# 2 Reflection, refraction, diffraction, and scattering

In this chapter we will describe how radiation is reflected off a surface, transmitted into a medium at a different angle, how it is influenced by passing an edge or through an aperture, and how it is scattered and spread out by interactions with particles at various scales.

We will focus on the propagation of light. In order to describe the direction in which light propagates, it is often convenient to represent a light wave by rays rather than by wave fronts. In fact, rays were used to describe the propagation of light long before its wave nature was established. Geometrically, the duality is easy to handle: Whenever we indicate the direction of propagation of light by a ray, the wave front is always locally perpendicular to the ray.

## *2.1 Reflection*

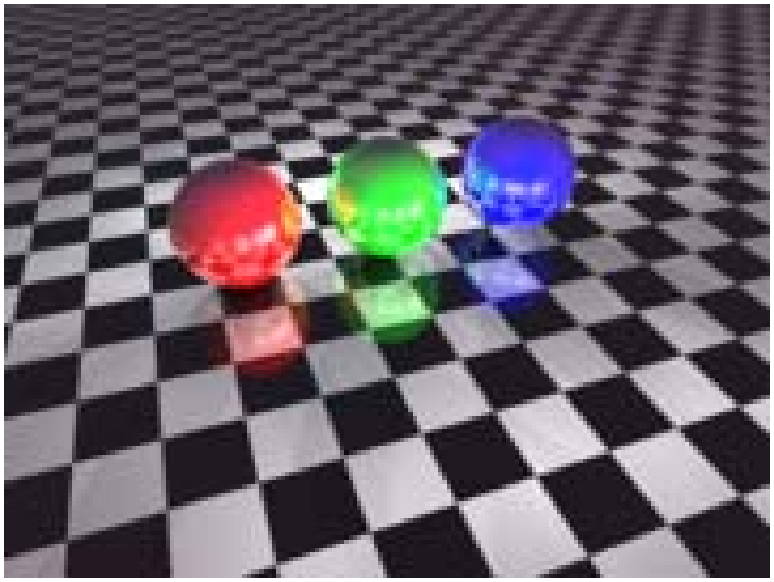


*Figure 2-1. Spheres reflected in the floor and in each other.*

Reflection occurs when a wave hits the interface between two dissimilar media, so that all of or at least part of the wave front returns into the medium from which it originated. Common examples are reflection of light, as shown in figure 2-1, as well as reflection of surface waves that may be observed in a pool of water, or sound waves reflected as echo from a wall.

Reflection of light may be specular or diffuse. The first occurs on a blank mirroring surface that retains the geometry of the beams of light. The second occurs on a rougher surface, not retaining the imaging geometry, only the energy.

### 2.1.1 Reflectance

R**eflectance** is the ratio of reflected power to incident power, generally expressed in decibels or percent. Most real objects have some mixture of diffuse and specular qualities, and surface reflectance is therefore often divided into diffuse and specular reflectance. In climatology, reflectivity is called albedo.

### 2.1.2 Diffuse reflection

When light strikes a rough or granular surface, it bounces off in all directions due to the microscopic irregularities of the interface, as illustrated in figure 2-2. Thus, an image is not formed. This is called *diffuse reflection*. The exact form of the reflection depends on the structure of the surface.

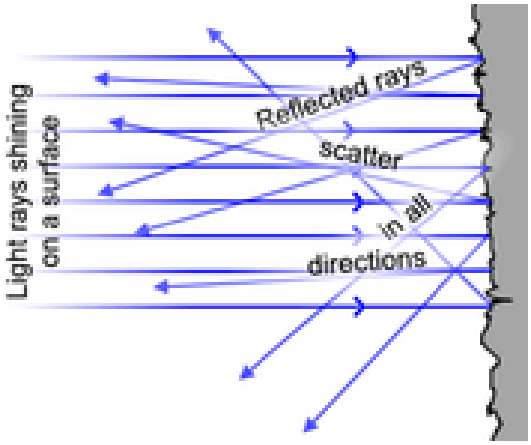



*Figure 2-2. Diffuse reflection from a rough surface.*

A common model for diffuse reflection is Lambertian reflectance, in which the light is reflected in accordance with Lambert's cosine law, see figure 2-3.

**Lambert's cosine law** says that the total radiant power observed from a "Lambertian" surface is directly proportional to the cosine of the angle θ between the observer's line of sight and the surface normal. The law is also known as the cosine emission law or Lambert's emission law (Johann Heinrich Lambert, *Photometria*, 1760).

When an area element is radiating as a result of being illuminated by an external source, the irradiance (energy/time/area) landing on that area element will be proportional to the cosine of the angle between the illuminating source and the normal. A Lambertian reflector will then reflect this light according to the same cosine law as a Lambertian emitter.

This means that although the radiance of the surface under diffuse reflection depends on the angle from the normal to the illuminating source, it will not depend on the angle from the normal to the observer.

An important consequence of Lambert's cosine law is that when an area element is viewed from any angle, it has the same radiance. Although the emitted power from an area element is reduced by the cosine of the emission angle, the observed size of the area element is also reduced by that same amount, so that while the area element appears smaller, its radiance is the same.

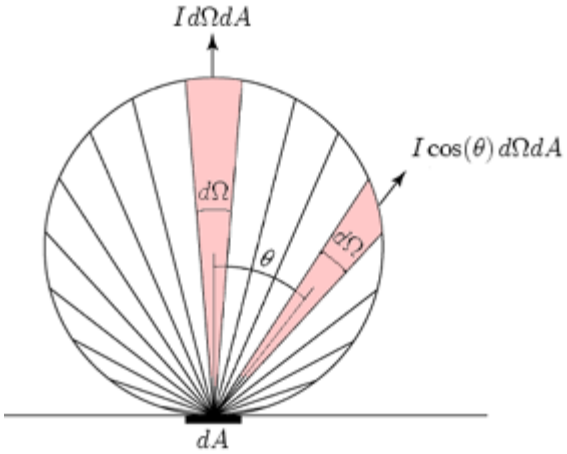



*Figure 2-3. Emission rate (photons/s) in a normal and off-normal direction. The number of photons/sec directed into any wedge is proportional to the area of the wedge.*

### 2.1.3 Specular (mirror-like) reflection

We describe the directions of the incident ray hitting an optical surface and the ray being reflected from the surface in terms of the angle they make with the normal (perpendicular) to the surface at the point of incidence, as shown in figure 2-4. In specular reflection, the angle of incidence $\theta_i$ equals the angle of reflection $\theta_r$.

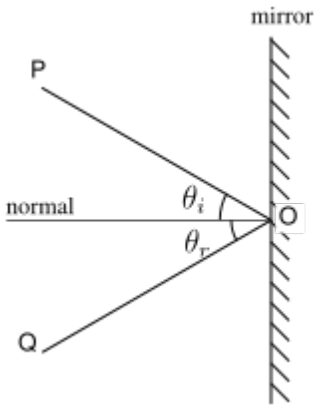



*Figure 2-4. Reflection from a mirror.*

A mirror is the most common example of specular light reflection. It usually consists of a glass sheet in front of a metallic coating where the reflection actually occurs. Reflection also occurs from the surface of transparent media, such as water or glass.

In fact, reflection of light may occur whenever light travels from a medium of a given refractive index into a medium with a different refractive index. In the most general case, a certain fraction of the light is reflected from the interface, and the remainder is refracted as it passes into the transparent medium.

When light reflects off a material denser than the external medium, it undergoes a 180° phase reversal. In contrast, a less dense, lower refractive index material will reflect light in phase. This is an important principle in the field of thin-film optics.

Specular reflection at a curved surface forms an image which may be magnified or demagnified. We will shortly return to the subject of curved mirrors and geometric optics.

### 2.1.4 Perfect mirrors

A **perfect mirror** is a theoretical mirror that reflects light (and electromagnetic radiation in general) perfectly, and doesn't transmit it. Domestic mirrors are not perfect mirrors as they absorb a significant portion of the light which falls on them.
Dielectric mirrors are glass or other substrates on which one or more layers of dielectric material are deposited, to form an optical coating. A very complex dielectric mirror can reflect up to 99.999% of the light incident upon it, for a narrow range of wavelengths and angles. A simpler mirror may reflect 99.9% of the light, but may cover a broader range of wavelengths.

### 2.1.5 Image formation by a plane mirror

As illustrated by figure 2-5, a plane mirror will form a virtual image point located exactly opposite the real image point, and as far behind the mirror as the object point is from the front of the mirror.

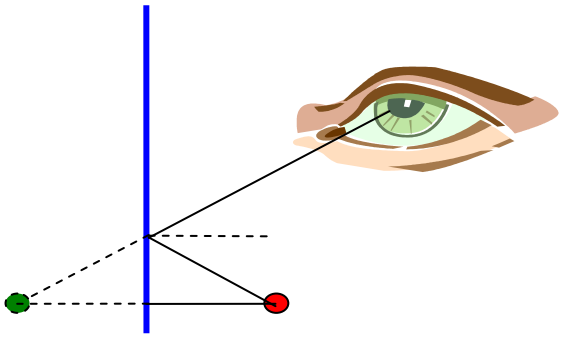


*Figure 2-5. The location of the virtual image formed by a plane mirror.*

Evidently, a “full size mirror” for your wardrobe doesn’t have to more than half the height of your body, no matter where on your body your eyes are placed, and no matter how far away from the mirror you are standing. The image of an extended object seen in a plane mirror is exactly the same size as the object, but it is perceived to be located at a negative image distance behind the mirror, equal to the positive object distance from the mirror. And since the object and image sizes are the same, the lateral magnification is unity.

### 2.1.6 Retro reflection

Retro reflectors are used both in a purely specular and a diffused mode. A simple retro reflector can be made by placing three ordinary mirrors mutually perpendicular to one another (a corner reflector). The image produced is the inverse of one produced by a single mirror. The principle is illustrated for two mirrors in figure 2-6. Applications also include simple radar reflectors for ships, and the Lunar Laser Range program based on four retro reflectors placed at the Apollo 11, 14, and 15 sites and the Lunakhod 2 rover. In such applications, it is imperative that the corner angle is exactly 90°. A deviation by an angle δ will cause the returned beam to deviate by 2δ from the incoming beam.

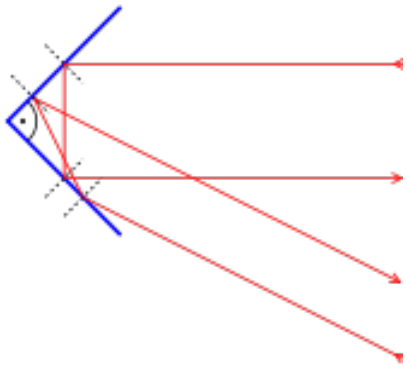


*Figure 2-6. The principle of a retro reflector.*

Surfaces may also be *retroreflective*. The structure of these surfaces is such that light is returned in the direction from which it came.

A surface can be made *partially retroreflective* by depositing a layer of tiny refractive spheres on it or by creating small pyramid like structures (cube corner reflection). In both cases internal reflection causes the light to be reflected back to where it originated. This is used to make traffic signs and automobile license plates reflect light mostly back in the direction from which it came. In this application a completely perfect retro reflection is not desired, since the light would then be directed back into the headlights of an oncoming car rather than to the driver's eyes.

The description of retro reflectors may illustrate an important property of all images formed by reflecting or refracting surfaces: An image formed by one surface can serve as the object for a second imaging surface. In the case of two orthogonal mirrors as shown in figure 2-7, mirror $M_1$ forms a virtual image $P_1$' of an object point P, and mirror $M_2$ forms another virtual image $P_2$'. Image $P_1$' serves as an object for mirror $M_2$, forming a virtual image $P_3$'. Likewise, image $P_2$' serves as an object for mirror $M_1$, forming the same virtual image $P_3$'.

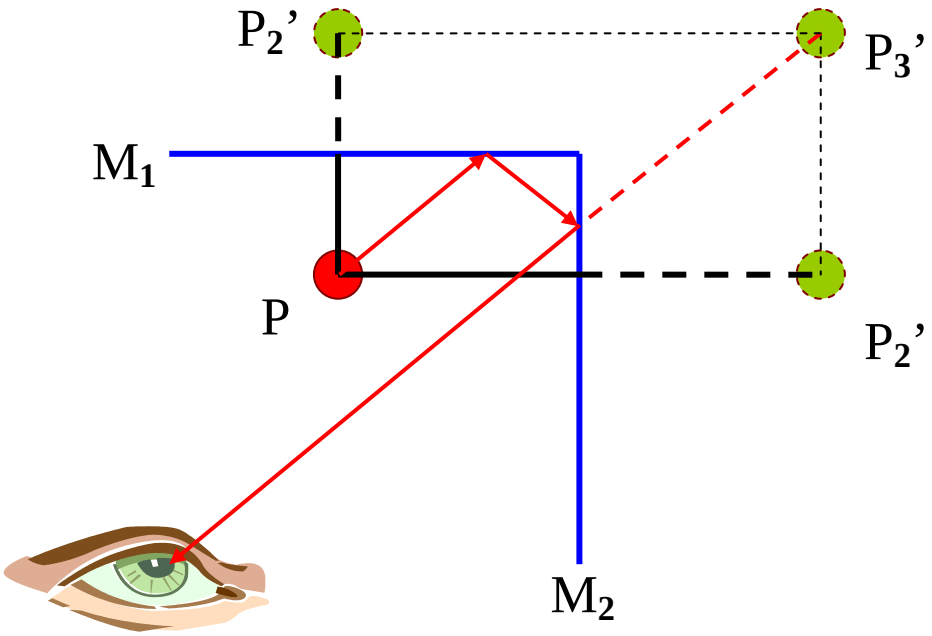



*Figure 2-7. Virtual images $P_1$' and $P_2$' are formed by a single reflection of the object P in mirrors $M_1$ and $M_2$, respectively. The virtual image $P_3$' is formed by a reflection of the virtual image $P_1$' in mirror $M_2$ or image $P_2$' in mirror $M_1$.*

This principle will be used extensively when locating the final image formed by several successive reflecting surfaces in a mirror-based telescope, or several refracting surfaces in a microscope or a refracting telescope.

### 2.1.7 Sign rules for image formation

In the case of an image formed by a few plane mirrors, the geometry is quite simple, and we do not have to worry about the sign of object and image distances. However, we want to state some general sign rules that will be applicable to all imaging situations that we will encounter later, when both real and virtual images are formed in front of or behind curved surfaces:

- **The object distance**: When the object is on the same side of the reflecting or refracting surface as the incoming light, the object distance s is positive; otherwise, it is negative.
- **The image distance**: When the image is on the same side of the reflecting or refracting surface as the outgoing light, the image distance s' is positive; otherwise, it is negative.
- **Radius of curvature**: When the centre of curvature C is on the same side as the outgoing light, the radius of curvature is positive; otherwise, it is negative.

## 2.1.8 Image formation by curved mirrors

### 2.1.8.1 Spherical mirrors

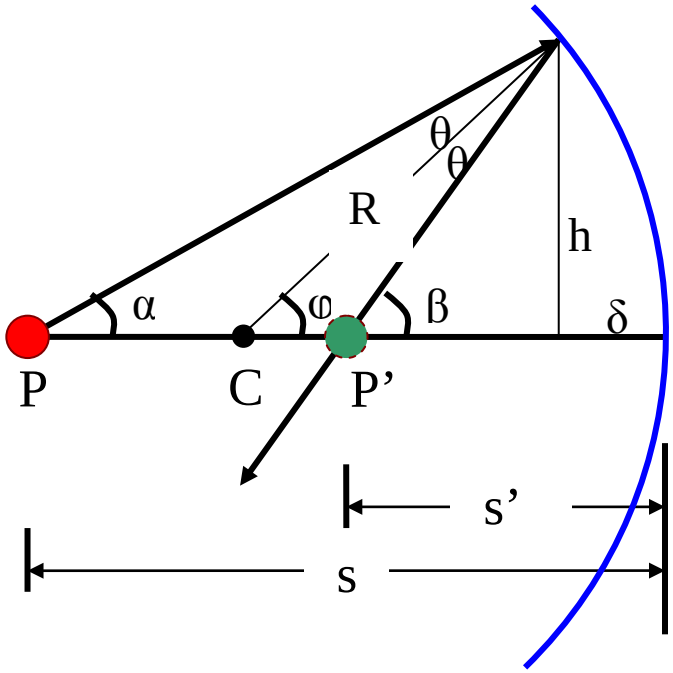



*Figure 2-8.Finding the image point formed by a concave mirror.*

Figure 2-8 shows a spherical mirror with radius of curvature R. The concave side is facing the incident light originating from an object point P on the optical axis. A ray from P at an angle α to the axis is reflected by the mirror. The angle of incidence and reflection are both θ, and the reflected ray crosses the optical axis at an angle β. In fact, all rays from P will intersect the axis at the same point P', provided that the angle α is small.

We have the following relations: $\varphi = \alpha + \theta$ and $\beta = \varphi + \theta$, which implies that $\alpha + \beta = 2\varphi$. The expressions for the tangents of α, β, and φ are simply $tg(\alpha) = h/(s-\delta)$, $tg(\beta) = h/(s'-\delta)$, $tg(\varphi) = h/(R-\delta)$. If the angle α is small, so are β and φ. If α is so small that the ray is almost parallel to the optical axis (the paraxial approximation), the tangent of an angle is equal to the angle itself (given in radians), and δ may be neglected compared to s, s', and R. So for small angles we have the following approximations: $\alpha = h/s$, $\beta = h/s'$, $\varphi = h/R$. Substituting this into $\varphi = \alpha + \theta$ we get the general object-image relation for a spherical mirror:

$$\frac{1}{s} + \frac{1}{s'} = \frac{2}{R}$$

If the radius of curvature becomes infinite ($R \to \infty$), the mirror becomes plane, and the relation above reduces to $s = -s'$ for a plane reflecting surface.

If the object is very far from the mirror ($s \to \infty$), the incoming rays are parallel, and the image will be formed at a distance $s' = R/2$ from the mirror. This is the focal length, f.

A simple sketch may illustrate how the size y' and position s' of the real image of an extended object is determined, once the size y of the object , the distance s from the spherical mirror to the object and the radius R of curvature of the mirror is known, see figure 2-9.

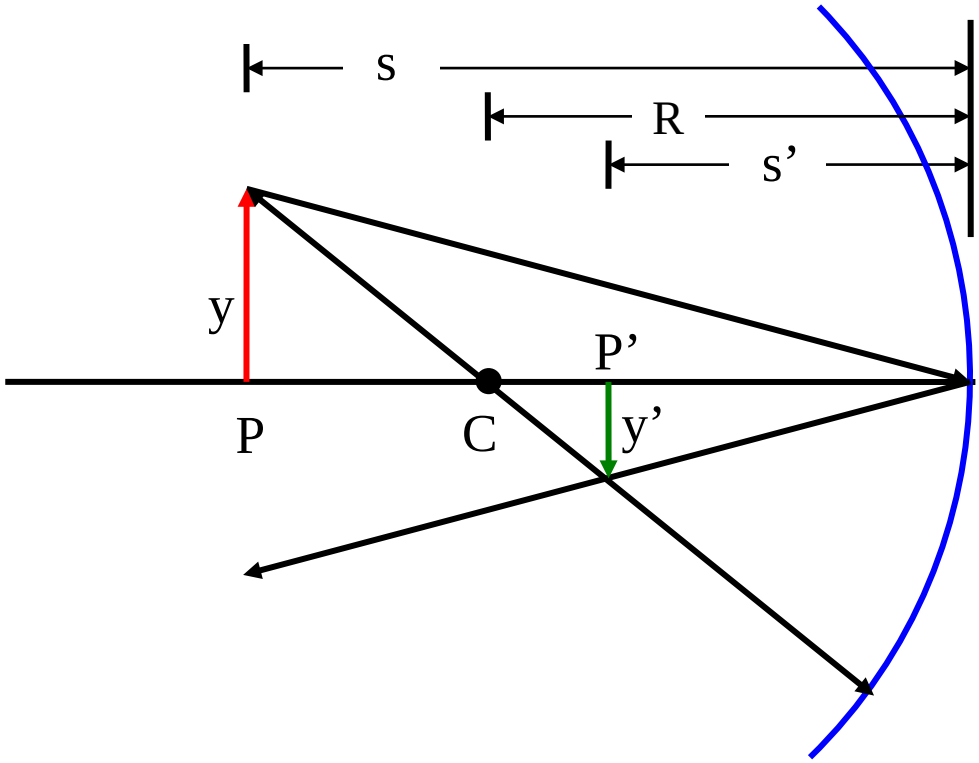



*Figure 2-9. Determining the position, orientation and size of an image.*

We see that y/s = - y'/s'. The negative sign indicates that the image is inverted relative to the object. The magnification is m= y'/y = - s'/s. The size of the image is |y'| = ys'/s. Substituting s' from the general object-image relation, we get a more useful expression: |y'| = yf /(s-f), as the object distance s and the focal length f are often readily available.

When the object is far from the mirror, the image of the object is smaller than the object, inverted, and real. If the object for all practical purposes is infinitely far away, the image is formed in the focal plane. As the object is moved closer to the mirror, the image moves farther from the mirror and increases in size. When the object is in the focal plane, the image is at infinity. If the object is inside the focal point, the image becomes larger than the object, erect, and virtual.

For a convex spherical mirror, R is negative. A virtual image is formed behind the mirror at a negative image distance. But using the sign rules, the same object-image relation is valid, and the expressions for the magnification and the size of the image are the same. However, if the incoming rays are parallel to the optical axis, the rays will not converge through a focal point. They will diverge as if they had come from a virtual focal point at negative focal length f = R/2 behind the mirror. In summary, the basic relationships for image formation by a spherical mirror are valid for both concave and convex mirrors, provided that the sign rules are used consistently.

## 2.1.8.2 Optical aberrations caused by curved mirrors

### *2.1.8.2.1 Spherical aberration*

A simple mirror with a spherical surface suffers from the optical imperfection called spherical aberration: Light striking nearer the periphery focuses closer to the mirror, while light striking near the center focuses farther away. While spherical mirrors are the most convenient to grind and polish, spherical aberration can be eliminated by grinding and polishing the surface to a paraboloid of revolution. Then parallel rays striking all parts of the mirror are reflected to the same focus.

### *2.1.8.2.2 Coma*

The principal disadvantage of a parabolic mirror is that it produces good images over only a relatively small field of view, that is, for light that strikes the mirror very nearly parallel to the optical axis. In a photograph of a field of stars, the star images near the center of the picture will appear as sharp, round dots, whereas those near the corners, which are formed by light coming in at an angle to the optical axis, are distorted into tiny "tear drops" or "commas" pointing towards the center of the photograph. The shape of these images accounts for the name "coma" given to this type of aberration that distorts images formed by light that strikes the mirror off-axis.

### *2.1.8.2.3 Astigmatism*

Astigmatism occurs when rays of light in different planes do not focus at the same distance from the mirror. Such aberrations may be caused by mechanical distortions of large mirrors.

### *2.1.8.2.4 Curvature of field*

This is an aberration in which the image is sharp, but different parts of it are formed at different distances from the mirror, so that the whole image cannot be captured by a flat detector.

### *2.1.8.2.5 Distortion*

This is an aberration in which the image may be sharp, but its shape is distorted, e.g. if straight lines in the object plane are imaged as curved lines. Distortion may vary within the field of view, being most noticeable near the edges of the field of view.

### *2.1.8.2.6 Vignetting*

This is an aberration that causes a darkening of the image towards the corners of the field of view.

### 2.1.8.3 Reflecting telescopes

A **telescope** (from the Greek *tele* = 'far' and *skopein* = 'to look or see'; *teleskopos* = 'far-seeing') is an instrument designed for the observation of remote objects. There are three main types of optical astronomical telescopes:

- The refracting (dioptric) telescope which uses an arrangement of lenses.
- The reflecting (catoptric) telescope which uses an arrangement of mirrors.
- The catadioptric telescope which uses a combination of mirrors and lenses.

The Italian monk Niccolo Zucchi is credited with making the first reflector in 1616, but he was unable to shape the concave mirror accurately and he lacked of means of viewing the image without blocking the mirror, so he gave up the idea. The first practical reflector design is due to James Gregory. In *Optica Promota* (1663) he described a reflector using two concave mirrors. A working example was not built until 10 years later by Robert Hooke. Sir Isaac Newton is often credited with constructing the first "practical" reflecting telescope after his own design in 1668. However, because of the difficulty of precisely producing the reflecting surface and maintaining a high and even reflectivity, the reflecting telescope did not become an important astronomical tool until a century later.

Telescopes increase the apparent angular size of distant objects, as well as their apparent brightness. If the telescope is used for direct viewing by the human eye, an eyepiece is used to view the image. And most, if not all eyepiece designs use an arrangement of lenses. However, in most professional applications the image is not viewed by the human eye, but is captured by photographic film or digital sensors, without the use of an eyepiece. In this configuration, telescopes are used as pure reflectors.

We will return to the effects of an eyepiece in the sections on refraction.

In a prime focus design in large observatory telescopes, the observer or image detector sits *inside* the telescope, at the focal point of the reflected light. In the past this would be the astronomer himself, but nowadays CCD cameras are used. Radio telescopes also often have a prime focus design.

### 2.1.8.4 Telescope mountings

The telescope tube must be mounted so that it can point to any direction in the sky. This is achieved by using two orthogonal axes of motion.

- The simplest is an altitude-azimuth mount, where one axis points towards the zenith, and the instrument is pointed at a given altitude angle above the horizon, measured along a vertical circle, and a given azimuthal angle, measured eastward from the north point. Following a celestial object over time requires repeated recalculation of altitude and azimuth if this mount is used.
- In the equatorial mount, one axis is parallel to the earth’s axis. This allows the telescope to be pointed directly in right ascension and declination, the coordinates in which astronomical positions are generally given. In addition, a simple clockwork about a single axis can compensate for the earth’s rotation when observing celestial objects.

### 2.1.8.5 Optical reflecting telescope designs

- The **Newtonian** usually has a paraboloid primary mirror but for small apertures, a spherical primary mirror is sufficient for high visual resolution. A flat diagonal secondary mirror mounted on a "spider" reflects the light to a focal plane at the side of the top of the telescope tube. Thus, the observer does not block the light entering the telescope tube. However, the observer may experience some awkward viewing positions.

