Chapter 4
Diffraction

Diffraction, like interference, is a wave phenomenon. From a mathematical point of view, the difference between interference and diffraction lies in the number of sources that generate the interference waves. In interference there is a discrete number of sources, whereas in diffraction there is a continuous number of sources. In terms of the behavior of the optical field, diffraction is considered the deviation of the rectilinear path (of light) that is not due to reflection or refraction.

In this chapter, diffraction is limited to the paraxial range, i.e., Fresnel diffraction and Fraunhofer diffraction. Detailed examples of diffraction by a circular aperture and by a rectangular aperture are given. With diffraction through a circular aperture, the formation of the image is analyzed taking into account the wave nature of light; with diffraction through a rectangular aperture, the basic mathematics for one-dimensional diffraction gratings are developed.

The image of a point object (monochromatic) generated by an optical system that models a human eye with myopia, astigmatism, and spherical aberration is shown in Fig. 4.1. The effect of diffraction and aberrations reduces visual acuity in the human eye and generally reduces resolution in imaging systems.

Note on calculated diffraction patterns
Except for Section 4.5.2, which deals with image resolution, calculated diffraction patterns are shown in this chapter as grayscale images that represent the square root of the irradiance distribution. This allows regions of lower intensity to be highlighted. Plots of the irradiance profiles are shown at scale.

4.1 Huygens–Fresnel Principle

Huygens' principle, discussed in Section 1.1.3, states that *every point on a wavefront can be considered as a source of secondary spherical waves that propagate with the same speed as the wavefront. After a while, the propagated*

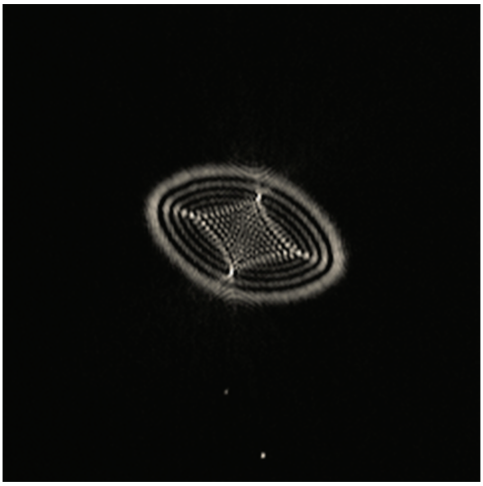


Figure 4.1 Experimental image of a monochromatic point source generated by an optical system with astigmatism, spherical aberration, and defocus.

wavefront will be the envelope of the secondary spherical waves [1]. With this principle, the (unobstructed) propagation of a wavefront can be derived, where Σ′ is obtained from Σ, as shown in Fig. 4.2, and the laws of reflection and refraction can be derived (Fig. 1.8). On the other hand, Fresnel establishes that the optical field at a point P is obtained from the interference of secondary waves [2]. In this way, Fresnel gives a satisfactory explanation of the phenomenon of diffraction. The combination of the Huygens principle and the interference of secondary Fresnel waves is called the *Huygens–Fresnel principle*.

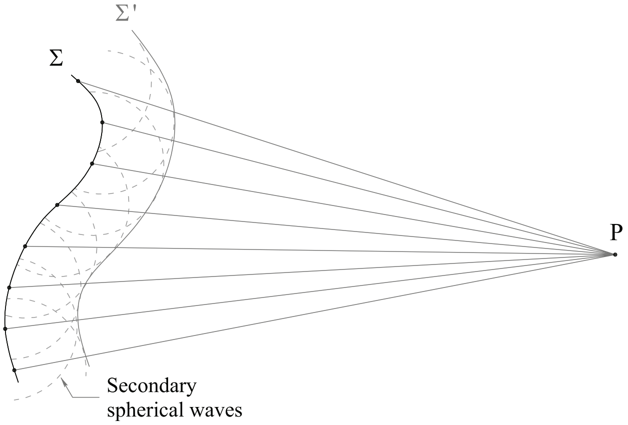


Figure 4.2 The Huygens–Fresnel principle states that the field at P is the superposition (interference) of the secondary spherical waves emitted by the virtual sources located in the wavefront Σ.

The mathematical formalization of the Huygens–Fresnel principle would be carried out a century later by Kirchhoff and later refined by Rayleigh and Sommerfeld [3]. Although the study of diffraction can be started from Kirchhoff's mathematical formalism, it is worth following Fresnel's ideas to gain a further conceptual understanding of diffraction. The treatment in this book follows that described by Born and Wolf [4].

Let us start with the simplest situation: the propagation in vacuum of a spherical wavefront emitted by a point source. There is a point source S depicted in Fig. 4.3. According to the optical field expression for a spherical wave, omitting the time phase term $e^{-i\omega t}$, the electrical field at point P would be

$$E(\mathrm{P}) = E_0^{\dagger}\frac{e^{ikd}}{d}, \tag{4.1}$$

where $d = \overline{\mathrm{SP}}$ and $E_0^{\dagger}$ is the amplitude of the field multiplied by the unit of length.

Using the Huygens–Fresnel principle to calculate the electric field at P, the result should be the same as in Eq. (4.1). Let Σ be the spherical wavefront of radius r_0 emitted by the point source S in a given time. From this wavefront, the sources that emit the secondary Fresnel waves will be located at the points that form Σ. In particular, at point Q of the wavefront Σ there will be a secondary emitter whose contribution to the field at P will be of the form $E(Q)e^{iks}/s$, where $E(Q) = E_0^{\dagger}e^{ikr_0}/r_0$. To obtain $E(P)$ from the sum of all the (infinite) secondary waves, another Fresnel hypothesis is included: the amplitude of the secondary waves varies with the direction defined by the angle χ, which is the angle between the normal of the wavefront Σ in Q and the line joining Q and P (Fig. 4.3). Therefore, the amplitude of the secondary waves will be of the form $E(Q)K(\chi)$, where $K(\chi)$ is the function that determines how the amplitude variation occurs. The angle χ is called the *inclination angle*,

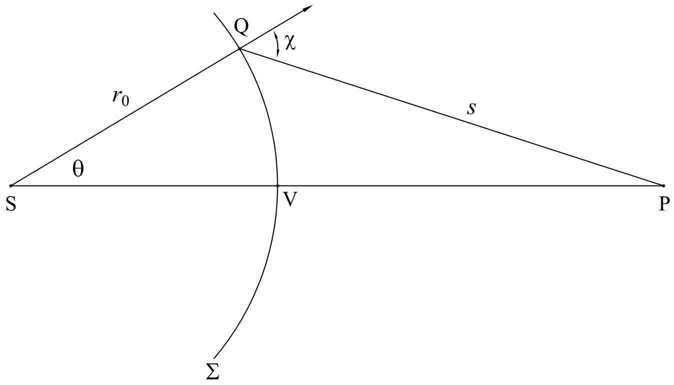


Figure 4.3 Propagation of a spherical wave. The contribution of the secondary source Q to the field at P depends on the angle χ.

and the function K is called the *inclination factor*. With the function K, the Fresnel hypothesis is described as follows: *given that the impulse communicated in any part of the primary wavefront Σ follows the normal of the wavefront, the effect on the medium*[*] *must be more intense in the direction of the normal, so the rays from Q to P will be less intense as they deviate from the normal* [2]. Fresnel mentions that determining the explicit form of the function K is a "very difficult matter"; it should not be an issue in many practical situations given that the rays from Q to P deviate little from the normal, so a constant value can remain for the function K. Following Fresnel, K is maximum when $\chi = 0$ and disappears when the line from Q to P is tangent to the wavefront Σ, i.e., $\chi = \pi/2$. This implies that not all of the spherical wavefront Σ contributes to the sum at P. The validity of these conditions is considered later with the analytical treatment developed by Kirchhoff. According to this, the field at P would be given by

$$E(\mathrm{P}) = \frac{E_0^{\dagger} e^{ikr_0}}{r_0} \iint_{\Sigma} \frac{e^{iks}}{s} K(\chi) d\sigma, \tag{4.2}$$

where $d\sigma$ describes the differential element of area in Q. This integral is the mathematical version of the Huygens–Fresnel principle.

4.1.1 Fresnel zones

The Huygens–Fresnel integral can be solved by dividing the domain into regions where the inclination factor approaches a constant value. This procedure proposed by Fresnel gives surprising results, which occur in practice, as shown in some later sections in this chapter. The regions into which the domain is divided are called *Fresnel zones* and for a spherical wavefront they are constructed as shown in Fig. 4.4. The spheres of radius $b + j\lambda/2$, with $j = 0,1, \ldots, N$, and $b = \overline{\mathrm{VP}}$, are drawn from the point P (thus, $d = r_0 + b$). The jth zone (Z_j) is the annular region of Σ contained between the spheres of radius $b + (j-1)\lambda/2$ and $b + j\lambda/2$.

If $b \gg \lambda$ and $r_0 \gg \lambda$, then the following approximations are made in the inclination factor:

- $K(\chi) \approx$ constant, for a given Fresnel zone, and changes very little between consecutive zones.
- $K_j(\chi) \approx [K_{j+1}(\chi) + K_{j-1}(\chi)]/2$.

Also,

- $K_N(\chi) = 0$, if $\chi = \pi/2$.

[*] In Fresnel's time, a hypothetical substance called Ether, or luminiferous Ether, was believed to occupy all of space and was supposed to act as a medium for the propagation of electromagnetic waves.

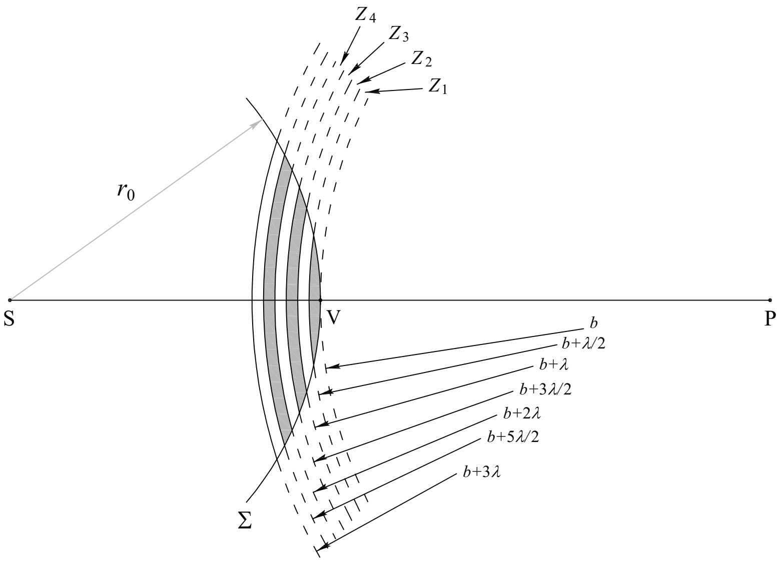


Figure 4.4 Fresnel zones in a spherical wavefront.

From the geometry of Fig. 4.3,

$$s^2 = r_0^2 + (r_0 + b)^2 - 2r_0(r_0 + b)\cos\theta,$$

where θ is the polar angle. By partial differentiation, then

$$2sds = 2r_0(r_0 + b)\sin\theta d\theta.$$

Because the differential element $d\sigma = r_0^2 \sin\theta d\theta d\phi$, where ϕ is the azimuth angle, substituting $\sin\theta d\theta$ leads to

$$d\sigma = \frac{r_0}{(r_0 + b)} sdsd\phi. \tag{4.3}$$

Taking into account the previous results, the diffraction integral [Eq. (4.2)] can be approximated as

$$E(\mathrm{P}) = \sum_{j=1}^{N} E_j(\mathrm{P}), \tag{4.4}$$

where

$$E_j(\mathrm{P}) = \frac{E_0^{\dagger} e^{ikr_0}}{r_0} 2\pi K_j \int\limits_{b+(j-1)\lambda/2}^{b+j\lambda/2} \frac{e^{iks}}{s} \frac{r_0}{(r_0 + b)} sds, \tag{4.5}$$

where the approximation $K_j(\chi) = K_j$ (constant for the jth zone) has been used. Evaluating the integral of Eq. (4.5), the optical field at P will be

$$E(\mathrm{P}) = \sum_{j=1}^{N} E_j(\mathrm{P}) = i2\lambda \frac{E_0^{\dagger} e^{ikd}}{d} \sum_{j=1}^{N} (-1)^{j+1} K_j. \tag{4.6}$$

Thus, the value of the integral depends on the sum of the inclination factors in each Fresnel zone. Taking into account the second approximation on the average value of the inclination factor of the adjacent zones for a given zone, i.e., $K_j(\chi) = [K_{j+1}(\chi) + K_{j-1}(\chi)]/2$, the sum

$$\sum_{j=1}^{N} (-1)^{j+1} K_j = K_1 - K_2 + K_3 - K_4 + \ldots + (-1)^{N+1} K_N \tag{4.7}$$

can be written as

$$\sum_{j=1}^{N} (-1)^{j+1} K_j = \frac{K_1}{2} + \left(\frac{K_1}{2} - K_2 + \frac{K_3}{2}\right) + \left(\frac{K_3}{2} - K_4 + \frac{K_5}{2}\right) + \ldots + \begin{cases} K_N/2 & ; N \rightarrow \text{odd} \\ K_{N-1}/2 - K_N & ; N \rightarrow \text{even.} \end{cases} \tag{4.8}$$

Because the average of the zones adjacent to zone Z_j is approximately equal to the value of zone Z_j, the sum reduces to

$$\sum_{j=1}^{N} (-1)^{j+1} K_j = \frac{K_1}{2} \pm \frac{K_N}{2} \rightarrow \begin{cases} +; N \rightarrow \text{odd} \\ -; N \rightarrow \text{even} \end{cases}. \tag{4.9}$$

Therefore,

$$E(\mathrm{P}) = i2\lambda \frac{E_0^{\dagger} e^{ikd}}{d} \left(\frac{K_1}{2} \pm \frac{K_N}{2}\right) \tag{4.10}$$

or

$$E(\mathrm{P}) = \frac{1}{2}\left[E_1(\mathrm{P}) \pm E_N(\mathrm{P})\right]. \tag{4.11}$$

When the wavefront Σ propagates unobstructed, the total number of zones N is obtained when $\chi = \pi/2$ and $E_N(P) = 0$. The field at P will be

$$E(\mathrm{P}) = \frac{1}{2} E_1(\mathrm{P}). \tag{4.12}$$

Taking into account the result of Eq. (4.1) for $E(P)$ and the field expression for the first zone $E_1(\mathrm{P}) = i2\lambda K_1 E_0^\dagger e^{ikd}/d$, Eq. (4.12) is satisfied if

$$K_1 = -\frac{i}{\lambda} = \frac{e^{-i\pi/2}}{\lambda}. \tag{4.13}$$

In this way, it is possible to find the explicit value of the inclination factor for the first zone.

