power of a Grating The angular width of a principal fringe, that is, the separation between the peak and the adjacent minimum, is found by setting the *change* of the quantity $N\gamma$ equal to π, that is, $\Delta\gamma = \pi/N = \frac{1}{2}kh\cos\theta\,\Delta\theta$, or

$$\Delta\theta = \frac{\gamma}{Nh\cos\theta} \tag{5.32}$$

Thus if N is made very large, then $\Delta\theta$ is very small, and the diffraction pattern consists of a series of sharp fringes corresponding to the different orders $n = 0, \pm1, \pm2$, and so forth [Figure 5.18(b), (c)]. On the other hand for *a given order* the dependence of θ on the wavelength [Equation (5.31)] gives by differentiation

$$\Delta\theta = \frac{n\,\Delta\lambda}{h\cos\theta} \tag{5.33}$$

This is the angular separation between two spectral lines differing in wavelength by $\Delta\lambda$. Combining Equation (5.32) and (5.33), we obtain the *resolving power* of a grating spectroscope according to the Rayleigh criterion, namely,

$$\mathrm{RP} = \frac{\lambda}{\Delta\lambda} = Nn \tag{5.34}$$

In words, the resolving power is equal to the number of grooves N multiplied by the order number n.

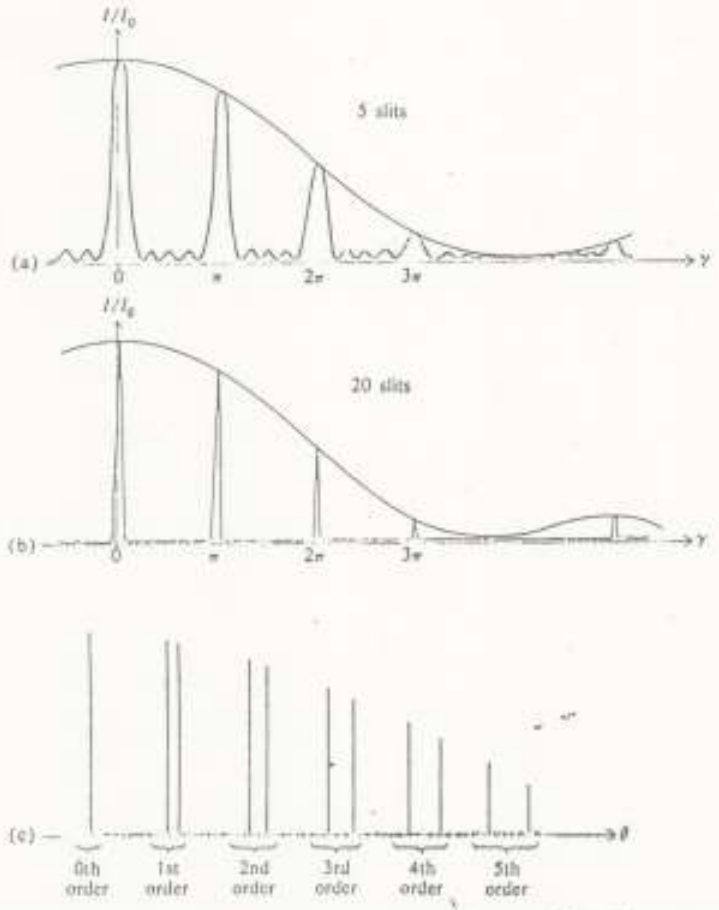


Figure 5.18. Fraunhofer diffraction pattern of a multiple-slit aperture. Graphs (a) and (b) are for monochromatic light. Graph (c) shows the pattern for a many-line grating illuminated with two different wavelengths.

Diffraction gratings used for optical spectroscopy are made by ruling grooves on a transparent surface (transmission type) or on a metal surface (reflection type). A typical grating may have, say, 600 lines/mm ruled over a total width of 10 cm. This would give a total of 60,000 lines and a theoretical resolving power of 60,000 n, where n is the order of diffraction used. In practice, resolving powers up to 90 percent of the theoretical values are obtainable with good gratings.

If the grooves are suitably shaped, usually of a sawtooth profile, most of the diffracted light can be made to appear in one order, thus increasing the efficiency of the grating. The essential requirement is that the spacing be uniform, within a fraction of a wavelength. This places extreme requirements on the mechanical rigidity of the ruling machine. High-quality replica gratings can be produced by a plastic molding process. These are much less expensive than original gratings.

Most of the gratings used in practical spectroscopy are of the reflection type. Reflection gratings are made with the ruled surface either plane or concave (Figure 5.19). Plane gratings require the use

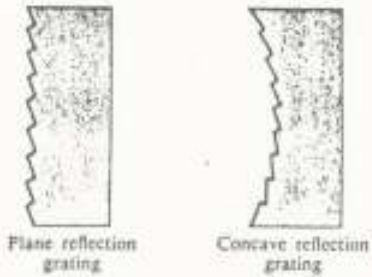


Figure 5.19. Reflection gratings.

of collimating and focusing lenses or mirrors, whereas concave gratings can perform the collimating and focusing functions as well as disperse the light into a spectrum. For more information on the subject of diffraction gratings and their use, the reader should consult References [17] and [35].

5.5 Fresnel Diffraction Patterns

According to the criteria discussed in Section 5.3, diffraction is of the Fresnel type when either the light source or the observing screen, or both, are so close to the diffracting aperture that the curvature of the wave front becomes significant. Since one is no longer dealing with plane waves, Fresnel diffraction is mathematically more difficult to treat than Fraunhofer diffraction but is actually simpler to observe experimentally because all that is needed is a source of light, an observing screen, and the diffracting aperture. The previously mentioned fringe effects seen around shadows are examples of Fresnel diffraction. In this section we shall discuss only a few relatively simple cases of Fresnel diffraction, which can be handled by elementary mathematical methods.

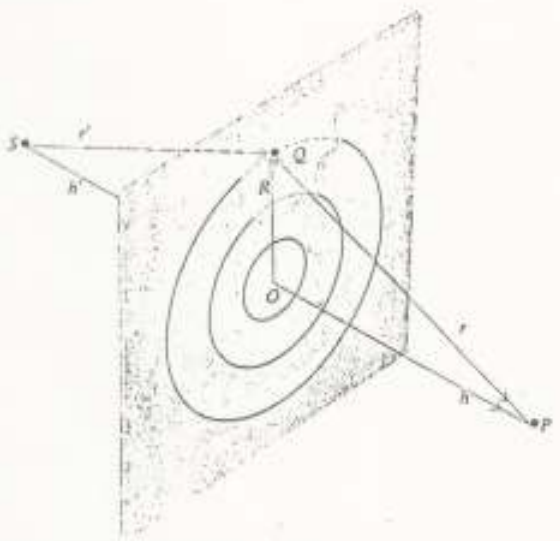


Figure 5.20. Fresnel zones in a plane aperture.