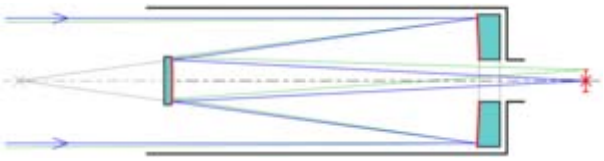


*Figure 2-10. A "classic" Cassegrain telescope design.*

- The "classic" **Cassegrain** has a parabolic primary mirror, and a convex hyperbolic secondary mirror that reflects the light back, either through a hole in the primary, or reflected out to the side of the tube by a third, flat mirror, mounted diagonally just in front of the primary (the Nasmyth-Cassegrain). Folding the optics makes this a compact design. On smaller telescopes, and camera lenses, the secondary is often mounted on a clear, optically flat, glass plate that closes the telescope tube. This eliminates the diffraction effects caused by the secondary supports in the Newtonian. It also protects the mirrors from dust.
- The **Ritchey-Chrétien** is a specialized Cassegrain reflector which has two hyperbolic mirrors (instead of a parabolic primary). It is free of coma and spherical aberration at a flat focal plane, making it well suited for wide field and photographic observations. Almost every professional reflector telescope in the world is of the Ritchey-Chrétien design.
- The **Dall-Kirkham** Cassegrain design uses a concave elliptical primary mirror and a convex spherical secondary. While this system is easier to grind than a classic Cassegrain or Ritchey-Chretien system, it does not correct for off-axis coma and field curvature so the image degrades quickly off-axis. Because this is less noticeable at longer focal ratios, Dall-Kirkhams are seldom faster than f/15.
- The **Schiefspiegler** uses tilted mirrors to avoid the secondary mirror casting a shadow on the primary. However, while eliminating diffraction patterns this leads to several other aberrations that must be corrected.
- The **Maksutov** telescope uses a full aperture corrector lens, a spherical primary, and a spherical secondary which is an integral part of the corrector lens.
- The **Gregorian** has a concave, not convex, secondary mirror and in this way achieves an upright image, useful for terrestrial observations.
- A **Schmidt** camera is an astronomical camera designed to provide wide fields of view with limited aberrations. It has a spherical primary mirror, and an aspherical correcting lens, located at the center of curvature of the primary mirror. The film or other detector is placed inside the camera, at the prime focus. The Schmidt camera is typically used as a survey instrument, for research programs in which a large amount of sky must be covered. There are several variations to the design:

- Schmidt-Väisälä: a doubly-convex lens slightly in front of the film holder.
- Baker-Schmidt: a convex secondary mirror, which reflected light back toward the primary. The photographic plate was then installed near the primary, facing the sky.
- Baker-Nunn: replaced the Baker-Schmidt camera's corrector plate with a small triplet corrector lens closer to the focus of the camera. A 20 inch Baker-Nunn camera installed near Oslo was used to track artificial satellites from the late 1950s.
- Mersenne-Schmidt :consists of a concave paraboloidal primary, a convex spherical secondary, and a concave spherical tertiary mirror.
- Schmidt-Newtonian: The addition of a flat secondary mirror at 45° to the optical axis of the Schmidt design creates a Schmidt-Newtonian telescope.
- Schmidt-Cassegrain: The addition of a convex secondary mirror to the Schmidt design directing light through a hole in the primary mirror creates a Schmidt-Cassegrain telescope. The last two designs are popular because they are compact and use simple spherical optics.

- The **Nasmyth** design is similar to the Cassegrain except no hole is drilled in the primary mirror; instead, a third mirror reflects the light to the side and out of the telescope tube, as shown in figure 2-11.
- Adding further optics to a Nasmyth style telescope that deliver the light (usually through the declination axis) to a fixed focus point that does not move as the telescope is reoriented gives you a **Coudé** focus. This design is often used on large observatory telescopes, as it allows the use of heavy observation equipment.

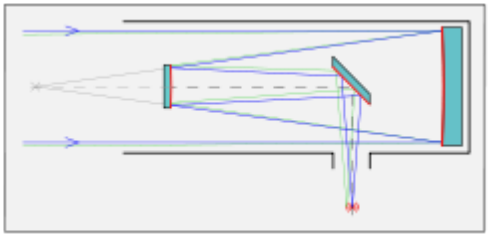


*Figure 2-11. A Nasmyth Cassegrain design.*

### 2.1.8.6 Non-optical telescopes

Single-dish radio telescopes are often made of a conductive wire mesh whose openings are smaller than the reflected wavelength. The shape is often parabolic, and the detector is either placed in prime focus, or a secondary reflector is used.

Multi-element radio telescopes are constructed from pairs or groups of antennae to synthesize apertures that are similar in size to the separation between the telescopes. Large baselines are achieved by utilizing space-based Very Long Baseline Interferometry (VLBI) telescopes such as the Japanese HALCA (Highly Advanced Laboratory for Communications and Astronomy) VSOP (VLBI Space Observatory Program) satellite. The VLBI technique is also using radio telescopes on different continents simultaneously.

## 2.2 Refraction

The reason why a glass lens may focus light is that light travels more slowly within the glass than in the medium surrounding it. The result is that the light rays are bent. This is described by the refraction index of the glass lens, which is equivalent to the ratio of the speed of light in air and glass.

We may illustrate this phenomenon by a simple example. Consider a ray of light passing through the air and entering a plane parallel slab of glass as shown in figure 2-12. Then Snell's law gives the relation between the angle of incidence ($\theta_1$), the angle of refraction ($\theta_2$), the refraction index of the glass slab ($n_2$) and the surrounding medium ($n_1$):

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 .$$

The angle of incidence and the refraction angle are the angles between the light beams and the normal to the surface at the point where the beam crosses the interface.

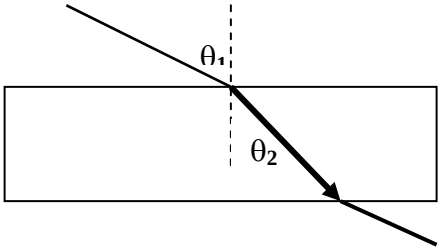



*Figure 2-12. A light beam is bent as it enters a slab of glass – refraction.*

### 2.2.1 A simple derivation of Snell's law

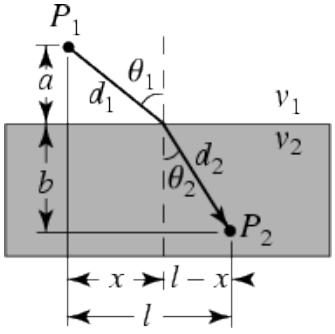
If we consider a beam of light passing at an angle from air into a plane-parallel slab of glass as illustrated in figure 2-12, then Snell's law gives a relationship between the angle of incidence, ($\theta_1$), the refraction angle ( $\theta_2$), the refraction index of the glass slab ($n_2$), and the refraction index of the surrounding air ($n_1$):

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 .$$

The angles of incidence and refraction are the angles between the light beams and the surface normal at the entry point.

Now we will give a simple proof of Snell's law using the sketch to the right, together with Fermat's principle:

**"A light ray will follow the path between two points that takes the shortest time."**




The velocity of light in the two media is given in the sketch. And as always: distance = velocity × time.

*Consider the time it takes to get from $P_1$ to $P_2$ in the sketch. This time is given by*

$$t = \frac{d_1}{v_1} + \frac{d_2}{v_2} = \frac{\sqrt{a^2 + x^2}}{v_1} + \frac{\sqrt{(l-x)^2 + b^2}}{v_2}$$

*$v_1$ and $v_2$ are the velocities in the two media, and everything else is given in the sketch..*

*We want to adhere to Fermat's principle and obtain the least time consuming path. Finding the derivative of time with regard to the x-coordinate of the entry point, and setting this derivative to zero, we find the x-value that gives a minimum value of the time.*

$$\frac{\partial t}{\partial x} = \frac{\frac{1}{2}\left(a^2 + x^2\right)^{-1/2} 2x}{v_1} + \frac{\frac{1}{2}\left[(l-x)^2 + b^2\right]^{-1/2} 2(l-x)(-1)}{v_2} = 0$$

$$\frac{x\left(a^2 + x^2\right)^{-1/2}}{v_1} = \frac{(l-x)\left[(l-x)^2 + b^2\right]^{-1/2}}{v_2}$$

*Rearranging a bit, we see that*

$$\frac{1}{v_1}\frac{x}{\sqrt{a^2 + x^2}} = \frac{1}{v_2}\frac{l-x}{\sqrt{(l-x)^2 + b^2}} \Rightarrow \frac{1}{v_1\sqrt{\left(\frac{a}{x}\right)^2 + 1}} = \frac{1}{v_2\sqrt{\left(\frac{b}{l-x}\right)^2 + 1}}$$

*From the sketch we see that tg $\theta_1$ =x/a, while tg $\theta_2$ = (l-x)/b, which gives us*

$$\frac{1}{v_1}\frac{x}{\sqrt{\cot g^2\theta_1 + 1}} = \frac{1}{v_2}\frac{l-x}{\sqrt{\cot g^2\theta_2 + 1}}$$

*And since*

$$\sin\theta = \frac{1}{\sqrt{\cot g^2\theta + 1}}$$

*this is equivalent to*

$$\frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2}$$

*We then use the fact that the index of refraction is defined as $n_i \equiv c/v_i$, where c is the velocity of light, and arrive at Snell's law (for planar waves):*

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

*Which was discovered 400 years ago by the Dutch Willebrord Snell (1591- 1626).*

---

Usually, the index of refraction is given relative to the surrounding medium, and we get the expression

**sin α / sin β = n**

where α is the angle of incidence and β is the angle of refraction.

## 2.2.2 The refractive index

Many materials have a well-characterized refractive index, but these indices depend strongly upon the wavelength of light. Therefore, any numeric value for the index is meaningless unless the associated wavelength is specified.[1]

There are also weaker dependencies on temperature, pressure, and stress, as well on the precise material compositions (including the presence of impurities and dopants). However, these variations are usually at the percent level or less. Thus, it is especially important to cite the source for an index measurement if high precision is claimed. http://www.luxpop.com/ lists the index for a number of materials.

| Material | *n* at λ=589.3 nm |
|---|---|
| Vacuum | 1 (exactly) |
| Air @ STP | 1.0002926 |
| Water (20°C) | 1.333 |
| Acrylic glass | 1.490 - 1.492 |
| Crown glass (pure) | 1.50 - 1.54 |
| Flint glass (pure) | 1.60 - 1.62 |
| Crown glass (impure) | 1.485 - 1.755 |
| Flint glass (impure) | 1.523 - 1.925 |
| Diamond | 2.419 |
| Gallium(III) arsenide | 3.927 |

*Table 2-1. The refractive index of some materials.*

In general, an index of refraction is a complex number with both a real and imaginary part, where the latter indicates the strength of the relative absorption loss at a particular wavelength. The imaginary part is sometimes called the extinction coefficient *k*.

### 2.2.2.1 The Sellmeier equation

Deduced in 1871 by W. Sellmeier, this equation models dispersion in a refractive medium, and is an empirical relationship between the refractive index $n$ and the wavelength $\lambda$ for a particular transparent medium. The usual form of the Sellmeier equation for glasses is:

$$n(\lambda) = \left[ 1 + \sum_i \frac{B_i \lambda^2}{\lambda^2 - C_i} \right]^2$$

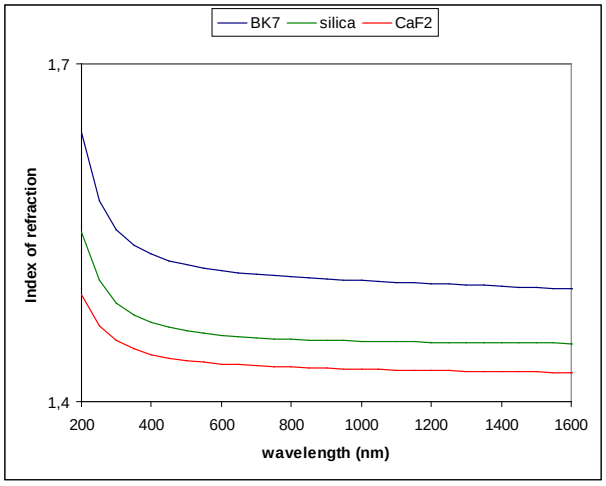



*Figure 2-13. The refractive index of BK7, silica, and $CaF_2$ as a function of wavelength.*

The coefficients $B_{1,2,3}$ and $C_{1,2,3}$ are usually quoted for λ in micrometers. Note that this λ is the vacuum wavelength; not that in the material itself, which is $\lambda/n(\lambda)$.

[1] The refractive index value is given for the D line of sodium at 589 nm at 25° C unless otherwise specified.

### 2.2.2.2 The Fresnel equations

When light moves from a medium with refractive index $n_1$ into a second medium with refractive index $n_2$, both reflection and refraction of the light may occur, as shown in figure 2-18. The relationship between the angles that the incident, reflected and refracted rays make to the normal of the interface (given as $\theta_i$, $\theta_r$ and $\theta_t$, in figure 2-14), is given by the law of reflection and Snell's law, respectively

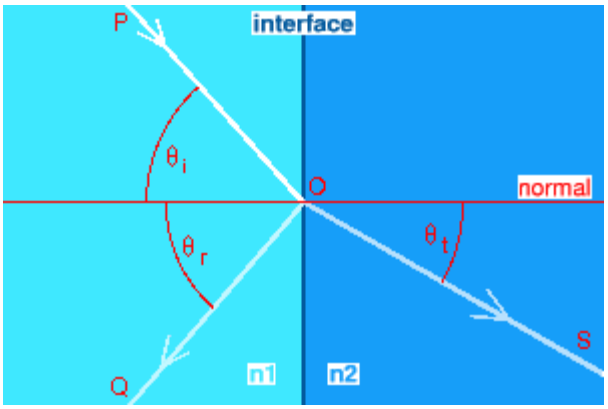



*Figure 2-14. Reflection and refraction when passing from one medium into another.*

Based on the assumption that the two materials are both *non-magnetic*, the Fresnel equations may be used to calculate the *reflection coefficient R*, i.e. the fraction of the intensity of incident light that is reflected from the interface. The reflection of light that these equations predict is known as **Fresnel reflection**.

The calculations of *R* depend on the polarization of the incident ray. If the light is polarized with the electric field of the light perpendicular to the plane of the diagram above (*s*-polarized), the reflection coefficient is given by:

$$R_S = \left[\frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t + \theta_i)}\right]^2 = \left[\frac{n_1 \cos(\theta_i) - n_2 \cos(\theta_t)}{n_1 \cos(\theta_i) + n_2 \cos(\theta_t)}\right]^2$$

since $\sin(x \pm y) = \sin(x)\cos(y) \pm \cos(x)\sin(y)$, and $n_1 \sin(\theta_i) = n_2 \sin(\theta_t)$.

If the incident light is polarized in the plane of the diagram (*p*-polarized), the *R* is given by:

$$R_P = \left[\frac{tg(\theta_t - \theta_i)}{tg(\theta_t + \theta_i)}\right]^2$$

The transmission coefficient in each case is given by $T_s = 1 - R_s$ and $T_p = 1 - R_p$.
If the incident light is unpolarized (containing an equal mix of *s*- and *p*-polarizations), the reflection coefficient is $R = (R_s + R_p)/2$.

At one particular incidence angle for a given $n_1$ and $n_2$, the value of $R_p$ goes to zero and a *p*-polarized incident ray is purely refracted. This angle is known as Brewster's angle, and is around 56° for a glass medium in air or vacuum. It is simply given by $\theta_i = \text{arctg}(n_2/n_1)$.

When moving from a more dense medium into a less dense one (i.e. $n_1 > n_2$), above an incidence angle known as the *critical angle*, all light is reflected and $R_s = R_p = 1$, as illustrated in figure 2-15. This phenomenon is known as total internal reflection. The critical angle is approximately 41° for glass in air.

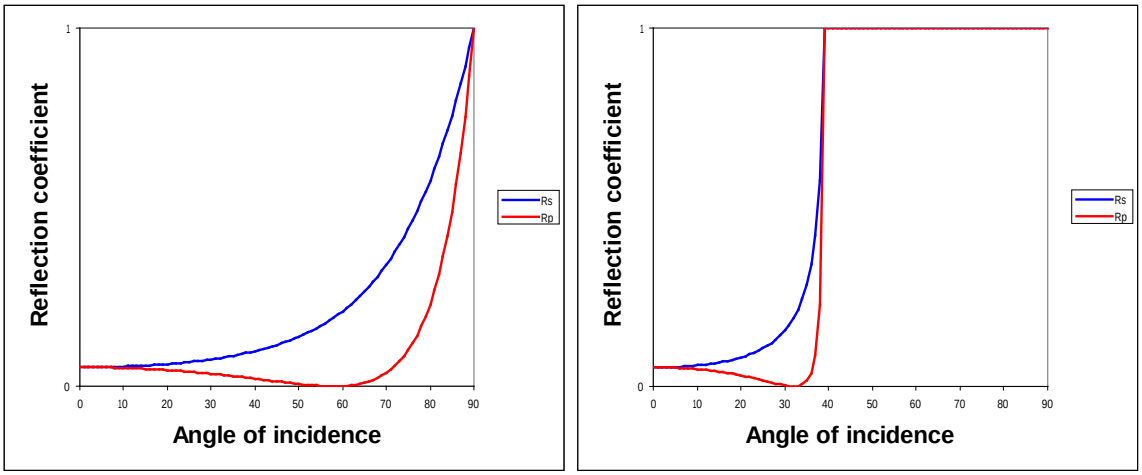



*Figure 2-15. The reflection coefficients from air to glass (left) and glass to air (right).*

When the light is at near-normal incidence to the interface ($\theta_i \approx \theta_t \approx 0$), the reflection and transmission coefficient are given by:

$$R = R_S = R_P = \left[\frac{n_1 - n_2}{n_1 + n_2}\right]^2$$

$$T = T_S = T_P = 1 - R = \frac{4 n_1 n_2}{(n_1 + n_2)^2}$$

For common glass, the reflection coefficient is about 4%. Note that reflection by a window pane is from the front side as well as the back side, and that some of the light bounces back and forth a few times between the two sides. The combined reflection coefficient for this case is $2R/(1 + R)$, when interference can be neglected.

In reality, when light makes multiple reflections between two parallel surfaces, the multiple beams of light generally interfere with one another, and the surfaces act as a Fabry-Perot interferometer. This effect is responsible for the colors seen in oil films on water, and it is utilized to make antireflective optical coatings and optical filters.

## *2.3 Optical prisms*

Optical prisms are often associated with dispersion of a white light beam into a spectrum. But the most common use of optical prisms is in fact to reflect light without dispersing it, in order to alter the direction of a beam, shift it by a certain amount in a given direction, and eventually rotate and / or flip the image at the same time.

### 2.3.1 Reflective prisms

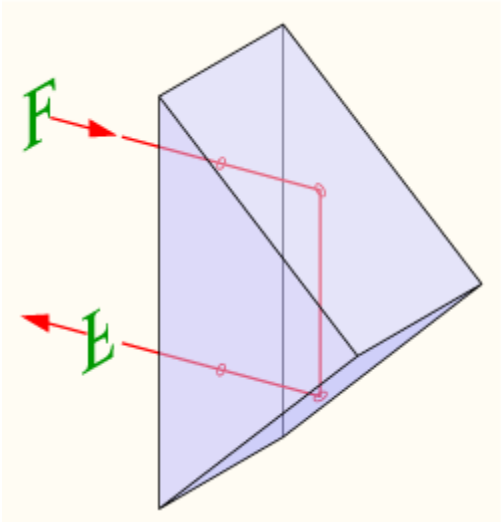
Reflective prisms utilize the internal reflection at the surfaces. In order to avoid dispersion, light must enter and exit the prism orthogonal to a prism surface. If light inside the prism hits one of the surfaces at a sufficiently steep angle, there is total internal reflection, and *all* of the light is reflected in accordance with the law of reflection (angle of incidence = angle of reflection). This makes a prism a very useful substitute for a planar mirror in some situations. Thus, reflective prisms may be used to alter the direction of light beams, to offset the beam, and to rotate or flip images.

#### 2.3.1.1 Single right-angle triangular prism

The right-angle triangular prism is the simplest type of optical prism. It has two right-angled triangular and three rectangular faces. As a reflective prism, it has two modes of operation.

1. The light beam enters orthogonal to one of the small rectangular faces, is reflected by the large rectangular face, and exits orthogonal to the other small rectangular face. The direction of the beam is altered by 90°, and the image is reflected left-to-right, as in an ordinary plane mirror.

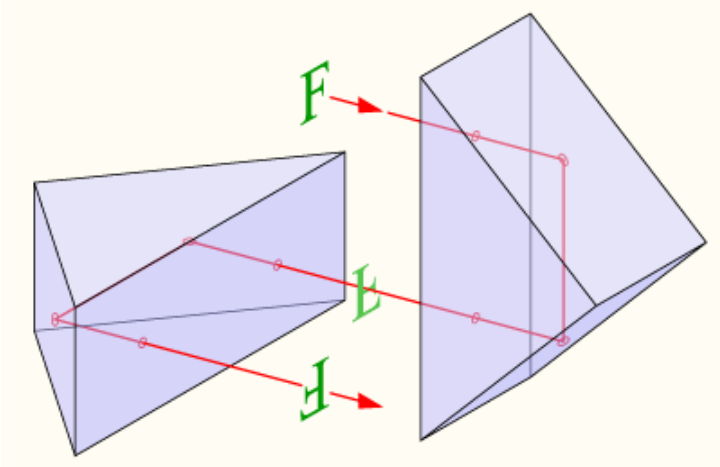
2. The light beam enters orthogonal to the large rectangular face, is reflected twice by the small rectangular faces, and exits again orthogonal to the large rectangular face. The beam exits in the opposite direction and is offset from the entering beam. The image is rotated 180°, and by two reflections the left-to-right relation is not changed (see figure to the right).



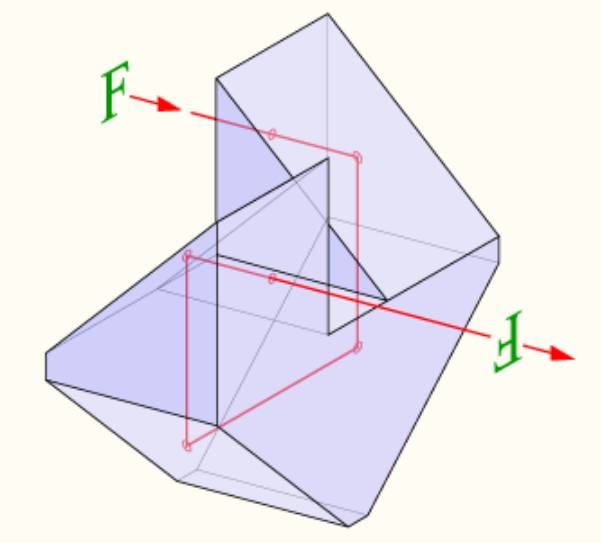

In both modes, there is no dispersion of the beam, because of normal incidence and exit. As the critical angle is approximately 41° for glass in air, we are also guaranteed that there will be total internal reflection, since the angles of incidence is always 41°.

### 2.3.1.2 Combinations of right-angle triangular prisms

In the simplest configuration, the two prisms are rotated 90° with respect to each other, and offset so that half of their large rectangular faces coincide. Often, the two prisms are cemented together, and the sharp ends of the prisms may be truncated to save space and weight. The net effect of the configuration is that the beam will traverse both prisms through four reflections. The net effect of the prism system is a beam parallel to but displaced from its original direction, both horizontally and vertically.



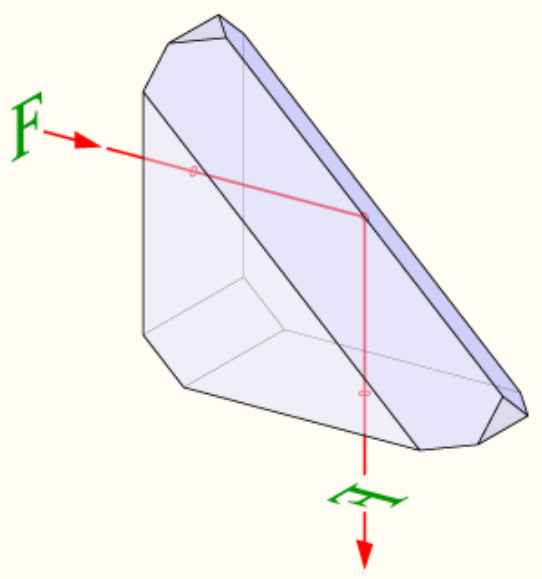

An alternative configuration is shown in the figure to the right. The exits beam again travels in the same direction as the input beam, but is offset in the horizontal direction.




In both configurations, the image rotated 180°, and given the even number of reflections, the handedness of the image is unchanged.

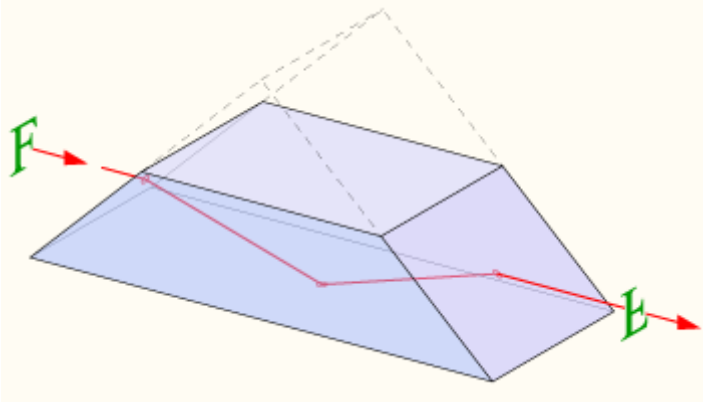
These prism systems are used in small telescopes to re-orient an inverted image. They are even more common in binoculars where they both erect the image and provide a longer, folded distance between the objective lens system and the eyepieces.

A r**oof prism** is a combination of two right-angled triangular prisms. It consists of a simple right-angled prism with a right-angled “roof” prism on its longest side, as shown in the figure to the right. Total internal reflection from the roof prism flips the image laterally, while the handedness of the image is unchanged. It is used to deviate a beam of light by 90° and simultaneously inverting the image, e.g., as an image erection system in the eyepiece of a telescope.



### 2.3.1.3 Truncated right-angle prism

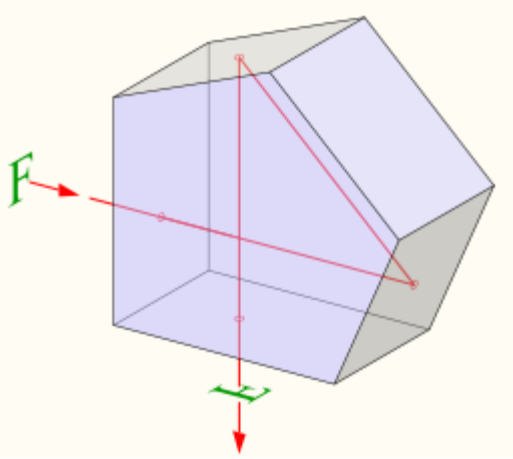
A truncated right-angle (Dove) prism may be used to invert an image.