4.1.2 Fresnel treatment results

Equation (4.12) can be interpreted as shown in Fig. 4.5. Figure 4.5(a) represents the free propagation of the wavefront Σ; in Fig. 4.5(b), the wavefront is obstructed by a circular aperture that allows the passage of the field delimited by the middle of the first Fresnel zone. In both cases, the field at P will be given by $E_0^\dagger e^{ikd}/d$ and the irradiance will be given by $I(\mathrm{P}) = I_0 = (\epsilon_0 c/2)(E_0^\dagger/d)^2$. From the point of view of geometrical optics, the result does not depend on the radius of the aperture since the energy that reaches P propagates along the ray that joins S with P. Therefore, the irradiance at P corresponds to the expected result. So, what is the gain of the Fresnel wave treatment?

In the Fresnel treatment, the field at P depends on the size of the aperture. Let us consider the following cases:

- **Aperture for** $N = 1$. If the radius of the aperture is such that it allows the passage of the field delimited by the first Fresnel zone, from Eq. (4.6),

$$E(\mathrm{P}) = i2\lambda \frac{E_0^\dagger e^{ikd}}{d} K_1.$$

Then the irradiance at P will be

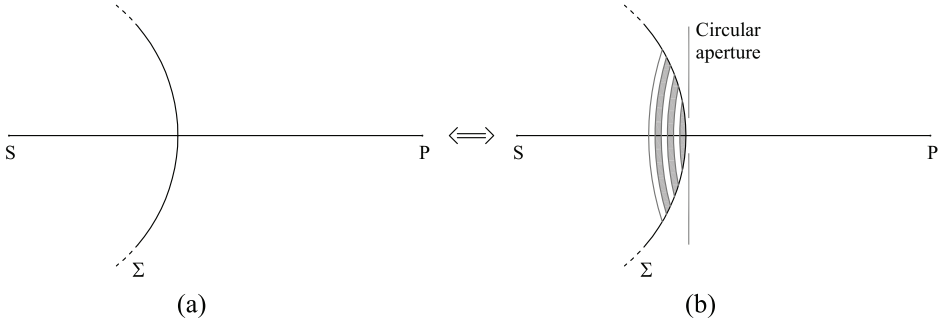


Figure 4.5 The field at P (a) due to the free propagation of a spherical wavefront is equal to (b) the field bounded by a circular aperture that allows only the field corresponding to half of the first Fresnel zone to pass.

$$I(P) = 4I_0.$$

This result is no longer predictable by geometrical optics. The increase in irradiance with increasing aperture radius seems reasonable. However, the Fresnel treatment also tells us that this is not always the case, since a further increase in the radius of the aperture decreases the irradiance, even to zero.

- **Aperture for** $N = 2$. If the radius of the aperture is increased, such that the aperture coincides with the outer edge of the second Fresnel zone, from Eq. (4.6),

$$E(\mathrm{P}) = i2\lambda\frac{E_0^\dagger e^{ikd}}{d}(K_1 - K_2).$$

Taking into account the Fresnel hypothesis, i.e., that the inclination factor changes very little between consecutive zones, then $K_1 \approx K_2$. Thus, the irradiance at P will be

$$I(\mathrm{P}) \approx 0.$$

This result is even more surprising, but it is explained by considering the interference. Generally speaking, we can say that the field of the second zone is out of phase by π with respect to the field of the first zone. This is because of the way Fresnel zones have been constructed: with spheres whose radii increase by $\lambda/2$.

Based on the previous results, it can be anticipated that when the aperture allows the passage of M Fresnel zones, where M is odd, the consecutive zones grouped in pairs cancel each other and the irradiance at P will be given only by the remaining zone ($I \sim 4I_0$). When the aperture allows M Fresnel zones to pass, where M is even, the consecutive zones grouped by pairs cancel each other and the irradiance at P will be zero ($I \sim 0$).

- **Opaque disk for** $N = 1$. Another interesting situation is if instead of an aperture in an opaque screen, like the one shown in Fig. 4.5(b), an opaque disk whose radius is equal to the edge of the first zone is placed, as in Fig. 4.6. Then, the passage of the field limited by the first Fresnel zone is blocked, and the passage of the field from the second Fresnel zone (up to the last zone where $\chi = \pi/2$) is allowed. The result for the field at P is

$$E(\mathrm{P}) = \frac{1}{2}E_1(\mathrm{P}) - E_1(\mathrm{P}) = -\frac{1}{2}E_1(\mathrm{P}).$$

Therefore, the irradiance at P would be given by

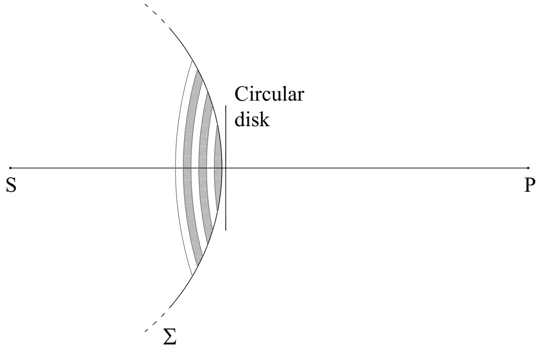


Figure 4.6 Poisson spot. Despite the circular disk that hinders light propagation within the first Fresnel zone, there is a bright spot at P behind the obstacle.

$$I(P) = I_0.$$

In other words, at P there will be a bright spot even though the ray from S to P is blocked. This point is called the *Poisson spot.*[*]

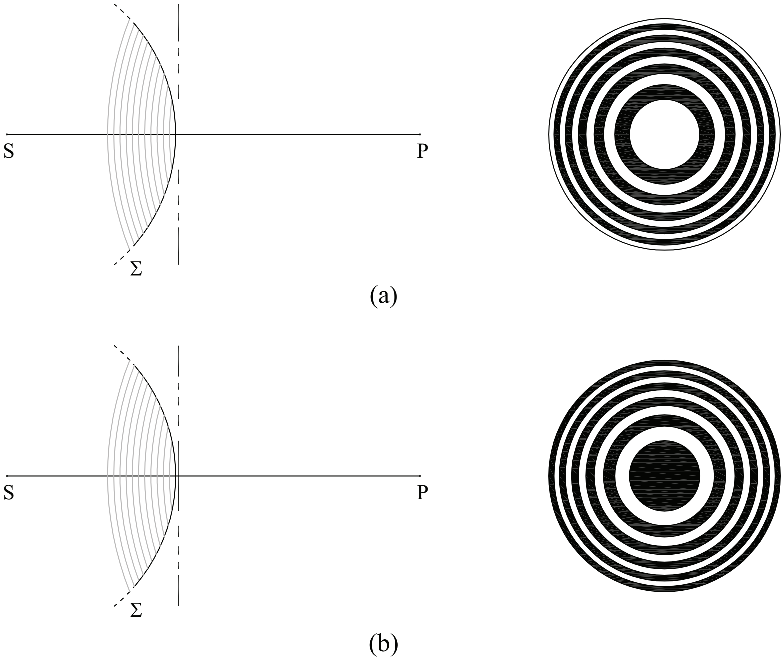
- **Fresnel zone plate**. Finally, let us consider the situation where only odd or even zones are blocked. As two consecutive zones have a phase shift of π and every two zones will be in phase (phase shift of 2π), an obstacle with apertures equivalent to the even or odd annular zones considerably increases the irradiance value at point P. This situation is illustrated in Fig. 4.7(a), with a zonal plate blocking even zones, and Fig. 4.7(b), with a zonal plate blocking odd zones. In both cases, if M zones are allowed to pass, the field at P is approximated by $E(P) \approx ME_1(P)$ and the irradiance would be

$$I(P) = 4M^2 I_0.$$

A further improvement to the zonal plate is achieved if instead of blocking the odd or even zones, an offset of π is introduced in the odd or even zones. This can be done by depositing on a glass substrate a thin film of transparent material whose optical thickness is equal to $\lambda/2$. The thin film is deposited only in the annular regions that correspond to the odd or even Fresnel zones. To do this, a mask is used that obstructs the deposit of the material (as in a lithographic process).

A Fresnel phase zone plate for the even zones is shown in Fig. 4.8. In this way, the irradiance at P increases even more. The increase in irradiance at P occurs at the expense of the decrease in irradiance at points neighboring P, which guarantees energy conservation.

[*] The Fresnel treatment predicts that there may be light behind an obstacle. This fact, pointed out by Poisson as erroneous, actually occurs.



(a)

(b)

Figure 4.7 Fresnel zonal plates to block (a) even and (b) odd zones.

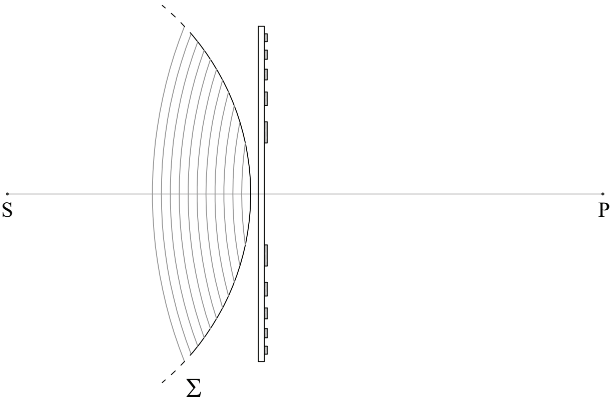


Figure 4.8 The phase zone plate takes advantage of the entire optical field Σ and constructively interferes with the fields contained in the odd and even Fresnel zones.

4.2 Diffraction Integral

Diffraction involves finding the optical field at any point in space generated by a source with boundary conditions. The typical geometry in diffraction is illustrated in Fig. 4.9. In region I, the source S (point-like or extended) is located; in region II, the volume is limited by the closed surface $\Sigma = \Sigma_1 + \Sigma_2$, in which the optical field is measured. Region II is called the *diffraction region*.

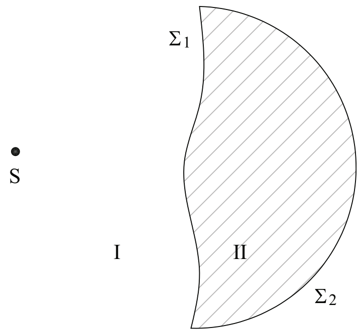


Figure 4.9 General geometry in the diffraction problem.

The surface Σ_1 that separates regions I and II is an opaque surface with apertures that allow the passage of part of the optical field emitted by the source S.

A first approach to the problem of finding the optical field in the diffraction region is to solve the homogeneous wave equation for the scalar optical field (ignoring polarization). Using a monochromatic wave,

$$E'(x, y, z, t) = E(x, y, z)e^{-i\omega t}, \tag{4.14}$$

from the wave equation in vacuum [Eq. (2.5)],

$$(\nabla^2 + k^2)E = 0, \tag{4.15}$$

with $k = \omega/c = 2\pi/\lambda$. To determine E at any point in the diffraction region, Green's theorem can be used, which in turn follows from Gauss' theorem. Gauss' theorem states that if $\mathbf{F}$ is a vector function of the position, then

$$\iint_\Sigma \mathbf{F} \cdot \hat{\mathbf{n}} d\sigma = \iiint_V \nabla \cdot \mathbf{F} dv, \tag{4.16}$$

where $\hat{\mathbf{n}}$ is the unit vector normal to the closed surface Σ (outwards), V is the volume enclosed by the surface, and $d\sigma$ and dv denote the differential elements of area and volume, respectively. If the function $\mathbf{F}$ can be obtained as

$$\mathbf{F} = E\nabla U, \tag{4.17}$$

where E and U are scalar functions defined on Σ and V, then

$$\iint_\Sigma (E\nabla U \cdot \hat{\mathbf{n}}) d\sigma = \iiint_V (E\nabla^2 U + \nabla E \cdot \nabla U) dv. \tag{4.18}$$

If E and U are exchanged, a similar relationship is obtained:

$$\iint_{\Sigma}(U\nabla E\cdot\hat{\mathbf{n}})d\sigma=\iiint_{V}(U\nabla^{2}E+\nabla U\cdot\nabla E)dv. \tag{4.19}$$

Subtracting Eq. (4.19) from Eq. (4.18) leads to

$$\iint_{\Sigma}(E\nabla U\cdot\hat{\mathbf{n}}-U\nabla E\cdot\hat{\mathbf{n}})d\sigma=\iiint_{V}(E\nabla^{2}U-U\nabla^{2}E)dv; \tag{4.20}$$

taking into account the directional derivatives $\partial E/\partial n=\nabla E\cdot\hat{\mathbf{n}}$ and $\partial U/\partial n=\nabla U\cdot\hat{\mathbf{n}}$, Green's theorem is obtained:

$$\iint_{\Sigma}\left(E\frac{\partial U}{\partial n}-U\frac{\partial E}{\partial n}\right)d\sigma=\iiint_{V}(E\nabla^{2}U-U\nabla^{2}E)dv. \tag{4.21}$$

If the function U satisfies the time-independent wave equation, [Eq. (4.15)], $(\nabla^{2}+k^{2})U=0$, then the right-hand side of Eq. (4.21) vanishes; therefore,

$$\iint_{\Sigma}\left(E\frac{\partial U}{\partial n}-U\frac{\partial E}{\partial n}\right)d\sigma=0. \tag{4.22}$$

With this integral, given the field E (and its derivative $\partial E/\partial n$) on the surface Σ, it is possible to calculate the field E at a point $\mathrm{P}(x', y', z')$ inside the surface Σ with the help of the function U (and its derivative $\partial U/\partial n$).

4.2.1 Kirchhoff integral theorem

Kirchhoff uses Green's theorem with the function

$$U(x, y, z)=\frac{e^{iks}}{s}, \tag{4.23}$$

where s is the distance between the point $\mathrm{P}(x', y', z')$ and the point (x, y, z) on the surface. This function generates a singularity in the diffraction region that must be removed (since U must be defined anywhere in V). The singularity can be eliminated by constructing a sphere Σ_{ε} of radius $\varepsilon\rightarrow 0$ centered at point P, as illustrated in Fig. 4.10.