Fresnel Zones Consider a plane aperture illuminated by a point source S (Figure 5.20) such that a straight line connecting S to the receiving point P to perpendicular to the plane of the aperture. Let O be the point of intersection of the line SP with the aperture plane, and call R the distance from O to any point Q in the aperture. Then the distance $PQS = r + r'$ can be expressed in terms of R as follows:

$$\begin{aligned} r + r' &= (h^2 + R^2)^{1/2} + (h'^2 + R^2)^{1/2} \\ &= h + h' + \frac{1}{2} R^2 \left(\frac{1}{h} + \frac{1}{h'}\right) + \cdots \end{aligned} \tag{5.35}$$

where h and h' are the distances OP and OS, respectively. Now suppose that the aperture is divided up into regions bounded by concentric circles, $R = constant$, defined such that $r + r'$ differs by $\frac{1}{2}$ wavelength from one boundary to the next. These regions are called *Fresnel zones*. From (5.35) the successive radii are $R_1 = \sqrt{\lambda L}$, $R_2 = \sqrt{2\lambda L}$, . . . $R_n = \sqrt{n\lambda L}$, where λ is the wavelength, and

$$L = \left(\frac{1}{h} + \frac{1}{h'}\right)^{-1} \tag{5.36}$$

If R_n and R_{n+1} are the inner and outer radii of the nth zone, then the area is $\pi R_{n+1}^2 - \pi R_n^2 = \pi R_1^2$. This is independent of n. The areas of the complete zones are therefore all equal.

Typically, the radii of the low-order Fresnel zones are very small.

For example, if $h = h' = 50$ cm and $\lambda = 600$ nm, then we find $R_1 = (\lambda L)^{1/2} = 0.4$ mm, approximately. Also, since R_n is proportional to $n^{1/2}$, we see that the radius of the hundredth zone is only about 4 mm.

The optical disturbance at P can be evaluated in terms of the contributions from the various Fresnel zones, U_1, U_2, U_3, ... Since the mean phase changes by exactly 180 degrees from one zone to the next, the sum of the contributions to the amplitude $|U_a|$ can be expressed as

$$|U_a| = |U_1| - |U_2| + |U_3| - \cdots \tag{5.37}$$

Consider, for example, the case of a circular aperture centered at O. If the aperture includes precisely n complete zones, then, since the areas are equal, the $|U|$s are all approximately the same. Hence the sum will be very nearly zero if n is even, and approximately the value of $|U_1|$ alone if n is odd.

Consideration of the obliquity factor and the radial distance factor in the Fresnel-Kirchhoff formula [Equation (5.11)] shows that the value of $|U_n|$ decreases slowly with increasing n. As a result, as $n \to \infty$ the total optical disturbance at P for the case of an infinitely large aperture, that is, *no aperture at all*, is approximately one half the contribution from the first Fresnel zone alone. To show this (at least qualitatively) we group the terms in Equation (5.37) in the following way:

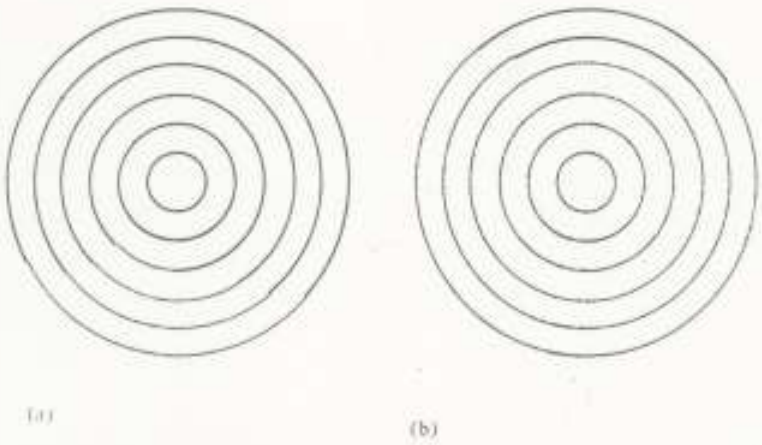
$$\begin{aligned} |U_a| = \tfrac{1}{2}|U_1| + (\tfrac{1}{2}|U_1| - |U_2| + \tfrac{1}{2}|U_3|) + \\ (\tfrac{1}{2}|U_3| - |U_4| + \tfrac{1}{2}|U_5|) + \cdots \end{aligned} \tag{5.38}$$

If the decrease with increasing n is very slow, the value of any $|U_n|$ is approximately equal to the mean value of the two adjacent $|U|$s, so that the terms in parentheses very nearly cancel. Thus $\frac{1}{2}|U_1|$ is the optical disturbance at P when there is no aperture at all.

Suppose we have a circular obstacle instead of an aperture. The construction of the Fresnel zones is now started at the edge of the obstacle. The value of $|U_a|$ is then, as above, just half the contribution from the first unobstructed zone. As a result the center of the shadow of a circular opaque object shows a bright spot.[2] The irradiance at the bright spot is very nearly the same as it would be in the absence of the obstacle.

In the case of an irregular obstacle or aperture, the appearance of the Fresnel zones as seen from the receiving point P is shown in Figure 5.21. In the illuminated region (a) the outer zones are partially blocked. Thus the higher terms in Equation (5.37) diminish more rap-

[2] The existence of such a bright spot was first predicted by Poisson in 1818 and later confirmed experimentally by Arago and Fresnel.



(a) (b)

Figure 5.21. Fresnel zones of a point source behind an irregular obstacle. (a) Outside geometrical shadow; (b) inside geometrical shadow.

idly than if there were no obstacle, but the beginning terms are unaffected. As a result the value of $|U_p|$ is hardly changed. On the other hand, in (b) the central zones are completely blocked and the outer zones are partially obstructed. Accordingly, the terms in the summation diminish at both ends and the result is almost complete cancellation. Thus if P is in the illuminated region, the presence of the obstacle makes little or no difference; but if it is in the shadow region, the optical disturbance is very small, which is roughly in agreement with geometrical optics. Diffraction fringes appear around the shadow only if the irregularities at the edge of the obstacle are small compared to the radius of the first Fresnel zone.

Zone Plate If an aperture is constructed so as to obstruct alternate Fresnel zones, say the even-numbered ones, then the remaining terms in the summation are all of the same sign. Thus

$$|U_p| = |U_1| + |U_3| + |U_5| + \cdots \tag{5.39}$$

Such an aperture is called a *zone plate*. It acts very much like a lens, because $|U_p|$, and hence the irradiance at P, is now much larger than if there were no aperture. The equivalent focal length is L in (5.36). It is given by

$$L = \frac{R_1^2}{\lambda} \tag{5.40}$$

Zone plates can be made by photographing a drawing similar to that of Figure 5.22. The resulting photographic transparency can

Figure 5.22. A zone plate.

focus light and form images of distant objects. It is a very chromatic lens, however, since the focal length is inversely proportional to the wavelength.