A beam of light entering one of the sloped faces of the prism at an angle of incidence of 45°, undergoes total internal reflection from the inside of the longest (bottom) face, and emerges from the opposite sloped face. Images passing through the prism are flipped, and because only one reflection takes place, the image's handedness is changed to the opposite sense.

Dove prisms have an interesting property that when they are rotated along their longitudinal axis, the transmitted image rotates at twice the rate of the prism. This property means they can rotate a beam of light by an arbitrary angle, making them useful in *beam rotators*, which have applications in fields such as interferometry, astronomy, and pattern recognition.

### 2.3.1.4 Pentaprism

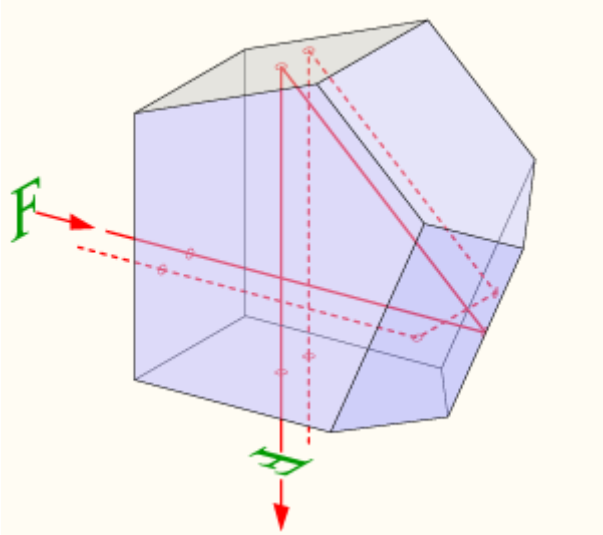



In a pentaprism the light beam enters orthogonal to one of the two orthogonal rectangular faces, is reflected by the two neighboring faces, and exits orthogonal to the face that is orthogonal to the entry face. Thus, the direction of the beam is altered by 90°, and as the beam is reflected twice, the prism allows the transmission of an image through a right angle without inverting it.

During the reflections inside the prism, the angles of incidence are less than the critical angle, so there is no total internal reflection. Therefore, the two faces have to be coated to obtain mirror surfaces. The two orthogonal transmitting faces are often coated with an antireflection coating to reduce reflections.

The fifth face of the (90°, 112.5°, 112,5°, 112.5°) prism is not used, but truncates what would otherwise be a sharp angle of 25° joining the two mirror faces. This fifth face is usually smaller than the two mirror faces in order to let the mirror faces receive all beams entering the input face.

### 2.3.1.5 Roofed pentaprism



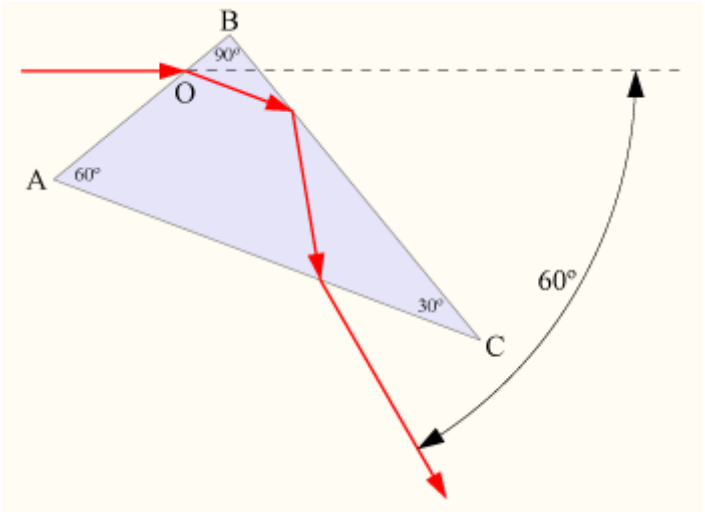

A roofed pentaprism is a combination of an ordinary pentaprism and a right-angle triangular prism. The triangular prism substitutes one of the mirror faces of the ordinary pentaprism. Thus, the handedness of the image is changed.

This construction is commonly used in the viewfinder of single-lens reflex cameras.

## 2.3.2 Dispersive prisms

A light beam striking a face of a prism at an angle is partly reflected and partly refracted. The amount of light reflected is given by Fresnel's equations, and the direction of the reflected beam is given by the law of reflection (angle of incidence = angle of reflection). The refracted light changes speed as it moves from one medium to another. This speed-change causes light striking the boundary between two media *at an angle* to proceed into the new medium at a *different* angle, depending on the angle of incidence, and on the ratio between the refractive indices of the two media (Snell's law). Since the refractive index varies with wavelength, light of different colors is refracted diffferently. Blue light is slowed down more than red light and will therefore be bent more than red light.

### 2.3.2.1 Abbe prism

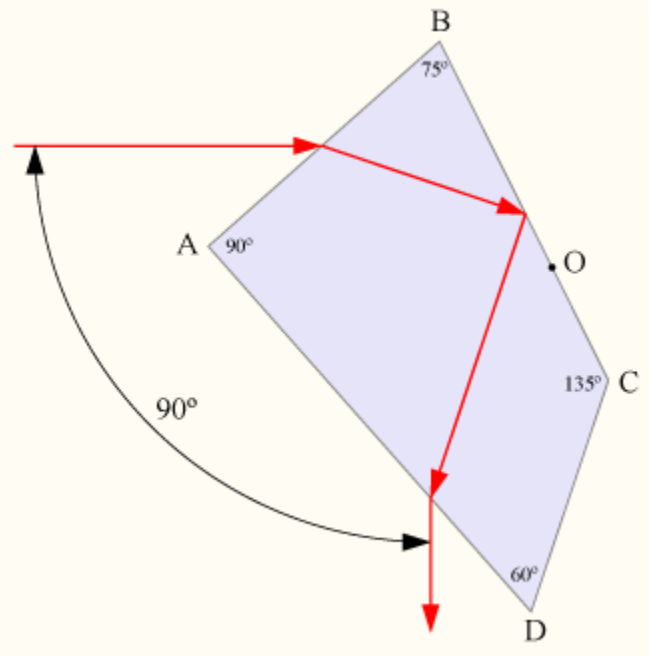



This is a right-angled prism with 30°-60°-90° triangular faces. A beam of light is refracted as it enters face AB, undergoes total internal reflection from face BC, and is refracted again on exiting face AC.

One particular wavelength of the light exits the prism at a deviation angle of exactly 60°. This is the minimum possible deviation of the prism, all other wavelengths being deviated by greater angles. By rotating the prism around any point O on the face AB, the wavelength which is deviated by 60° can be selected. Thus, the Abbe prism is a type of constant deviation dispersive prism.

### 2.3.2.2 Pellin-Broca prism

This is similar to the Abbe prism but consist of a four-sided block of glass with 90°, 75°, 135°, and 60° angles on the end faces. Light enters the prism through face AB, undergoes total internal reflection from face BC, and exits through face AD. The refraction of the light as it enters and exits the prism is such that one particular wavelength of the light is deviated by exactly 90°.

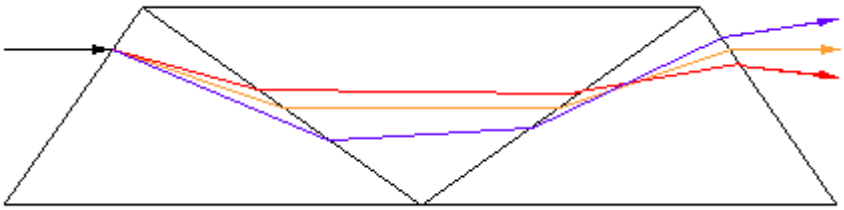



As the prism is rotated around a point O, one-third of the distance along face BC, the selected wavelength which is deviated by 90° is changed without changing the geometry or relative positions of the input and output beams.

It is commonly used to separate a single wavelength from a light beam containing multiple wavelengths. Thus, this is a type of constant deviation dispersive prism.

### 2.3.2.3 Direct vision spectroscope

This three prism arrangement, known as a **double Amici prism**, consists of a symmetric pair of right angled prisms of a given refractive index, and a right angled prism of a different refractive index in the middle. It has the useful property that the centre wavelength is refracted back into the direct line of the entrance beam. The prism assembly is thus a *direct-vision prism*, and is commonly used as such in hand-held spectroscopes.



### 2.3.2.4 Critical angle and total internal reflection

When light travels from a medium with a higher refractive index $n_1$ into a medium with a lower refractive index $n_2$ - for example when passing from glass to air – the ratio $n_1/n_2$ is greater than unity, sin $\theta_2$ is larger than sin $\theta_1$, and the ray is bent away from the normal. Thus, there must be a value of $\theta_1$ less than 90° for which Snell's law gives sin $\theta_2 = 1$ and $\theta_2 = 90°$.

The angle of incidence for which the refracted ray will be grazing the surface at an angle of refraction of 90°, is called the critical angle, given by $\theta_{crit} = \arcsin(n_2/n_1)$, where $n_2$ is the refractive index of the less dense medium, and $n_1$ is the refractive index of the denser medium.

- If $\theta < \theta_{crit}$, the ray will split. Some of the ray will reflect off the boundary, and some will refract as it passes through.
- If the incident ray is precisely at the critical angle, the refracted ray is tangent to the boundary at the point of incidence.
- If $\theta > \theta_{crit}$, all of the ray reflects from the boundary. None passes through.

This is illustrated in figure 2-16 for a semi-circular bowl of water. For a light-ray from a 632.8 nm laser entering perpendicular to the surface 4/10 of the radius from the centre of the bowl, grazing refraction and reflection will occur, provided that the refractive index of the water is raised by increasing the salinity of the water to about 16%.

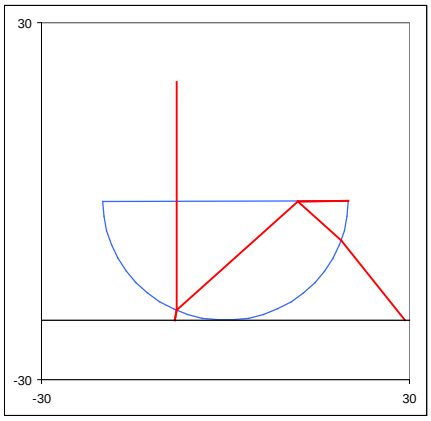



*Figure 2-16. Reflection and gracing refraction of laser light on water-air interface.*

Total internal reflection is the physical property that makes optical fibers useful. It is also what gives diamonds their distinctive sparkle, as diamond has an extremely high refractive index and therefore a low critical angle - about 24.4° - which means that light is much more likely to be internally reflected within a diamond than in glass, where the critical angle is about 41.5°. The "brilliant cut" is designed to achieve high total reflection of light entering the diamond, and high dispersion of the reflected light.

### 2.3.3 Non-imaging refraction

This text is supposed to be about imaging. So why de we dwell on the subject of non-focusing refraction, which does not produce images?

First, we need to understand simple refraction in order to perform ray-tracing through image-forming lenses. And the simplest refraction is found in plane-parallel slabs of glass and in glass prisms. Thereby we get a practical handle on parallel displacement of beams and images, as well as the concept of dispersion.

Secondly, a brief review of refraction and reflection in circular discs will take us along the path traveled by many – from the Persian astronomer Qutb al-Din al-Shirazi and his student Kamal al-din al-Farisi to Roger Bacon, René Descartes and Isac Newton – to realize the explanation for the common rainbow phenomenon.

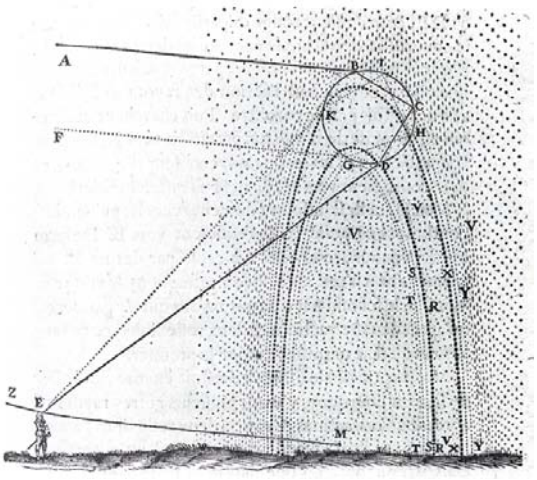



*Figure 2-17. René Descartes' sketch from 1637 of how primary and secondary rainbows are formed (*Discours de la Méthode Pour Bien Conduire Sa Raison et Chercher la Vérité dans les Sciences).

Isaac Newton was the first to demonstrate that white light was composed of the light of all the colors of the rainbow, which a glass prism could separate into the full spectrum of colors. He also showed that red light gets refracted less than blue light, which led to the first scientific explanation of the major features of the rainbow.

Newton's corpuscular theory of light was unable to explain supernumerary rainbows – the faint band of often multiple rainbows seen inside the primary rainbow. These are not possible to explain using classical geometric optics. The very existence of supernumerary rainbows was historically a first indication of the wave nature of light. A satisfactory explanation was not found until Thomas Young realized that light behaves as a wave under certain conditions, and can interfere with itself.

Finally, the strength of the various colors of the rainbow depends on the spectral distribution of the light source, as well as on the spectral characteristics of the absorption and scattering of light as it travels through the atmosphere.

So, although rainbows are not a result of imaging, they provide insight into very important aspects of imaging. Besides, some insight doesn't hurt!

### 2.3.3.1 Parallel displacement by refraction in plane-parallel slab

As a practical example of refraction, let us look at a beam of monochromatic light passing into a plane-parallel slab of glass at a certain angle of incidence. For a given refractive index we may use Snell's law to compute the angle of refraction, and trace the refraction into the slab, see figure 2-18, which illustrates light at 550 nm entering a slab of BK7 (n=1.513614) at a 45° angle of incidence. The reflections at the glass/air interface are also traced, as well as the refractions of these beams out of the slab.

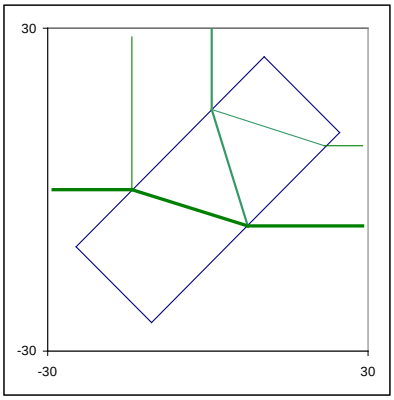



*Figure 2-18. Refraction and reflection of light in a plane parallel slab.*

When the beam exits the glass slab, the angle of incidence equals β, the relative index of refraction is 1/n, and the exit angle φ is given by n sin β= sin φ. Thus, the exit angle will be the same as the angle of incidence when the light beam entered the glass slab, and the net effect is that the plane parallel glass slab has caused a parallel displacement of the light beam. The displacement, d, given relative to the thickness of the slab, is

$$d = \sin\alpha\left(1-\frac{\cos\alpha}{\sqrt{n^2-\sin^2\alpha}}\right)$$

At the exit interface, most of the light will pass through, but a small fraction (say 4 %) is reflected. Thus, a twice reflected beam having a 1.6 x $10^{-3}$ relative intensity, will be displaced relative to the first by an amount l, given in units of the thickness of the slab:

$$l = 2\frac{\sin\alpha\cos\alpha}{\sqrt{n^2-\sin^2\alpha}}$$

This is illustrated in figure 2-19 for n = 1.5 for angles of incidence between 0 and 60 degrees. As shown, the displacement is an approximately linear function of angle of incidence for small angles. A simple application could be a single-detector scanning of an object by inserting step-motor controlled plane-parallel slabs into the light path.

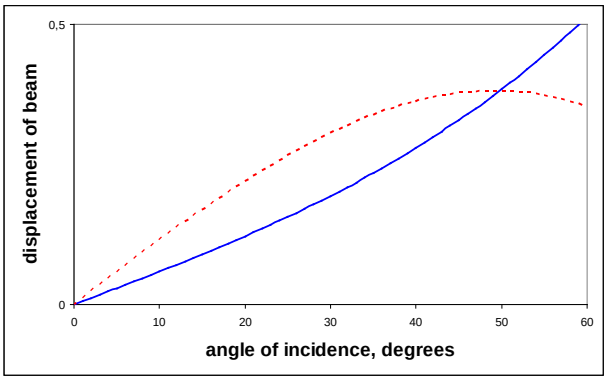



*Figure 2-19. Relative displacements of a light beam by a plane parallel slab.*

### 2.3.3.2 Monochromatic refraction by a prism

Let α be the angle between the two faces of a symmetric triangular prism. Let the edge A where the two faces meet be perpendicular to the plane which contains the incident, transmitted, and emergent rays. Let us first assume that the light is monochromatic.

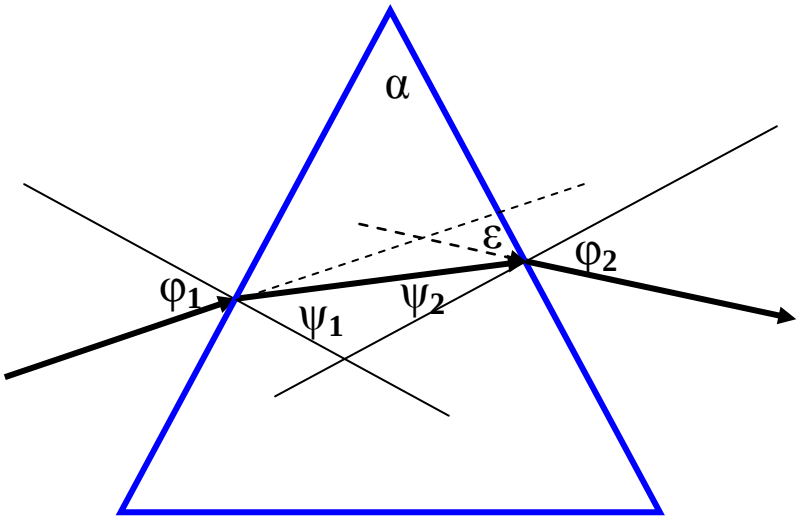



*Figure 2-20. Refraction of a light beam passing through a symmetric triangular prism.*

Assuming that the angles are as indicated in figure 2-20, if ε is the angle of deviation, i.e. the angle which the emergent ray makes with the incident ray, then

$$\phi_1 + \phi_2 = \varepsilon + \alpha\,, \qquad \psi_1 + \psi_2 = \alpha$$

and

$$\sin\phi_1 = n\,\sin\psi_1\,, \qquad \sin\phi_2 = n\sin\psi_2$$

where n is the refractive index of glass with respect to the surrounding air.
Now, how does the deviation ε vary with $\varphi_1$ for a given angle α and index n?
If an extremum value of ε exists, then $\phi_1 + \phi_2 = \varepsilon + \alpha$ implies that

$$\left(\frac{d\phi_2}{d\phi_1}\right)_{extr.} = -1$$

The three next equations above imply that

$$\frac{d\psi_1}{d\phi_1} = -\frac{d\psi_2}{d\phi_1}$$

$$\cos\phi_1 = n\cos\psi_1\frac{d\psi_1}{d\phi_1}, \quad \cos\phi_2\frac{d\phi_2}{d\phi_1} = n\cos\psi_2\frac{d\psi_2}{d\phi_1}$$

And hence

$$\left(\frac{d\phi_2}{d\phi_1}\right) = -\frac{\cos\phi_1\cos\psi_2}{\cos\psi_1\cos\phi_2}$$

From

$$\left(\frac{d\phi_2}{d\phi_1}\right)_{extr.} = -1$$

it follows that

$$\frac{\cos\phi_1 \cos\psi_2}{\cos\psi_1 \cos\phi_2} = 1$$

Squaring and using Snell's law we get

$$\frac{1-\sin^2\phi_1}{n^2-\sin^2\phi_1} = \frac{1-\sin^2\phi_2}{n^2-\sin^2\phi_2}$$

which is satisfied by $\phi_1 = \phi_2$, which implies that $\psi_1 = \psi_2$.
So the extremum value of ε (which is a minimum) is obtained when the passage of light through the prism is symmetrical. This minimum value of ε is given by

$$\varepsilon_{\min} = 2\phi_1 - \alpha$$

or

$$\psi_1 = \frac{1}{2}\alpha$$

so that

$$n = \frac{\sin\phi_1}{\sin\psi_1} = \frac{\sin\left[\frac{1}{2}\left(\varepsilon_{\min}+\alpha\right)\right]}{\sin\left(\frac{1}{2}\alpha\right)}$$

This general formula may be used to determine the refraction index n when α is known and $\varepsilon_{min}$ is determined experimentally. The simulation in figure 2-21 demonstrates that the shape of the deviation-versus-angle-of-incidence curve depends on the angle α between the two faces of the symmetric triangular prism.

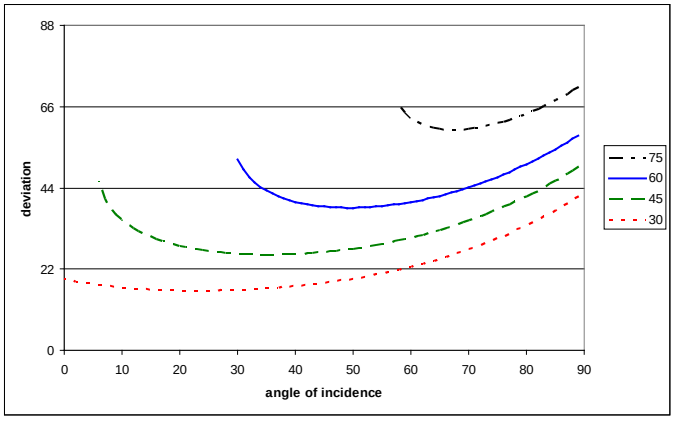



*Figure 2-21. Deviation angle versus angle of incidence for a symmetric triangular prism, for four different values of the angle α between the two faces of the symmetric triangular prism. Computed for BK7 borosilicate crown glass at 589 nm.*

### 2.3.3.3 Chromatic dispersion by regular triangular prism

Figure 2-26 shows a very rough sketch of a light beam being refracted through a regular triangular prism, and the dispersion of the white light into the colors of the spectrum.


*Figure 2-22. A qualitative illustration of chromatic refraction through a prism.*

Let us use CIE primary colors 435.8 nm (blue), 546.1 nm (green) and 700.0 nm (red) and the refractive index n given by the Sellmeier equation below, and the B and C coefficients for BK7 given by table 2-4, in order to trace the light through a prism.

$$n(\lambda)=\left[1+\sum_i \frac{B_i\,\lambda^2}{\lambda^2-C_i}\right]^2$$

So the refractive index for the three CIE primaries are n(blue) = 1.526688, n(green) = 1.518721, n(red) = 1.513064. Figure 2-23 illustrates the real dispersion of the RGB-colors through a prism. As we can see, the dispersion is tiny compared to the exaggerated sketch of figure 2-22.

| Coefficient | Value |
|---|---|
| $B_1$ | 1.03961212 |
| $B_2$ | $2.31792344\text{x}10^{-1}$ |
| $B_3$ | 1.01046945 |
| $C_1$ | $6.00069867\text{x}10^{-3}$ μm$^2$ |
| $C_2$ | $2.00179144\text{x}10^{-2}$ μm$^2$ |
| $C_3$ | $1.03560653\text{x}10^{2}$ μm$^2$ |

*Table 2-2. The Sellmeier equation coefficients for the BK7 borosilicate crown glass.*

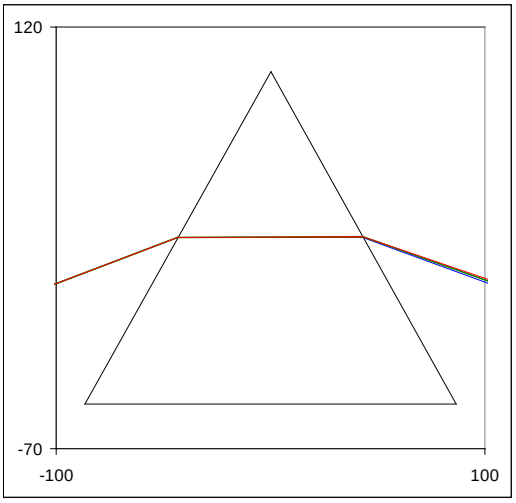



*Figure 2-23. A quantitative illustration of the refraction of RGB-colors through a prism.*

### 2.3.3.4 Two refractions and one reflection by a circular disc

When parallel rays of light enter a circular disc, they are refracted. A small fraction (4%) of the refracted light is reflected from the back of the disc, and refracted again upon exiting the disc, as illustrated in figure 2-24 for ten equidistant, parallel and monochromatic rays entering the upper half of a disc from the left hand side.

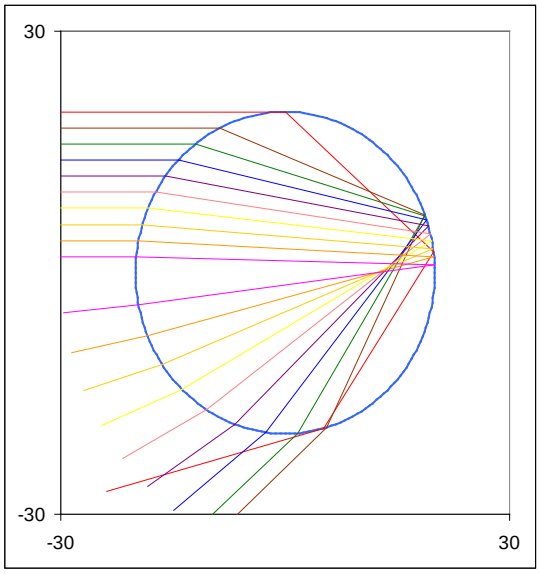



*Figure 2-24. Two refractions and a reflection of parallel rays in a circular disc.*

A light ray that enters the middle of the disc is reflected straight back towards the light source, while all other rays exit at an angle φ to that middle ray, ”piling up” at a maximum exit angle $\varphi_M$, as shown in figure 2-25. This angle obviously depends on the refraction index of the disc. For a refraction index of 1.333 (water), $\varphi_M \approx 42°$. Rene Descartes, in 1637, performed this experiment of passing rays of light through a large glass sphere filled with water, in order to explain the observed rainbow.

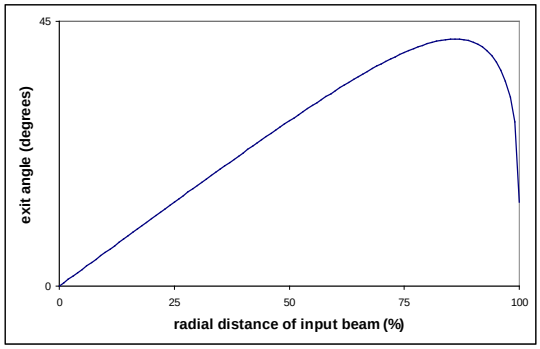



*Figure 2-25. The exit angle of 100 equidistant parallel rays entering the upper half of a water-filled circular disc.*

### 2.3.3.5 The primary rainbow

A rainbow may be observed whenever there are water drops in the air and a strong light source behind the observer. The luminance of the reflected light from the water drops must be sufficiently high to make the color sensitive cones of the retina to dominate over the rod-based gray-level vision. Thus colorful rainbows are often seen when you are facing falling rain while the sun is shining behind you.