With $\varepsilon\rightarrow 0$, the volume V enclosing Σ is maintained, but now the integration surface will be $\Sigma+\Sigma_{\varepsilon}$. Therefore, the diffraction integral over Σ becomes

$$\iint_{\Sigma}\left(E\frac{\partial U}{\partial n}-U\frac{\partial E}{\partial n}\right)d\sigma=-\iint_{\Sigma_{\varepsilon}}\left(E\frac{\partial U}{\partial n}-U\frac{\partial E}{\partial n}\right)d\sigma. \tag{4.24}$$

The directional derivative $\partial U/\partial n$ is equal to

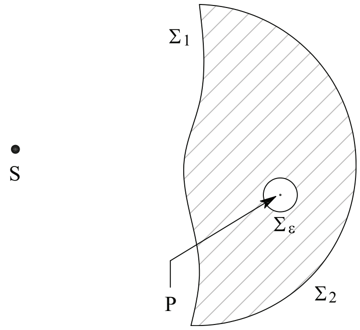


Figure 4.10 Geometry to calculate the diffraction at point P with the Green's function $U(x,y,z) = e^{-iks}/s$.

$$\begin{aligned}\nabla U \cdot \hat{\mathbf{n}} &= \frac{ike^{iks} s \nabla s - e^{iks} \nabla s}{s^2} \cdot \hat{\mathbf{n}} \\ &= \frac{e^{iks}}{s}\left(ik - \frac{1}{s}\right)(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}}),\end{aligned} \tag{4.25}$$

where $\nabla s = \hat{\mathbf{s}}$ is the unit vector in the direction radial from point P. For the sphere Σ_ε, the unit normal vector points toward point P, so $(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}}) = -1$.

In the integral to the right of Eq. (4.24),

$$\left.\frac{\partial U}{\partial n}\right|_{s=\varepsilon} = \left(\frac{1}{\varepsilon} - ik\right)\frac{e^{ik\varepsilon}}{\varepsilon}; \tag{4.26}$$

thus, the integral remains

$$\iint_{\Sigma_\varepsilon} \left(E\left(\frac{1}{\varepsilon} - ik\right)\frac{e^{ik\varepsilon}}{\varepsilon} - \frac{e^{ik\varepsilon}}{\varepsilon}\frac{\partial E}{\partial n}\right)\varepsilon^2 \sin\theta d\theta d\phi, \tag{4.27}$$

where the differential area has been written in spherical coordinates (for the sphere Σ_ε), with θ as the polar angle and ϕ as the azimuthal angle. In the limit $\varepsilon \rightarrow 0$, this last integral reduces to

$$\iint_{\Sigma_\varepsilon} E \sin\theta d\theta d\phi = 4\pi E(x', y', z'). \tag{4.28}$$

Consequently, the field at point P can be calculated as

$$E(x', y', z') = -\frac{1}{4\pi}\iint_{\Sigma} \frac{e^{iks}}{s}\left[E\left(ik - \frac{1}{s}\right)(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}}) - \frac{\partial E}{\partial n}\right] d\sigma, \tag{4.29}$$

where the unit vector $\hat{\mathbf{n}}$ now corresponds to surface Σ. This integral is called the *Kirchhoff integral theorem.*

4.2.2 Fresnel–Kirchhoff diffraction

Suppose we have an opaque flat screen with an aperture and want to determine the diffracted optical field when the light source S is point-like, as in Fig. 4.11(a). To solve the integral equation [Eq. (4.29)] for point P, the first thing to do is select the integration surface Σ. The surface that is usually proposed in this problem is shown in Fig. 4.11(b). The closed surface consists of three open surfaces: the flat surface A that fills the aperture, the flat surface Σ_1 behind the opaque screen, and the surface Σ_2, which is a spherical cap of radius R centered at point P.

Therefore, the Kirchhoff integral must be solved for the three surfaces, which together complete the closed surface $\Sigma = A + \Sigma_1 + \Sigma_2$. Because the surface is arbitrarily (but conveniently) chosen, if $R \rightarrow \infty$, then the surface Σ_1 will have infinite extent. In this case, the one-aperture diffraction problem assumes that the opaque screen has infinite extent. Thus, the contribution of the field emitted by the point source $E = E_0^{\dagger} e^{ikr}/r$, where r is the distance between source S and a point (x, y, z), on each surface will be:

- in A, assuming that the field in the aperture is equal to the field in the absence of the opaque screen with the aperture, then

$$E_A = E_0^{\dagger} e^{ikr}/r, \tag{4.30}$$

  where r is the distance between S and a point in the aperture;
- in Σ_1, assuming that the opaque screen does not transmit light, then

$$E_{\Sigma_1} = 0; \tag{4.31}$$

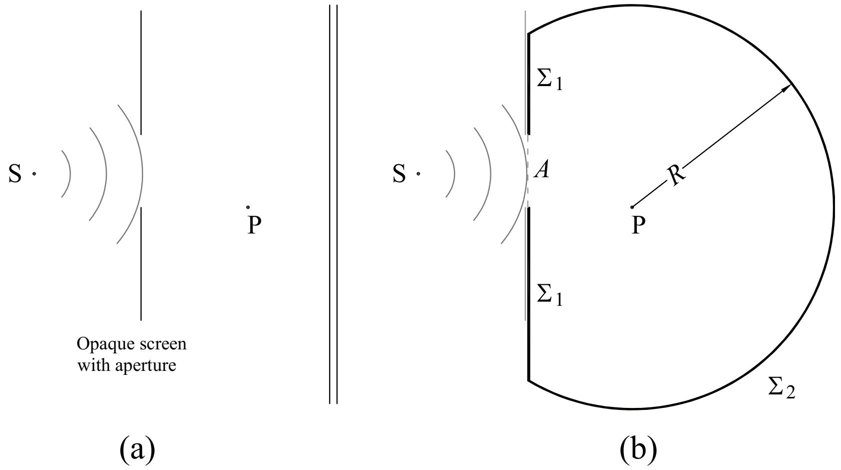


Figure 4.11 (a) Geometry of the diffraction of a spherical wave in an aperture. (b) Selection of the integration surface to solve the Kirchhoff integral.

- in Σ_2, with $R \to \infty$, E_{Σ_2} (and also U) decreases as $1/R$, so the field E is practically null. However, the area of integration grows as R^2 (in Σ_2, $d\sigma = R^2 \sin\theta d\theta d\phi$), so it is not obvious that the diffraction integral vanishes in Σ_2.

The first two assumptions also have some drawbacks. In the first one, the presence of the screen changes the field at the edge of the aperture; in the second assumption, the field extends behind the opaque screen in the vicinity of the aperture [5]. However, for practical problems where the size of the aperture is much larger than the wavelength, these two assumptions work very well. For the surface Σ_2, if $R \gg \lambda$, the integral of Eq. (4.29), with $s = R$, is approximated by

$$-\frac{1}{4\pi}\iint_{\Sigma_2} UR\left(ikE - \frac{\partial E}{\partial n}\right) R\sin\theta d\theta d\phi, \tag{4.32}$$

taking into account that now $(\hat{\mathbf{s}}\cdot\hat{\mathbf{n}}) = 1$. Because UR is finite valued with $R \to \infty$, the integral in Σ_2, Eq. (4.32), vanishes if

$$\lim_{R\to\infty} R\left(ikE - \frac{\partial E}{\partial n}\right) = 0. \tag{4.33}$$

This is called the *condition of radiation of Sommerfeld*, and is satisfied if $E \to 0$ as fast as $1/R$. This occurs for a point source, and the contribution of the integral on the surface Σ_2, in effect, is null. Therefore, the diffraction generated by an aperture when illuminated by a point source will be given by

$$E(x', y', z') = -\frac{1}{4\pi}\iint_A E_0^\dagger \frac{e^{ikr}}{r}\frac{e^{iks}}{s}\left[\left(ik - \frac{1}{s}\right)(\hat{\mathbf{s}}\cdot\hat{\mathbf{n}}) - \left(ik - \frac{1}{r}\right)(\hat{\mathbf{r}}\cdot\hat{\mathbf{n}})\right] d\sigma, \tag{4.34}$$

where $\hat{\mathbf{r}}$ is the unit vector from the source S to a point Q [of coordinates (x, y, z)] in the aperture A, $\hat{\mathbf{s}}$ is the unit vector from the observation point P [of coordinates (x', y', z')] to point Q, and $\hat{\mathbf{n}}$ is the unit normal vector to surface A at point Q, as illustrated in Fig. 4.12(a).

If the distances $r = \overline{\text{SQ}}$ and $s = \overline{\text{QP}}$ are much greater than the wavelength, the approximations $(ik - 1/s) \approx ik$ and $(ik - 1/r) \approx ik$ can be used in Eq. (4.34); therefore,

$$E(x', y', z') = -\frac{i}{\lambda}\iint_A E_0^\dagger \frac{e^{ikr}}{r}\frac{e^{iks}}{s}\left[\frac{(\hat{\mathbf{s}}\cdot\hat{\mathbf{n}}) - (\hat{\mathbf{r}}\cdot\hat{\mathbf{n}})}{2}\right] d\sigma. \tag{4.35}$$

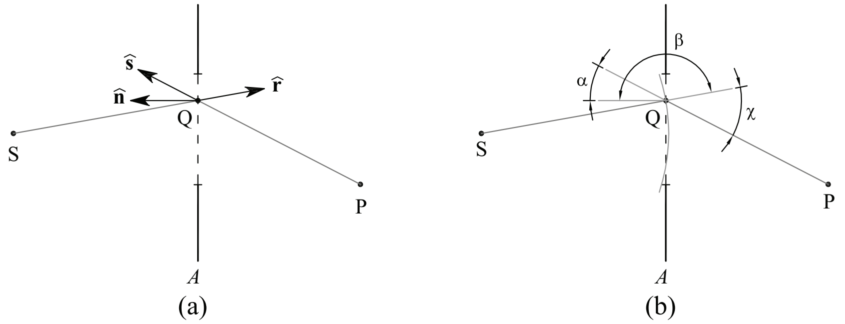


Figure 4.12 Unit vectors at the Q point of the diffraction aperture.

This integral is called the *diffraction integral of Fresnel–Kirchhoff*. In fact, the term $-i[(\hat{\mathbf{s}}\cdot\hat{\mathbf{n}})-(\hat{\mathbf{r}}\cdot\hat{\mathbf{n}})]/2\lambda$ formally defines the inclination factor $K(\chi)$ from Eq. (4.2) in the Fresnel treatment. By defining the angles between the unit vectors as shown in Fig. 4.12(b), the inclination factor can be written as

$$\begin{aligned} K(\chi) &= -\frac{i}{\lambda}\frac{[\cos\alpha-\cos\beta]}{2} \\ &= \frac{i}{\lambda}\frac{[\cos\beta+\cos(\chi+\beta)]}{2}. \end{aligned} \tag{4.36}$$

This definition leads to an analogous equation to Eq. (4.2); thus,

$$E(\mathrm{P}) = \iint_A E_0^{\dagger}\frac{e^{ikr}}{r}\frac{e^{iks}}{s}K(\chi)d\sigma. \tag{4.37}$$

The inclination factor defined in Eq. (4.36) does not depend only on the angle χ, as the Fresnel formulation suggests. This is because the integration surface in Eq. (4.37) is flat, whereas the integration surface in Eq. (4.2) is the spherical wavefront of radius r_0, in which the amplitude of the secondary sources is constant and equal to $E_0^{\dagger}e^{ikr_0}/r_0$. Instead, in Eq. (4.37), not only is the amplitude of the secondary sources variable at the aperture A, but the secondary sources $E(Q)=E_0^{\dagger}e^{ikr}/r$ are not always on the same wavefront. In this sense, Eq. (4.37) is a generalized version of the Huygens–Fresnel principle.

When the source is at infinity, Eqs. (4.2) and (4.37) coincide, since the angle $\beta=\pi$ (and $\alpha=\chi$). In this case, the inclination factor turns out to be

$$K(\chi) = -\frac{i}{\lambda}\left(\frac{1+\cos\chi}{2}\right). \tag{4.38}$$

This situation is illustrated in Fig. 4.13.

A further simplification occurs in the paraxial approximation where $\chi\approx 0$. This last situation is very common in practice and is analyzed in the next section.

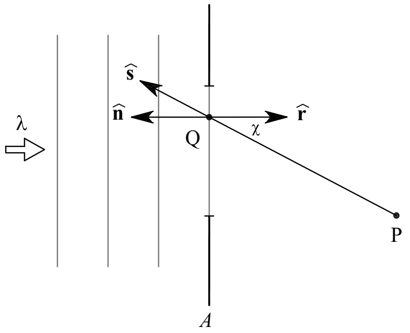


Figure 4.13 Unit vectors when the incident wavefront at the aperture is flat.

4.2.3 Sommerfeld diffraction

To obtain the Fresnel–Kirchhoff integral, the spherical wave of unit amplitude e^{iks}/s was used as a Green's function. The selection of the function is arbitrary, but it should facilitate the calculation of the diffraction. This function has some issues with the E field inside the aperture, at the edges of the aperture, and just behind the aperture near the edges. Sommerfeld proposes another Green's function, such that the aperture boundary problems are solved, while maintaining the assumption that the optical field within the aperture is equal to the optical field in the same region when there is no opaque screen defining the aperture.

The new function U is constructed with two unit spherical waves, one originating from the observation point P [as in Eq. (4.23)] and the other originating from the point P′, which is the mirror image of P with respect to the plane of the aperture A, as illustrated in Fig. 4.14.

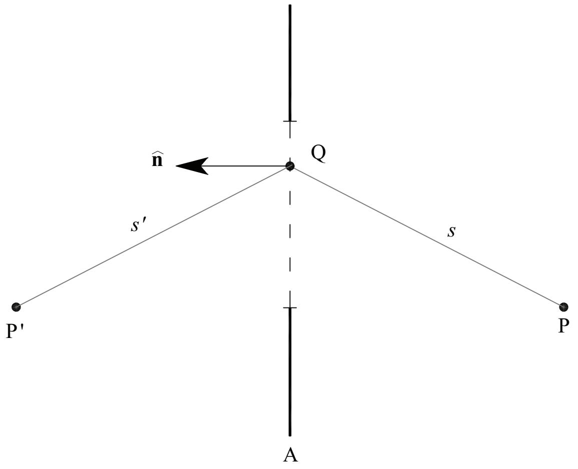


Figure 4.14 P and P′: origin of the two auxiliary unit spherical waves for the diffraction calculation.