Rectangular Aperture Fresnel diffraction by an aperture of rectangular shape is treated by using the Fresnel-Kirchhoff formula [Equation (5.11)]. We shall employ Cartesian coordinates x,y in the aperture plane as shown in Figure 5.23. Then $R^2 = x^2 + y^2$, and therefore,

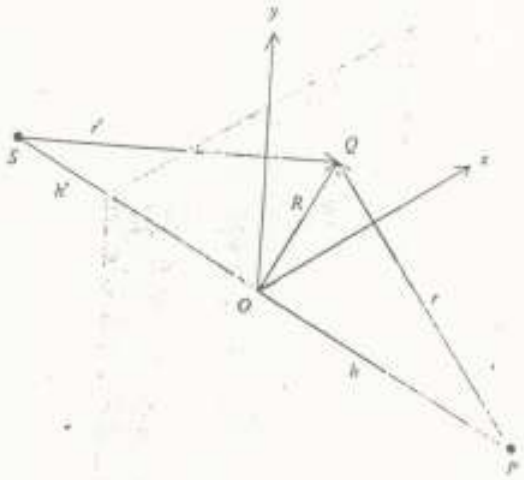


Figure 5.23. Geometry of a rectangular aperture.

referring to Equations (5.35) and (5.36) we have approximately

$$r + r' = h + h' + \frac{1}{2L}(x^2 + y^2) \tag{5.41}$$

Again, as in the treatment of Fraunhofer diffraction, we shall assume that the obliquity factor $\cos(\mathbf{n},\mathbf{r}) - \cos(\mathbf{n},\mathbf{r}')$ and the radial factor $1/rr'$ vary so slowly compared to the exponential factor $e^{ik(r+r')}$ that they can be taken outside the integrand. The Fresnel-Kirchhoff formula then becomes

$$\begin{aligned} U_p &= C \int_{x_1}^{x_2} \int_{y_1}^{y_2} e^{ik(x^2+y^2)/2L}\, dx\, dy \\ &= C \int_{x_1}^{x_2} e^{ikx^2/2L}\, dx \int_{y_1}^{y_2} e^{iky^2/2L}\, dy \end{aligned} \tag{5.42}$$

where C includes all other factors. Upon introducing the dimensionless variables u and v defined as

$$u = x\sqrt{\frac{k}{\pi L}} \qquad v = y\sqrt{\frac{k}{\pi L}}$$

or equivalently

$$u = x\sqrt{\frac{2}{\lambda L}} \qquad v = y\sqrt{\frac{2}{\lambda L}} \tag{5.43}$$

where L is defined by Equation (5.36) and λ is the wavelength, we can write

$$U_p = U_1 \int_{u_1}^{u_2} e^{i\pi u^2/2}\, du \int_{v_1}^{v_2} e^{i\pi v^2/2}\, dv \tag{5.44}$$

where $U_1 = C\pi L/k$.

The integrals in Equation (5.44) are evaluated in terms of the integral

$$\int_0^s e^{i\pi w^2/2}\, dw = C(s) + iS(s) \tag{5.45}$$

in which the real and the imaginary parts are given by

$$\begin{aligned} C(s) &= \int_0^s \cos(\pi w^2/2)\, dw \\ S(s) &= \int_0^s \sin(\pi w^2/2)\, dw \end{aligned} \tag{5.46}$$

These are known as *Fresnel integrals*. A short table of numerical values is presented in Table 5.2, and a graph showing $C(s)$ versus $S(s)$, called the *Cornu spiral*, is shown in Figure 5.24.

Table 5.2. FRESNEL INTEGRALS

s	$C(s)$	$S(s)$
0.0	0.000	0.000
0.2	0.200	0.004
0.4	0.398	0.033
0.6	0.581	0.111
0.8	0.723	0.249
1.0	0.780	0.438
1.2	0.715	0.623
1.4	0.543	0.714
1.6	0.366	0.638
1.8	0.334	0.451
2.0	0.488	0.343
2.5	0.457	0.619
3.0	0.606	0.496
3.5	0.533	0.415
4.0	0.498	0.420
∞	0.500	0.500

The Cornu spiral is useful for graphical evaluation of the Fresnel integrals. The limit points s_1 and s_2 are marked on the spiral. A straight line segment drawn from s_1 to s_2 [Figure 5.24(b)] then gives the value of the integral $\int_{s_1}^{s_2} e^{i\pi w^2/2}\, dw$. The length of the line segment is the magnitude of the integral, and the projections on the C and S axes are the real and imaginary parts, respectively. Also, from Equations (5.46) we see that $(dC)^2 + (dS)^2 = (ds)^2$, hence ds represents an element of arc. The total arc length on the Cornu spiral is equal to the difference between the two limits, namely, $s_2 - s_1$. This difference is proportional to the size of the aperture, that is,

$$s_2 - s_1 = u_2 - u_1 = (x_2 - x_1)\sqrt{\frac{2}{\lambda L}}$$

for the x dimension, and

$$s_2 - s_1 = v_2 - v_1 = (y_2 - y_1)\sqrt{\frac{2}{\lambda L}}$$

for the y dimension.*

The limiting case of an infinite aperture, that is, *no diffraction screen at all*, is obtained by setting $u_1 = v_1 = -\infty$ and $u_2 = v_2 = +\infty$. Since $C(\infty) = S(\infty) = \frac{1}{2}$, and $C(-\infty) = S(-\infty) = -\frac{1}{2}$, we obtain the value $U_1(1 + i)^2$ for the unobstructed optical disturbance. On the Cornu

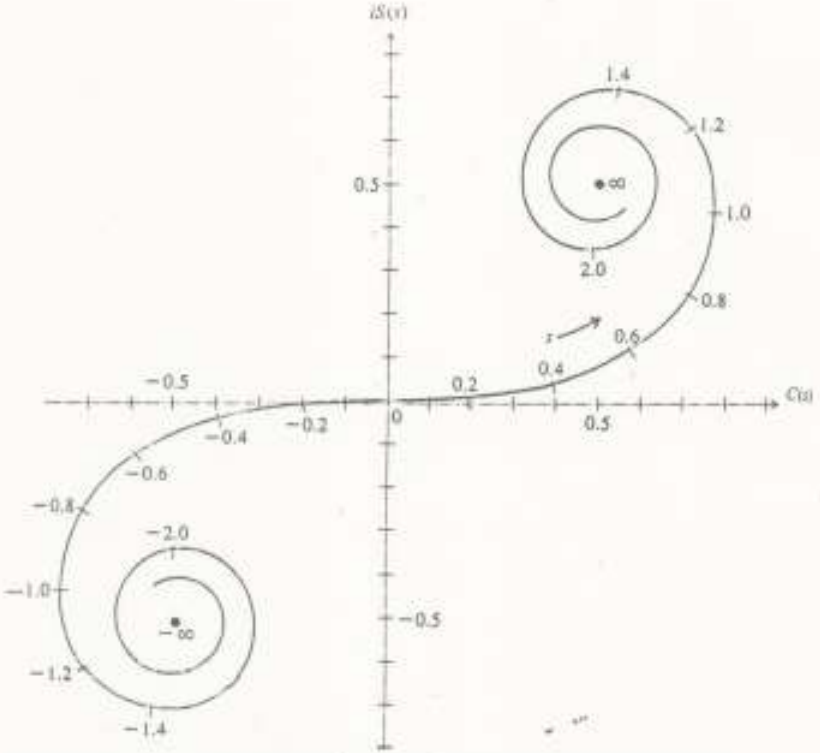


Figure 5.24. (a) The Cornu spiral. The scale of w is marked on the curve.