As a consequence of the distribution of exit angles shown in figure 2-25, what we see is a disc of light having angular radius $\varphi_M$ centered on the point in the sky opposite the light source, e.g. the sun. Due to the ”piling up” of exit angles, the disc is brightest around its rim. And because no light reaches our eyes from angles greater than $\varphi_M$, the sky looks darker outside this disc, as seen in figure 2-26.

*Figure 2-26. A primary and a secondary rainbow.*

The value of the angle $\varphi_M$ depends on the index of refraction of the water that makes up the raindrops. The refraction index again depends on the wavelength of the light (see e.g. http://www.luxpop.com). As blue light is refracted more than red light, the overall effect is that the bright disc of red light is slightly larger than that for orange light, which in turn is slightly larger than that for yellow light, and so on.

This is illustrated by the computed exit angle for some equidistant input rays, for three different wavelengths (400, 589 and 700 nm), at 20 degrees C, shown in figure 2-27.

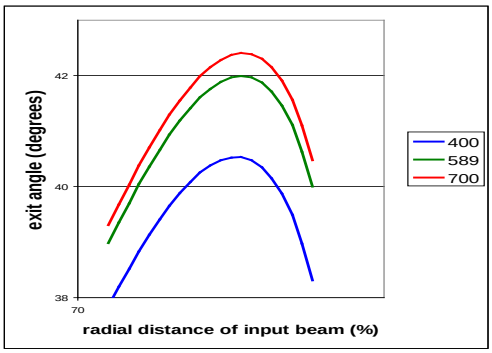



*Figure 2-27. The maximum exit angle after two refractions and one reflection in a water-filled circular disc, at three different wavelengths (400, 589 and 700 nm), at 20 °C.*

As a net result, we see a rainbow as a band with the visual color spectrum from violet to red spread out over 40.8° – 42.5°, as illustrated in figure 2-28. The strength of the various colors of the rainbow depends on the spectral distribution of the light source, and the absorption and scattering as light travels through the atmosphere.

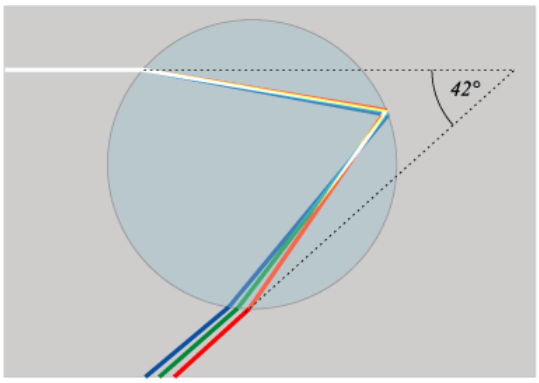



*Figure 2-28. Dispersion of light as a result of wavelength dependent refraction in water.*

The rays of the primary rainbow form a cone with its tip at the observer's eye and its axis directed towards the antisolar point, as indicated in Decartes sketch in figure 2-17. Drops near the cone's surface send sunlight into the eye to produce the rainbow. Distance does not matter, because the rainbow is a collection of rays with particular directions, it is not located at any particular point in space. Drops *inside* the cone brighten the sky inside the rainbow. Drops *outside* the cone send no light in the direction of the eye.

### 2.3.3.6 Two refractions and two reflections by a circular disc

As we have seen, rays of light entering a circular disc are refracted. A small fraction (0.16%) of the refracted light is reflected twice from the back of the disc, and refracted again upon exiting the disc, as illustrated in figure 2-29 for ten equidistant, parallel and monochromatic rays entering the lower half of a disc from the left hand side.

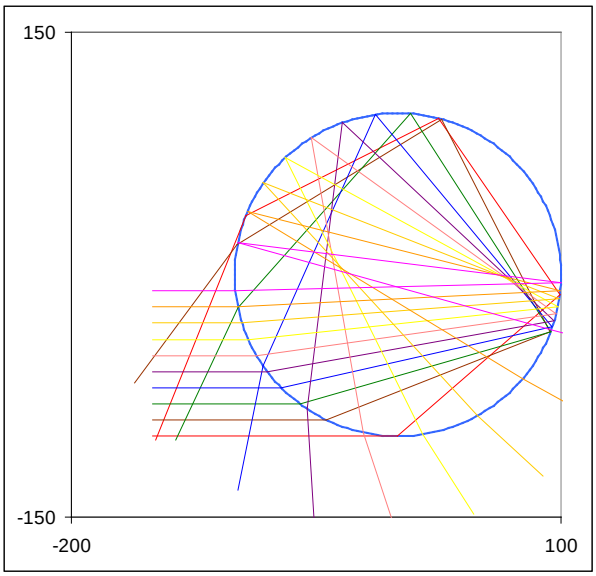



*Figure 2-29. Two refractions and two reflections of parallel rays in a circular disc.*

A light ray that enters the middle of the disc is reflected straight back twice to continue in its original direction, at an angle of 180° to the direction back to the light source. All other rays exit at an angle φ to the backwards direction, ”piling up” at a minimum exit angle $\varphi_{M2}$, as shown in figure 2-30.

This angle again depends on the refraction index of the disc. For a refraction index of 1.333 (water), $\varphi_{M2} \approx 52°$.

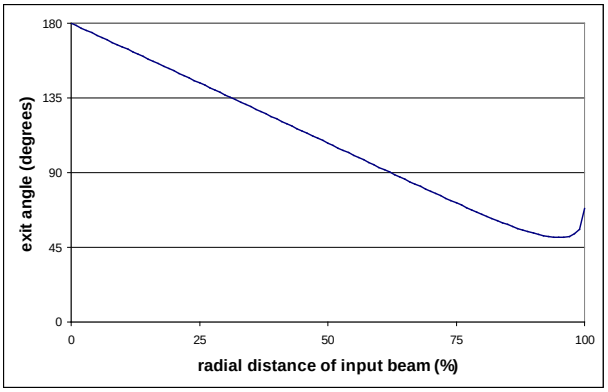



*Figure 2-30. The exit angle of 100 equidistant parallel rays entering one half of a water-filled circular disc.*

### 2.3.3.7 The secondary rainbow

The most frequently observed rainbow variation is a dimmer secondary rainbow seen outside the primary bow. Secondary rainbows are caused by the double reflection of sunlight inside the raindrops, and appear at an angle of 50°–53° to the direction of the point in the sky opposite the light source, as we have seen in the photograph in figure xxx, where we can also see that the secondary rainbow is fainter than the primary bow.

Notice that now the result of the distribution of exit angles is not a bright disc centered on the point in the sky opposite the light source, but a faint disc centered on the light source behind the observer (φ = 180°) stretching all the way round to a bright rim around a darker “hole” centered on the point in the sky opposite the light source. The angular radius of this “hole” is equal to the minimum exit angle $\varphi_{M2}$, or about 52°. Together with the bright disc associated with the primary bow, this gives rise to the dark so-called “Alexander-band” (after Alexander of Afrodisias) that is easily visible between the two bows in figure 2-26.

The value of the angle $\varphi_{M2}$ depends on the index of refraction of water, which is a function of the wavelength (see e.g. http://www.luxpop.com). As we now have an additional reflection compared to the ordinary rainbow, the order of the colors is reversed, with blue on the outside and red on the inside, as seen in the photograph in figure 2-26. This is illustrated by the computed exit angle for some equidistant input rays, for three different wavelengths (400, 589 and 700 nm), at 20 degrees C, shown in figure 2-31. Notice also that the secondary bow is wider than the primary bow.

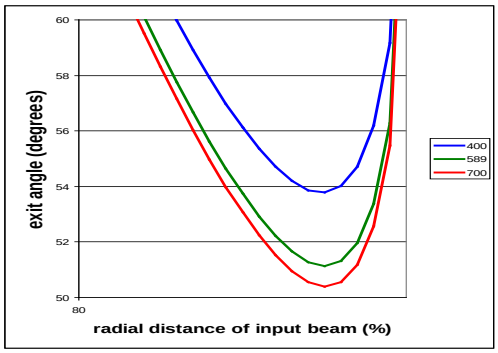



*Figure 2-31. The minimum exit angle after two refractions and two reflections in a water filled circular disc, for three different wavelengths (400, 589 and 700 nm), at 20 °C.*

### 2.3.3.8 Rainbow variations

**Supernumerary rainbows**:
These are a series of several faint rainbows that can be observed on the inner side of the primary rainbow, and very rarely also outside the secondary rainbow. It is not possible to explain their existence using classical geometric optics. They are caused by patterns of interference between rays of light following slightly different paths with slightly varying lengths within the raindrops. The patterns of interference are slightly different for rays of different colors, so each bright band is differentiated in color, creating a small rainbow.

Supernumerary rainbows can be simulated to produce images like the one in figure 2-32. However, crisp and distinct supernumeraries like these are not seen in nature, as the real ones are blurred by the finite angular size of the sun and variations in raindrop size.

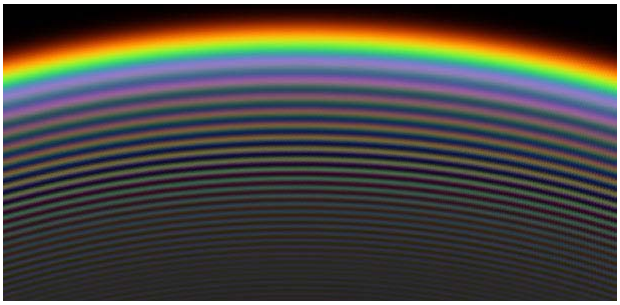
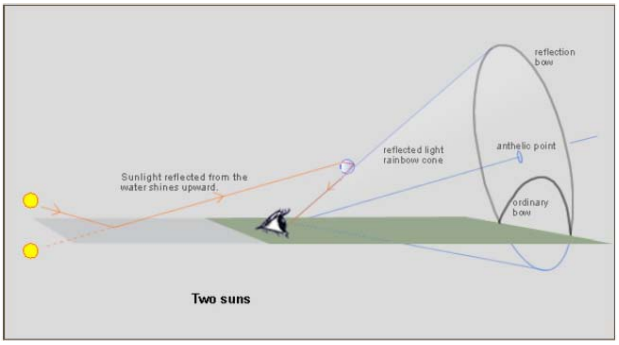
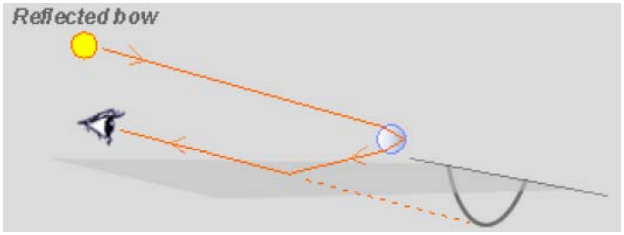


*Figure 2-32. AirySim simulation of supernumerary rainbows (http://www.atoptics.co.uk/rainbows/airysim.htm).*

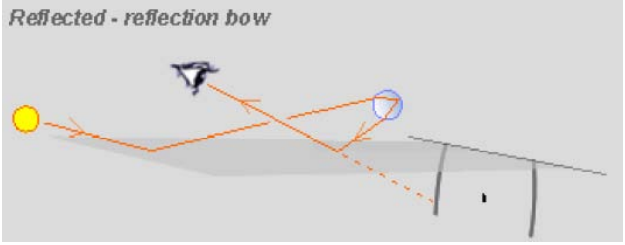
**Reflection rainbows:**
Reflection rainbows are produced when a calm water surface behind the observer sends sunlight upwards, as if there was a second sun shining from below the horizon. From the observers point of view, this creates a reflection rainbow centered on an "antisolar point" at the same height above the horizon as the true antisolar point is below it. Such rainbows share the same endpoints as normal rainbows but may describe a greater arc. Both primary and secondary reflection rainbows can be observed (see image on top of next page).



**Reflected rainbows:**
A reflected rainbow, on the other hand, is produced when light from a rainbow is reflected off a body of water before reaching the observer. A reflected rainbow is not a mirror image of the ordinary rainbow seen at the same time, but a reflection of the rainbow that an observer at an equal distance below the surface would have seen. Both primary and secondary reflected rainbows may be seen.




Obviously, we may also observe reflected reflection rainbows (below left): Sunlight is reflected off the water and travels upwards to be refracted by to form primary and secondary reflection rainbows. These reflect again off the water into the eye. However, the reflected reflection rainbows are not mirror images of the reflection rainbows that are observed simultaneously, since rainbows are not physical objects that may be reflected. In this case the rainbow below the horizon is formed by different raindrops from the ones that produced the bow above the horizon.



**Salt water rainbows:**
These are seen in seaspray and have a slightly smaller radius, because the refraction index of water depends on the salinity. The index increases by 0.001 per 5 g of salt per kg of water, giving n = 1.341 for a salinity of 3.5%, instead of n = 1.33394 at 589 nm and 20°C. The difference in the rainbow radius is clearly noticeable when a raindrop rainbow occurs above a seaspray bow.

**Moonlit rainbows:**
The luminance of the reflected light from the water drops must be sufficiently high to make the color sensitive cones of the retina to dominate over the rodbased graylevel vision. Thus, ”moonbows” are perceived as a pale white phenomenon.

**Colorful optical phenomena that are not rainbows:**

- Halos are a class of phenomena caused by ice crystals in the atmosphere. The most common is the 22º radius bright halo around the sun (or the moon) caused by thin high cirrostratus clouds. The sharp inner edge may be red. The sky inside the halo is darker.
- Parhelia are among the most frequent halo phenomena, caused by planar crystals lying mostly horizontal in the atmosphere. Thus, when the sun is low, light passes through the crystal side faces inclined 60º to each other, and we see bright halos on both sides of the sun, about 22º off. These “sundogs” are often brightly colored because of differential refraction in the ice crystals.
- Pillars are narrow columns of light *apparently* beaming directly up and sometimes downwards from the sun, most frequently seen near to sunset or sunrise. They can be 5 -10º tall and occasionally even higher. Pillars are not actually vertical rays; they are instead the collective glints of millions of ice crystals. As they take on the colors of the sun and clouds, they can appear white and at other times shades of yellow, red or purple.
- The circumzenithal arc is also a beautiful halo phenomenon, like an upside down rainbow.
- Coronae are halo-like, but smaller colored rings around the sun or moon that are produced by scattering in water droplets rather than ice crystals.
- The glory phenomenon is a colorful set of rings scattered from cloud or fog droplets. The rings are centered on the shadow of your head.
- Fogbows are also a scattering phenomenon. They are almost as large as rainbows, and much broader.
- Heiligenschein is caused by the focusing of dew drops, and backscattering of the focused light through the drops.
- Iridescence in clouds most often occurs close to the sun, but is best seen if the sun is hidden. It is caused by diffraction in small same-size droplets in thin tropospheric clouds.
- Nacreous or mother-of-pearl clouds are a much rarer manifestation of iridescence. They can glow very brightly and are far higher than ordinary tropospheric clouds.

### 2.3.3.9 Atmospheric refraction

The biggest refractor in our neighborhood is the atmosphere. Due to the variation in air density as a function of altitude, it causes light or other electromagnetic waves to deviate from a straight line during a passage through the atmosphere.

The atmospheric refraction is zero in the zenith, is less than 1' at 45° altitude, still only 5' at 10° altitude, but then quickly increases towards the horizon. Atmospheric refraction causes all astronomical objects to appear slightly higher in the sky than they are in reality. It affects not only light rays but all electromagnetic radiation, although to varying degrees. For example in visible light, blue is more affected than red. This may cause astronomical objects to be spread out into a spectrum in high-resolution images.

On the horizon itself refraction is about 34', which just a little bit larger than the apparent size of the sun. So when the setting sun appears to be just above the horizon, it has in reality already set, as seen in the simulation in figure 2-33. The image of the setting sun is nonlinearly flattened, because the atmospheric refraction is 34' on the horizon, but only 29' half a degree (one solar diameter) above it. So the setting or rising sun seems to be flattened by about 5' or 1/6 of its apparent diameter. Thermal inversion layers may create interesting additional phenomena.

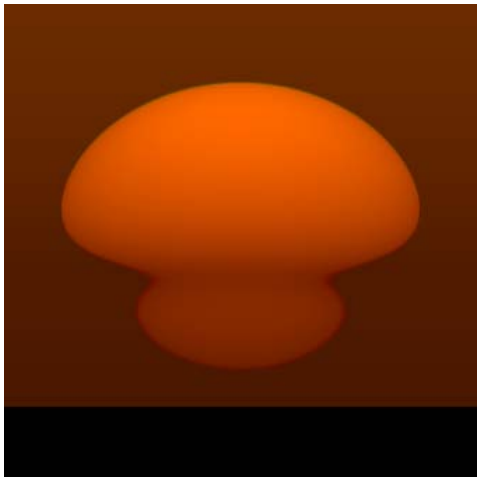


*Figure 2-33. The sun, its centre actually 30' below the horizon, lifted and flattened, on top of a mirror image of its lower limb caused by a temperature inversion (http://mintaka.sdsu.edu/GF ).*

The refraction is also a function of temperature and pressure. Hot air is less dense, and will therefore have less refraction. The values given above are for 10 °C and 1003 mbar. Add 1% to the refraction for every 3° C colder, subtract if hotter. Add 1% for every 9 mbar higher pressure, subtract if lower. Evidently day to day variations in the weather will affect the exact times of sunrise and sunset as well as moonrise and moonset. For that reason, almanacs never give these times more accurately than to the nearest whole minute.

Atmospheric refraction near the ground produces mirages and can make distant objects appear to shimmer or ripple, as we can easily observe in objects seen at a low angle over a hot surface.

When light rays coming from point sources like stars travel through an inhomogeneous and often turbulent atmosphere, the light rays have to pass through layers and bubbles of air having different densities, causing a continuous altering of direction. This causes stars to appear as twinkling. The twinkling will vary with the atmospheric turbulence.

## 2.3.4 Imaging by refraction

### 2.3.4.1 Refraction at a planar surface

In the same way that a virtual image is formed by a planar mirror, a planar refracting surface may form a virtual image, causing a straight object that passes through the surface to appear bent at the surface, as shown in figure 2-34.

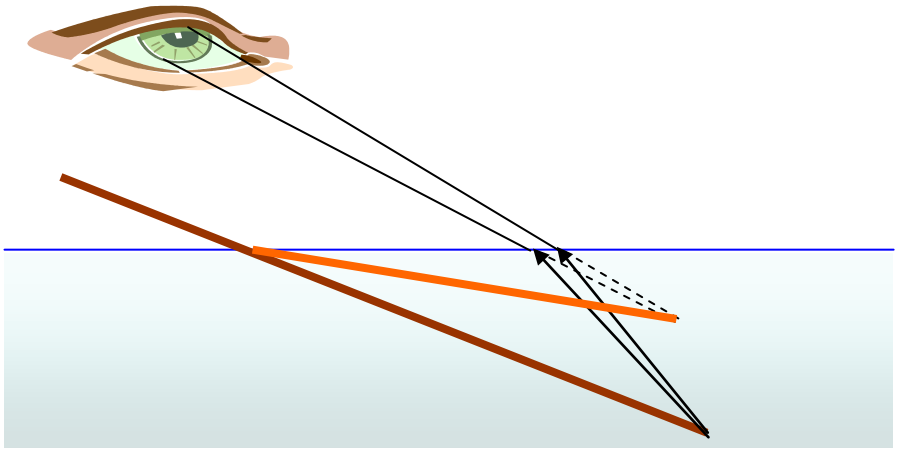


*Figure 2-34. Displacement caused by refraction in water.*

### 2.3.4.2 Refraction at a single spherical surface

Let us consider refraction at a single spherical surface before we look at refraction by simple lenses, which have two spherical surfaces.

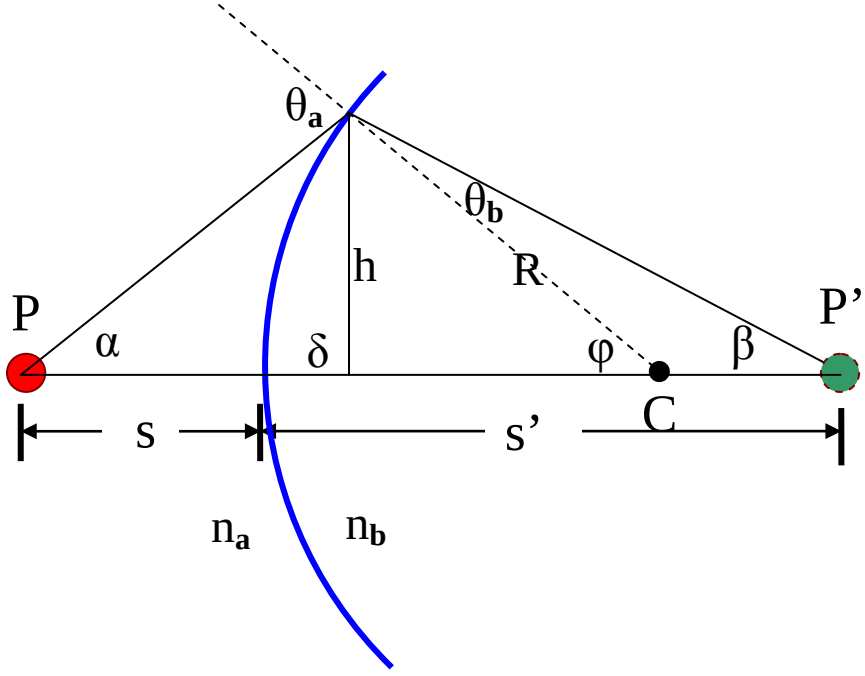



*Figure 2-35. Finding the image point of a point object formed by refraction at a spherical surface.*

In figure 2-35 we assume that $n_a < n_b$. Snell's law of refraction gives: $n_a \sin\theta_a = n_b \sin\theta_b$. If the angle α is small (the paraxial approximation), Snells law becomes $n_a\theta_a = n_b\theta_b$. Combining this with $\theta_a = \alpha + \varphi$ gives $\theta_b = (\alpha + \varphi)\, n_a/n_b$.
Substituting this into $\varphi = \beta + \theta_b$ we get $(n_a\alpha + n_b\beta) = (n_b - n_a)\varphi$.
The expressions for the tangents of α, β, and φ are simply

$$\mathrm{tg}(\alpha) = h/(s+\delta),\ \ \mathrm{tg}(\beta) = h/(s'-\delta),\ \ \mathrm{tg}(\varphi) = h/(R-\delta).$$

If the angle α is small, so are β and φ. If α is so small that the ray is almost parallel to the optical axis (the paraxial approximation), the tangent of an angle is equal to the angle itself (given in radians), and δ may be neglected compared to s, s', and R. So for small angles we have the following approximations: $\alpha = h/s$, $\beta = h/s'$, $\varphi = h/R$.
Substituting this into $(n_a\alpha + n_b\beta) = (n_b - n_a)\varphi$ we get the general object-image relation for a single spherical surface of radius R, given the two refractive indices $n_a$ and $n_b$.

$$\frac{n_a}{s} + \frac{n_b}{s'} = \frac{n_b - n_a}{R}$$

This expression is the same no matter if $n_a$ is greater or less than $n_b$. Besides, the expression does not contain the angle α, which in itself proves that under the paraxial assumption all light rays coming from P will intersect at P'.

A simple sketch may again illustrate how the size y' and position s' of the real image of an extended object is determined, once the size y of the object, the distance s from the spherical surface to the object and the radius R of curvature of the surface is known, see figure 2-36.

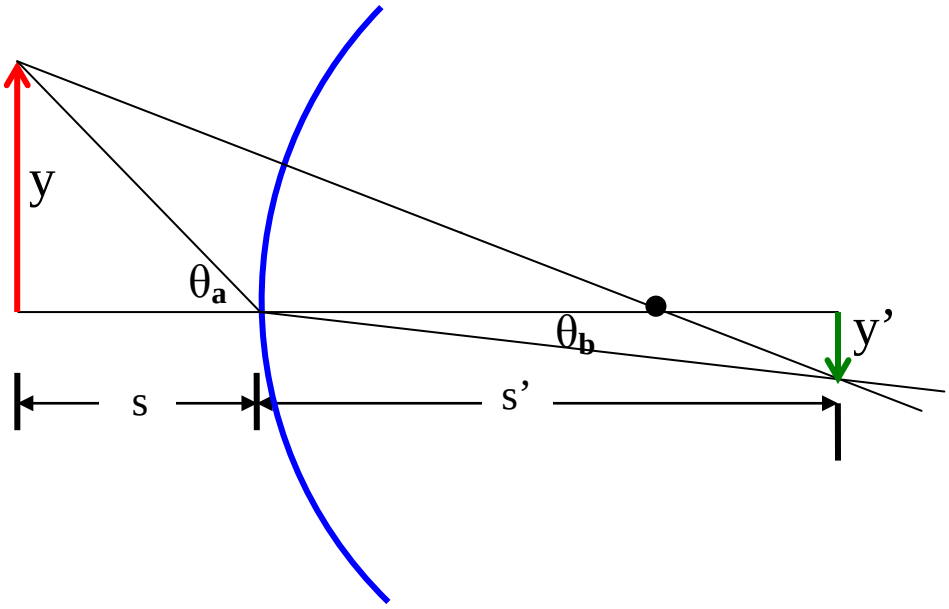



*Figure 2-36. Finding the size and position of an extended image formed by refraction at a spherical surface.*

We see that $\mathrm{tg}\,\theta_a = y/s$ and $\mathrm{tg}\,\theta_b = -y'/s'$. Using the refraction law $n_a \sin\theta_a = n_b \sin\theta_b$ and assuming that the angles are small, we get $n_a\, y/s = -n_b\, y'/s'$. So the lateral magnification is $m = y'/y = -(n_a\, s')/(n_b\, s)$. The negative sign indicates that the image is inverted relative to the object.

### 2.3.4.3 Image formation by thin lenses

Lenses are the most widely used image forming devices. The simplest lens has two refracting surfaces. These are often spherical, and placed sufficiently close together that we can neglect the distance between them. Such lenses are therefore called thin lenses.

The centers of curvature of the two spherical surfaces lie on and define the optical axis. A double-convex lens will focus a beam of rays parallel to the optical axis to a focal point, and form a real image there. Likewise, rays emerging from or passing through that point will emerge from the lens as a beam of parallel rays. The two points on either side of the lens are called the first and second focal point. The distance from the focal point to the middle of the lens is called the focal length. The focal length of a convex lens is defined to be positive, and such a lens is also called a positive lens.