With this configuration, the Green's function becomes

$$U = \frac{e^{iks}}{s} - \frac{e^{iks'}}{s'}. \tag{4.39}$$

Using this function in the diffraction problem described in Fig. 4.11(b) when $R \to \infty$, for any point in Σ_1 and in A, $s' = s$ and, therefore, $U = 0$ and $\partial U / \partial n = 0$. Thus, it is not necessary to make any assumptions about the boundary conditions of the field E at Σ_1 and at the edge of the aperture, eliminating the inconsistencies of the entire Green's function chosen by Kirchhoff. Using Eq. (4.39), leads to [6]

$$E(x', y', z') = -\frac{i}{\lambda} \iint_A E_0^{\dagger} \frac{e^{ikr}}{r} \frac{e^{iks}}{s} (\hat{\mathbf{s}} \cdot \hat{\mathbf{n}}) d\sigma. \tag{4.40}$$

This integral is called the *first Rayleigh–Sommerfeld solution.*

The function $U = e^{iks}/s + e^{iks'}/s'$ can also be chosen, which gives rise to the *second Rayleigh–Sommerfeld solution* [6]:

$$E(x', y', z') = \frac{i}{\lambda} \iint_A E_0^{\dagger} \frac{e^{ikr}}{r} \frac{e^{iks}}{s} (\hat{\mathbf{s}}' \cdot \hat{\mathbf{n}}) d\sigma. \tag{4.41}$$

The inclination factor will be different in each case:

- $K(\chi) = -i[(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}}) - (\hat{\mathbf{r}} \cdot \hat{\mathbf{n}})]/2\lambda$ in Fresnel–Kirchhoff;
- $K(\chi) = -i(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}})/\lambda$ in the first Rayleigh–Sommerfeld solution; and
- $K(\chi) = i(\hat{\mathbf{s}}' \cdot \hat{\mathbf{n}})/\lambda$ in the second Rayleigh–Sommerfeld solution.

When the aperture is illuminated by a plane wave (source S at infinity) and in the paraxial approximation, the inclination factors coincide at $K(\chi) = -i/\lambda$.

4.3 Fresnel and Fraunhofer Diffraction

The problem of diffraction by an aperture in a flat opaque screen, as illustrated in Fig. 4.15, is considered in this section.

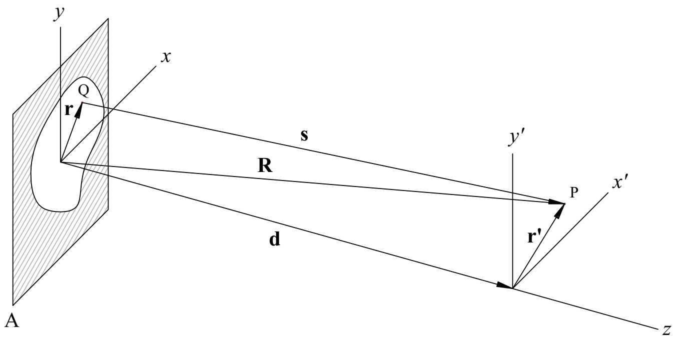


Figure 4.15 Geometry for the calculation of diffraction by a plane aperture.

According to Eq. (4.37), the Huygens–Fresnel diffraction integral can be written in general as

$$E(\mathrm{P}) = \iint_A E(\mathrm{Q})K(\chi)\frac{e^{iks}}{s}d\sigma. \tag{4.42}$$

$E(Q)$ is the complex amplitude of the field at the aperture. The aperture is in the xy plane, and the diffraction pattern observation screen (spatial distribution of irradiance) is in the $x'y'$ plane. These planes are separated by the distance $d = |\mathbf{d}|$. In particular, we will limit ourselves to the paraxial approximation, which is defined under the following conditions:

1. The separation d between the plane of the aperture and the plane of the diffraction pattern satisfies $d \gg \lambda$.
2. The dimensions of the aperture are much smaller than the distance d.
3. The dimensions of the region of observation of the diffraction pattern are much smaller than the distance d.
4. The inclination factor for any point in the aperture is approximated by the inclination factor of the first Fresnel zone, i.e., $K(\chi) \approx -i/\lambda$.[*]

The vectors indicated in Fig. 4.15 are:

- $\mathbf{d}$, separation vector between the planes of the aperture and the observation screen;
- $\mathbf{r} = \{x,y\}$, position vector of the point Q in the aperture;
- $\mathbf{r}' = \{x',y'\}$, position vector of point P on the observation screen;
- $\mathbf{R} = \mathbf{d} + \mathbf{r}'$, relative position vector of point P with respect to the origin of coordinates of the aperture; and
- $\mathbf{s}$, relative position vector of point P with respect to point Q.

Taking into account the paraxial condition, it is fulfilled that $|\mathbf{r}| \ll |\mathbf{R}|$, $|\mathbf{r}'| \ll |\mathbf{R}|$ and $|\mathbf{R}| \approx |\mathbf{d}|$.

From the law of cosines,

$$s^2 = R^2 + r^2 - 2\mathbf{R}\cdot\mathbf{r} \tag{4.43}$$

or

$$s = R\sqrt{1 + \left(\frac{r}{R}\right)^2 - \frac{2\mathbf{R}\cdot\mathbf{r}}{R^2}}. \tag{4.44}$$

[*] This inclination factor value is obtained from $K(\chi) = -i(1 + \cos\chi)/2\lambda$ when in Eq. (4.36) the angle $\beta = \pi$, i.e., when illuminated by a plane wave. In practice, the wavefront at the aperture may have small deviations, such that $\beta \approx \pi$. In this case, it could still be assumed that $K(\chi) = -i/\lambda$.

Because $\mathbf{R} \cdot \mathbf{r} = (\mathbf{d} + \mathbf{r}') \cdot \mathbf{r} = \mathbf{r}' \cdot \mathbf{r}$, then

$$s = R\sqrt{1 + \left(\frac{r}{R}\right)^2 - \frac{2\mathbf{r}' \cdot \mathbf{r}}{R^2}}. \tag{4.45}$$

The paraxial approximation implies that the square root can be approximated to second order; thus,

$$s = R\left[1 + \frac{1}{2}\left(\frac{r^2 - 2\mathbf{r}' \cdot \mathbf{r}}{R^2}\right)\right]. \tag{4.46}$$

Completing the square binomial for the term that is in parentheses in Eq. (4.46),

$$s = R\left[1 - \frac{r'^2}{2R^2} + \frac{1}{2}\left(\frac{r'^2 - 2\mathbf{r}' \cdot \mathbf{r} + r^2}{R^2}\right)\right]. \tag{4.47}$$

With $r' \ll R$, we have $1 - r'^2/2R^2 \approx 1$, and replacing R with d, Eq. (4.47) simplifies to

$$s = d + \frac{|\mathbf{r}' - \mathbf{r}|^2}{2d}. \tag{4.48}$$

Therefore, the paraxial form of Eq. (4.42) is obtained by replacing s with the expression given by Eq. (4.48). The distance s is in the argument of the exponential function and in the denominator of Eq. (4.42). The quantity $|\mathbf{r}' - \mathbf{r}|^2/2d$ describes small variations with respect to distance d. In the exponential function, these small variations are comparable with the wavelength; thus, the relationship between these two quantities gives important phase changes in the interference process. On the other hand, in the denominator, these small variations are only compared with the distance d; therefore, there is no appreciable change in the denominator, and there s can be exchanged for d. Finally, the paraxial version of Eq. (4.42) becomes

$$E(\mathbf{r}') = -i\frac{e^{ikd}}{\lambda d}\iint_A E(\mathbf{r})e^{ik|\mathbf{r}' - \mathbf{r}|^2/2d}d^2r. \tag{4.49}$$

From this integral, one can see what the secondary waves of the Huygens–Fresnel principle look like in the paraxial approximation. Consider Fig. 4.16, in which $\mathbf{r} = y\mathbf{j}$ and $\mathbf{r}' = y'\mathbf{j}$. At Q($y$), there is a source of amplitude $E(y)$ that emits a parabolic wave. The field at P(y'), due to Q, is equal to the attenuated field $E(y)/d$, and the phase shift of the wave at P is given by $ik[d + (y' - y)^2/2d]$. Note that the term $(y' - y)^2/2d$ is the sag at distance $y' - y$ from the vertex V of the front parabolic waveform centered on Q. Thus, in the paraxial approximation, the secondary waves of the Huygens–Fresnel

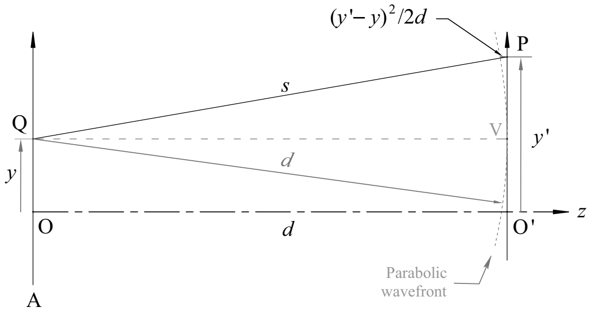


Figure 4.16 Approximation of the distance s as $d+(y'-y)^2/2d$.

principle are paraboloids centered on the sources located in the primary wavefront Σ.

Writing the vectors of Eq. (4.49) in Cartesian components,

$$E(x', y') = -i\frac{e^{ikd}e^{ik(x'^2+y'^2)/2d}}{\lambda d}\iint_A E(x, y)e^{ik(x^2+y^2)/2d}e^{-ik(x'x+y'y)/d}dxdy. \tag{4.50}$$

This version of the diffraction integral allows us to observe the influence of the Fresnel zones in the field that fills the opening A. Focusing on the plane of the aperture, the field $E(x,y)$ is modulated by the term $e^{ik(x^2+y^2)/2d}$. Now, the term $(x^2+y^2)/2d$ will be the sag at the distance $(x^2+y^2)^{1/2}$ from the origin of coordinates O of a paraboloid with center at O′ (Fig. 4.16). If the distance $(x^2+y^2)/2d$ is divided by $\lambda/2$, we will count the number of Fresnel zones contained in a circular opening of radius $(x^2+y^2)^{1/2}$. Therefore,

$$e^{ik(x^2+y^2)/2d} = e^{i\pi N}, \tag{4.51}$$

where $N=(x^2+y^2)/\lambda d$ is the number of Fresnel zones.

4.3.1 Fraunhofer diffraction

Let $\rho_{\max}$ be the radius of the circle circumscribing the diffraction aperture and $N_{\max}=\rho^2_{\max}/\lambda d$ be the number of Fresnel zones subtended by the circle with respect to the diffraction pattern observation plane. If $N_{\max}\ll 1$ ($N_{\max}\approx 0$), then the term $e^{ik(x^2+y^2)/2d}$ inside the integral can be neglected and the diffraction pattern will be given by

$$I(x', y') = \frac{(\epsilon_0 c/2)}{\lambda^2 d^2}\left|\iint_A E(x, y)e^{-i2\pi(x'x+y'y)/\lambda d}dxdy\right|^2. \tag{4.52}$$

This integral is called the *Fraunhofer diffraction integral.*

4.3.2 Fresnel diffraction

Let $\rho_{\max}$ be the radius of the circle circumscribing the diffraction aperture and $N_{\max} = \rho^2_{\max}/\lambda d$ be the number of Fresnel zones subtended by the circle with respect to the diffraction pattern observation plane. If $N_{\max}$ is of the order of a Fresnel zone, then the term $e^{ik(x^2+y^2)/2d}$ cannot be negligible within the integral and, in this case, the diffraction pattern will be given by

$$I(x', y') = \frac{(\epsilon_0 c/2)}{\lambda^2 d^2} \left| \iint_A [E(x, y) e^{ik(x^2+y^2)/2d}] e^{-i2\pi(x'x+y'y)/\lambda d} dxdy \right|^2. \tag{4.53}$$

This integral is called the *Fresnel diffraction integral.*

The diffraction integrals given by Eqs. (4.52) and (4.53) have the form of a two-dimensional Fourier transform whose spatial frequencies are $x'/\lambda d$ and $y'/\lambda d$. Therefore, it is easy to compute these integrals numerically.

4.3.3 Some examples

A circular aperture

A very common aperture in diffraction is the circular aperture. This has very important practical applications because the aperture diaphragms or lens edges of an optical system are usually circular.

Suppose we have a circular aperture of radius w illuminated by a flat homogeneous wavefront (of amplitude E_0 and wavelength λ) orthogonal to the plane of the aperture. The field at the aperture can be described as

$$E(x, y) = E_0 \mathrm{circ}\left(\frac{\sqrt{x^2 + y^2}}{w} \right), \tag{4.54}$$

where circ( ) is called the circular function and is defined as 1 for $\sqrt{x^2 + y^2} \leq w$, and 0 for other values of (x,y)

In this case, the aperture coincides with the circle that circumscribes the diffracting aperture. Then the number of Fresnel zones is given by $N = w^2/\lambda d$. The Fresnel diffraction patterns for a circular aperture of radius $w = 0.5$ mm are shown in Fig. 4.17,[*] obtained when the observation screen distance corresponds to (a) the first Fresnel zone ($d = 395.1$ mm), (b) the first two Fresnel zones ($d = 197.5$ mm), (c) the first three Fresnel zones ($d = 131.7$ mm), and (d) the first four Fresnel zones ($d = 98.7$ mm), using

[*] To make the rings or regions of lower intensity visible, instead of plotting the irradiance, the square root of the irradiance is drawn. This is done for the simulated patterns in this section, as well as in Sections 4.4 and 4.5. But in Section 4.5.2, the irradiance is drawn because it better illustrates resolution for a two-point image and corresponds to what is observed in practice.

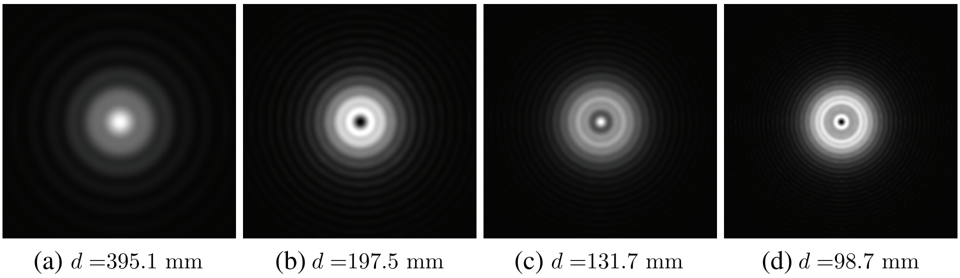


Figure 4.17 Fresnel diffraction patterns of a circular aperture of radius 0.5 mm, for a wavelength of 632.8 nm. The size of the image box in all cases is 3 mm on each side.

$\lambda = 632.8$ nm. Of course, the distance that the observation screen must be in each case is calculated from $d = w^2/N\lambda$.

In the case of the Fraunhofer diffraction, the calculation is even easier because it does not include the term $e^{ik(x^2+y^2)/2d}$ within the integral. This implies that $d \to \infty$. In practice, the observation screen must be placed at a finite distance such that $d \gg w$, e.g., the Fraunhofer pattern when $d = 3951$ mm is shown in Fig. 4.18(a). This particular value of the distance d corresponds to the distance by which the circular aperture subtends 0.1 Fresnel zones. The profile of this pattern, Fig. 4.18(b), corresponds to the square of a Bessel function divided by its argument, as shown below.