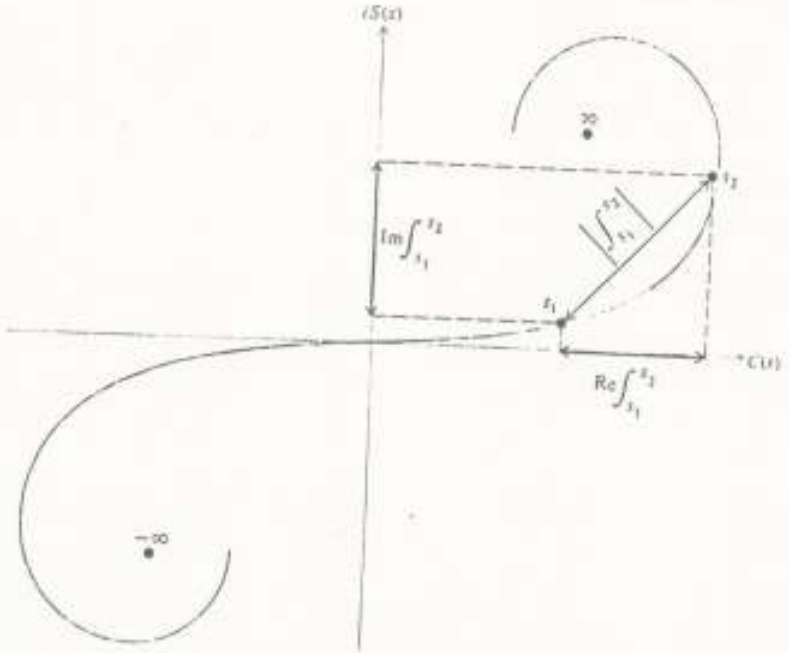


Figure 5.24. (b) Evaluation of Fresnel integrals with the Cornu spiral.

spiral this would be U_1 times the length of the line from $-\infty$ to ∞ [Figure 5.24(b)]. Setting this equal to U_0, we can express the general case in the normalized form

$$U_p = \frac{U_0}{(1+i)^2}\,[C(u)+iS(u)]_{u_1}^{u_2}[C(v)+iS(v)]_{v_1}^{v_2} \tag{5.47}$$

Strictly speaking, very large values of the parameters u, v, or s would be inconsistent with the approximation expressed by Equation (5.41). However, in normal cases of interest most of the contribution to U_p comes from the lower-order Fresnel zones in the aperture, corresponding to low values of the above parameters, hence the approximation is still valid.

Slit and Straightedge Fresnel diffraction by a long slit is treated as a limiting case of a rectangular aperture, namely, by letting $u_1 = -\infty$ and $u_2 = +\infty$ in Equation (5.47). This yields the formula

$$U_p = \frac{U_0}{1+i}\,[C(v)+iS(v)]_{v_1}^{v_2} \tag{5.48}$$

for the slit where v_1 and v_2 define the slit edges.

The straightedge is similarly taken as a limiting case of a slit: $v_1 = -\infty$. This gives

$$\begin{aligned} U_p &= \frac{U_0}{1+i}\,[C(v)+iS(v)]_{-\infty}^{v_2} \\ &= \frac{U_0}{1+i}\left[C(v_2)+iS(v_2)+\frac{1}{2}+\frac{1}{2}\,i\right] \end{aligned} \tag{5.49}$$

which is a function of only the one variable v_2. This variable specifies the position of the diffracting edge. If the receiving point P is exactly at the geometrical shadow edge, then $v_2 = 0$. We have then $U_p = [U_0/(1+i)]\,(\frac{1}{2}+\frac{1}{2}i) = \frac{1}{2}U_0$. Hence the amplitude at the shadow edge is one half, and the irradiance is one fourth the unobstructed value. A plot of $I_p = |U_p|^2$ as given by Equation (5.49) is shown in Figure 5.25. Here I_p is plotted as a function of v_2. This is equivalent to having a fixed position for the receiving point and varying the posi-

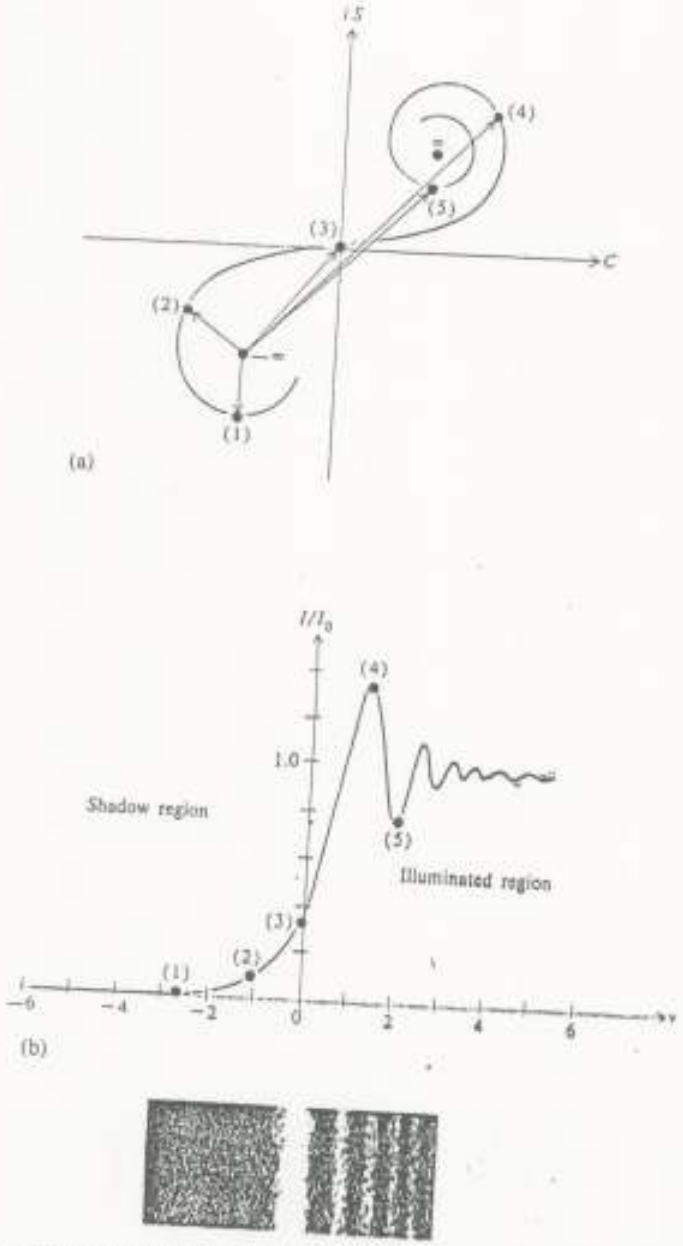


Figure 5.25. Fresnel diffraction by a straightedge. (a) Points on the Cornu spiral; (b) corresponding points on the intensity curve; $v = 0$ defines the geometrical shadow edge. A photograph of the diffraction pattern is shown below.

tion of the diffracting edge. The result is virtually the same as a diffraction pattern. From the graph it can be seen that the irradiance falls off rapidly and monotonically in the shadow zone ($v_z < 0$) as $v_z \to -\infty$. On the other hand, in the illuminated zone ($v_z > 0$) the irradiance oscillates with diminishing amplitude about the unobstructed value U_0 as $v_z \to +\infty$. The highest irradiance occurs just inside the illuminated region at the point $v_z = 1.25$, where I_p is 1.37 times the irradiance of the unobstructed wave. This is seen as a bright fringe next to the geometrical shadow.

5.6 Applications of the Fourier Transform to Diffraction

Let us return to the discussion of Fraunhofer diffraction. We now consider the general problem of diffraction by an aperture having not only an arbitrary shape, but also an arbitrary transmission including phase retardation, which may vary over different parts of the aperture.