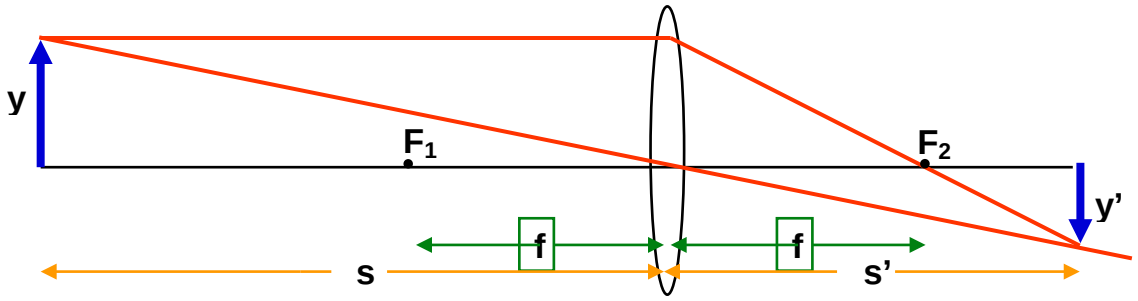



*Figure 2-37. Construction of an image y' of an object y by a thin, convex lens.*

By comparing two similar triangles on either side of the lens in figure 2-37, we see that

$$\frac{y}{s} = \frac{y'}{s'} \Rightarrow \frac{y'}{y} = \frac{s'}{s}$$

In the same way, two similar triangles on either side of the second focal point $F_2$ give

$$\frac{y}{f} = \frac{y'}{s'-f} \Rightarrow \frac{y'}{y} = \frac{s'-f}{f}$$

Now we have two expressions for y'/y, giving

$$\frac{s'}{s} = \frac{s'-f}{f}$$

Rearranging this we obtain what is known as the "*object-image relation*" for a thin lens:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

The lateral magnification is given by: m = y'/y = s'/s, so that y' = ys'/s. Substituting from the object-image relation, we get a useful expression for the size of the image in the focal plane:

$$y' = \frac{y\,f}{(s-f)}$$

When the object is far from the lens, the image of the object is smaller than the object, inverted, and real. If the object for all practical purposes is infinitely far away, the image is formed in the focal plane. As the object is moved closer to the lens, the image moves farther from the lens and increases in size. When the object is in the focal plane, the image is at infinity. If the object is inside the focal point, the image becomes larger than the object, erect, and virtual; and located on the same side of the lens as the object.

### 2.3.4.4 The size of the image

A practical example: You are fascinated by a beautiful Moon low on the horizon one cloudless night. You run inside, grab your camera featuring a ”normal” lens of f =50 mm. You take the picture, but the result may be a bit disappointing. Could you have estimated the size of the image of the Moon, knowing that the lunar diameter is 3476 km, and that the distance from the Earth to the Moon is 384 405 km?

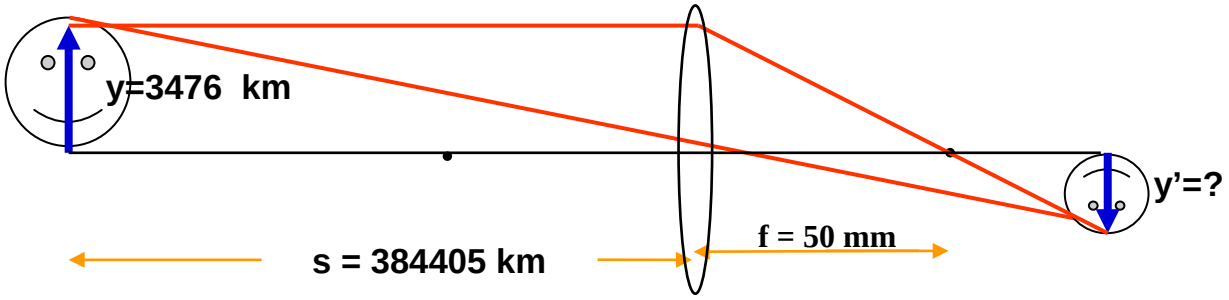



*Figure 2-38. Making an image of the Moon by a f =50mm lens.*

From the previous equation we get the diameter of the image of the Moon:

y’ = yf / (s-f) = 3476 km x 50 mm/ (384405 km - 50 mm) **= 0.45 mm.**

This rather modest image will only fill less than 0.02 % of the 24 x 36 mm film area!

### 2.3.4.5 ”The lensmakers equation”

We have already established that

$$\frac{n_a}{s} + \frac{n_b}{s'} = \frac{n_b - n_a}{R}$$

Obviously, if the image formed by refraction by one spherical surface is imaged by a second refracting surface, this expression has to be applied twice:

$$\frac{n_a}{s_1} + \frac{n_b}{s_1'} = \frac{n_b - n_a}{R_1}$$

$$\frac{n_b}{s_2} + \frac{n_a}{s_2'} = \frac{n_a - n_b}{R_2}$$

Where $s_1$ is the distance to the object and $s_2$' is the distance to the final image. For a thin lens, the distance between the two surfaces is small. This implies that $s_2$ and $s_1$' have the same magnitude but opposite signs: $s_2 = -s_1$'. Usually, the lens is used in air or in vacuum, so we may set $n_a = 1$, and the refractive index $n_b$ of the lens may simply be called n. Substituting all this we get

$$\frac{1}{s_1} + \frac{n}{s_1'} = \frac{n-1}{R_1}$$

$$-\frac{n}{s'_1} + \frac{1}{s_2'} = \frac{1-n}{R_2}$$

Adding the two equations we eliminate the term $n/s_1$', and get the inverse of the focal length of the thin lens:

$$d = \frac{1}{f} = \frac{1}{s_1} + \frac{1}{s'_2} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

We note that:

- The two focal lengths on either side of a thin lens are always equal, even when the two sides of the lens have different curvatures.
- For a symmetric convex lens – where $R_1$ and $R_2$ have equal magnitudes but opposite signs – the focal length is f = R/2(n-1).
- A very curved lens where the surfaces have a small radius of curvature will have a short focal length, while a thin lens with a large radius of curvature has a long focal length.

### 2.3.4.6 Correcting lens aberrations

Even under the paraxial approximation lenses suffer from aberrations that we do not find when imaging by curved mirrors (see section 2.1.8.2). The most important additional aberration is caused by the fact that for most transparent materials, the index of refraction depends on wavelength – as we saw in figure 2-13, so images will be formed at different distances from the lens, depending on wavelength, as illustrated in the left hand panel of figure 2-39.


*Figure 2-39. A compound lens to correct for chromatic aberration.*

The chromatic aberration can be minimized by using an **achromatic doublet** (or simply called an **achromat)** in which two materials with differing dispersion (usually crown and flint glass, see Table 2-1) are bonded together to form a single lens, as illustrated to the

right in figure 2-39. This reduces the amount of chromatic aberration over a certain range of wavelengths, though it doesn't produce perfect correction, see figure 2-40. By combining more than two lenses of different chemical composition, the degree of correction can be further increased, as seen in an *apochromatic lens* or *apochromat*, see figure 2-40. *Achromatic* lenses are corrected to bring two wavelengths (typically red and blue) into focus in the same plane. *Apo*chromatic lenses are designed to bring three wavelengths (typically red, green, and blue) into focus in the same plane. The residual color error (secondary spectrum) can be up to an order of magnitude less than for an achromatic lens of equivalent aperture and focal length. Apochromats are also corrected for spherical aberration at two wavelengths, rather than one as in an achromat. Obviously, the use of achromats and later apochromats were important steps in the development of telescopes and microscopes.

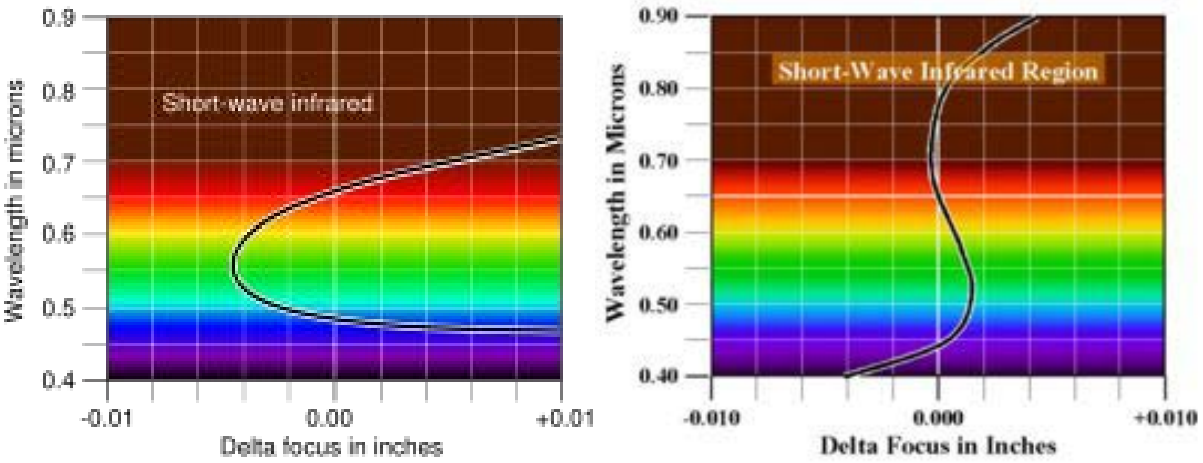



*Figure 2-40. Difference in focal length (x-axis) as a function of wavelength (y-axis). Left: An achromatic doublet brings two wavelengths to a common focus. Right: Apochromatic lens brings 3 wavelengths to a common focal plane. (From http://en.wikipedia.org/wiki/Chromatic_aberration and http://en.wikipedia.org/wiki/Apochromatic )*

The *Abbe number V* of a transparent material is a measure of the material's dispersion (variation of refractive index with wavelength), and is defined as:

$$V = \frac{n_D - 1}{n_F - n_C}$$

where $n_D$, $n_F$ and $n_C$ are the refractive indices of the material at the wavelengths of the Fraunhofer D-, F- and C- spectral lines (589.2 nm, 486.1 nm and 656.3 nm respectively). Materials with low dispersion have high values of *V*.

Abbe numbers are used to classify transparent materials. For example, flint glasses have $V < 50$ and crown glasses have $V > 50$. Typical values of *V* range from around 20 for very dense flint glasses, up to 65 for very light crown glass, and up to 85 for fluor-crown glass.

Abbe numbers are only a useful measure of dispersion for wavelengths in the visible range of the electromagnetic spectrum [2].

Abbe numbers are used to calculate the necessary focal lengths of achromatic doublet lenses to minimize chromatic aberration. For a doublet consisting of two thin lenses in contact, the Abbe number of the lens materials is used to calculate the correct focal length of the lenses to ensure correction of chromatic aberration. If the focal lengths of the two lenses for light at the yellow Fraunhofer sodium D-line (589.2 nm) are $f_1$ and $f_2$, then best correction occurs for the condition:

$$f_1 \cdot V_1 + f_2 \cdot V_2 = 0$$

where $V_1$ and $V_2$ are the Abbe numbers of the materials of the first and second lenses, respectively. Since Abbe numbers are positive, one of the focal lengths must be negative, i.e. a diverging lens, for the condition to be met.

The overall focal length of the doublet $f$ is given by the standard formula for thin lenses in contact:

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}$$

and the above condition ensures this will be the focal length of the doublet for light at the blue and red Fraunhofer F and C lines (486.1 nm and 656.3 nm respectively). The focal length for light at other visible wavelengths will be similar but not exactly equal to this.

Although chromatic aberration is most visible in color images, one should realize that it also affects so-called “black and white” photography – whether it is gray-tone or truly binary black and white. Although there are no colors in the photograph, chromatic aberration will blur the image because the light used is panchromatic. The chromatic blurring can be reduced by using a narrow band color filter, or by converting a single color channel to black and white. Obviously, using a narrow band filter will result in longer exposure times.

### 2.3.4.7 The camera

A camera consists of a light-tight box, a lens – often consisting of a number of elements, an adjustable aperture or diaphragm and a controllable shutter to regulate the exposure time interval. In the focal plane we find either a light sensitive photographic film or an array of electronic detectors.

To get a sharp image of the object, we have to make the image formed by the lens to coincide with the fixed focal plane of the camera. As the image distance increases with decreasing distance to the object, this means that we have to move the lens closer to the detector for a distant object and farther away from the detector for a nearby object.

[2] Alternative definitions of the Abbe number are used in some contexts, see http://en.wikipedia.org/wiki/Abbe_number.

In a zoom lens, several separate lens elements are mechanically assembled so that both the focus and the focal length can be varied, maintaining a fixed physical focal plane. Zoom lenses are often described by the ratio of their longest to shortest focal length: a 20 mm to 200 mm zoom is described as a 10:1 or "10×" zoom. Some digital cameras allow cropping and enlarging of a captured image, in order to emulate the effect of a longer focal length zoom lens. This is commonly known as digital zoom and always results in a lower quality image than optical zoom, as no optical resolution is gained.

A simple scheme for a zoom lens divides the lens assembly into two parts: a focusing lens similar to a standard, fixed-focal-length photographic lens, preceded by an *afocal zoom lens system*. The latter is an arrangement of fixed and movable lens elements that does not focus the light, but alters the size of a beam of light traveling through it, and thus the overall magnification of the lens system.

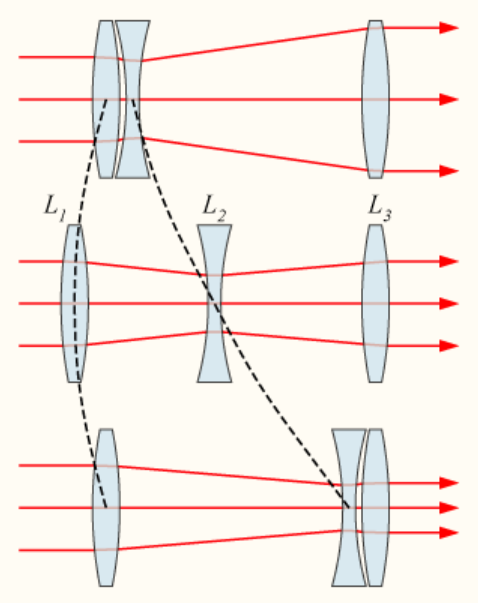



*Figure 2-41. The movement of three lenses in a simple afocal zoom lens system.*

In the simple afocal zoom lens system shown in figure 2-45 lens $L_3$ is fixed, while $L_1$ and $L_2$ are moved axially in a fixed, non-linear relationship, usually performed by a complex arrangement of gears and cams in the lens housing, although some modern zoom lenses use computer-controlled servos to perform this positioning.

An important issue in zoom lens design is the correction of optical aberrations (such as chromatic aberration and field curvature) across the whole range of focal lengths of the zoom lens. This is considerably harder in a zoom lens than a fixed lens, which need only correct the aberrations for one focal length. Modern optical design techniques have enabled the construction of zoom lenses with good aberration correction over widely variable focal lengths and apertures.

For a given size of the detector (film or CCD) the focal length will decide the field of view. If we are using a 24 x 36 mm film, the field of view measured along the diagonal will be 47º if we use a f = 50 mm lens. A focal length f = 28 mm will give a wide angle field of view (75º) which is nice for landscapes, f = 105 mm is ideal for portraits (25º), while a focal length f = 300 mm zooms in the field of view to a mere 8º (the diameter of the full moon is 1/2º). The field of view depends both on the focal length and the size of the detector in the focal plane. If we replace the 24 x 36 mm film with a smaller digital detector array, we may get a smaller registered field of view, see figure 2-42.

*Figure 2-42. Three different images of the same scene using focal lengths f = 18, 70, and 200 mm. The fields of view of this digital SLR camera illustrated in blue, green and red in the lower right hand aerial image correspond to approximately 1.5 times longer focal lengths in cameras using standard 35 mm photographic film (image area 24 x 36 mm).*

The focal length may alter the perspective. This is plainly visible if you use a wide angle lens to make a portrait. You will have to move in close to the object to fill the whole frame, and in an *en face* portrait the nose will be significantly closer than the rest of the face – and will look too big. Telephoto lenses, on the other hand, will compress depth within the scene.

So called ”normal objectives” give approximately the same perspective as our eyes. To obtain this, the focal length must be approximately equal to the diagonal in the image plane. For a 24 x 36 mm film the diagonal is 43 mm, and 45 – 50 mm are regarded as ”normal” focal lengths. Using a smaller detector chip in the focal plane of a digital camera you may get a normal perspective using a small lens having a shorter focal length, but the angular resolution will be poorer.

Zoom lenses give the possibility of manipulating the perspective in time sequences of images. If the camera is pulled away from the object while the lens zooms in so that the

field of view is maintained, or vice versa, the size of the foreground objects will be constant, but background details will change size relative to the foreground. This special effect was invented by Irmin Roberts, and was used by Alfred Hitchcock in Vertigo, hence the term Vertigo or Hitchcock zoom. The continuous perspective distortion is highly counter-intuitive, as the human visual system uses both size and perspective cues to judge the relative sizes of objects. Seeing a perspective change without a size change is a highly unsettling effect. The visual appearance for the viewer is that either the background suddenly grows in size and detail, overwhelming the foreground; or the foreground becomes immense and dominates its previous setting, depending on which way the zoom is executed.

If the film or the electronic detector is to record the image properly, the light energy per unit area in the focal plane must fall within certain limits. This is regulated by the shutter and the lens aperture. The f-number of the camera lens is simply the ratio of the focal length to the aperture diameter, f/D. The intensity of light reaching the focal plane is proportional to the square of the inverse of this ratio. Increasing the diameter by $\sqrt{2}$ changes the f-number by $1/\sqrt{2}$ and increases the intensity in the focal plane by a factor of 2. The f-numbers are often related by $\sqrt{2}$, such as f/2, f/2.8, f/4, f/5.6, f/8, f/11, f/16.

The total exposure (amount of light reaching the focal plane during the time interval that the shutter is open) is proportional to both the aperture area and the length of the exposure time interval. So 1/500 s at f/4, 1/250 s at f/5.6, and 1/125 s at f/8 all correspond to the same exposure. Shorter exposure times will minimize motion-blurring. A larger effective lens aperture will give better resolution of details in the image. And as we shall see, the depth of focus will increase with decreasing aperture.

## 2.3.4.8 Depth of field

The **depth of field** (DOF) is the distance in front of and beyond the object that we have focused on that appears to be in focus. Imaging a landscape, one may want a large DOF in order to have both the foreground and the background in focus. In a close-up image one may want a very shallow DOF to isolate an interesting object from a distracting background.

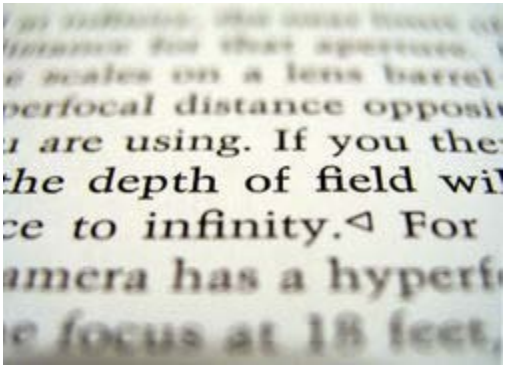



*Figure 2-43. Example of small DOF.*

For a given image format, the depth of field is determined by three factors: the focal length of the lens, the f-number of the lens aperture, and the camera-to-object distance. Increasing the f-number (reducing the aperture diameter) increases the DOF; however, it also reduces the amount of light transmitted, and increases diffraction, placing a practical limit on the extent to which the aperture size may be reduced.

### *2.3.4.8.1 Near and far limits of DOF*

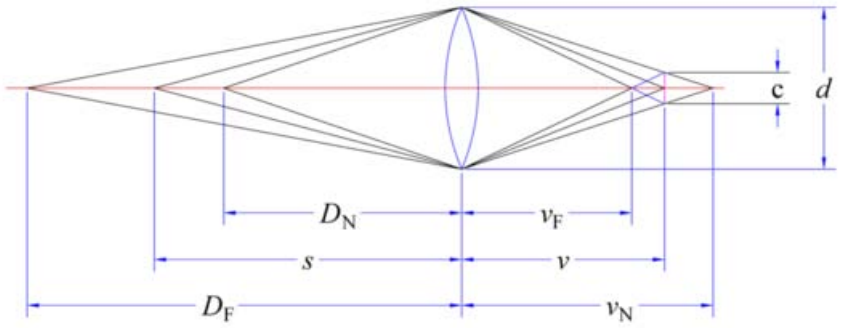



*Figure 2-44. Geometry of DOF, given the diameter of acceptable circle of confusion, c.*

Figure 2-44 gives the geometry needed to derive simple formulae for the near and far limits of the DOF given a thin lens having an aperture diameter d. An object at a distance *s* from the lens is in focus at an image distance *v*. Point objects at distances $D_F$ and $D_N$ would be in focus at image distances $v_F$ and $v_N$, respectively. However, at image distance *v*, they are both imaged as blurred spots. When the blur spot diameter is equal to the diameter of the acceptable circle of confusion *c (often abbreviated COC)*, the near and far limits of DOF are at $D_N$ and $D_F$. From similar triangles we see that

$$\frac{v_N - v}{v_N} = \frac{c}{d} \qquad \frac{v - v_F}{v_F} = \frac{c}{d}$$

When handling a camera, we usually don't compute the aperture diameter, but rather use the f-number of the lens, knowing its focal length. The *f*-number *N* is related to the lens focal length *f* and the aperture diameter *d* by N = f/d. Substituting d into the previous equations we get the focus limits on the image side of the lens

$$v_N = \frac{f\,v}{f - N\,c} \qquad v_F = \frac{f\,v}{f + N\,c}$$

The image distance *v* is related to a given object distance *u* by the thin-lens equation

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

Substituting this into the two previous equations give the near and far limits of DOF:

$$D_N = \frac{s\,f^2}{f^2 + N\,c\,(s - f)} \qquad D_F = \frac{s\,f^2}{f^2 - N\,c\,(s - f)}$$

The DOF beyond the object is always greater than the DOF in front of the object, but for longer focal lengths the fractions of the DOF in front of and behind the focus distance

tend towards being equal. For manual work, the DOF-limits are usually indicated on the camera objective; on zoom lenses as a function of both focal length and f-number. This is based on a fixed value of the COC diameter c. However, different applications may require different circles of confusion. In figure 2-45 one may find the far and near limits of the DOF for color coded values of the COC for a given focal length and f-number. For the 35-mm film format, a typical value for the acceptable COC is 30 µm. For this format the DOF range is bounded by the yellow color. When focusing on an object at 10 meters, the DOF thus ranges from 6 m to 30 m. The near DOF extends 4 meters and the far DOF 20 meters from the focus distance. A smaller/larger COC gives a larger/smaller DOF.

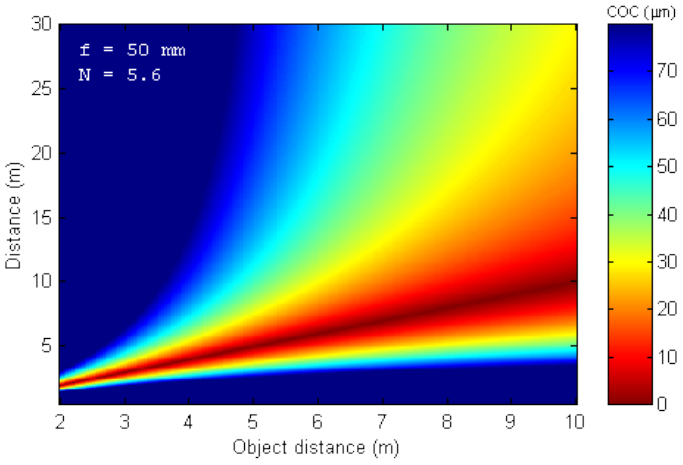



*Figure 2-45. Color coded DOF for different COC's. f = 50 mm and N = 5.6 (from [http://www.vanwalree.com/optics/dof.html](http://www.vanwalree.com/optics/dof.html) ) © Paul van Walree 2007.*

### *2.3.4.8.2 The hyperfocal distance*

The hyperfocal distance is the nearest focus distance at which the far limit of the DOF extends to infinity. Focusing the camera at the hyperfocal distance gives the largest possible DOF for a given *f*-number. Setting the far limit $D_F$ to infinity and solving for *s* gives

$$s = H = \frac{f^2}{N\,c} + f \approx \frac{f^2}{N\,c}$$

where *H* is the hyperfocal distance. The approximation is based on the fact that for any practical value of H, the focal length is very small in comparison.

Substituting the approximate expression for the hyperfocal distance into the formulae for the near and far limits of DOF gives

$$D_N = \frac{s\,H}{H + (s - f)} \qquad D_F = \frac{s\,H}{H - (s - f)}$$

Combining, the length of depth of field $D_F - D_N$ is given by

$$DOF = \frac{2sH(s - f)}{H^2 - (s - f)^2}, \quad s < H$$

#### *2.3.4.8.3 DOF at moderate-to-large object distances*

When the object distance is large in comparison with the focal length of the lens, the previous equations are simplified to

$$D_N \approx \frac{sH}{H+s} \qquad D_F \approx \frac{sH}{H-s}$$

$$DOF \approx \frac{2s^2H}{H^2-s^2}, \quad s<H$$

We notice that if the lens is focused at the hyperfocal distance, i.e. s = H, then

$$D_N \approx \frac{H}{2} \qquad D_F \approx \infty$$

For $s \geq H$, the far limit of DOF is at infinity and the DOF is infinite. However, only objects at or beyond the near limit of DOF will be recorded with acceptable sharpness.

Substituting $H \approx f^2/Nc$ into the above expression for the DOF, we get

$$DOF \approx \frac{2Ncf^2s^2}{f^4-N^2c^2s^2}$$

So, for a given image format (and thereby a given diameter of the circle of confusion, c), the DOF is determined by three factors: the focal length f of the lens, the f-number N of the lens opening, and the distance s to the object.

#### *2.3.4.8.4 Close up DOF*

When the distance *s* approaches the lens focal length, the focal length no longer is negligible, and the approximate formulae above cannot be used without introducing significant error. We therefore have to go back to

$$D_N = \frac{sf^2}{f^2+Nc(s-f)} \qquad D_F = \frac{sf^2}{f^2-Nc(s-f)}$$

From the object-image relation for a thin lens we may express s and (s-f) in terms of the magnification, m, and the focal length f:

$$s = \frac{m+1}{m}f \qquad (s-f) = \frac{f}{m}$$

Substituting this into the formula for DOF and rearranging gives

$$DOF = D_F - D_N = \frac{2f(m+1)/m}{mf/Nc - Nc/mf}$$

If the distance s is small, s<<H, the second term of the denominator becomes small in comparison with the first, and the DOF is independent of the focal length for a given m:

$$DOF \approx \frac{2Nc(m+1)}{m^2}$$

### 2.3.4.9 The eye

The eye is a complex anatomical unit consisting of several parts. The human eye is nearly spherical and about 2.5 cm in diameter. The front is somewhat more curved and is covered by the outer, transparent *cornea* that both protects the eye and performs much of the focusing. Behind that the *iris* acts as a stop, controlling how much light will be let through the *pupil* to the *lens*. The lens produces a sharp image of whatever we are looking at on the *retina*, where we find two types of photo detectors that in turn will send their signals via the optic nerve to the *visual cortex* located in the posterior part of the brain.