The Fraunhofer integral for a circular aperture has a well-known analytical solution. Because the aperture is circular, it is convenient to solve the integral in cylindrical coordinates. Let $r = (x^2 + y^2)^{1/2}$, $\tan\phi = y/x$; $r' = (x'^2 + y'^2)^{1/2}$, and $\tan\varphi = y'/x'$. Hence, $x = r\cos\phi$, $y = r\sin\phi$, $x' = r'\cos\varphi$, and $y' = r'\sin\varphi$. By changing the variables in the diffraction integral,

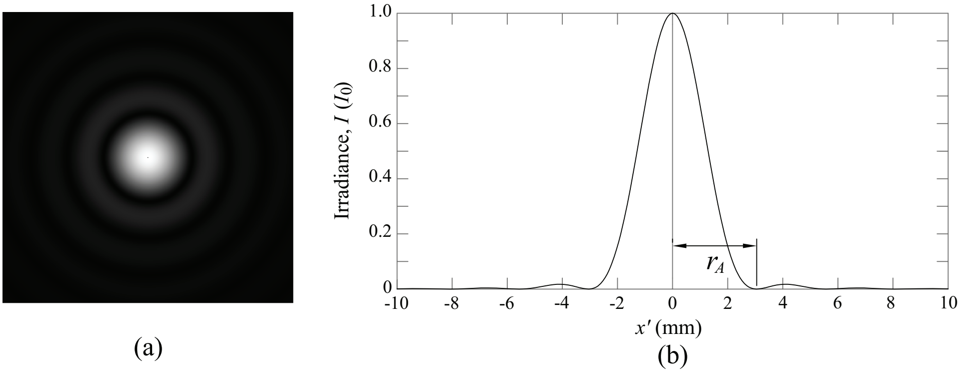


Figure 4.18 (a) Fraunhofer diffraction pattern on a screen observation at a distance of 3951 mm from a circular aperture of radius 0.5 mm. The size of the image box is 20 mm on a side. (b) Profile of the diffraction pattern.

$$I(r') = \frac{E_0^2(\epsilon_0 c/2)}{\lambda^2 d^2}\left|\int_0^w \int_0^{2\pi} e^{-i2\pi r' r[\cos(\phi-\varphi)]/\lambda d} r dr d\phi\right|^2. \tag{4.55}$$

Because the problem has symmetry for φ, it can be solved for any φ. In particular, for $\varphi = 0$, the integral becomes

$$\int_0^w \int_0^{2\pi} e^{-i2\pi r' r \cos\phi/\lambda d} r dr d\phi, \tag{4.56}$$

which is equal to

$$2\pi \int_0^w J_0(2\pi r' r/\lambda d) r dr. \tag{4.57}$$

Using the recurrence relation $\int_0^u J_0(u')u' du' = uJ_1(u)$, then

$$I(r') = I_0\left|\frac{2J_1(2\pi w r'/\lambda d)}{2\pi w r'/\lambda d}\right|^2, \tag{4.58}$$

with $I_0 = (\epsilon_0 c/2)(\pi w^2 E_0/\lambda d)^2$.

The first dark ring of the pattern in Fig. 4.18 is obtained for the first zero of the Bessel function, i.e., when $2\pi w r'/\lambda d = 3.8317$. Because the energy contained in the region enclosed by the first dark ring is around 84%, the first ring plays a very important role in the image of a point source. The distribution of irradiance contained within the first ring is called the *Airy disk*. The radius of the Airy disk is given by

$$r'_A = 1.22\frac{\lambda d}{(2w)}. \tag{4.59}$$

In the example of Fig. 4.18(b), the Airy radius is 3.0541 mm. The second dark ring of the pattern is obtained for the second zero of the Bessel function, i.e., $2\pi w r'/\lambda d = 7.0156$, which gives $r' = 2.23\lambda d/(2w)$. In the example in Fig. 4.18(b), this radius is 5.5825 mm. The energy in the region enclosed by the second dark ring is 91%. The function $2J_1(u)/u$ is called the Jinc(u) function.

A rectangular aperture

Another aperture, also widely used in practical diffraction problems, is the rectangular aperture. Suppose we have a rectangular aperture with sides $2w_x$ and $2w_y$ illuminated by a homogeneous plane wavefront (of amplitude E_0 and wavelength λ) orthogonal to the plane of the aperture. The field at the aperture can be described as

$$E(x, y) = E_0 \mathrm{rect}\left(\frac{x}{2w_x}\right) \mathrm{rect}\left(\frac{y}{2w_y}\right), \tag{4.60}$$

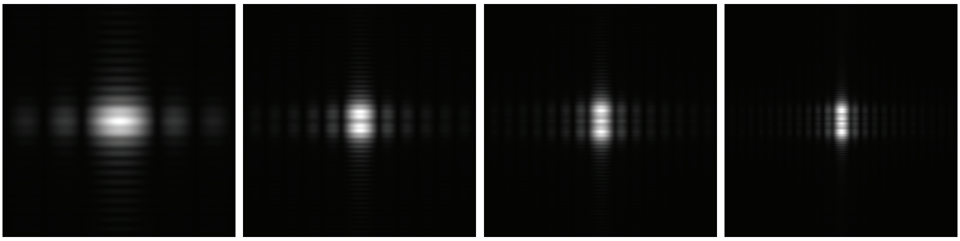
where rect(x/2w_x) is called the *rectangle function* and is defined as 1 for $|x| \leq w_x$ and 0 for other values of x. Similarly, the function rect(y/2w_y) is defined for the coordinate y.

Fresnel diffraction patterns for a rectangular aperture with sides $2w_x = 1$ mm and $2w_y = 4$ mm are shown in Fig. 4.19, obtained when the distance to the observation screen corresponds to (a) the first Fresnel zone ($d = 6321$ mm), (b) the first two Fresnel zones ($d = 3161$ mm), (c) the first three Fresnel zones ($d = 2107$ mm), and (d) the first four Fresnel zones ($d = 1580$ mm), using $\lambda = 632.8$ nm. Because $w_y > w_x$, the circle that circumscribes the rectangular aperture will have a radius close to w_y. Therefore, the number of Fresnel zones in this case is calculated as $N = w_y^2/\lambda d$ and, consequently, the distance at which the observation screen should be at is $d = w_y^2/N\lambda$.

The Fraunhofer diffraction when $d = 63211$ mm is shown in Fig. 4.20(a). In this example, this value of d also corresponds to the distance by which the circle circumscribing the rectangular aperture subtends 0.1 Fresnel zones.

The Fraunhofer integral for the rectangular aperture also has a well-known analytical solution. In this case,

$$\begin{aligned} I(x', y') &= \frac{E_0^2(\epsilon_0 c/2)}{\lambda^2 d^2} \left| \int_{-w_x}^{w_x} \int_{-w_{yx}}^{w_y} e^{-i2\pi(x'x+y'y)/\lambda d} dxdy \right|^2 \\ &= \frac{E_0^2(\epsilon_0 c/2)}{\lambda^2 d^2} \left| \int_{-w_x}^{w_x} e^{-i2\pi(x'x)/\lambda d} dx \right|^2 \left| \int_{-w_y}^{w_y} e^{-i2\pi(y'y)/\lambda d} dy \right|^2, \end{aligned} \tag{4.61}$$



(a) d =6321 mm (b) d =3161 mm (c) d =2107 mm (d) d =1580 mm

Figure 4.19 Fresnel diffraction patterns of a rectangular aperture with sides of 1 mm and 4 mm, for a wavelength of 632.8 nm. The size of the image in all cases is 24 mm on each side.

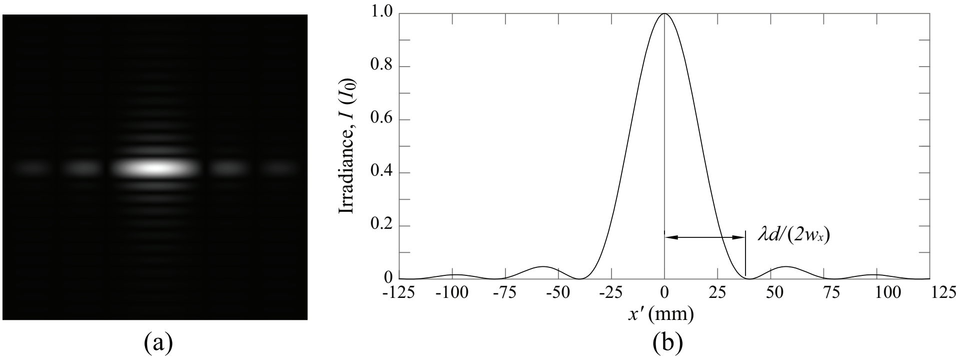


Figure 4.20 (a) Fraunhofer diffraction pattern on the observation screen at a distance of 63211 mm from the rectangular aperture with sides 1 mm and 4 mm. Image size is 250 mm on each side. (b) Profile of the diffraction pattern.

i.e.,

$$I(x', y') = I_0 \left[\frac{\sin(2\pi w_x x'/\lambda d)}{(2\pi w_x x'/\lambda d)}\right]^2 \left[\frac{\sin(2\pi w_y y'/\lambda d)}{(2\pi w_y y'/\lambda d)}\right]^2, \tag{4.62}$$

with $I_0 = (\epsilon_0 c/2)(4w_x w_y E_0/\lambda d)^2$. The function sin($u$)/$u$ was introduced in Section 3.1.3, i.e., the function sinc(u). The profile of this pattern in the horizontal direction (x') is shown in Fig. 4.20(b). The zeros of the function $\text{sinc}(2\pi w_x x'/\lambda d)$ are obtained at

$$x'_m = m\frac{\lambda d}{(2w_x)}, \tag{4.63}$$

with $m = \pm 1, \pm 2, \pm 3, \ldots$. In the vertical direction (y'), there is a similar behavior, with the zeros at $y'_m = \pm m\lambda d/(2w_y)$. Most of the energy is in the region bounded by the leading zeros: $x'_{\pm 1} = \pm\lambda d/(2w_x)$ and $y'_{\pm 1} = \pm\lambda d/(2w_y)$. Therefore, it is convenient to define the width of the sinc( ) as $\Delta x' = 2x'_1$ in the direction x' and $\Delta y' = 2y'_1$ in the direction y'.

Note that the diffraction pattern has a greater dispersion in the direction in which the rectangular aperture has a shorter length [Fig. 4.20(a)]. This topic is treated in Section 4.6 on diffraction gratings. One-dimensional gratings are such that each diffraction element satisfies $w_x \ll w_y$, which makes the diffraction pattern look like a one-dimensional irradiance distribution.

On the other hand, when $w_x = w_y = w$, the aperture geometry is a square of side $2w$. In such a case, the Fraunhofer diffraction pattern is symmetric, as shown in Fig. 4.21.

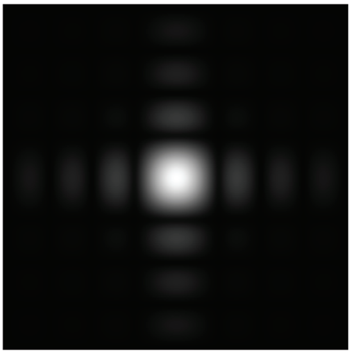


Figure 4.21 Fraunhofer diffraction pattern generated by a square aperture of side $2w = 1$ mm, at a distance $d = 3951$ mm ($N = 0.1$), for a wavelength of 632.8 nm. The image size is 20 mm on each side.

4.4 Young Interferometer II

In Section 3.8, the Young interferometer is analyzed as a system of two mutually coherent point sources, S_1 and S_2. This section deals with two practical aspects of the Young interferometer. The first has to do with the finite size of the two sources (apertures) S_1 and S_2, while the second deals with the size and coherence of the light source. With this, an enhanced description of Young's original experiment from the 19th century [7] is presented in this book.

A diagram of Young's experiment that will be discussed in this section is shown in Fig. 4.22. The primary source S with which the apertures S_1 and S_2 are illuminated is an incoherent, monochromatic extended source of wavelength λ and lateral size σ. The two apertures are circles of radius w and are separated from each other (from their centers) by a distance a along the x direction. The distance between the source S and the apertures is z_p, and the distance between the apertures and the observation screen is d.

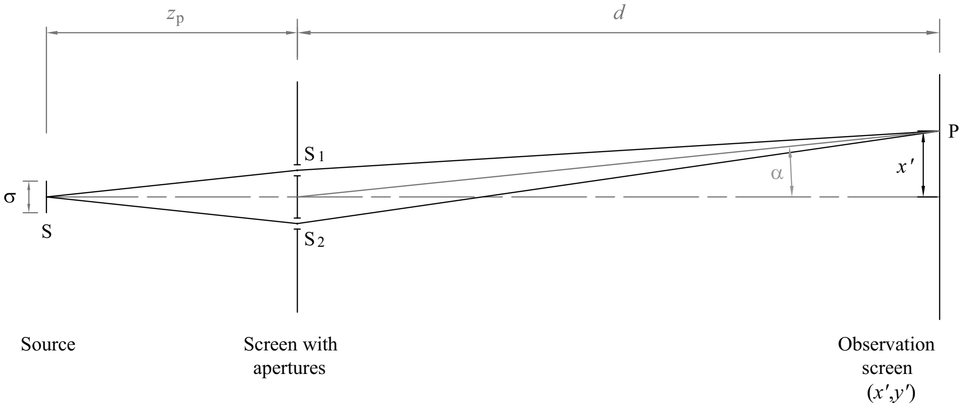


Figure 4.22 Young's experiment, or diffraction through two apertures, in an opaque screen.

Taking into account the conditions under which Young's experiment is performed, $d \gg a \gg \lambda$, the diffraction on the screen corresponds to the Fraunhofer diffraction.

4.4.1 Effect of the size of the diffraction aperture

First, let us assume that the primary source is a point source (on the optical axis) and that the field amplitude at each aperture is uniform and of constant phase. Therefore, the irradiance will be

$$I(x', y') = \frac{\epsilon_0 c}{2\lambda^2 d^2} \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E_0 \left(\mathrm{circ}\left[\frac{\sqrt{x^2+y^2}}{w}\right] * \left[\delta(x-a/2)+\delta(x+a/2)\right] \right) \right.$$
$$\left. \times e^{-i2\pi(x'x+y'x)/\lambda d} dxdy \right|^2 . \tag{4.64}$$

The function circ( ) describes the geometry of each aperture of radius w, and the Dirac delta functions locate the aperture at $x = -a/2$ and $x = a/2$. The symbol $*$ denotes the convolution operation. Taking into account that the Fourier transform of the convolution of two functions is equal to the product of the Fourier transforms of each function,

$$I(x', y') = 4I_0 \left[\frac{2J_1(2\pi w r'/\lambda d)}{2\pi w r'/\lambda d}\right]^2 \left[\cos\left(\frac{\pi a}{\lambda d} x'\right)\right]^2, \tag{4.65}$$

with $I_0 = (\epsilon_0 c/2)(\pi w^2 E_0/\lambda d)^2$ and $r' = \sqrt{x'^2 + y'^2}$. Note that $\cos^2(\pi a x'/\lambda d)$ is the Fourier transform of $[\delta(x-a/2)+\delta(x+a/2)]$. In fact, the result given in Eq. (4.65) is the modulated version of the result given in Eq. (3.117). Thus, the modulation of the pattern is determined by the size of the aperture, whereas the spacing between the fringes is determined by the spacing between the apertures. The simulation of the diffraction pattern by two identical circular apertures of radius $w = 0.5$ mm separated by $a = 5$ mm, with $\lambda = 632.8$ nm, is shown in Fig. 4.23(a) when the observation screen is at a distance $d = 3951$ mm (as in Fig. 4.18). The horizontal profile of the pattern is shown in Fig. 4.23(b); the gray segmented curve describes the modulation of the interference pattern due to the diffraction pattern.