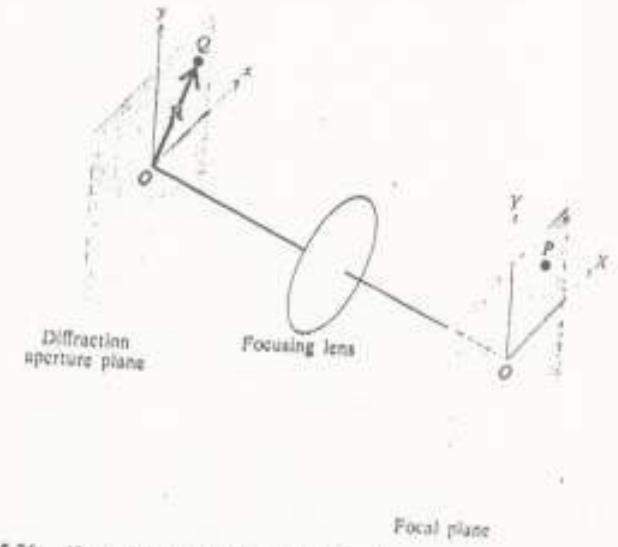


Figure 5.26. Geometry of the general diffraction problem.

We choose coordinates as indicated in Figure 5.26. The diffracting aperture lies in the xy plane, and the diffraction pattern appears in the XY plane, which is the focal plane of the focusing lens. According to elementary geometrical optics, all rays leaving the diffracting aperture in a given direction, specified by direction cosines α, β, and γ, are

brought to a common focus. This focus is located at the point $P(X,Y)$ where $X = L\alpha$ and $Y = L\beta$, L being the focal length of the lens. The assumption is made here that α and β are small, so that $\alpha = \tan\alpha$ and $\beta = \tan\beta$. We also assume that $\gamma = 1$.

Now the path difference δr, between a ray starting from the point $Q(x,y)$ and a parallel ray starting from the origin O, is given by $\mathbf{R} \cdot \hat{\mathbf{n}}$, (Figure 5.27), where $\mathbf{R} = \mathbf{i}x + \mathbf{j}y$ and $\hat{\mathbf{n}}$ is a unit vector in the direction

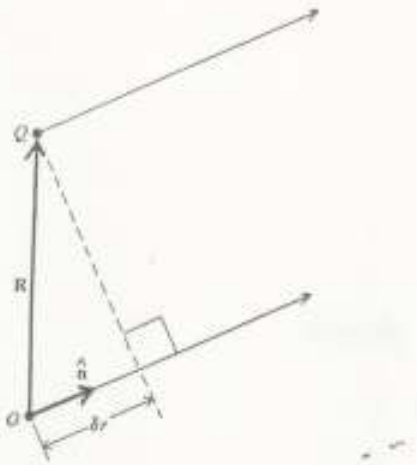


Figure 5.27. Path difference between two parallel rays of light originating from points O and Q in the xy plane.

of the ray. Since $\hat{\mathbf{n}}$ can be expressed as $\hat{\mathbf{n}} = \mathbf{i}\alpha + \mathbf{j}\beta + \mathbf{k}\gamma$, then

$$\delta r = \mathbf{R} \cdot \hat{\mathbf{n}} = x\alpha + y\beta = x\frac{X}{L} + y\frac{Y}{L} \tag{5.50}$$

It follows that the fundamental diffraction integral [Equation (5.16)] giving the diffraction pattern in the XY plane is, aside from a constant multiplying factor, expressible in the form

$$U(X,Y) = \iint e^{ikr}\,d\mathscr{A} = \iint e^{ik(xX+yY)/L}\,dx\,dy \tag{5.51}$$

This is the case for a uniform aperture.

For a uniform rectangular aperture the double integral reduces to the product of two one-dimensional integrals. The result is stated earlier in Section 5.4.

For a nonuniform aperture we introduce a function $g(x,y)$ called the *aperture function*. This function is defined such that $g(x,y)\,dx\,dy$

is the amplitude of the diffracted wave originating from the element of area $dx\,dy$. Thus instead of Equation (5.51), we have the more general formula

$$U(X,Y) = \iint g(x,y)\,e^{ik(xX+yY)}\,dx\,dy \tag{5.52}$$

It is convenient at this point to introduce the quantities

$$\mu = \frac{kX}{L} \quad \text{and} \quad \nu = \frac{kY}{L} \tag{5.53}$$

μ and ν are called *spatial frequencies*, although they have the dimensions of reciprocal length, that is, wavenumber. We now write Equation (5.52) as

$$U(\mu,\nu) = \iint g(x,y)\,e^{i(\mu x+\nu y)}\,dx\,dy \tag{5.54}$$

We see that the functions $U(\mu,\nu)$ and $g(x,y)$ constitute a two-dimensional Fourier transform pair. The diffraction pattern, in this context, is actually a Fourier resolution of the aperture function.

Consider as an example a grating. For simplicity we treat it as a one-dimensional problem. The aperture function $g(y)$ is then a periodic step function as shown in Figure 5.28. It is represented by a Fourier series of the form

$$g(y) = g_0 + g_1\cos(\nu_0 y) + g_2\cos(2\nu_0 y) + \cdots \tag{5.55}$$

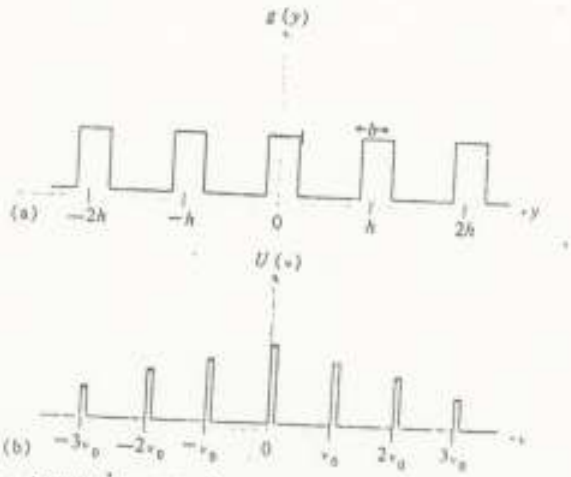


Figure 5.28. Aperture function for a grating and its Fourier transform.

The fundamental spatial frequency ν_0 is given by the periodicity of the grating, namely,

$$\nu_0 = \frac{2\pi}{h} \tag{5.56}$$

where h is the grating spacing. This dominant spatial frequency appears in the diffraction pattern as the first-order maximum, the amplitude of which is proportional to g_1. Maxima of higher order correspond to higher Fourier components of the aperture function $g(y)$. Thus if the aperture function were of the form of a cosine function $g_0 + g_1 \cos(\nu_0 y)$ instead of a periodic step function, then the diffraction pattern would consist only of the central maximum and the two first-order maxima. Second or higher diffraction orders would not appear.

Apodization Apodization (literally "to remove the feet") is the name given to any process by which the aperture function is altered in such a way as to produce a redistribution of energy in the diffraction pattern. Apodization is usually employed to reduce the intensity of the secondary diffraction maxima.

It is perhaps easiest to explain the theory of apodization by means of a specific example. Let the aperture consist of a single slit. The aperture function in this case is a single step function: $g(y) = 1$ for $-b/2 < y < b/2$ and $g(y) = 0$ otherwise (Figure 5.29). The corresponding diffraction pattern, expressed in terms of spatial frequencies, is

$$U(\nu) = \int_{-b/2}^{+b/2} e^{i\nu y}\, dy = b\,\frac{\sin(\frac{1}{2}\nu b)}{(\frac{1}{2}\nu b)} \tag{5.57}$$

This is equivalent to the normal case already discussed in Section 5.5.