The refractive index of the cornea is 1.376, so most of the refraction in the eye takes place at the interface between the cornea and the surrounding air. The watery fluid behind the cornea (aqueous humor) has an index of refraction of 1.336 – similar to water – so its focusing effect will disappear if the eye is under water.

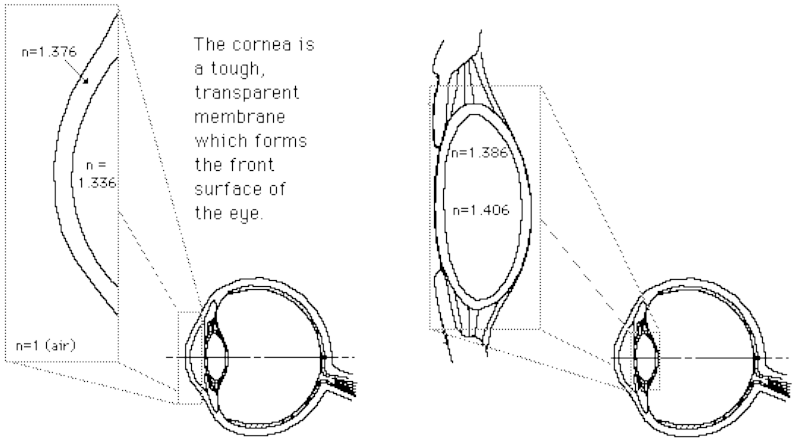



*Figure 2-46. The cornea and the eye lens.*
*(From [http://hyperphysics.phy-astr.gsu.edu/hbase/vision](http://hyperphysics.phy-astr.gsu.edu/hbase/vision) )*

The eye lens is responsible for approximately 20 % of the refraction in the eye. It has a diameter of about 9 mm and a thickness of 4 mm, and consists of a multitude of layers of transparent fibers, resulting in a variable refractive index – from n = 1.406 in the middle to 1.386 in the outer layers. It is soft, and fixed to muscles that can alter its shape, as shown in figure 2-46. The sphere behind the lens is filled with a watery gel, *vitreous humor*, having a refractive index like water, n = 1.336.

The lens system of the human eye has a focal length, f, of about 1.5 cm. The focusing power of the lens is often given as ”dioptres”, d, where d = 1/f and the focal length f is given in meters. The eye lens is usually 67 dioptres ($1/1.5 \times 10^{-2}$ m ≈ 67 $m^{-1}$), of which the cornea is responsible for 45 dioptres.

In their relaxed mode, the radial ligaments around the lens will stretch it to a flattened disc, so that it focuses on far-away objects. If the eye is to create a sharp image of an object closer than this, the ring-shaped muscle around the radial fibers is relaxed, the lens becomes more spherical and its focal length is shortened. The distance from the lens to the retina is preserved, contrary to lenses having a fixed focal length.

The ability to alter the focal length of the eye lens – *accommodation* – is quite automatic, but is affected by ageing. For a young and healthy individual, accommodation may alter the focal power of the eye by up to 4 dioptres. The extremes of the range of distances over which distinct vision is possible are known as the *far point* and the *near point* of the eye. The far point of a normal eye is at infinity. The position of the near point depends on the ability of the ciliary muscle to increase the curvature of the lens. And because the lens grows with age the muscles are less able to alter its curvature, and the near point recedes as one grows older (*presbyopia*), from 10 cm at age 20 to about 200 cm at age 60.

If the eyeball is too long from front to back in comparison with the radius of curvature of the cornea, rays from an object at infinity are focused in front of the retina. The far point of such a nearsighted (*myopic*) eye is then nearer than infinity. In a farsighted *(hyperopic*) eye, the eyeball is too short or the cornea is not curved enough. The sharpest image of an infinitely distant object is behind the retina. Both defects can be corrected by eyeglasses or contact lenses, as shown in figure 2-47.

To correct *myopic* vision we use a diverging lens to move the virtual image of a distant object at or inside the far point of the eye, so that the eye lens will give a sharp image. *Hyperopic* vision is corrected by a converging lens. This will form a virtual image of a nearby object at or beyond the eye's near point, so that a sharp retinal image may be formed. Thus, the effective near point of a *presbyopic* or *hyperopic* eye may be corrected to a normal reading distance of 25-30 cm. *Astigmatism* is caused by a cornea that is not spherical. If it is more curved in one plane than in another, it will not focus sharply to the same distance in the two planes. This is remedied by lenses that are not rotationally symmetric.

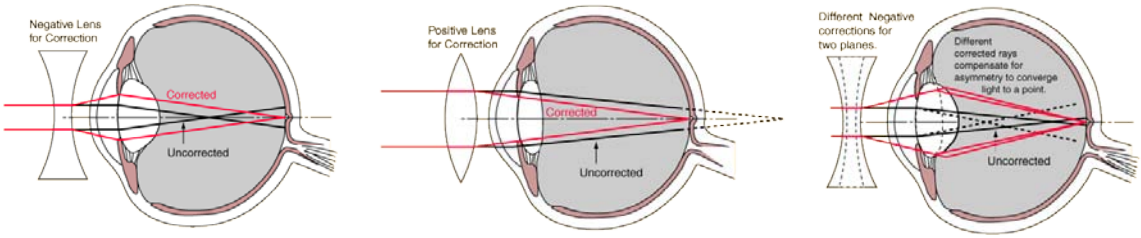



*Figure 2-47. Nearsightedness (myopia), farsightedness (hyperopia) and astigmatism. (From [http://hyperphysics.phy-astr.gsu.edu/hbase/vision](http://hyperphysics.phy-astr.gsu.edu/hbase/vision) )*

We are actually using the concept of dioptres in everyday conversations about eyeglasses. If farsightedness is corrected by a +3.0 eyeglass, we are using a converging lens having a focal length f =1/d =1/3 m. A negative sign indicates that it is a divergent lens. Modern eyeglasses may also be *bifocal*; i.e. divided into an upper half that corrects for distance vision, and a lower half that corrects for near vision, or *progressive*; i.e. a gradual blending of focal lengths.

### 2.3.4.10 The magnifier

The apparent size of an object depends on the angle θ subtended by the object at the eye. To get a better look at an object, we could bring it closer to the eye, making its angular size and the retinal image larger. But as objects inside the near point are not focused sharply on the retina, the largest possible viewing angle is obtained at a distance of approximately 25 cm.

A converging lens may be used to form a virtual image that is larger and farther from the eye than the object. Then the object may be moved inside the near point, and the angular size of the virtual image may be much larger than the angular size of the real object seen by the unaided eye at 25 cm, as shown in figure 2-48. A lens used in this way is called a magnifier. The angular magnification M of a magnifier is given by the ratio of the angle θ' (with the magnifier) to the angle θ (without the magnifier). As it is most comfortable to view a virtual image with a completely relaxed ciliary muscle, i.e. when it is placed at infinity, the magnification given often assumes that this is done.

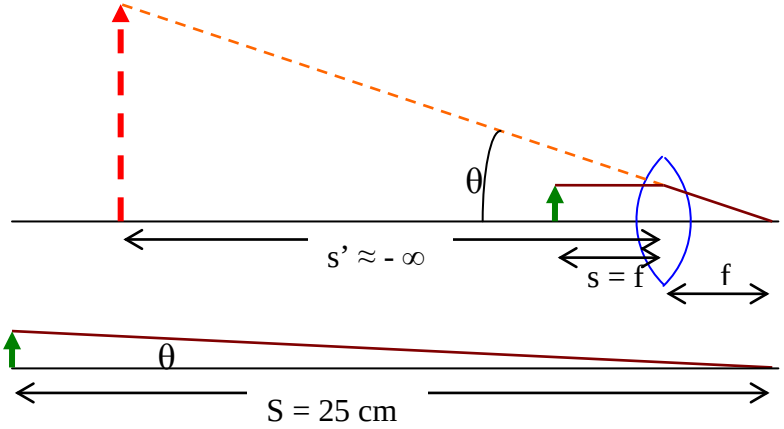



*Figure 2-48. Geometry of a magnifier (top) compared to the naked eye (bottom).*

Assuming that the angles are small enough that an angle (given in radians) is equal to its sine and its tangent, we see that

$$\theta = \frac{y}{25\,cm} \qquad \theta' = \frac{y}{f}$$

$$M = \frac{\theta'}{\theta} = \frac{y/f}{y/25cm} = \frac{25cm}{f}$$

From this expression we may get the impression that we can make the angular magnification as large as we desire by simply decreasing the focal length of the magnifier. In practice, optical aberrations of simple double-convex lenses set a limit to M of about 4 or 5. If these aberrations are properly corrected, we may use a focal length as short as f = 1.25 cm, giving an angular magnification M = 20. If we need more magnification than this, we will have to use a more complex arrangement of lenses.

### 2.3.4.11 The eyepiece

Both concave mirrors and concave lenses form real images. If we are to inspect this image visually, we may use a second lens, an eyepiece or ocular, as a magnifier to produce an enlarged virtual image. We find this application of a magnifier both in telescopes and microscopes.

### 2.3.4.12 Microscopes

The essential elements of a compound microscope are shown in figure 2-49. The object to be viewed is placed just outside the first focal point of the objective, a converging lens that forms a real, inverted and enlarged image. This image is placed just inside the first focal point of the second converging lens – the eyepiece or ocular – forming a final virtual image. The position of the final image may be anywhere between the near and far points of the eye. The objective and ocular of an actual microscope are compound lenses, corrected for chromatic and geometric aberrations.

The overall angular magnification of a compound microscope is the product of two factors; the lateral magnification $m_1$ of the objective determines the linear size of the real, inverted image, and the angular magnification $M_2$ of the ocular. The lateral magnification of the objective is $m_1 = -s_1'/s_1$. If we place the object close to the focal point, we have $m_1 = -s_1'/f_1$. The angular magnification of the eyepiece is $M_2 = (25\ \text{cm})/f_2$, assuming that the real image is placed close to the focal plane. Thus, the total angular magnification M (ignoring the negative sign) is the product of the two magnifications:

$$M = m_1 M_2 = \frac{(25cm)s_1'}{f_1\ f_2}$$

Usually, a number of objectives of different focal lengths are mounted in a rotating turret, so that an object may be viewed at different magnifications.

What if the real image formed by the objective is not placed so that the final virtual image formed by the eyepiece is at infinity, but at an average viewer's near point?

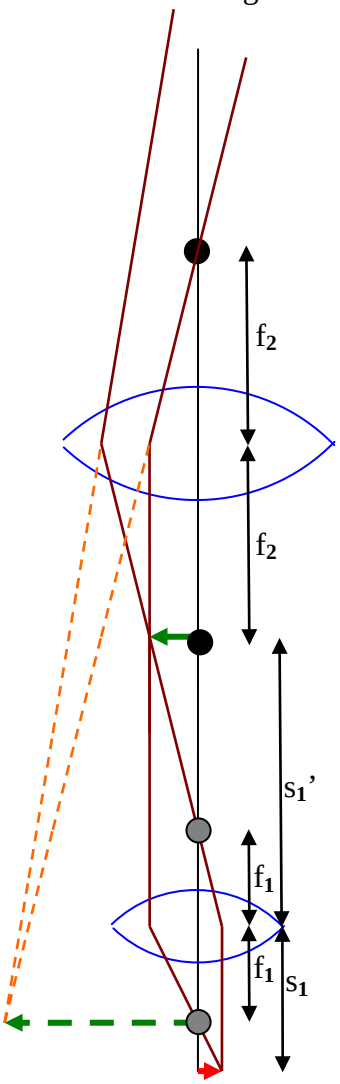



*Figure 2-49. Schematic geometry of a microscope.*

### 2.3.4.13 Telescopes

The arrangement of the lenses in a telescope are similar to that of a compound microscope, except for the fact that telescopes are used to view large objects at large distances, while microscopes are used to view very small objects at very short distances. In addition, a telescope may use either a concave mirror or a convex lens as objective, while a microscope uses lenses.

In astronomical telescopes, it doesn't matter that the image is inverted. Figure 2-50 shows a refractor where the objective lens forms a real, inverted and reduced image that is viewed through an eyepiece lens, producing an enlarged virtual image. If the object is sufficiently distant, the real image is formed at the second focal point of the objective lens. If we want the final virtual image to be formed at infinity – for most comfortable viewing by a normal eye – this image must also be placed at the first focal point of the eyepiece, as in figure 2-50. Then the distance between the objective and the eyepiece lenses is the sum of the two focal lengths.

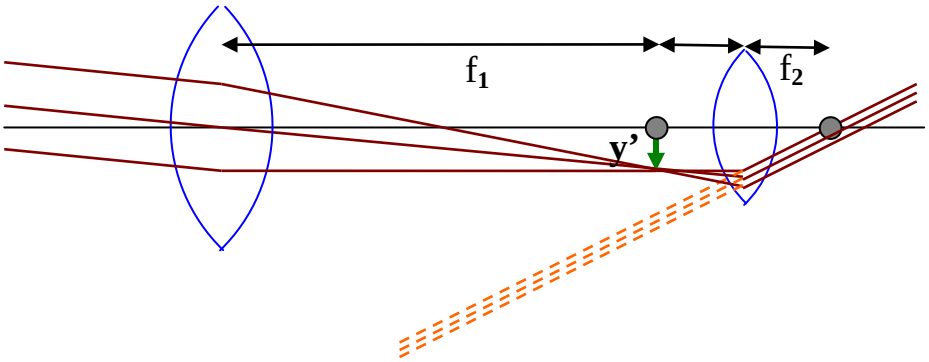



*Figure 2-50. Geometry of a refracting telescope.*

The angular magnification of a telescope is given by the ratio of the angle subtended at the eye by the final virtual image to the angle of the object when viewed by the unaided eye. As with the magnifier, the former is given by $\theta' = y'/f_2$, while the latter is given by $\theta = -y'/f_1$ (the angles are so small that we approximate them by their tangents). The angular magnification of the telescope is then

$$M = \frac{\theta'}{\theta} = -\frac{y'/f_2}{y'/f_1} = -\frac{f_1}{f_2}$$

The negative sign indicates that the image is inverted. So while a microscope should have a short objective focal length, a telescope should have a long objective focal length. And while we change objectives in the microscope to alter the magnification, we use different eyepieces with different focal lengths to vary the magnification of a telescope.

In reflecting telescopes, the objective lens is replaced by a concave mirror. In large telescopes this has several advantages: mirrors are inherently free from chromatic aberrations, the spherical aberrations associated with the paraxial approximation are easier to correct than with a lens, and the problems of production and mechanical support of a large mirror are much easier to handle than for a lens of the same size.

### 2.3.4.14 Multiple lens systems

As we saw earlier, lenses may be combined to form more complex optical systems. The simplest case is when lenses are placed in contact, as we saw when discussing corrections for chromatic aberrations (section 2.3.4.6). If two lenses of focal lengths $f_1$ and $f_2$ are "thin", the combined focal length $f$ of the lenses is given by:

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \quad \Rightarrow \quad f = \frac{f_1 f_2}{f_1 + f_2}$$

However, if two thin lenses are separated by some distance $d$, the distance from the second lens to the focal point of the combined lenses is called the *back focal length* (BFL). This is given by:

$$BFL = \frac{f_2(d - f_1)}{d - (f_1 + f_2)}$$

Note that as $d$ tends to 0, i.e. two lenses in contact, the value of the BFL tends to the value of $f$ given above for two thin lenses in contact.

If the separation distance is equal to the sum of the focal lengths ($d = f_1 + f_2$), the BFL is infinite. This corresponds to a pair of lenses that transform a parallel (collimated) beam into another collimated beam. This type of system is called *afocal*, since it produces no net convergence or divergence of the beam, as we saw when discussing the frontal part of a zoom lens (section 2.3.4.7). Two lenses at this separation also form the simplest type of optical telescope, as we saw in the previous section.

Although such a system does not alter the divergence of a collimated beam, it does alter the width of the beam. The magnification is given by:

$$M = -\frac{f_1}{f_2}$$

which is also equal to the ratio of the input beam width to the output beam width. Note that a telescope with two convex lenses ($f_1 > 0$, $f_2 > 0$) produces a negative magnification, indicating an inverted image. A convex plus a concave lens ($f_1 > 0 > f_2$) produces a positive magnification and the image is upright.

Assume that the converging lens has a focal length $f_1$ while the diverging lens has a negative focal length $f_2$. The two lenses are separated by a variable distance $d$ that is alwas less than $f_1$, and the magnitude of the focal length of the diverging lens satisfies the inequality $|f_2| > ( f_1 - d )$. The effective focal length is given by $f = f_1 |f_2| / (|f_2| - f_1 + d)$. Zoom lenses are often described by the ratio of their longest to shortest focal length. In this case, the maximum effective focal length occurs when the lenses are in contact, giving a "zoom factor" that is simply $f_{max}/f_{min} = 1 + (d / (|f_2| - f_1))$.

## 2.4 Diffraction

We observe the phenomenon of waves bending around corners every day, e.g. when we hear sound from sources that are out of sight around a corner. Light can also bend around corners. When light from a point source falls on a straight edge and casts a shadow, the edge of the shadow is not a perfectly sharp step edge. Neither is the shadow of the edge just smeared out. There is some light in the area that we expected to be in the shadow, and we find alternating bright and dark fringes in the illuminated area close to the edge. This is the result of interference between many light waves (Huygens' Principle). Such effects are referred to as diffraction.

In the following we will describe the intensity patterns that are observed when light waves pass through a single slit, through two parallel slits, and through multiple parallel slits. We will also describe the pattern observed as a result of light passing through a circular and an annular aperture.

It is customary to distinguish between two descriptions of diffraction:

- **Fresnel (near-field)** diffraction occurs when both the light source and the observation plane are close to the aperture. Since the wave fronts arriving at the aperture and observation plane are not planar, the curvature of the wave fronts must be taken into account. We will discuss Fresnel diffraction at a later stage.
- **Fraunhofer (far-field)** diffraction is observed when the wave fronts arriving at the aperture and the observation plane may be considered planar. This is usually taken to imply that both the light source and the observation plane must be sufficiently far away from the slit to allow all light rays to be parallel. However, a thin lens having the light source in its primary focal point will collimate the beam before it reaches the aperture; and similarly a lens behind the aperture may collimate the light beam traveling towards the observation plane.

The nearfield/farfield limit is the same as the hyperfocal distance (see section 2.3.4.2).

### 2.4.1 Fraunhofer diffraction pattern from a single slit

The intensity of the Fraunhofer (far-field) diffraction pattern formed by monochromatic light passing through a long, narrow slit consists of a central bright band, which may be much wider than the width of the slit, bordered by parallel alternating dark and bright bands with rapidly decreasing intensity. The mathematical expression is

$$I(\theta \mid a, \lambda) = \left( \frac{\sin\left[\pi a (\sin\theta)/\lambda\right]}{\pi a (\sin\theta)/\lambda} \right)^2 I_0$$

where $a$ is the width of the slit, $\lambda$ is the wavelength, and $\mathbf{I_0}$ is the intensity at $\theta = 0$. About 85% of the power of the transmitted beam is found in the central bright band. For visible light, the wavelength $\lambda$ is usually smaller than the slit width $a$, and $\sin(\theta) \approx \theta$. With this approximation, the first minimum of the Fraunhofer diffraction pattern occurs at $\theta = \lambda/a$. So the width of the central bright band is inversely proportional to the ratio of the width of the slit $a$ to the wavelength $\lambda$, as illustrated in figure 2-51 for a ± 22.5° range of $\theta$.

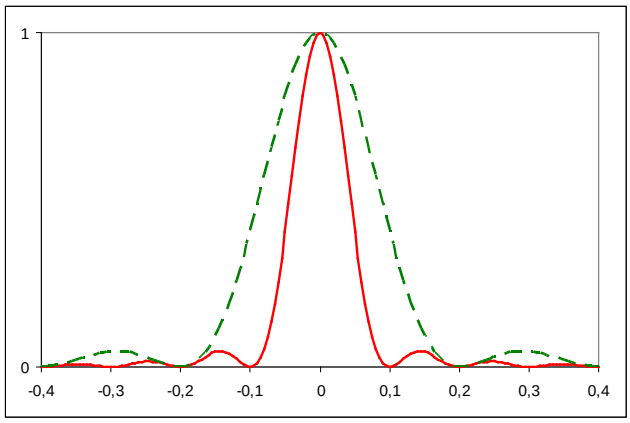



*Figure 2-51. Fraunhofer diffraction pattern of single slit, for -π/8 < θ <π/8.*
*For two slit width to wavelength ratios: a = 10 λ (solid line) and a = 5 λ (dashed line)*

### 2.4.2 Fraunhofer diffraction pattern from two slits

The intensity of the diffraction pattern from two slits of width a, separated by a distance d, formed by monochromatic light at a wavelength λ, is given by the expression

$$I(\theta \mid a,d,\lambda) = \cos^2\left(\pi\, d(\sin\theta)/\lambda\right)\left(\frac{\sin\left[\pi a(\sin\theta)/\lambda\right]}{\pi a(\sin\theta)/\lambda}\right)^2 I_0$$

Figure 2-52 illustrates this for a ± 11.25° range of θ, for a slit width to wavelength ratio a/λ = 10 and a slit distance (center-center) d = 4a. The single-slit pattern of figure 2-51 acts as an envelope on the cosine interference pattern of the two slits. Since d = 4a in this case, every fourth maximum outside the central maximum in the interference pattern is missing because it coincides with the single slit diffraction minima. Again, the width of the envelope is inversely proportional to the ratio of the slit width *a* to the wavelength λ.

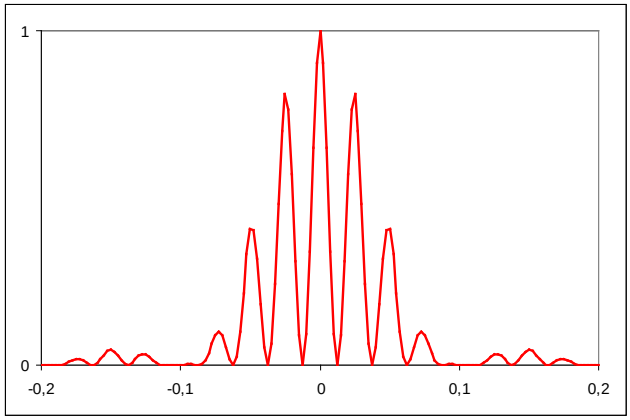



*Figure 2-52. Fraunhofer diffraction pattern of two slits, for -π/16 < θ <π/16.*
*The slit width to wavelength ratio is a/λ = 10 and the distance between the slits is d = 4a.*

### 2.4.3 Diffraction from several parallel slits

If the slits are narrow in comparison to the wavelength, constructive interference occurs for rays at an angle θ to the normal that arrive with a path length difference between adjacent slits equal to an integer number of wavelengths

$$d \cdot \sin(\theta) = m\lambda, \quad m = 0, \pm 1, \pm 2, \ldots$$

The normalized intensity of the diffraction pattern from N slits of width a, separated by a distance d, formed by monochromatic light at a wavelength λ, is given by the expression

$$I(\theta \mid a, d, \lambda, N) = \left[ \frac{\sin\left(\frac{N\pi d}{\lambda}\sin\theta\right)}{\sin\left(\frac{\pi d}{\lambda}\sin\theta\right)} \right]^2 \cdot \left( \frac{\sin\left[\frac{\pi a}{\lambda}\sin\theta\right]}{\frac{\pi a}{\lambda}\sin\theta} \right)^2 I_0$$

For N = 8, this results in a pattern as illustrated in figure 2-53 for a ± 2.3° range of θ, for a slit width to wavelength ratio a/λ = 10 and a slit distance (center-center) d = 4a.

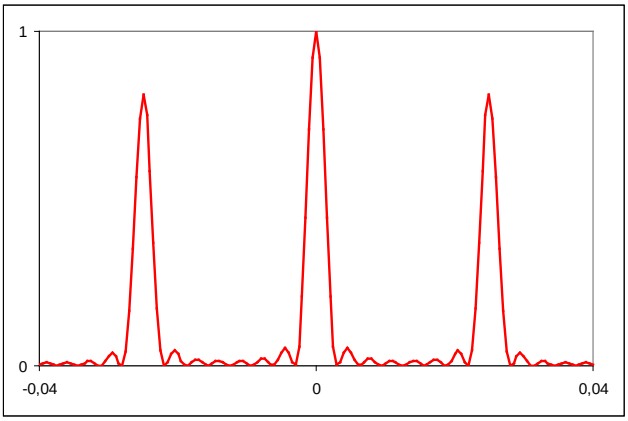



*Figure 2-53. Fraunhofer diffraction pattern of eight slits, for -π/80 < θ <π/80. The slit width to wavelength ratio is a/λ = 10 and the distance between the slits is d = 4a.*

The principal maxima are in the same positions as in the two-slit case, but get narrower as the number of slits, N, is increased. There are N-1 minima between each pair of principal maxima, and the small secondary maxima get smaller as N increases. The height of each principal maximum is proportional to $N^2$, so from simple energy conservation the width of each principal maximum is proportional to 1/N.

Because the principal maxima are in the same position but get sharper and narrower as N is increased, their position, and hence the wavelength, can be determined with very high precision. This is utilized in diffraction gratings.

### 2.4.4 The diffraction grating

A diffraction grating is an assembly of narrow slits or grooves, which by diffracting light produces a large number of beams which can interfere in such a way as to produce spectra. Since the angles at which constructive interference are produced by a grating depend on the wavelengths of the light being diffracted, the various wavelengths in a beam of light striking the grating will be separated into a number of spectra, produced in various orders of interference on either side of an undeviated central image.

Reflection gratings are usually designated by their *groove density*, expressed in grooves or lines per millimeter. The dimension and period of the grooves must be on the order of the wavelength in question. Gratings for visible light (λ from 400 to 700 nm) usually have about 1000 lines per mm, corresponding to d on the order of 1/1000 mm = 1000 nm.