4.4.2 Effect of light source size

Now let us see how the extent σ of the source S affects the diffraction pattern. Let us assume that the source is monochromatic and spatially incoherent, i.e., the oscillations of the fields emitted by two (independent) point sources of S are uncorrelated and therefore these fields do not interfere with each other. This implies that if we consider two point sources of S, each will generate its

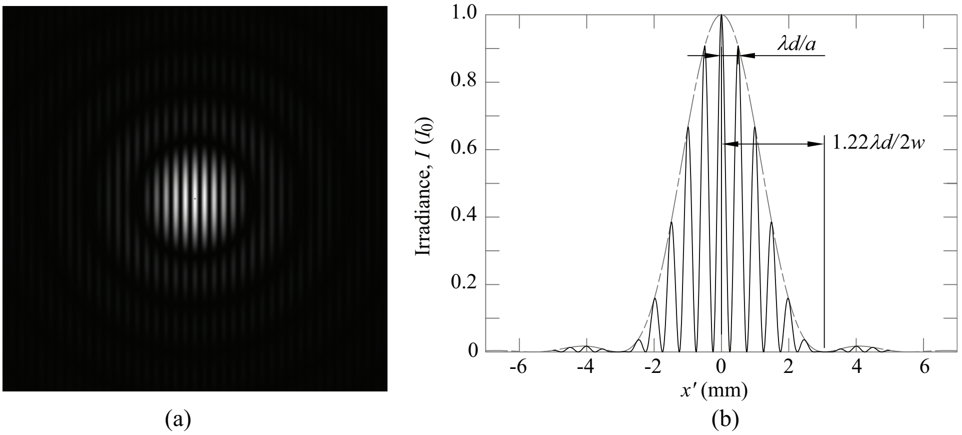


Figure 4.23 (a) Interference pattern generated by two circular apertures of 0.5 mm radius and 5 mm apart. The interference pattern is modulated by the diffraction pattern (in gray) of one of the apertures. (b) Interference pattern profile.

own Young interference pattern. The final result is the sum of the intensities produced by each of the point sources.

To qualitatively see the effect of the source size, let us consider a light source formed by two incoherent point sources separated from each other by the distance σ', as shown in Fig. 4.24. The angular size of the source will be $\alpha = \sigma'/z_p$. Each source generates its own interference pattern with an offset for the maximum of $\pm d\alpha/2$. The interference fringes in each pattern will be separated from each other by the distance $\lambda d/a$. Because the two patterns are identical, when the offset between the patterns is equal to $\lambda d/2a$, the minima of one pattern coincide with the maxima of the other pattern; therefore, the sum of the patterns eliminates the interference fringes. Let us denote by $\alpha_0 = \lambda/2a$ the angle that subtends the displacement $\lambda d/2a$ with respect to the midpoint between the apertures.

The evolution of the interference pattern, as the angular size of the source increases, is shown in Fig. 4.25. The parameters used are the same as those used

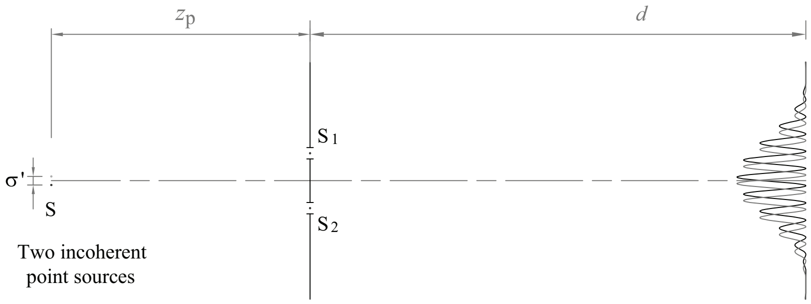


Figure 4.24 Sum of intensities of two incoherent sources.

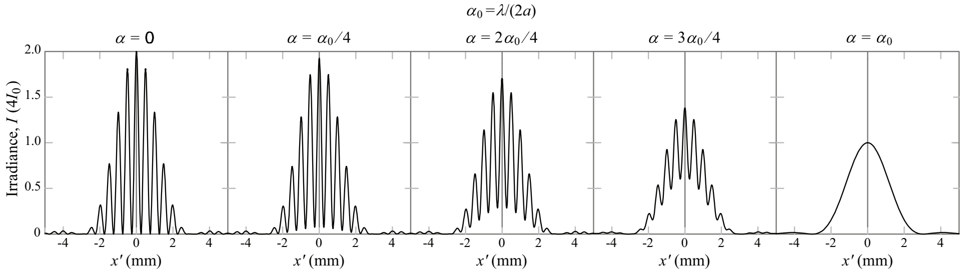
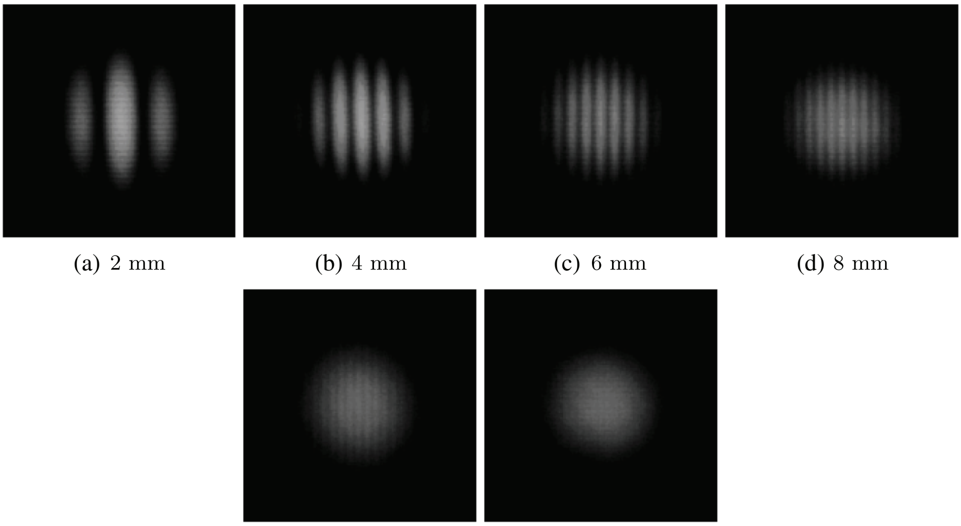


Figure 4.25 Decrease in the visibility of the Young-type interferogram generated by two circular apertures of radius $w = 0.5$ mm separated by $a = 5$ mm, when illuminated with light emitted by two point sources that are incoherent with each other depending on the angular separation α of the sources. The wavelength of light is $\lambda = 632.8$ nm.

in Fig. 4.18. When $\alpha = 0$, the two point sources coincide on the optical axis, and the intensity at each point is twice the intensity of the pattern generated by a single source. The visibility of the interferogram [Eq. (3.32)] is the maximum ($C = 1$). When $\alpha = \alpha_0/4$, the point sources are separated by $\sigma' = \lambda z_p/8a$. The intensity change is small compared with the first case, and the interferogram visibility is still high ($C = 0.93$). With $\alpha = \alpha_0/2$, the point sources are separated by $\sigma' = \lambda z_p/4a$ and now the intensity change is appreciable; the minima inside the envelope move away from zero. The latter notably reduces the visibility of the interferogram ($C = 0.71$). For $\alpha = 3\alpha_0/4$, the point sources are separated by $\sigma' = 3\lambda z_p/8a$ and the visibility decreases considerably ($C = 0.38$). Finally, when $\alpha = \alpha_0$, the point sources are separated by $\sigma' = \lambda z_p/2a$ and the visibility of the interferogram becomes zero; i.e., there are no interference fringes.

This example qualitatively illustrates what happens to a Young interference pattern when illuminated by an extended source (composed of an infinite number of incoherent point sources). According to the visibility of the interferogram generated by the optical fields emerging from the apertures S_1 and S_2, the degree of spatial similarity of these fields can be established. The visibility of the interferogram measures the degree of spatial coherence of the fields in the apertures. If the fields in the apertures are identical, which happens if the source is a point source, as in the case of $\alpha = 0$ in Fig. 4.25, the visibility is maximum and the fields are mutually or fully coherent. As the size of the extended source increases, the field oscillations at the apertures become less correlated, thus decreasing the visibility of the interferogram and the degree of spatial coherence.

Results similar to those shown in Fig. 4.25 can be obtained if instead of changing the size of the light source, the separation of the two apertures is changed. Suppose the size of the light source is σ. Based on what was seen above, the separation of the apertures has a limit value at which the visibility of the interferogram becomes zero. For smaller separations, interference will be observed. In the case of a source made up of two incoherent point sources,



(a) 2 mm (b) 4 mm (c) 6 mm (d) 8 mm

(e) 10 mm (f) 12 mm

Figure 4.26 Interference patterns in Young's experiment, with a partially coherent monochromatic light of wavelength 632.8 nm, as a function of the separation of two circular apertures of 0.5 mm radius each.

the separation limit is $a = \lambda z_p/2\sigma$. However, for a continuous source, this limit value is calculated from the van Cittert–Zernike theorem: the position of the first zero of the Fourier transform of the irradiance distribution of the incoherent source is taken as the maximum separation of the two apertures.*

If the amplitude of the optical fields at the apertures is approximately equal, the result of the van Cittert–Zernike theorem measures the visibility of the interferogram, which is equivalent to measuring the spatial coherence of the fields in the aperture; e.g., if the light source is a square of side σ, with a constant irradiance distribution, the visibility will be given by $C \sim |\sin(\pi\sigma a/\lambda z_p)/(\pi\sigma a/\lambda z_p)|$. If the light source is circular with a radius of $\sigma/2$, with a constant irradiance distribution, the visibility will be given by $C \sim |J_1(\pi\sigma a/\lambda d)/(\pi\sigma a/\lambda d)|$. Then, the limiting values for the aperture spacing will be: $a = \lambda z_p/\sigma$ with the square light source, and $a = 1.22\lambda z_p/\sigma$ with the circular light source.

The experimental interference patterns generated by two circular openings of radius $w = 0.5$ mm, when the illumination source is incoherent with an irradiance distribution that follows a Gaussian profile [8], are shown in Fig. 4.26. In this experiment, $C \sim e^{-(a/5.95)^2}$, where 5.95 is the width of the

*The van Cittert–Zernike theorem states that the region in the plane of the diffraction apertures within which the spatial coherence is not zero is determined by the Fourier transform of the irradiance distribution of the incoherent light source. A good explanation of this theorem is found in Born and Wolf [4].

Gaussian profile in millimeters and is taken as the radius of the coherence region. In the experiment, the opening gap varies from 2 to 12 mm in steps of 2 mm. By examining the profile of the interferograms in the horizontal direction (x'), curves similar to those shown in Fig. 4.25 are obtained. When $a = 12$ mm, there is no more interference. Note that in these images no interference rings are observed [as shown in the simulated pattern in Fig. 4.23(a)]. This occurs because in practice, the maximum irradiance of the first ring is very small with respect to the maximum of the central region, as can be seen in the profile of Fig. 4.23(b).

From the lessons learned in this section, one can imagine how careful Thomas Young was to look at the interference fringes, which must be colored if the primary source is the sun.

4.5 Image Formation with Diffraction

According to geometrical optics, the image of a point formed by an optical system free of optical aberrations is also a point. Suppose the point object is located on the optical axis. The spherical wavefront that diverges from the object when passing through the optical system will be truncated by the aperture diaphragm; i.e., the diaphragm plays the role of the aperture that diffracts the light. This implies that the image cannot be a point. On the other hand, the image of a large object will depend on the spatial coherence of the optical field in the object. This section briefly deals with the topic of imaging by taking diffraction into account in the paraxial approximation (Fresnel/Fraunhofer diffraction).*

4.5.1 Image of a point (source) object

Let us consider the system shown in Fig. 4.27. The thin lens represents the imaging optics, and the edge of the lens is the aperture diaphragm. The lens introduces a phase delay in the wavefront as it passes through the diaphragm. With this in mind, the lens can be modeled as a complex variable transmittance that changes the phase of the incident wavefront at the diaphragm. Thus, the process of image formation of a point object can be described as follows: a spherical wavefront that diverges from the point object is truncated by the aperture diaphragm and undergoes a phase shift due to the transmittance of the lens, then converges as a diffraction pattern in the Gaussian image plane.

In Fig. 4.27, $-s_o$ and s_i are the object and image distances from the thin lens in the plane of the aperture (diaphragm). The phase of the optical field (diverging from the point object) just before the aperture would be

*For a detailed discussion of the problem of imaging, by taking diffraction into account, see Goodman [6].

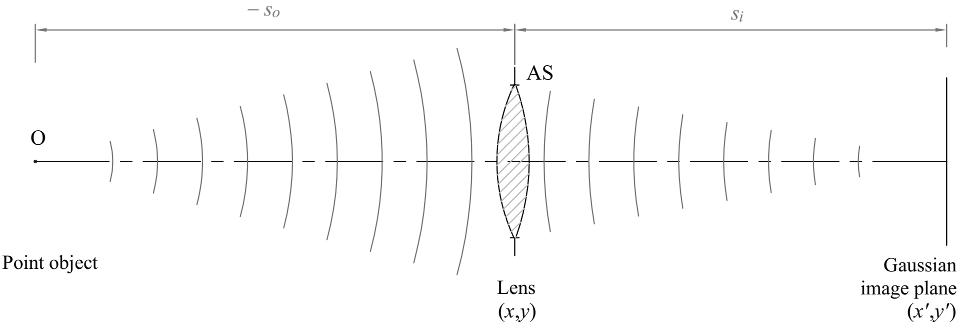


Figure 4.27 Schematic of an optical imaging system. The aperture diaphragm limits the wavefront and generates diffraction in the image.