Suppose now that the aperture function is altered by apodizing in such a way that the resultant aperture transmission is a cosine function: $g(y) = \cos(\pi y/b)$ for $-b/2 < y < b/2$ and zero otherwise, as shown in Figure 5.29. This could be accomplished, for example, by means of a suitably coated-glass plate placed over the aperture. The new diffraction pattern is given by

$$U(\nu) = \int_{-b/2}^{+b/2} \cos\left(\frac{\pi y}{b}\right) e^{i\nu y}\, dy$$

$$= \cos(\nu b/2)\left(\frac{1}{\nu - \pi/b} - \frac{1}{\nu + \pi/b}\right) \tag{5.58}$$

A comparison of the two diffraction patterns is shown graphically in the figure. The result of apodization in this case is a substantial reduction in the secondary maxima relative to the central maximum; in other words, apodization has suppressed the higher spatial frequencies.

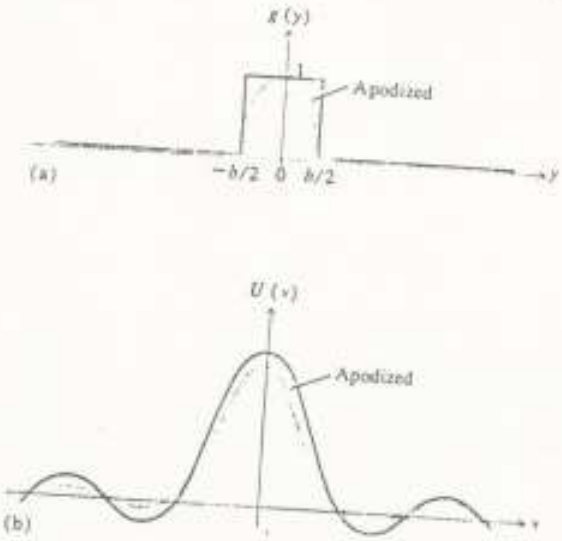


Figure 5.29. (a) Aperture functions for a slit and an apodized slit; (b) the Fourier transforms.

In a similar way it is possible to apodize the circular aperture of a telescope so as to reduce greatly the relative intensities of the diffraction rings that appear around the images of stars (discussed in Section 5.5). This enhances the ability of the telescope to resolve the image of a dim star near that of a bright one.

Spatial Filtering Consider the diagram shown in Figure 5.30. Here the xy plane represents the location of some *coherently* illuminated object.[9] This object is imaged by an optical system (not shown), the image appearing in the $x'y'$ plane. The diffraction pattern $U(\mu,\nu)$ of the object function $g(x,y)$ appears in the $\mu\nu$ plane. This plane is analogous to the XY plane in Figure 5.26. Hence, from Equation (5.54) $U(\mu,\nu)$ is the Fourier transform of $g(x,y)$. The image function $g'(x',y')$ that appears in the $x'y'$ plane is, in turn, the Fourier transform of $U(\mu,\nu)$. Now if *all* spatial frequencies in the range $\mu = \pm\infty$, $\nu = \pm\infty$ were transmitted equally by the optical system, then, from the properties of the Fourier transform, the image function $g'(x'y')$

[9] For a discussion of the theory of spatial filtering with incoherent illumination see Reference [10].

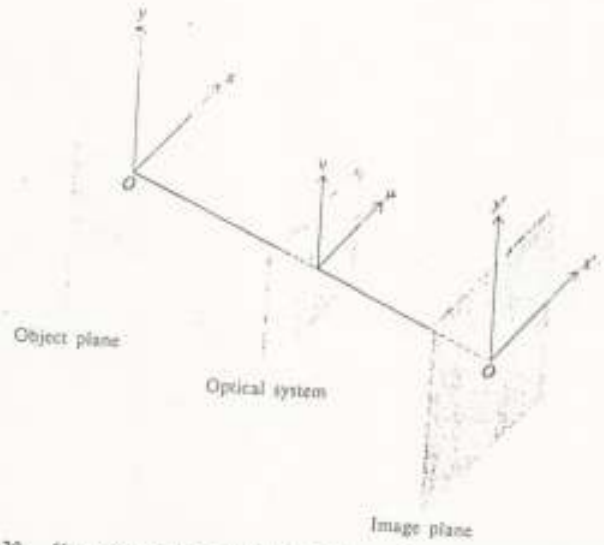


Figure 5.30. Geometry for the general problem of image formation by an optical system.

would be *exactly* proportional to the object function $g(x,y)$; that is, the image would be a true reproduction of the object. However, the finite size of the aperture at the $\mu\nu$ plane limits the spatial frequencies that are transmitted by the optical system. Furthermore there may be lens defects, aberrations, and so forth, which result in a modification of the function $U(\mu,\nu)$. All of these effects can be incorporated into one function $T(\mu,\nu)$ called the *transfer function* of the optical system. This function is defined implicitly by the equation

$$U'(\mu,\nu) = T(\mu,\nu)\, U(\mu,\nu)$$

Thus

$$g'(x', y') = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} T(\mu,\nu)U(\mu,\nu)\, e^{-i(\mu x'+\nu y')}\, d\mu\, d\nu \qquad (5.59)$$

that is, the image function is the Fourier transform of the product $T(\mu,\nu)\cdot U(\mu,\nu)$. The limits of integration are $\pm\infty$ in a formal sense only. The actual limits are given by the particular form of the transfer function $T(\mu,\nu)$.

The transfer function can be modified by placing various screens and apertures in the $\mu\nu$ plane. This is known as *spatial filtering*. The situation is quite analogous to the filtering of an electrical signal by means of a passive electrical network. The object function is the input signal, and the image function is the output signal. The optical system

acts like a filter that allows certain spatial frequencies to be transmitted but rejects others.

Suppose, for example, that the object is a grating so that the object function is a periodic step function. This case can be treated as a one-dimensional problem. The object function $f(y)$ and its Fourier transform $U(\nu)$ are then just those shown in Figure 5.28. Now let the aperture in the $\mu\nu$ plane be such that only those spatial frequencies that lie between $-\nu_{max}$ and $+\nu_{max}$ are transmitted. This means that we have low-pass filtering. From Equation (5.53) we have $\nu_{max} = kb/f$, where $2b$ is the physical width of the aperture in the $\mu\nu$ plane. The transfer function for this case is a step function: $T(\nu) = 1$, $-\nu_{max} < \nu < +\nu_{max}$, and zero otherwise. The image function is, accordingly,

$$g'(y') = \int_{-\nu_{max}}^{+\nu_{max}} U(\nu)e^{-i\nu y'}\, d\nu \qquad (5.60)$$

Without going into the details of the calculation of $g'(y')$, we show in Figure 5.31(a) a graphical plot for some arbitrary choice of ν_{max}. Instead of the sharp step function that constitutes the object, the image is rounded at the corners and also shows small periodic variations.

A high-pass optical filter is obtained by placing in the $\mu\nu$ plane a screen that blocks off the central part of the diffraction pattern. This part of the diffraction pattern corresponds to the low frequencies. The approximate form of the resulting image function is shown in Figure 5.31(b). Only the edges of the grating steps are now visible in the image plane. *The edge detail comes from the higher spatial frequencies.*

A practical example of spatial filtering is the *pinhole spatial filter* which is used in laser work to reduce the spurious fringe pattern that always occurs in the output beam of a helium–neon laser. The beam is brought to a sharp focus by means of a short-focal-length lens. A fine pinhole placed at the focal point constitutes the filter, which removes the higher spatial frequencies and hence improves the beam quality of the laser output. A second lens can be used to render the beam parallel.