When a beam is incident on a grating with an angle $\theta_i$ (measured from the normal of the grating), it is diffracted into several beams. The beam that corresponds to direct specular reflection is called the zero order, and is denoted $m = 0$. The other orders correspond to diffraction angles which are represented by non-zero integers $m$ in the **grating equation**. For a groove period $d$ and an incident wavelength $\lambda$, the **grating equation** gives the value of the diffracted angle $\theta_m(\lambda)$ in the order $m$:

$$d \cdot \left[\sin\theta_m(\lambda) + \sin\theta_i\right] = m\lambda, \quad m = 0, \pm 1, \pm 2, \ldots$$

### 2.4.5 A simple slit spectrograph design

An actual special-purpose spectrograph may be incredibly complex in order to avoid internal reflections and unwanted straylight, but the basic principle of a reflection grating slit spectrograph is very simple, as illustrated in figure 2-54. Light is focused onto the slit. At the rear end of the spectrograph a slightly tilted concave collimating mirror reflects the light onto a plane reflecting diffraction grating. Dispersed light of some order $m$ from the grating is focused by a second concave mirror onto a detector array or photographic film.

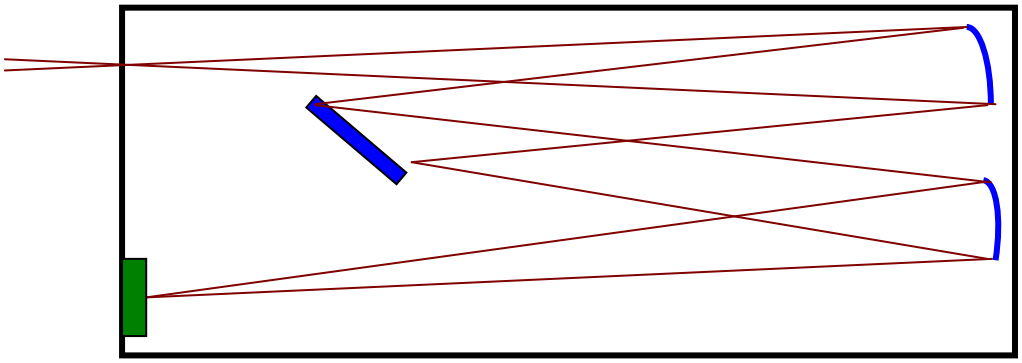


*Figure 2-54. A reflection grating slit spectrograph.*

### 2.4.6 A simple spectroheliograph

A spectroheliograph produces monochromatic images of the Sun. In the simplest form of the instrument, an image of the Sun from a solar telescope is focused on a plane where a narrow slit lets light into a spectrograph. At the rear end of the spectrograph a slightly tilted concave collimating mirror reflects the light coming from the slit back onto a plane reflecting diffraction grating. Part of the dispersed light from the grating is focused by a second concave mirror, identical to the first mirror, onto an exit slit identical to the entrance slit. By symmetry of the optical system, the portion of the solar disk imaged on the entrance slit is reimaged in the plane of the exit slit with the same image scale but in dispersed wavelength. The light imaged along the exit slit then corresponds to the portion of the solar image falling on the entrance slit, but in the light of only a narrow region of the spectrum, as determined by the spectrographic dispersion. So the spectral dispersion and the width of the exit slit determine how "monochromatic" the output will be. The dispersion is determined by the grating characteristics and the spectral order, while the particular wavelength sampled is set by the grating angle.

By letting the image of the Sun move in a uniform transverse motion the entrance slit is scanned across the solar image. By moving a photographic film behind the exit slit at the same speed as the image is moving across the entrance slit – or performing a corresponding sequential readout of a linear array of digital sensors behind the slit - a monochromatic image of the Sun is recorded. Such images are routinely made – both ground-based and from satellites – in order to study the solar magnetic field as well as active regions, flares and solar eruptions etc., as illustrated in figure 2-55.

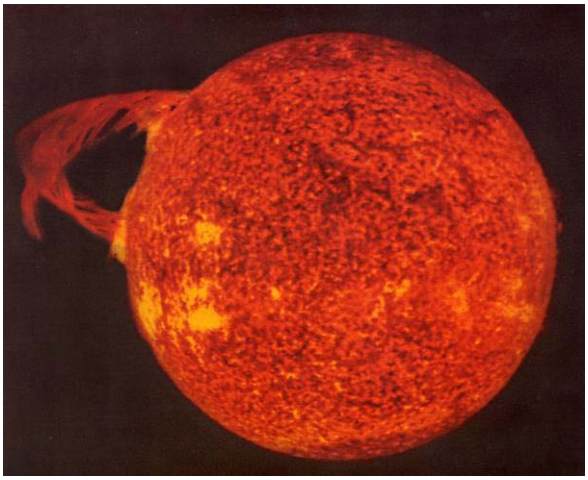


*Figure 2-55. Spectroheliograph image at the EUV He II emission line at 30.4 nm. US Naval Research Laboratory pseudo-color image from the 1973 Skylab mission.*

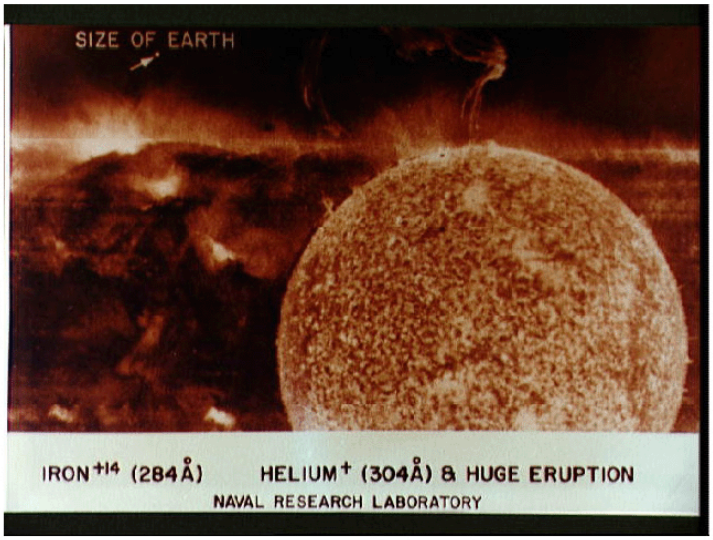
### 2.4.7 A slitless spectrograph

One of the difficulties of spectroscopy lies in obtaining co-temporal spectra of all parts of extended objects. An ordinary slit spectrograph will only give the spectrum of that particular section of the image of an extended object that falls on the slit. So if we are observing e.g. the Sun, we could obtain a large number of spectra by "stepping" a vertical slit along the x-axis of the solar image. Each image would contain the spectrum within a given wavelength region of that part of the solar atmosphere that was imaged onto the slit at that particular time, and the angular resolution in the y-direction would be given by the resolution of the telescope. However, the time it takes to acquire these images may be longer than the lifetime of some of the phenomena that we want to study.

The same problem is inherent in the spectroheliographic technique that we just described: the various parts of the image are not acquired at the same time.

One way of obtaining co-temporal spectra is to use a slit-less spectrograph. A concave grating produces a sharp image and a spectrum at the same time. Then we can get a co-temporal picture of how the intensity at each point on the Sun varies as a function of x, y, and $\lambda$.

But now we have a different problem. Each emission line in the spectrum will produce an image, so the spectral and spatial information are convolved into a complicated image that we may call an "overlappogram". And the fact that the emission obviously varies across the image – and that this variation is not at all the same from one wavelength to another can lead to some very complicated overlapping, as illustrated by figure 2-56.




Part of an "overlappograph" image taken by the S082A EUV spectro-heliograph on the Skylab Apollo Telescope Mount, January 1, 1974 (*JSC Digital Image Collection*)

*Figure 2-56. Overlapping EUV emission line images from slitless spectrograph. US Naval Research Laboratory pseudo-color image from the 1974 Skylab mission.*

## 2.4.8 The diffraction profile of rectangular apertures

We saw in section 2.4.1 that the Fraunhofer diffraction pattern formed by monochromatic light passing through a long, narrow slit consists of a central bright band bordered by parallel alternating dark and bright bands with rapidly decreasing intensity. If light is passing through a rectangular aperture, we get a product of two such orthogonal 1-D diffraction patterns given by

$$I(\theta,\varphi \mid a,\lambda) = \left(\frac{\sin[\pi a(\sin\theta)/\lambda]}{\pi a(\sin\theta)/\lambda}\right)^2 \left(\frac{\sin[\pi b(\sin\varphi)/\lambda]}{\pi b(\sin\varphi)/\lambda}\right)^2 I_0$$

where *a* and *b* are the width and length of the aperture, λ is the wavelength, and $I_0$ is the intensity at (θ,φ) = (0,0). The horizontal and vertical widths of the central bright band are inversely proportional to the ratio of the size of the aperture (*a,b)* to the wavelength λ, as illustrated in figure 2-57 for a ± 0.4 radians range of θ and φ.

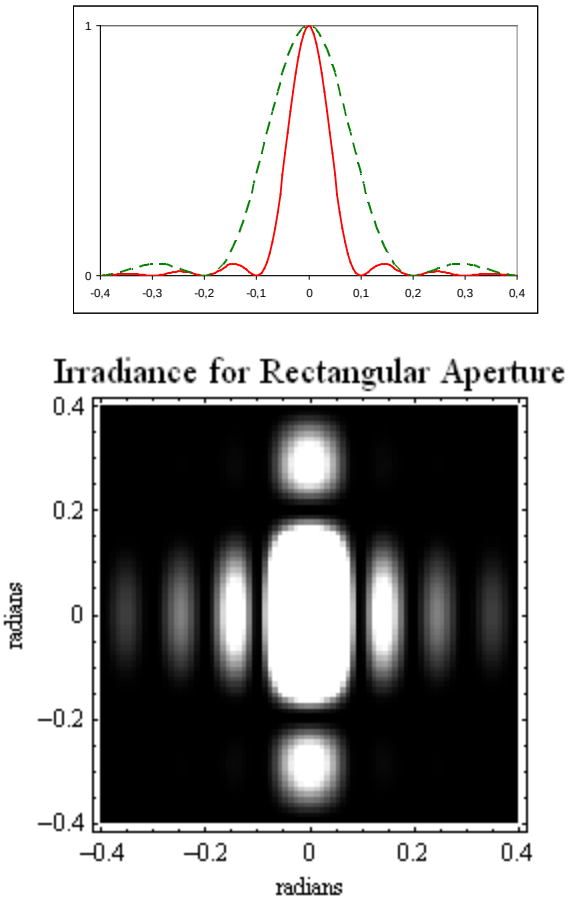


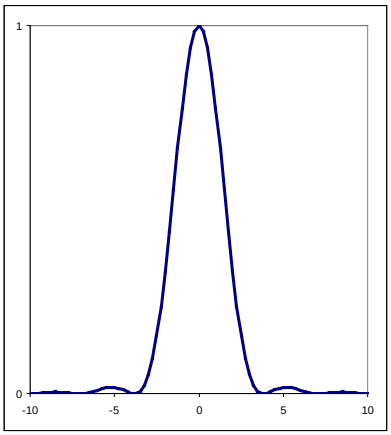

*Figure 2-57. Top: Two 1-D Fraunhofer diffraction patterns for two slit width to wavelength ratios: a = 10 λ (solid line) and b = 5 λ (dashed line) for -0.4 < θ <0.4. Bottom: The 2-D diffraction pattern from a rectangular aperture where a = 10 λ is the horizontal and b = 5 λ is the vertical size of the aperture.*

### 2.4.9 The diffraction profile of circular apertures

The phenomenon of optical diffraction sets a limit to the resolution and image quality that a telescope can achieve. Even with perfect optics, an image-forming lens or a mirror will not produce an infinitely small image of a point source. The diffraction pattern of the aperture will cause the point source to be imaged as a small disc, surrounded by alternating dark and bright rings, where the intensity of the rings decrease rapidly with distance from the central disc. The angular point spread function (PSF) of a circular aperture is given by the continuous circular symmetric function

$$F_0(\theta) = 4\left[ J_1\left(\pi \frac{\theta}{\beta_0}\right) \Big/ \pi \frac{\theta}{\beta_0} \right]^2$$

where θ is given in radians, $J_1$ is the first order Bessel function, $\beta_0$ is the ratio between the wavelength λ and the aperture diameter D: $\beta_0 = \lambda/D$, and $F_0(r)$ is normalized to $F_0(0) = 1$. Figure 2-58 shows a section through such a diffraction profile. The unit on the horizontal axis is $\pi D\theta/\lambda$, where θ is given in radians, while λ and D are given in meters.



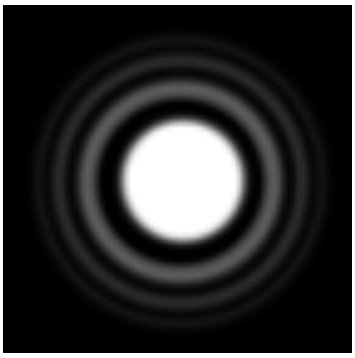



*Figure 2-58. Left: Section through the diffraction profile of a circular aperture. Right: The 2-D diffraction pattern of a circular aperture.*

#### 2.4.9.1 Airy disk and Rayleigh criterion

The angular radii of the first dark rings are given by

$$\sin\theta_1 = 1.22\frac{\lambda}{D} \quad \sin\theta_2 = 2.23\frac{\lambda}{D} \quad \sin\theta_3 = 3.24\frac{\lambda}{D}$$

Between these are the bright rings with angular radii given by

$$\sin\theta = 1.63\frac{\lambda}{D} \quad \sin\theta = 2.68\frac{\lambda}{D} \quad \sin\theta = 3.70\frac{\lambda}{D}$$

The central bright spot is called the Airy disk, in honor of George Airy who first derived the expression for the intensity of the pattern. About 85% of the light energy falls within the Airy disk. The peak intensity of the first ring is only 1.7% of the value at the center of the disk, and the peak intensity of the second ring is merely 0.4%.

Figure 2-59 shows a cross-section through a diffraction-limited image of two equally bright point sources at infinity (e.g. two stars) viewed at a wavelength λ, through a lens of diameter D, when the angular separation between the point sources is given by sin(θ) = 1.22 λ/D. This corresponds to overlaying two diffraction patterns so that the maximum of the first corresponds with the first minimum of the second. We can still observe that there are two sources, because there is a 27 % ”dip” in the intensity between the two peaks.

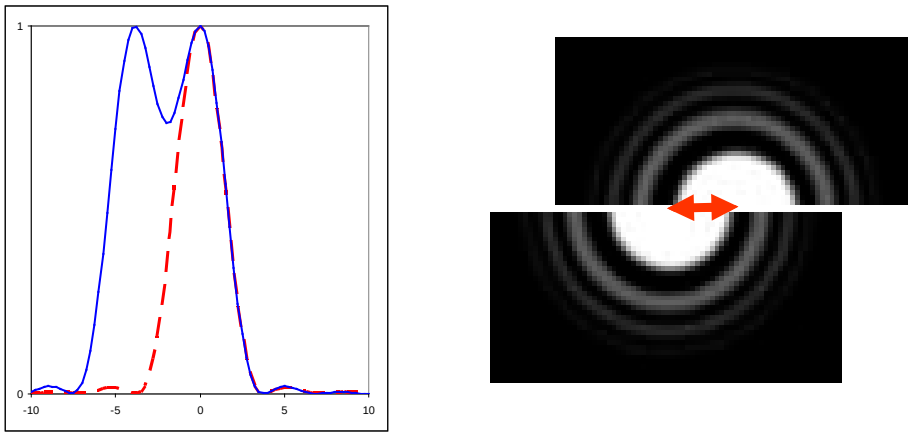



*Figure 2-59. Section through image of two point sources separated by sin(θ) = 1.22 λ/D.*

This corresponds to the so-called ”Rayleigh criterion”. According to this, two objects are barely resolved if the angular separation of the object centers is given by

$$\sin\theta = 1.22\frac{\lambda}{D}$$

### 2.4.9.2 The Sparrow criterion

If we move the point sources closer to each other than the angle given by the Rayleigh criterion, the “dip” between the two peaks will gradually become shallower, until the ”dip” becomes a flat plateau. This angular separation is called the Sparrow criterion, and may be seen as the limit when the images of two point sources “melt together”. Figure 2-60 shows a cross-section through the intensity profile of both one point source and two point sources separated by the Sparrow criterion of θ = 0.952 λ/D.

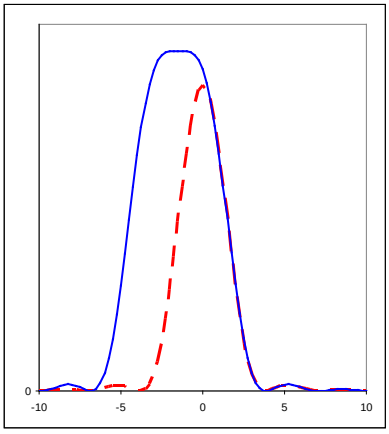



*Figure 2-60. The Sparrow criterion.*

### 2.4.9.3 Apertures having a central obstruction

Large mirror telescopes as well as large camera lenses have annular apertures, i.e. the secondary mirror constitutes a central obstruction. The diffraction profile of such apertures will differ somewhat from the profile of an unobstructed aperture.

The central obstruction is often given as a fraction δ of the full diameter D. Normal values of δ range from 0.15 to 0.33, while specialized instruments dedicated to deep-sky astrophotography can be obstructed more than 40% of the diameter.

One effect of an obstruction is that the amount of light received by the instrument is decreased. When a central fraction δ of the diameter is obstructed, the corresponding fraction of the surface being obstructed is obviously given by $\delta^2$. So an annular telescope aperture of diameter D having an obstruction δ will give the same amount of light to the focal plane as an unobstructed aperture with diameter $D' = D(1-\delta^2)^{1/2}$. The diffraction profile of an annular aperture is given by the continuous circular symmetric function

$$F(\theta \mid \delta) = \frac{1}{(1-\delta^2)^2}\left[\left(\frac{2J_1(v)}{v}\right) - \delta^2\left(\frac{2J_1(\delta v)}{\delta v}\right)\right]^2$$

where $J_1$ is the first order Bessel function, the angle θ is given in radians, the argument v is given by the aperture diameter D and the wavelength λ: $v = \pi\theta D/\lambda$, δ is the ratio between the diameter d of the central obstruction and the diameter D of the aperture, $\delta = d/D$, and $I(\theta|\delta)$ is normalized so that $I(0|\delta) = 1$. Figure 2-61 shows a radial section through such a diffraction profile for $\delta = 0.5$, compared to the profile of an unobstructed aperture as well as the limiting case of $\delta \rightarrow 1$. T he central obstruction results in a slightly narrower central disk; at $\delta = 0.5$ the radius of the first dark ring has moved from $\theta = 1.22\ \lambda/D$ to $\theta = \lambda/D$.

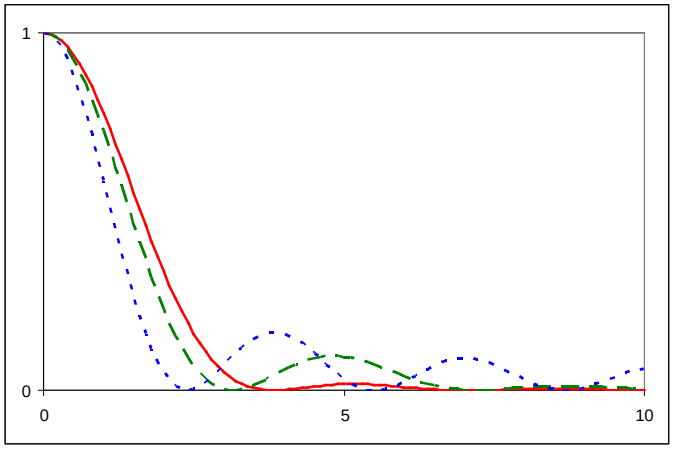



*Figure 2-61. Radial section through diffraction profile of unobstructed aperture (red -), and annular apertures with δ = 0.5 (green - -) and δ →1 (blue …).*

Figure 2-62 shows the appearance of the Airy pattern for obstructions of 0 %, 20 % and 33 %. The intensity is given on a logarithmic scale to fit human vision. From 0 to 1/3

obstruction, the maximum intensity of the first diffraction ring increases from 1.7 % to 5.4 % of the maximum intensity of the central disk.

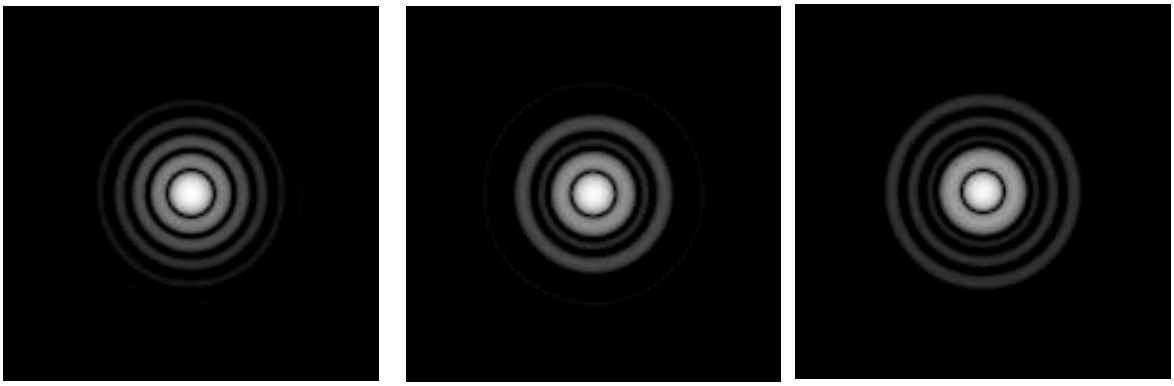


*Figure 2-62. Diffraction profiles of an unobstructed aperture (left), and annular apertures with δ = 1/5 (middle), and δ = 1/3 (right).*

As seen in figure 2-61 and verified in the logarithmic plot of figure 2-63, the obstruction affects both the position of the minima and the intensity of the maxima in the ring pattern.

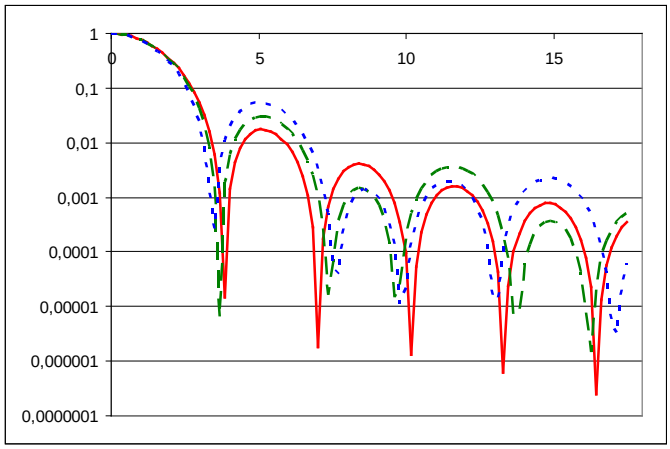



*Figure 2-63. Radial section through diffraction profile of unobstructed aperture (red-), and annular apertures with δ = 1/5 (green --), and δ = 1/3 (blue …). Logarithmic scale.*

The modulation transfer function (MTF, see next section) in figure 2-64 shows that the central obscuration will affect the image contrast in a relatively complex manner. In the low and medium frequencies the contrast is decreased as the size of the central obscuration increases. In the high frequencies, it is not decreased, but slightly increased.

The dotted curves in figure 2-64 show to which unobstructed apertures the annular apertures are equivalent in terms of contrast in the low frequencies. It appears that in these frequencies an aperture with an obstruction of 33 % is equivalent to an unobstructed aperture whose diameter is 33 % less. An aperture with an obstruction of 20 % is equivalent to an unobstructed aperture whose diameter is 15 % less.

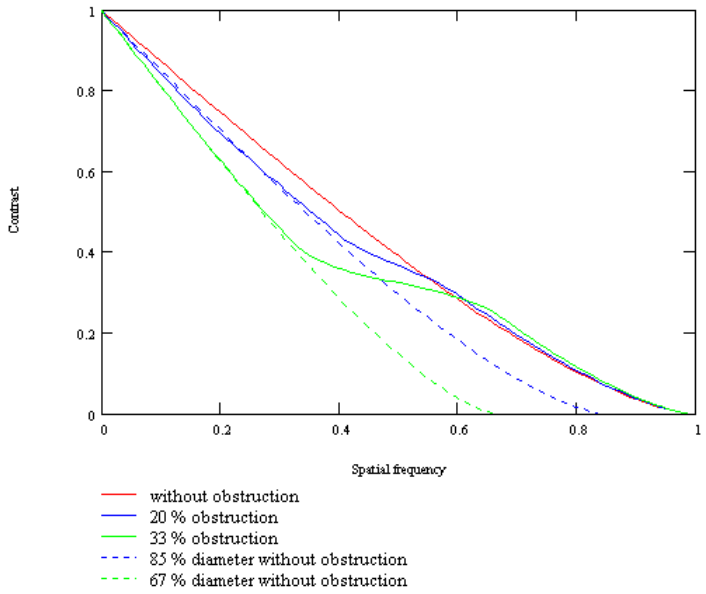



*Figure 2-64. MTF's of unit circular aperture without obstruction, 1/5 and 1/3 obstruction, together with MTF's of 0.85 and 0.67 unobstructed circular apertures.*

### 2.4.10 Optical transfer function

The angular resolution given by the Rayleigh criterion is a basic measure of the resolving power of a lens or a mirror. However, the quantity expressed by the Rayleigh criterion is the limiting resolution. A more complete understanding of the performance of the optical system is expressed by the **Optical Transfer Function** (OTF), of which the limiting angular resolution is but one point.

The Optical Transfer Function describes the spatial (angular) variation as a function of spatial (angular) frequency. The OTF may be broken down into the magnitude and phase components as follows:

$$\mathbf{OTF}(\xi,\eta) = \mathbf{MTF}(\xi,\eta) \cdot \mathbf{PTF}(\xi,\eta)$$

Where

$$\mathbf{MTF}(\xi,\eta) = |\mathbf{OTF}(\xi,\eta)|$$

$$\mathbf{PTF}(\xi,\eta) = e^{-i2\cdot\pi\cdot\lambda(\xi,\eta)}$$

and (ξ,η) are spatial frequency in the x- and y-plane, respectively.

The magnitude is known as the **Modulation Transfer Function (MTF)** and the phase portion is known as the **Phase Transfer Function (PTF)**. In imaging systems, the phase component is typically not captured by the sensor. Thus, the important measure with respect to imaging systems is the MTF.

The OTF is the Fourier transform of the Point Spread Function of the lens system

### 2.4.10.1 Integrating the diffraction profile over wavelength

The expressions for the diffraction profile above are valid for a given wavelength λ. Clearly the width of the profile may vary substantially over the bandwidth of a given broadband detector.