$e^{-iks_o}e^{-ik(x^2+y^2)/2s_o}$. The transmittance of the lens at the aperture is given by $t(x, y) = e^{-ik(x^2+y^2)/2f}$. This result is easily deduced and can be found in *Introduction to Fourier Optics* [6]. Therefore, if E_0 is the amplitude of the field at the aperture, the optical field just after the aperture will be

$$E(x, y) = E_0 e^{-iks_o}\text{circ}\left(\frac{\sqrt{x^2+y^2}}{w}\right)e^{-ik(x^2+y^2)/2s_o}e^{-ik(x^2+y^2)/2f}; \qquad (4.66)$$

this describes the edge of the lens, and w is the radius of the lens.

The diffraction between the diaphragm (lens) and the Gaussian image plane is

$$I(x', y') = \frac{(\epsilon_0 c/2)}{\lambda^2 s_i^2}\left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[E(x,y)e^{ik(x^2+y^2)/2s_i}]e^{-i2\pi(x'x+y'y)/\lambda s_i}dxdy\right|^2. \qquad (4.67)$$

Inside the integral are the following phase terms:

$$e^{-ik(x^2+y^2)/2s_o}e^{ik(x^2+y^2)/2s_i}e^{-ik(x^2+y^2)/2f} = e^{ik[1/s_i-1/s_o-1/f](x^2+y^2)/2} = 1. \qquad (4.68)$$

This follows from the thin lens equation $1/s_i - 1/s_o = 1/f$ [Eq. (1.42)]. Thus, the image of a point object would be given by

$$I(x', y') = \frac{E_0^2(\epsilon_0 c/2)}{\lambda^2 s_i^2}\left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\text{circ}\left(\frac{\sqrt{x^2+y^2}}{w}\right)e^{-i2\pi(x'x+y'y)/\lambda s_i}dxdy\right|^2. \qquad (4.69)$$

This integral was solved in the circular aperture example (Section 4.3.3). Thus, the image of a point object formed by a lens of diameter $2w$ depends on the diameter of the lens and the distance of the Gaussian image s_i, and is of the form

$$I(r') = I_0 \left| \frac{2J_1(2\pi w r'/\lambda s_i)}{(2\pi w r'/\lambda s_i)} \right|^2, \tag{4.70}$$

where $I_0 = (\epsilon_0 c/2)(\pi w^2 E_0/\lambda s_i)^2$.

In geometrical optics, the geometrical PSF (Section 1.9) was defined to describe the shape of the image of a point object. If the optical system is free of aberrations, the geometrical PSF would be a point. Taking diffraction into account, the image of a point object is no longer a point but a diffraction pattern. Analogously, in physical optics (when the wave nature of light is taken into account), the diffractive PSF is defined to describe the shape of the image of a point object. If the optical system is free of aberrations, the diffractive PSF will be given by Eq. (4.70). When an optical system is free of optical aberrations, the system is said to be *diffraction-limited.*

Pupil function and optical aberrations

Equation (4.69) indicates that in the image formation process, which is in the Fresnel diffraction domain, the lens compensates for the quadratic Fresnel phase term, which results in a Fraunhofer diffraction integral. This equation can be generalized to multiple lens optical systems with an aperture diaphragm separated from the lenses. The diffraction aperture would be the edge of the exit pupil, and the distance at which the Fresnel diffraction occurs would be the distance between the exit pupil and the Gaussian image plane, say s_{ps}. The pupil is described by a function $P(x,y)$ that includes the geometry of the pupil and a possible variation of the transmittance in the pupil. If, in addition, the optical system presents optical aberrations, these affect the phase of the wavefront in the pupil, which can be included by multiplying the function $P(x,y)$ by a term $e^{-ikW(x,y)}$, where $W(x,y)$ is the variation of the real wavefront with respect to the ideal spherical wavefront of radius s_{ps}. Thus, the PSF of an optical system in general will be given by

$$I(x', y') = \frac{E_0^2(\epsilon_0 c/2)}{\lambda^2 s_{ps}^2} \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x,y) e^{-ikW(x,y)} e^{-i2\pi(x'x+y'y)/\lambda s_{ps}} dxdy \right|^2. \tag{4.71}$$

The function $W(x,y)$ is a polynomial whose terms describe the optical aberrations present in the optical system [9]; e.g., primary aberrations such as astigmatism, coma, and spherical aberrations in the wavefront are given by $a_a(x^2 - y^2)$, $a_c y(x^2 + y^2)$, and $a_s(x^2 + y^2)^2$, respectively. Defocus can also be included as an aberration, given by $a_d(x^2 + y^2)$. The coefficients a_d, a_a, a_c, and a_s depend on the parameters of the optical system. The PSF of a diffraction-limited optical system is shown Fig. 4.28, in which the distance between the exit pupil and the Gaussian image plane is s_{ps} = 100 mm and the exit pupil

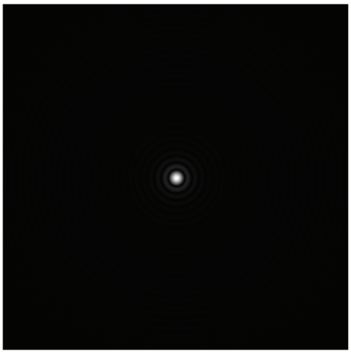
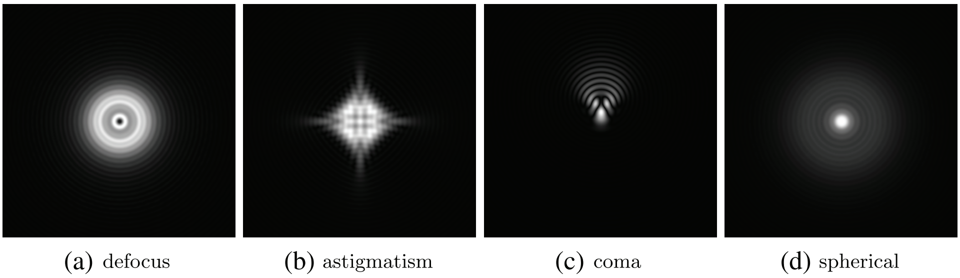


Figure 4.28 Diffraction pattern [function Jinc( )] without aberrations. The radius of the Airy disk is 7.7 μm.

diameter is $2w = 10$ mm, with $\lambda = 632.8$ nm. The radius of the Airy disk is $r_A = 7.7$ μm. The size of the box in the image is 316 μm on a side.

Figure 4.29 shows the diffraction patterns in the same optical system as Fig. 4.28. The wavefront at the exit pupil is affected by aberrations such as: defocus, with $a_d = 5 \times 10^{-5}$ mm^{-1}; astigmatism, with $a_a = 3.5 \times 10^{-5}$ mm^{-1}; coma, with $a_c = 1.25 \times 10^{-5}$ mm^{-2}; and spherical aberration, with $a_s = 2 \times 10^{-6}$ mm^{-3}. The size of the image box in all cases is 316 μm on a side.

The first thing to note is that the presence of any of these aberrations increases the size of the PSF compared with the PSF without aberrations shown in Fig. 4.28. Note that the defocus aberration corresponds to a Fresnel diffraction pattern because in this case, the term $e^{-ikW(x,y)}$ in the integral of Eq. (4.71) is equal to $e^{-ika_d(x^2+y^2)}$, which has the form of the quadratic Fresnel factor. The number of Fresnel zones introduced by the defocus in the pupil will be $N = 2a_d w^2/\lambda$. In the example given here, this is $N = 3.95$. The defocus coefficient is given by $a_d = -\Delta s/(2s_{ps}^2)$, where Δs is the defocus. In our example, $\Delta s = -1$ mm, and the negative sign means that defocus occurs when the image plane moves 1 mm closer to the exit pupil.



(a) defocus (b) astigmatism (c) coma (d) spherical

Figure 4.29 Diffraction patterns corresponding to the primary aberrations when the diameter of the exit pupil is 10 mm and the distance between the exit pupil and the Gaussian image plane is 100 mm, with $\lambda = 632.8$ nm. The size of the image box in all cases is 0.316 mm on a side.

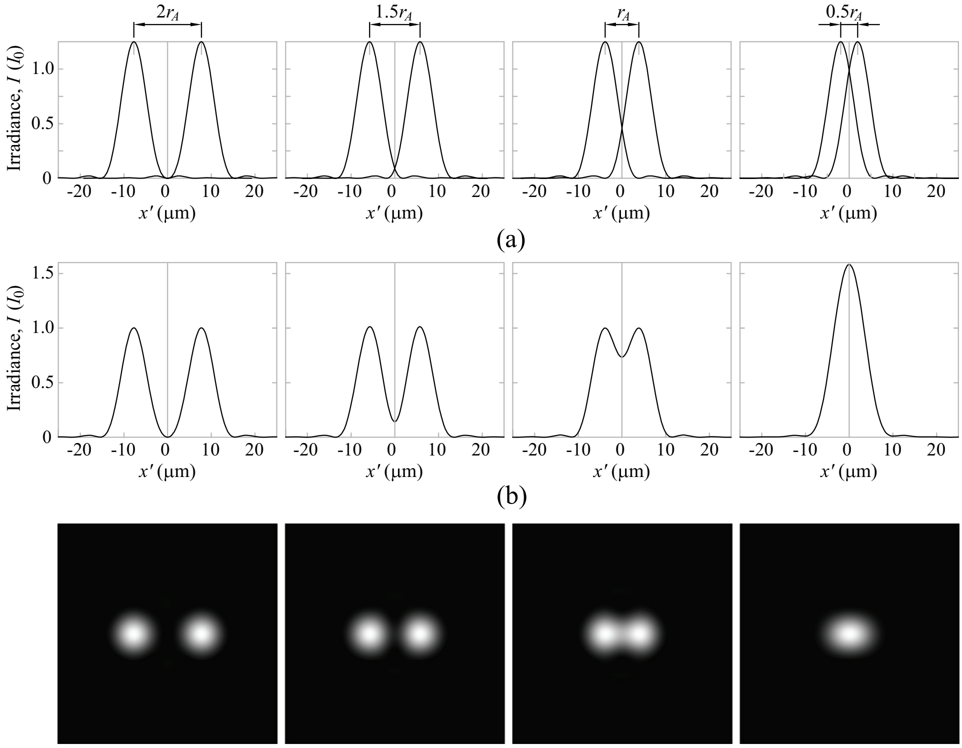
In practice, the PSF would be the result of a combination of different aberrations; e.g., the experimental PSF in Fig. 4.1 corresponds to an optical system (human eye model) affected by defocus (myopia), astigmatism, and spherical aberration.

4.5.2 Resolution in the image (two points)

From the point of view of geometrical optics, if an object is formed by two points separated by a certain distance, the image should have two points separated by a distance that depends on the magnification of the system. Regardless of the distance that separates the two points of the object, two points should be observed in the image. But as mentioned before, the image of a point object is an irradiance distribution called PSF. Therefore, the image of two spatially incoherent points would be two diffraction patterns that can overlap (in intensity) depending on the distance that separates them. This means that there may be a situation where the two diffraction patterns overlap, resulting in an irradiance distribution where the individual patterns cannot be distinguished. What is the minimum distance between these two diffraction patterns at which they can still be identified? This distance is called the *limit of spatial resolution*, and it depends on the diameter of the exit pupil of the optical system and the distance of the image.

If the imaging system is diffraction-limited, the size of the image of a point object is taken to be equal to that of the Airy disk. The incoherent superposition of the images of two identical point objects is shown in Fig. 4.30 when they are separated by $2r_A$, $1.5r_A$, r_A, and $0.5r_A$, where r_A is the radius of the Airy disk. The distance separating the images is measured from the center of each of the diffraction patterns, which corresponds to the distance of the point images according to geometrical optics. The calculations were made taking into account the same parameters as those shown in Fig. 4.28; i.e., when the distance between the exit pupil and the Gaussian image plane is $s_{ps} = 100$ mm and the exit pupil diameter is $2w = 10$ mm, with $\lambda = 632.8$ nm. The profiles of each of the diffraction patterns are shown in Fig. 4.30(a), the result of the incoherent superposition of the two diffraction patterns is shown in Fig. 4.30(b), and the diffractive images of the superposition are shown in Fig. 4.30(c). When the separation is $2r_A$ or $1.5r_A$, the images of the two points can be clearly identified. When the separation is r_A, part of the patterns overlap and the two diffractive images can still be resolved, but when they get closer, at the distance $0.5r_A$, it is no longer possible to identify the two point objects. Although there may be distances between r_A and $0.5r_A$ for which the images can still be resolved, the separation r_A is usually established as a resolution criterion. This is the *Rayleigh criterion* [1]:

> Images of two incoherent point sources are resolved when the center of the Airy pattern of one of the images falls on the first minimum of the Airy pattern of the other image.



(a)

(b)

(c)

Figure 4.30 Incoherent superposition of the images of two point sources when the images are separated by $2r_A$, $1.5r_A$, r_A, and $0.5r_A$. (a) Profiles of the irradiance of each image, (b) profiles of the image resulting from the superposition of the two images, and (c) diffraction patterns of the image of two point sources.

Given that $r_A = 1.22\lambda d/(2w)$, the size of the aperture diaphragm of the optical system is of great relevance for a fixed distance; the larger the diameter, the better the resolution. That is why it is desirable in astronomy to have large primary mirrors in telescopes. On the other hand, in other cases, such as lithography, it is possible to modify (decrease) the wavelength of the light coming from the object to increase the resolution. The same thing happens in electron microscopy, where the wavelength is of the order of 1 Å. Of course, the quality of the diffraction pattern depends on the optical aberrations, which increase with the diameter of the aperture stop. Therefore, increasing the diameter of the diaphragm does not necessarily improve resolution.

If the two point sources are coherent with each other, the result of the superposition of the images depends on the initial phase of the light in each of the sources; e.g., if the phase difference between the two sources is π, then the distance between the images can be reduced below r_A. But if the phase difference is 0, the separation of the images must be increased above r_A in order to resolve them [6].

Visual acuity

In visual optics (optometry and ophthalmology), instead of using the concept of resolution as explained above, the term *visual acuity* is used. Although in practice these two concepts are equivalent, a standard for the human eye has been established that determines the conditions in which a person is said to have good visual acuity. If at a distance of 20 feet (~6 m) a person can resolve two separate lines 1′ of arc, that person is said to have 20/20 visual acuity (emmetropic eye).

Following the Rayleigh criterion, the angular resolution limit is given by

$$(\Delta\theta)_{\min} = \frac{r_A}{s_{ps}} = 1.22\frac{\lambda}{2w}. \tag{4.72}$$

For the emmetropic eye, the resolution limit will be $(\Delta\theta)_{\min\_ojo} = 1'$. Taking 540 nm as the value of the wavelength in the center of the visible spectrum, the diameter of the pupil of the eye for which the value of the resolution limit is obtained will be $2w = 2.3$ mm. If we consider the Gullstrand–Emsley schematic eye (Fig. 1.68), where $f' = 22.05$ ($\sim s_{ps}$), the size of a point source in the retina would be a circular spot of 12.8 μm in diameter.