Phase Contrast and Phase Gratings The method of phase contrast was invented by the Dutch physicist Zernike. It is used to render visible a transparent object whose index of refraction differs slightly from that of a surrounding transparent medium. Phase contrast is particularly useful in microscopy for examination of living organisms, and so forth. In essence, the method consists of the use of a special type of spatial filter.

To simplify the theory of phase contrast, we shall treat the case of a so-called "phase grating" consisting of alternate strips of high- and

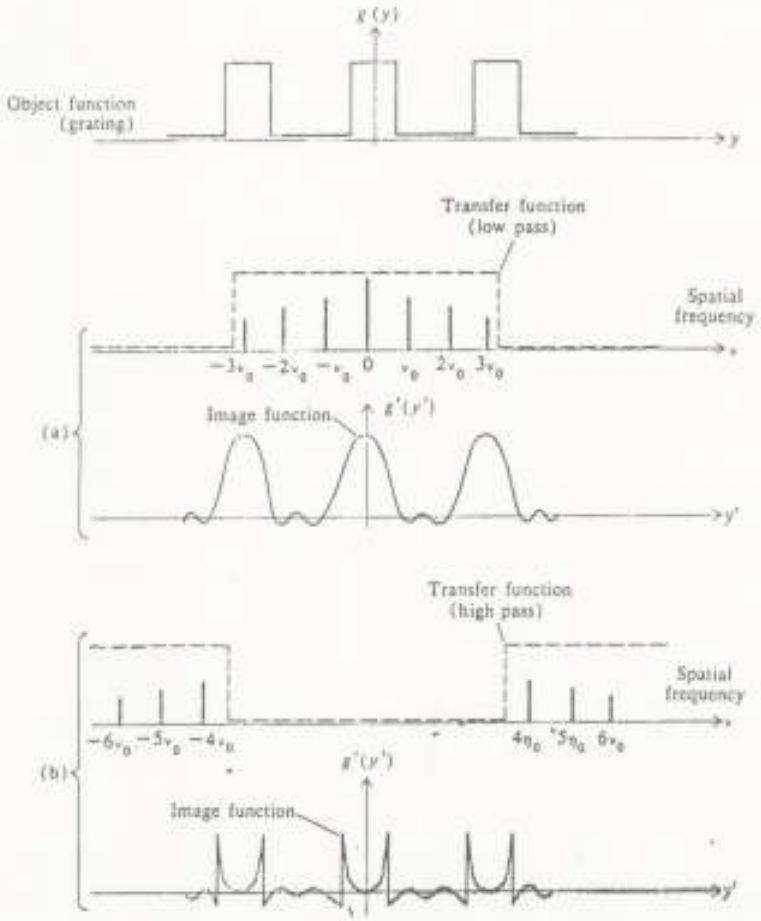


Figure 5.31. Graphs illustrating spatial filtering. (a) Low-pass filtering; (b) high-pass filtering.

low-index material, all strips being perfectly transparent. The grating is coherently illuminated and constitutes the object. The object function is thus represented by the exponential

$$g(y) = e^{i\phi(y)} \tag{5.61}$$

where the phase factor $\phi(y)$ is a periodic step function as shown in Figure 5.32(a). The "height" of the step is the optical-phase difference between the two kinds of strips; that is, $\Delta\phi = kz\,\Delta n$, where z is the thickness and Δn is the difference between the two indices of

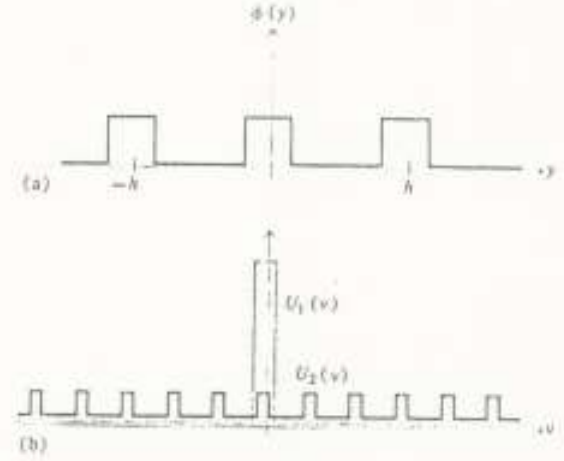


Figure 5.32. (a) The phase function of a periodic phase grating; (b) Fourier transforms of the aperture U_1 and the grating U_2.

refraction. If we assume that this phase difference is very small, then to a good approximation, we can write

$$g(y) = 1 + i\phi(y) \tag{5.62}$$

The Fourier transform of the above function is

$$U(\nu) = \int_{-\infty}^{\infty} [1 + i\phi(y)]e^{i\nu y}\,dy = \int_{-b/2}^{+b/2} e^{i\nu y}\,dy + i\int_{-b/2}^{+b/2} \phi(y)\,e^{i\nu y}\,dy$$
$$= U_1(\nu) + iU_2(\nu) \tag{5.63}$$

Here $U_1(\nu)$ represents the diffraction pattern of the whole-object aperture. It is essentially zero everywhere except for $\nu = 0$; that is, $U_1(\nu)$ contains only very low spatial frequencies. On the other hand, $U_2(\nu)$ represents the diffraction pattern of the periodic step function $\phi(y)$. The two functions are plotted in Figure 5.32(b).

By virtue of the factor i in the result, $U_1 + iU_2$, the two components U_1 and iU_2 are 90 degrees out of phase. The essential trick in the phase-contrast method consists of inserting a spatial filter in the $\mu\nu$ plane, which has the property of shifting the phase of iU_2 by an additional 90 degrees. In practice this is accomplished by means of a device known as a *phase plate*. The physical arrangement is shown in Figure 5.33. The phase plate is just a transparent-glass plate having a small section whose optical thickness is $\frac{1}{4}$ wavelength greater than the remainder of the plate. This thicker section is located in the central part of the $\mu\nu$ plane, that is, in the region of low spatial frequencies. The result of inserting the phase plate is to change the function

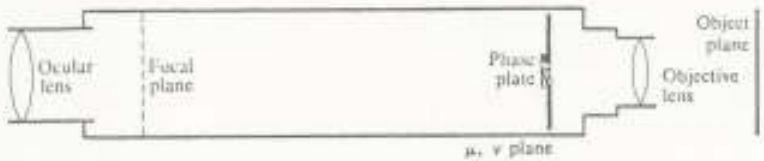


Figure 5.33. Physical arrangement of the optical elements for phase contrast microscopy.

$U_1 + iU_2$ to $U_1 + U_2$. The new image function is given by the Fourier transform of the new $U(\nu)$, namely,

$$g'(y') = \int U_1(\nu)e^{-i\nu y'}\,d\nu + \int U_2(\nu)e^{-i\nu y'}\,d\nu \qquad (5.64)$$
$$= g_1(y') + g_2(y')$$

Now the first function g_1 is just the image function of the whole-object aperture. It represents the constant background. The second function g_2 is the image function for a regular grating of alternate transparent and opaque strips. This means that the phase grating has been rendered visible. It appears in the image plane as alternate bright and dark strips. Although the above analysis has been for a periodic grating, a similar argument can be applied to a transparent-phase object of any shape.

The method of optical-phase contrast has a close analogy in electrical communications. A phase-modulated signal is converted into an amplitude-modulated signal by introduction of a phase shift of 90 degrees to the carrier frequency. This is essentially what the phase plate does in the phase-contrast method. The net result is that phase modulation in the object is converted into amplitude modulation in the image.