4.5.3 Image of an extended object

An object can be considered as an infinite set of points. From the discussion in the previous section, we already know how the image of a point object is formed; this is the PSF given by Eq. (4.71). The generalization to a set of points is not immediate but rather depends on the degree of coherence of the illumination of the object. For example, for the incoherent case, the image is given by

$$I_i(x') = I_o(x') * PSF, \tag{4.73}$$

where $I_o(x')$ represents the Gaussian image, *PSF* is the diffractive point spread function [Eq. (4.71)], and the symbol $*$ is the operation of convolution. This means that every point in the Gaussian image is affected by the PSF of the system in the same way.[*]

For coherent lighting, the situation is more complex because the convolution of the Gaussian image must be performed with a function that depends on the optical system (pupil geometry and optical aberrations) and the characteristics of the object (spatial frequency content). In other words, if the object is changed (keeping the same optical system), the function changes [10,11]. A similar situation occurs with partially coherent light. A generalization of the partially coherent light imaging process can be seen in Mejía and Suárez [12].

[*] This is valid only in the paraxial region. It should be noted that for larger fields of view the PSF varies with angle; e.g., the coma aberration increases as the chief ray increases its inclination.

4.6 Diffraction Gratings

A diffraction grating consists of a large number of identical elements that diffract light. These elements can be apertures in an opaque screen, steps or grooves in a substrate, or even an interference pattern of straight parallel fringes etched in amplitude or phase into a photosensitive material. The position of the irradiance maxima produced by diffraction gratings is a function of wavelength; thus, diffraction gratings find great application in the spectral measurement of the wavelength of light.

In this section, a basic configuration of diffraction gratings is analyzed, consisting of an array of N rectangular openings of width b and length c separated from each other by a distance $a\,(> b)$, as shown in Fig. 4.31. Assuming that $b \ll c$, it is sufficient to analyze the diffraction pattern in one dimension along the aperture distribution (axis x in Fig. 4.31). The distance at which the diffraction pattern produced by diffraction gratings is usually observed is such that $d \gg Na$; thus, the observed diffraction pattern corresponds to Fraunhofer diffraction.

In the one-dimensional case, the profile of each of the apertures can be described by the function rect( ), which is introduced in Section 4.3.3. On the other hand, the interference of N point sources [Fig. 3.47(b)] is described in Section 3.6. The interference of N sources with rectangular geometry is equivalent to the diffraction of the rectangular apertures array. By illuminating the set of apertures with a (presumable) plane wave of amplitude E_0 in the orthogonal direction, the optical field at the set of apertures can be written simply as

$$E(x) = E_0 \mathrm{rect}\left(\frac{x}{b}\right) * \sum_{j=1}^{N} \delta(x - ja), \tag{4.74}$$

where $\delta(x - ja)$ determines the position of the jth aperture; with the convolution operation, the rectangular shape of the aperture is copied at each point located at $x = ja$. The Fraunhofer diffraction would be

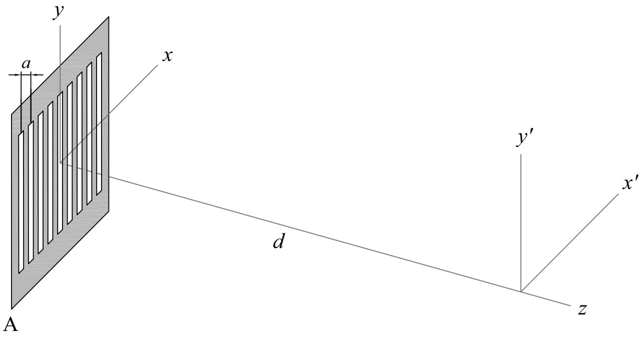


Figure 4.31 Array of *N* identical rectangular apertures.

$$I(x') = \frac{E_0^2(\epsilon_0 c/2)}{\lambda^2 d^2}\left|\int_{-\infty}^{\infty}\left[\text{rect}\left(\frac{x}{b}\right) * \sum_{j=1}^{N}\delta(x - ja)\right]e^{-i2\pi x' x/\lambda d}dx\right|^2, \tag{4.75}$$

which turns out to be

$$I(x') = I_0\left|\frac{\sin(\pi x' b/\lambda d)}{(\pi x' b/\lambda d)}\right|^2\left|\sum_{j=1}^{N} e^{-i2\pi x'(ja)/\lambda d}\right|^2. \tag{4.76}$$

I_0 absorbes $(\epsilon_0 c/2)(E_0/\lambda d)^2$ and the other constants that come out of the integral. The sum of the second factor from the right of the equality is solved in the same way as in Eq. (3.101). The end result is

$$I(x') = I_0\left|\frac{\sin(\pi x' b/\lambda d)}{(\pi x' b/\lambda d)}\right|^2\left|\frac{\sin(N\pi x' a/\lambda d)}{\sin(\pi x' a/\lambda d)}\right|^2. \tag{4.77}$$

This integral differs from the integral given in Eq. (3.104) in the term corresponding to the diffraction of a single aperture. Therefore, the diffraction pattern produced by an array of N identical apertures is equal to the interference pattern of N point sources (located in the center of the apertures) modulated by the diffraction pattern of one of the openings.

The diffraction pattern profile for an array of eight identical rectangular apertures separated by $a = 4b$ is shown in Fig. 4.32. The array is located

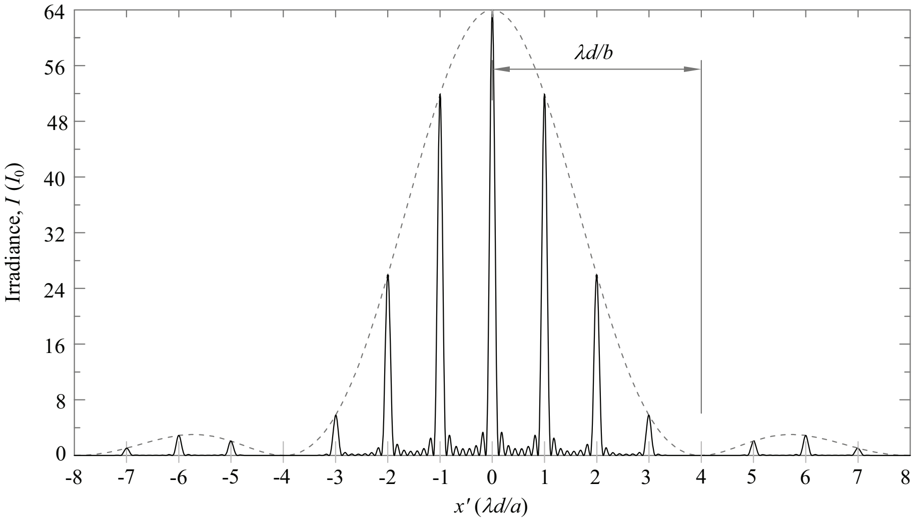


Figure 4.32 Diffraction pattern produced by an array of eight identical rectangular apertures spaced $a = 4b$, with b being the width of each aperture. The segmented curve corresponds to the diffraction pattern of a single aperture.

symmetrically with respect to the optical axis (axis z in Fig. 4.31). Figure 4.32 includes the diffraction profile (segmented curve) of an aperture as well as the modulation this produces in the interference pattern generated by eight point sources located in the center of the apertures. The unit of the horizontal scale is given as $\lambda d/a$, i.e., the separation between the principal maxima of the pattern without interference. On the other hand, at distances $\pm m'\lambda d/b$ ($m' = 1,2,3,\ldots$) from the center at 0 (optical axis) are the zeros of the diffraction pattern. Because $a = 4b$, the principal maximum of the interference pattern located at $4(\lambda d/a)$ falls right on the first zero of the diffraction pattern located at $\lambda d/b$, and therefore the maximum of the interference is not observed there. In the central lobe there will be a total of $2(a/b) - 1$ principal maxima, and in the side lobes there will be a total of $(a/b) - 1$ principal maxima. Principal maxima are at $\pm m\lambda d/a$ ($m = 0,1,2,3,\ldots$), with m used to label the maxima or the *diffraction order*. Thus, the central maximum will be the zero order of diffraction, the two maxima next to the central one will be diffraction orders $+1$ and -1, and so on.

As N increases, the energy in the secondary maxima decreases (approaching zero) and the principal maxima are in the form of very sharp peaks. The peak of order m subtends the angle θ_m with respect to the grating center (on the optic axis) and is given by

$$\tan\theta_m = m\frac{(\lambda d/a)}{d}. \tag{4.78}$$

In the Fraunhofer approximation, the function tan( ) can be changed to the function sin(); thus, the equation of the diffraction grating that angularly locates a diffraction order is

$$m\lambda = a\sin\theta_m. \tag{4.79}$$

When the lighting source is polychromatic, at zeroth order there will be a maximum of the same color as the source, but for $m \neq 0$, there will be spatially separated maxima of different colors, e.g., if the source is white light, a continuous spectrum will be seen for a given $m\,(\neq 0)$.

If the spectrum of the source consists of two wavelengths, what is the smallest difference in wavelength that can be resolved in the diffraction pattern? This problem can be treated similarly to the way spatial resolution of two-point images is handled in the previous section. Equation (3.112) allows the separation of the minima between the principal maxima of the interference pattern of N sources to be determined. The separation between two consecutive minima will be given by

$$\Delta x' = \frac{\lambda d}{Na}, \tag{4.80}$$

which in turn is half the width of the principal maxima. By applying the Rayleigh criterion, $\Delta x'$ would be the minimum separation in x' between the diffraction maxima for two wavelengths λ and $\lambda + \Delta\lambda$ for the diffraction order m. On the other hand, the angular separation between the two maxima can be calculated as follows: let x' be the position of the mth order, and then $x'/d = \tan\theta_m \approx \sin\theta_m$; thus,

$$\frac{\Delta x'}{d} = \Delta\theta_m \cos\theta_m. \tag{4.81}$$

By inserting Eq. 4.80 in Eq. 4.81, the angular separation is equal to

$$\Delta\theta_m = \frac{\lambda}{Na\cos\theta_m}. \tag{4.82}$$

On the other hand, the angular separation can also be evaluated from Eq. (4.79). Differentiating,

$$m\Delta\lambda = a\Delta\theta_m \cos\theta_m, \tag{4.83}$$

i.e.,

$$\Delta\theta_m = \frac{m\Delta\lambda}{a\cos\theta_m}. \tag{4.84}$$

The resolving power of the diffraction grating[*] is defined as

$$\mathfrak{R} = \frac{\lambda}{(\Delta\lambda)_{\min}}, \tag{4.85}$$

where $(\Delta\lambda)_{\min}$ is the difference in wavelength that can be resolved around the wavelength λ. By equating Eqs. (4.82) and (4.84), the resolving power of the diffraction grating can also be written as

$$\mathfrak{R} = Nm; \tag{4.86}$$

i.e., the resolving power increases with N and the diffraction order m.

In a diffraction grating, the density of diffraction elements (apertures) is usually in the hundreds or thousands per millimeter. This density is the parameter that is usually used to characterize the grating, and its value is given in lines per millimeter. A line is equivalent to a diffracting element. For example, in an optical store catalog, a 12.5 mm wide diffraction grating has 500 lines/mm. From this information, it can be concluded that $N = 6250$. If the grating is completely illuminated with a plane wave, the resolving power of the order $m = 1$ would be $\mathfrak{R} = 6250$. Thus, for a light signal around $\lambda = 540$ nm, two wavelengths can be separated whose difference is $(\Delta\lambda)_{\min} = 0.086$ nm.

[*] The resolving power on a diffraction grating is a measure of the ability to spatially separate two wavelengths.

If the analysis is done for the second order of diffraction, the wavelength difference would be $(\Delta\lambda)_{\min} = 0.043$ nm.

In practice, gratings are designed so that the amount of light in the diffraction orders can be controlled. In the case of the orthogonally illuminated planar grating, most of the light is in the zero order, where the spectrum cannot be resolved. Thus, it would be convenient to have most of the light in the first or second orders. This can be done, e.g., by etching steep steps into the surface of a mirror. If the mirror is also concave, it is possible to focus the different orders of diffraction, which is common in spectrometers. For an extension to this topic, *Diffraction Gratings and Application* [13] and *Diffraction Grating Handbook* [14] can be consulted.

References

[1] E. Hecht, *Optics*, Global Edition, 5th ed., Pearson, Harlow, England (2017).

[2] A. Fresnel, "The diffraction of light," Chap. 8 in *Great Experiments in Physics: Firsthand Accounts from Galileo to Einstein*, 2nd ed., M. H. Shamos, Ed., Dover Publications, Mineola, New York, 108–120 (1987).

[3] J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley, New York (1999).

[4] M. Born and E. Wolf, *Principles of Optics*, 6th ed., Pergamon Press, Oxford, England (1993).

[5] A. Sommerfeld, *Lectures on Theoretical Physics*, Vol. 4: *Optics*, Academic Press, New York (1964).

[6] J. W. Goodman, *Introduction to Fourier Optics*, Roberts and Company Publishers, Englewood, Colorado (2005).

[7] T. Young, "The Interference of Light," Chap. 7 in *Great Experiments in Physics: Firsthand Accounts from Galileo to Einstein*, 2nd ed., M. H. Shamos, Ed., Dover Publications, Mineola, New York, 93–107 (1987).

[8] Y. Mejía and A. I. González, "Measuring spatial coherence by using a mask with multiple apertures," *Opt. Commun.* **273**(2), 428–434 (2007).

[9] Y. Mejía, "El frente de onda y su representación con polinomios de Zernike," *Cien. Tecnol. Salud Vis. Ocul.* **9**(2), 145–166 (2011).

[10] H. H. Hopkins, "On the diffraction theory of optical images," *Proc. R Soc. Lond. A Math. Phys. Sci.* **217**(1130), 408–432 (1953).

[11] B. J. Thompson, "Image formation with partially coherent light," Chap. 4 in *Progress in Optics*, Vol. 7, E. Wolf, Ed., Elsevier, Amsterdam, 169–230 (1969).

[12] Y. Mejía and D. Suárez, "Optical transfer function with partially coherent monochromatic illumination," *Optik* **193**(163021), 1–4 (2019).

[13] E. G. Loewen and E. Popov, *Diffraction Gratings and Applications*, CRC Press, Boca Raton, Florida (2018).

[14] C. A. Palmer and E. G. Loewen, *Diffraction Grating Handbook*, 6th ed., Newport Corporation, Rochester, New York (2005).