5.7 Reconstruction of the Wave Front by Diffraction. Holography

An unusual and interesting method of producing an image—known as the method of *wave-front reconstruction*—has recently become of importance in the field of optics. Although the basic idea was originally proposed by Gabor in 1947 [12], it attracted little attention until the highly coherent light of the laser became available.

In this method a special diffraction screen, called a *hologram*, is used to reconstruct in detail the wave field emitted by the subject. To make the hologram the output from a laser is separated into two beams, one of which illuminates the subject. The other beam, called the *reference beam*, is reflected onto a fine-grained photographic film

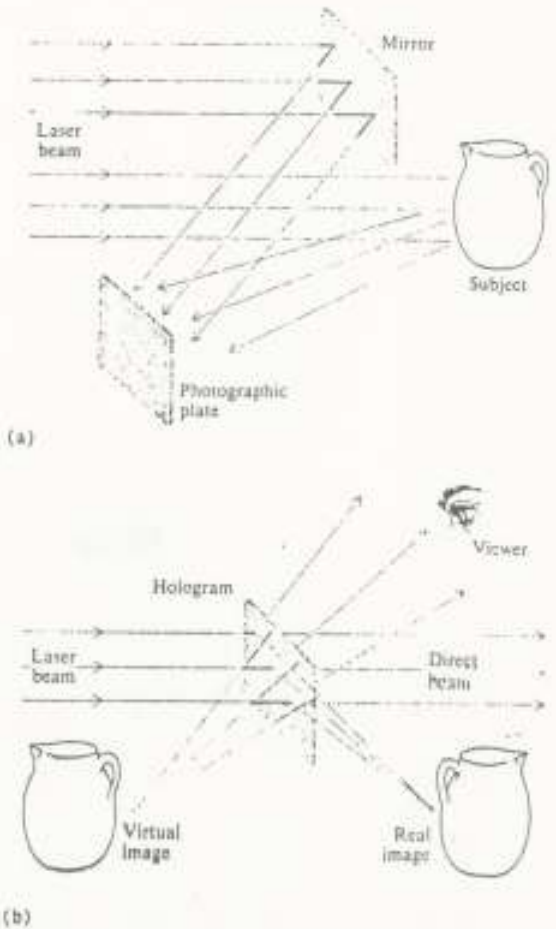


Figure 5.34. (a) Arrangement for producing a hologram; (b) use of the hologram in producing the real and virtual images.

by means of a mirror. The film is exposed simultaneously to the reference beam and the reflected laser light from the subject [Figure 5.34(a)]. The resulting complicated interference pattern recorded by the film constitutes the hologram. It contains all the information needed to reproduce the wave field of the subject.

In use the developed hologram is illuminated with a single beam from a laser as shown in Figure 5.34(b). Part of the resulting diffracted wave field is a precise, three-dimensional copy of the original wave reflected by the subject. The viewer looking at the hologram

sees the image in depth and by moving his head can change his perspective of the view.

In order to simplify the discussion of the theory of holography, we shall assume that the reference beam is collimated; that is, it consists of plane waves, although this is not actually necessary in practice. Let x and y be the coordinates in the plane of the recording photographic plate, and let $U(x,y)$ denote the complex amplitude of the reflected wave front in the xy plane. Since $U(x,y)$ is a complex number, we can write it as

$$U(x,y) = a(x,y)e^{i\phi(x,y)} \tag{5.65}$$

where $a(x,y)$ is real.

Similarly, let $U_0(x,y)$ denote the complex amplitude of the reference beam. Since this beam is plane, we can write

$$U_0(x,y) = a_0\, e^{i(\mu x + \nu y)} \tag{5.66}$$

where a_0 is a constant and μ and ν are the spatial frequencies of the reference beam in the xy plane. They are given by

$$\mu = k \sin \alpha \qquad \nu = k \sin \beta \tag{5.67}$$

in which k is the wave number of the laser light, and α and β specify the direction of the reference beam.

The irradiance $I(x,y)$ that is recorded by the photographic film is thus given by the expression

$$\begin{aligned} I(x,y) &= |U + U_0|^2 = a^2 + a_0^2 + aa_0 e^{i[\phi(x,y) - \mu x - \nu y]} + aa_0 e^{-i[\phi(x,y) - \mu x - \nu y]} \\ &= a^2 + a_0^2 + 2aa_0 \cos\,[\phi(x,y) - \mu x - \nu y] \end{aligned} \tag{5.68}$$

This is actually an interference pattern. It contains information in the form of amplitude and phase modulations of the spatial frequencies of the reference beam. The situation is somewhat analogous to the impression of information on the carrier wave of a radio transmitter by means of amplitude or phase modulation.

When the developed hologram is illuminated with a single beam U_3 similar to the reference beam, the resulting transmitted wave U_T is proportional to U_3 times the transmittance of the hologram at the point (x,y). The transmittance will be proportional to $I(x,y)$. Hence, except for a constant proportionality factor that we ignore,

$$\begin{aligned} U_T(x,y) &= U_3 I = a_3(a^2 + a_0^2)e^{i(\mu x + \nu y)} + a_3 a\, e^{i\phi} + a_3 a_0 e^{-i(\phi - 2\mu x - 2\nu y)} \\ &= (a^2 + a_0^2)\,U_3 + a_0^2 U + a^2 U^{-1} U_0^2 \end{aligned} \tag{5.69}$$

The hologram acts somewhat like a diffraction grating. It produces a direct beam and two first-order diffracted beams on either side of the direct beam [Figure 5.34(b)]. The term $(a^2 + a_0^2)U_3$ in Equation (5.69) comprises the direct beam. The term $a_0^2 U$ represents one of the diffracted beams. Since it is equal to a constant times U, this beam is the one that reproduces the reflected light from the subject and forms the virtual image. The last term represents the other diffracted beam and gives rise to a real image.

We shall not attempt to prove the above statements in detail. They can be verified by considering a very simple case, namely, that in which the subject is a single white line on a dark background. In this case the hologram turns out to be, in fact, a simple periodic grating. The zero order of the diffracted light is the direct beam, whereas the two first orders on either side comprise the virtual and the real images.

In holography the viewer always sees a positive image whether a positive or a negative photographic transparency is used for the hologram. The reason for this is that a negative hologram merely produces a wave field that is shifted 180 degrees in phase with respect to that of a positive hologram. Since the eye is insensitive to this phase difference, the view seen by the observer is identical in the two cases.

Remarkable technical advances have been made in the field of holography in recent years. Holography in full color is possible by using three different laser wavelengths instead of just one, the holographic record being on black-and-white film. The holographic principle has been extended to include the use of acoustic waves for imaging in optically opaque media and to microwaves for long-distance holography.

Holographic Interferometry One of the most notable applications of holography is in the field of interferometry. In this application the surface to be tested can be irregular and diffusely reflecting instead of smooth and highly polished as is required for the ordinary Michelson and Twyman-Green type of interferometric work. In *double-exposure* holographic interferometry two separate exposures are made on a single recording film. If the surface under study undergoes any deformation or movement during the time interval between exposures, such movement is revealed on the reconstructed image in the form of interference fringes. In *double-pulse* holography the two exposures are produced by short, intense laser pulses from a high-power pulsed laser. These pulses are closely spaced in time so that the holographic image fringes can show motion, vibration patterns, and so on. The method is especially useful for nondestructive testing. For more information on the subject of holography, the reader is encouraged to consult a text such as *An Introduction to Coherent Optics and Holography* by G. W. Stroke [38].