Let us look at a practical example in some detail. The panchromatic detector of the QuickBird satellite-based remote sensing telescope covers the wavelength interval from 450 to 900 nm. The diffraction profiles of the D=60 cm telescope at 450 and 900 nm are very different, as shown by the blue and red curves in figure 2.65. The average of the two curves, as well as the diffraction curve of the middle wavelength, 675 nm, is also shown. The unit on the horizontal axis is in microradians (1 μradian = $10^{-6}$ radian), corresponding to 0.45 m on the ground as seen from a height of 450 km.

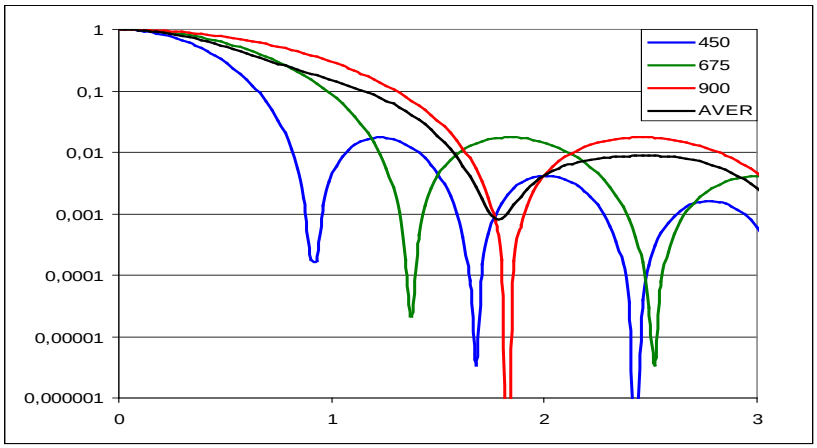



*Figure 2-65. Radial section through diffraction profile of 60 cm unobstructed aperture at 450 nm (blue), 675 nm (green) and 900 nm (red), together with average profile of 450 and 900 nm (black). Horizontal scale in microradians. Logarithmic scale on vertical axis.*

However, the proper way of obtaining the panchromatic diffraction profile is to perform a numerical integration over the wavelength interval for each value of the angle θ. A continuous function f(y) may be integrated over the interval (a,b) by Gauss' formula:

$$\int_a^b f(y)dy \approx \frac{b-a}{2}\sum_{i=1}^{n} w_i\left[\left(\frac{b-a}{2}\right)x_i + \left(\frac{b+a}{2}\right)\right]$$

Where $x_i$ are the roots of the corresponding Legendre polynomials, and $w_i$ are the corresponding weights. Both are tabulated in Abramowitz and Stegun: "Handbook of mathematical functions", pp. 916-919. In this way, polynomials up to degree (2n-1) will be integrated exactly. Figure 2.66 shows the diffraction profiles at the 8 wavelengths needed for a weighted summation equivalent to a 15th order polynomial integration. We assume that the spectral response function S(λ) of the panchromatic detector is independent of wavelength, while the spectral irradiance I(λ,T) of the Sun at the top of our atmosphere may be approximated by a Planck curve:

$$I(\lambda,T) = \frac{2hc^2}{\lambda^5}\cdot\frac{\pi}{e^{\frac{hc}{\lambda kT}}-1}\left(\frac{r}{d}\right)^2$$

where h (Planck's constant) = 6.6260693•10$^{-34}$ Js, c (speed of light) = 2.99792458•10$^{8}$ m/s, k (Boltzmann's constant) = 1.3806505 •10$^{-23}$ J/K, r = 6.96 •10$^{8}$ m, d = 1.5 •10$^{11}$ m, as discussed in Chapter 1. The integrated diffraction profile over the wavelength interval 450 – 900 nm is then given by the red curve of figure 2.65:

$$F\left(\theta\middle|\lambda_1,\lambda_2,D,T\right)=\int_{\lambda_1}^{\lambda_2} I(\lambda,T)\left[\left(\frac{2J_1\left(\frac{\pi\theta D}{\lambda}\right)}{\left(\frac{\pi\theta D}{\lambda}\right)}\right)\right]^2 d\lambda \Bigg/ \int_{\lambda_1}^{\lambda_2} I(\lambda,T)\,d\lambda$$

We notice that the resulting panchromatic diffraction profile for the wavelength interval from 450 to 900 nm is very different from all the profiles in figure 2.64. The central part of the profile actually corresponds closely to a Gaussian having a standard deviation of 0.45 μradians, i.e. a 1/e-width of 60 cm as seen from a height of 450 km. The spectral irradiance values I(λ,T) used as weights in the integration are given in figure 2.67.

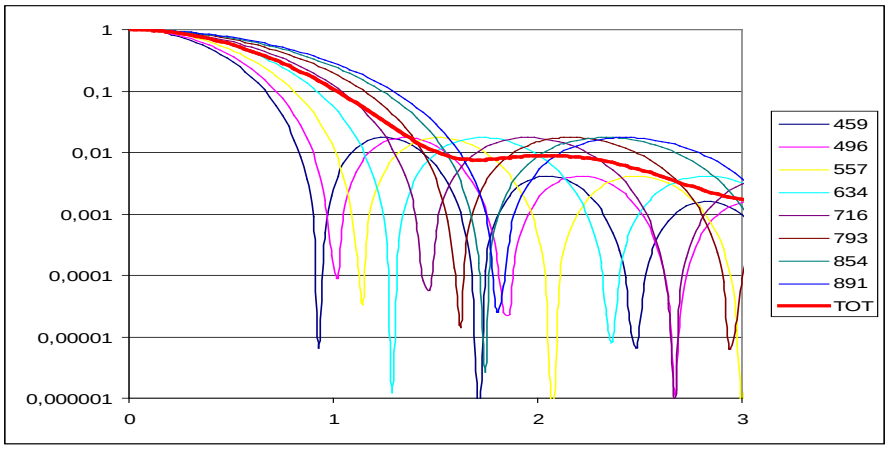



*Figure 2-66. Radial section through diffraction profile of unobstructed 60 cm aperture at 8 different wavelengths to integrate a panchromatic profile over the wavelength interval of 450 and 900 nm, together with the integrated profile weighted by the spectral irradiance given by a Planck curve at T=6000K. Horizontal scale in microradians.*

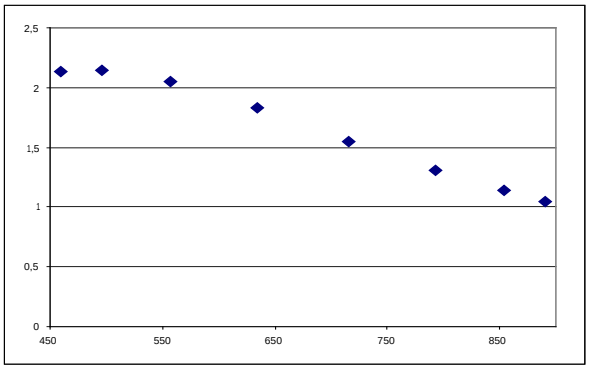



*Figure 2-67. Solar spectral irradiance (kW/m$^2$/μm) from a black body at T=6000K at the eight wavelengths used to obtain a panchromatic diffraction profile on the wavelength interval from 450 to 900 nm in figure 2-65. Horizontal scale in nanometers.*

### 2.4.11 The smallest detail visible to the naked eye

Iris – the colored part of the eye – acts as a stop that regulates how much light will be let through the pupil into the eye. It closes down to a diameter of about 2 mm in bright light, and opens up to about 8 mm when the light intensity is low.

As a rule of thumb the angular resolution of the human eye is about 60 lines per degree. The lines should be black stripes on a white background, so that there are 120 alternating black and white lines of equal thickness.

If we place a standard A4 sheet of paper in ”landscape” format 30 cm in front of one eye and close the other , the sheet will cover 50º horizontally and 40º vertically. If the sheet is filled with 3 000 black and 3 000 white vertical stripes, then a person having normal vision should be able so see the individual stripes. Similarly 2 400 individual horizontal black stripes on a white background should be resolved.

Printer resolutions are often given in dots per inch (dpi). 3 000 black stripes on a white background on an A4 sheet of paper corresponds to 6 000 dots in 11 inches, or about 550 dpi. As this is the resolution limit of the eye, it may perhaps explain the popularity of 600 dpi printers.

Obviously, a higher number of dots per inch are needed if we are to inspect printed matter at a closer range than 30 cm. But there is a practical limit to how close we may inspect anything with the unaided eye and still be able to focus sharply. The *accommodation distance*, or ”near point” is about 7 cm for a 10 year old, 10 cm for a 20 year old, 14 cm for a 30 year old, 22 cm for a 40 year old, 40 cm for a 50 year old, and about 100 cm for a 60 year old. For a 47 year old it is about 30 cm, as we have used in the example above, an a printer better than 600 dpi is a waste. But a student having an accommodation distance of 10 cm can inspect the printout at a much closer range, and will be put off by anything less than 1 200 dpi – which happens to be very common for printing high quality images.

Assuming that the diameter of the pupil is 2.5 mm, and considering that the maximum sensitivity of the retinal cones is at 550 nm, we may estimate the resolution of the eye. The Rayleigh criterion is given by

$$\sin\theta = \frac{1.22\,\lambda}{D} = \frac{1.22\,x550\,x10^{-9}}{2.5\,x10^{-3}} = 2.7\,x10^{-4}$$

For small angles we have **sin(θ) = θ**, when the angle θ is given in radians. Converting from radians to degrees, we find that angular resolution is

$$\theta = 2.7\,x10^{-4}\,\frac{180}{\pi} \approx \frac{1}{60}^{\circ}$$

just as stated by the rule-of-thumb.

## 2.4.12 What is the size of the smallest visible detail …

### 2.4.12.1 … in an image from an f = 35 mm lens, f/D = 3 ?

Assume that we use a lens having a focal length f=35 mm with f/D = 3. We are making an image of an object 3.5 meters away. We assume that the wavelength is λ=500 nanometers (green light).

**Q:** What is the distance y between two points on the object that can just be resolved?
**A:** The aperture is D = f/3 = 35 mm/3 = 11.67 mm = $1.167 \times 10^{-2}$ m.
The angle between to just resolvable points on the object, as seen from the center of the lens, is given by the Rayleigh criterion: $\sin(\theta) = 1.22\ \lambda/D$.
In this case $\sin(\theta) = 1.22 \times 500 \times 10^{-9} / 1.167 \times 10^{-2}$.
However, the tangent of this angle must be (y/3.5).

For small angles we have $\sin(\theta) = \tan(\theta) = \theta$, when θ is given in radians.

We find that $y = 3.5 \times 1.22 \times 500 \times 10^{-9} / 1.167 \times 10^{-2} = 1.83 \times 10^{-4}$ m ≈ **0.2 mm.**

**Q:** What is the corresponding distance y' in the focal plane?
**A:** We remember that y'=yf/(s-f) (see section 2.2.4.3)
Which gives us: y' = 0.2 x 35 /(3500–35) mm ≈ 0.002 mm = **2 µm.**

### 2.4.12.2 … in an image from a compact digital camera ?

The compact digital camera is supplied with a lens having a much shorter focal length, f = 5.8 mm, but the f-number is approximately the same, f/D = 3.1 .

**Q:** What is now the distance y between two just resolvable points on the object?
**A:** Now the lens aperture is D=f/3.1 = 5.8 mm/3.1 = 1.87 mm = $1.87 \times 10^{-3}$ m .
By the same computation as above we get
$y = 3.5 \times 1.22 \times 500 \times 10^{-9} / 1.87 \times 10^{-2} = 1.14 \times 10^{-3}$ m ≈ **1.1 mm.**

So the more modern camera gives 5.5 times less resolution of detail on the object.
The corresponding focal plane distance is y' = 1.1 x 5.8 /(3500–5.8) mm ≈ **2 µm.**
So the demands on physical sampling in the focal plane are the same.

### 2.4.12.3 … in an image from a cellular phone camera ?

Take for instance Nokia 5140. It is equipped with an f = 4 mm lens with an f-number f/D ≈ 4, implying that D = f/4 = 1 mm. At a distance of 3.5 meters the smallest resolvable detail will be
$y = 3.5 \times 1.22 \times 500 \times 10^{-9} / 1.0 \times 10^{-3}$ => $y = 2.1 \times 10^{-3}$ m ≈ **2.1 mm.**

So the cellular phone camera gives a factor of 2 less resolution compared to the compact digital camera, and a factor of 10 less resolution compared to the old-fashioned 35 mm lens.

### 2.4.13 An example of extremely high angular resolution

The angular resolution will improve if we increase the diameter of the lens/mirror, or do the imaging at a shorter wavelength. Extreme examples are found in astronomy, where large optical telescopes are employed to resolve objects at fantastic distances.

A 1 meter diameter telescope located on a mountain top in the Canary Islands has acquired some of the very best images of the solar atmosphere. Assuming a wavelength of 600 nanometers and a diameter of 1 meter, the Rayleigh criterion implies an angular resolution of $4.2 \times 10^{-5}$ °. The (tangent of) this angle must be equal to the ratio of the size of the smallest resolvable detail on the Sun to the distance from the Earth to the Sun – 150 million kilometers. The theoretical resolution turns out to be 110 km on the Sun.
Figure 2-68 shows a portion of an image of a sunspot recorded with this instrument. There are tick marks for every 1000 km around the borders of the image.

To achieve such a high resolution one would have to place the telescope outside the atmosphere, since the inhomogeneities in the atmosphere would cause the image to be unstable and blurred. In the present case, however, an "active mirror" has been used. The wave front has been analyzed entering the telescope, and small pistons on the rear surface of the flexible mirror have been used to adjust its shape in close to real time, to compensate for atmospheric disturbances and create a very sharp image.

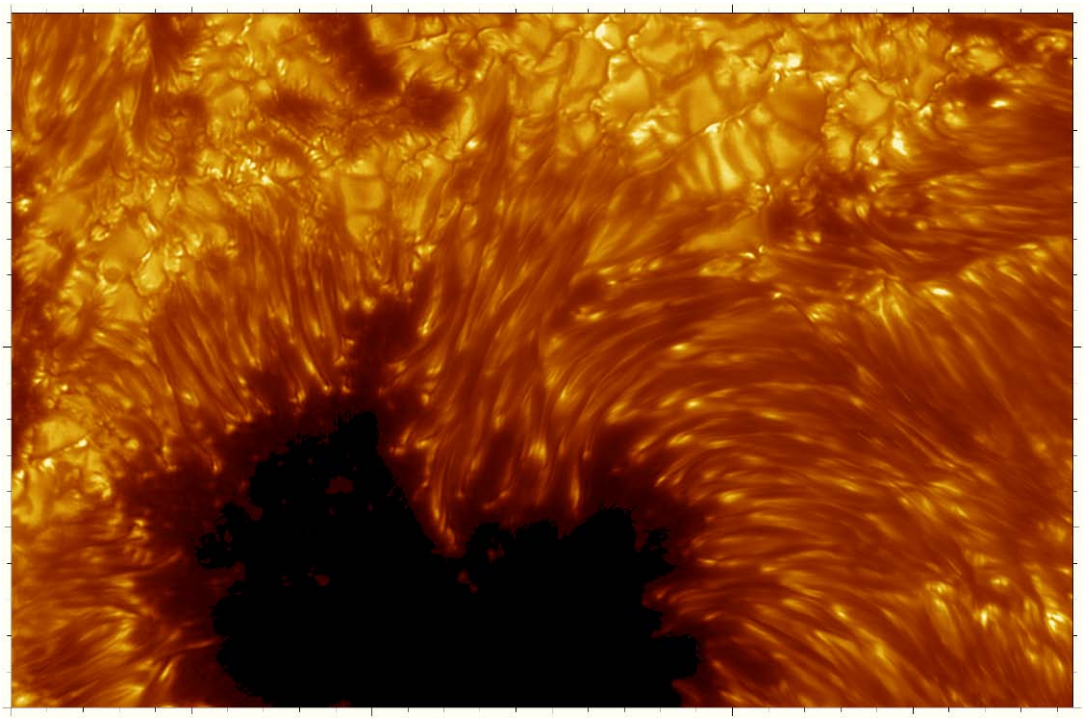


*Figure 2-68. A sunspot image recorded by a 1 meter active mirror solar telescope. © Swedish 1-meter Solar Telescope (SST), Royal Swedish Academy of Sciences.*

### 2.4.14 Depth of focus in diffraction limited optics

The intensity of the diffraction pattern in the vicinity of focus is given by

$$I(u,v)=\left(\frac{2}{u}\right)^2\left[1+V_0^2(u,v)+V_1^2(u,v)-2V_0(u,v)\cos\left\{\frac{1}{2}\left(u+\frac{v^2}{u}\right)\right\}-2V_1(u,v)\sin\left\{\frac{1}{2}\left(u+\frac{v^2}{u}\right)\right\}\right]I_0$$

where (u,v) = (0,0) is the focal point, u is measured along the optical axis and v is measured orthogonal to the optical axis. The Lommel functions U and V are related to the Bessel functions J by

$$U_n(u,v)=\sum_{s=0}^{\infty}(-1)^s\left(\frac{u}{v}\right)^{n+2s}J_{n+2s}(v), \qquad V_n(u,v)=\sum_{s=0}^{\infty}(-1)^s\left(\frac{v}{u}\right)^{n+2s}J_{n+2s}(v)$$

For points on the optical axis, v = 0, and the two Lommel V-functions in the expression for I(u,0) are reduced to

$$V_0(u,0)=1, \qquad V_1(u,0)=0$$

so that

$$I(u,0)=\left(\frac{2}{u}\right)^2\left[2-2\cos\left\{\frac{u}{2}\right\}\right]I_0=\left(\frac{\sin(u/4)}{u/4}\right)^2 I_0$$

In the neighborhood of the focus the diffraction intensity distribution is symmetric about the focal plane, and symmetric about the u-axis. We see that the intensity along the optical axis is characterized by a $[\sin(x)/x]^2$ function (see figure 2-69). The first minimum along the optical axis is at u/4 = ± π, i.e. at a distance z = ± $2\lambda(f/D)^2$ from the focus. For an f/D = 5.6 lens at λ = 500 nm this corresponds to z = ± 0.03 mm, while at a smaller aperture, f/D = 22, we get a quadratic increase in the depth of focus, giving z = ± 0.5 mm. Increasing the wavelength will give a linear increase the depth of focus: an f/D = 5.6 lens at λ=1000 nm gives z = ± 0.06 mm. Note that the physical depth of focus is independent of the physical dimensions of the optics. It only depends on the ratio of the focal length to the lens diameter, and on the wavelength of the light used.

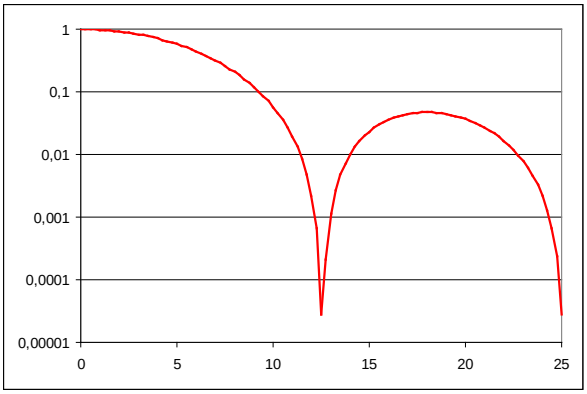



*Figure 2-69. An (u,log(I(u,0))-plot of the along-axis intensity of the diffraction pattern.*

Figure 2-70 shows a contour plot of the intensity I(u,v) in a meridional plane near the focus of a converging spherical wave diffracted by a spherical aperture. The horizontal u-axis is the optical axis, and the vertical v-axis is in the focal plane. If the figure is rotated about the horizontal axis, the maxima and minima along the v-axis will generate the bright and dark rings of the Airy disc (see figure 2-58).

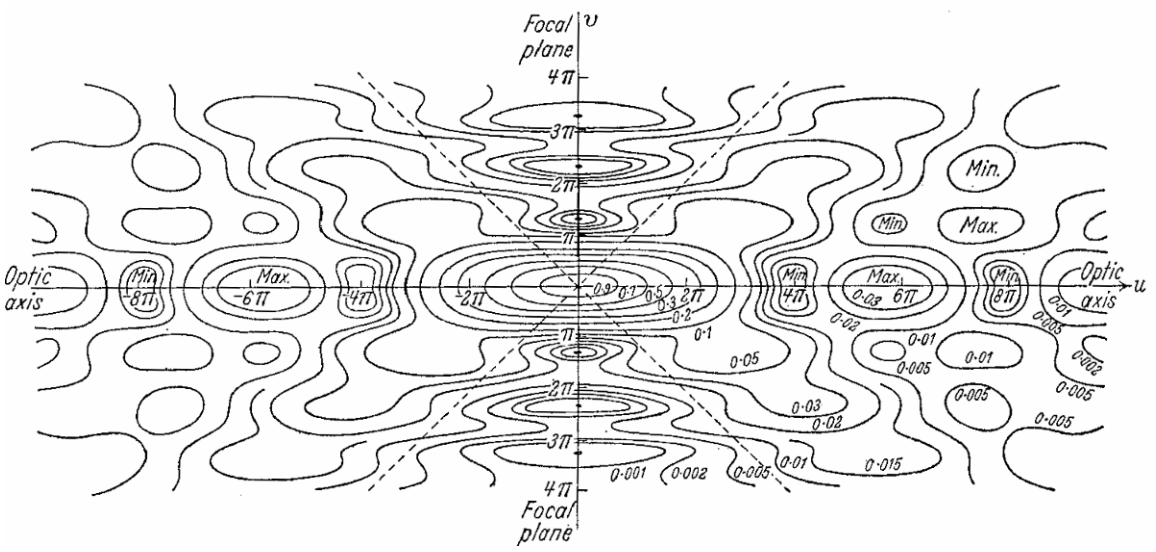



*Figure 2-70. A contour plot (isophotes) of the intensity I(u,v) in a meridional plane near the focus of a converging spherical wave diffracted by a circular aperture without central obstruction. The intensity is normalized at the centre of the focal plane.*
*From M. Born and E. Wolf: "Principles of Optics", Pergamon Press, 4th. Ed., 1970.*

It may be instructive to look at the effect of a central aperture obstruction on the depth of focus. The only effect is that the previous expression is scaled along the optical axis:

$$I(u,0 \mid \varepsilon) = \left( \frac{\sin(u(1-\delta^2)/4)}{u/4} \right)^2 I_0$$

where $\delta$ = d/D is the ratio between the diameter d of the central obstruction and the diameter D of the aperture.

So the distance from the focal plane at which I(u,0) drops to a certain value is increased by a factor $1/(1-\delta^2)$. Hence, the focal depth is increased by $1/(1-\delta^2)$ while at the same time the aperture area is decreased by a factor $(1-\delta^2)$. The conventional method of increasing the focal depth is to stop down the aperture by decreasing the aperture diameter, which will result in an increase in the radial scale of the diffraction pattern. Using a central obscuration that decreases the aperture area by the same factor, on the other hand, will increase the focal depth as we have just seen, but will result in a *narrowing* of the radial diffraction pattern, as we saw in section 2.4.9.3. As we also saw in that section, the price paid for a narrower Airy disc lies in higher intensities of the surrounding ring pattern.

### 2.4.15 Convolving point spread function and sampling aperture

We have seen that the diffraction-limited angular point spread function (PSF) of a circular aperture is given by the continuous circular symmetric function

$$F_0(r) = 4\left[ J_1\left(\pi \frac{r}{\beta_0}\right) \Big/ \pi \frac{r}{\beta_0} \right]^2$$

where $J_1$ is the first order Bessel function, the angle $\beta_0$ is given by the aperture diameter D and the wavelength λ: $\beta_0 = \lambda/D$, and $F_0(r)$ is normalized so that $F_0(0) = 1$.

A sampling of this continuous function will always imply integration over a small aperture in the focal plane. The sampling can either be performed by a spatial grid of often quadratic sensors, or by a single aperture that is scanning the focal plane and subject to a temporally discrete sampling. Figure 2-71 illustrates this for a circular aperture of radius ρ, positioned at a distance r from the centre of the PSF. Thus, we do not get $F_0(r)$, but a smeared-out modification that can be described by the integral

$$F_\rho(r) = C(\rho) \int_A \left[ J_1\left(\pi \frac{r}{\beta_0}\right) \Big/ \pi \frac{r}{\beta_0} \right]^2 dA$$

When the sampling area is a circular aperture of radius ρ, the normalization constant C(ρ) is simply

$$C(\rho) = \frac{\pi}{1 - J_0^2\left(\pi \frac{\rho}{\beta_0}\right) - J_1^2\left(\pi \frac{\rho}{\beta_0}\right)} \cdot \frac{1}{\beta_0^2}$$

where $J_0$ denotes the Bessel function of order zero.

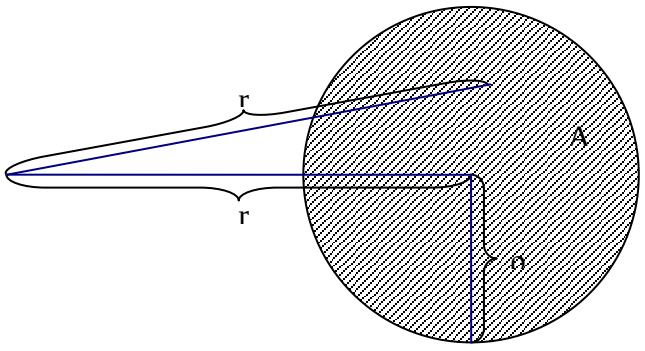



*Figure 2-71. Geometry of integration over a circular aperture in the focal plane.*

The function $F_\rho$ is the point spread function of the combination of a diffraction limited aperture (mirror or lens) and a scanning pinhole in the focal plane. This is also called "the instrumental profile". If we are observing an extended object, the observed intensity distribution will be the convolution of $F_\rho$ and the true intensity distribution of the object.

The ring pattern of the diffraction by a circular aperture will vanish rapidly as the scanning aperture radius increases. Using the asymptotic form of the Bessel function

$$J_1(x) \approx \sqrt{\frac{2}{\pi x}} \cos\left( x - \frac{3\pi}{4} \right)$$

We find that the wings of $F_\rho$ may be expressed by

$$F_\rho(r) \approx C(\rho) \rho^2 \left( \pi \frac{r}{\beta_0} \right)^{-3}$$

We also get

$$F_0(r) \approx \frac{4}{\pi} \left( \pi \frac{r}{\beta_0} \right)^{-3}$$

These approximations work well when r exceeds a few times $\beta_0$ or $\rho$, depending on which is the largest. Figure 2-72 shows the core of the instrumental profile of the combination of a diffraction limited aperture (mirror or lens) and a scanning pinhole in the focal plane, for a few values of the pinhole radius, expressed in terms of $\beta_0 = \lambda/D$. Note that $\rho = 1.5\,\beta_0$ will give a slight ringing effect.

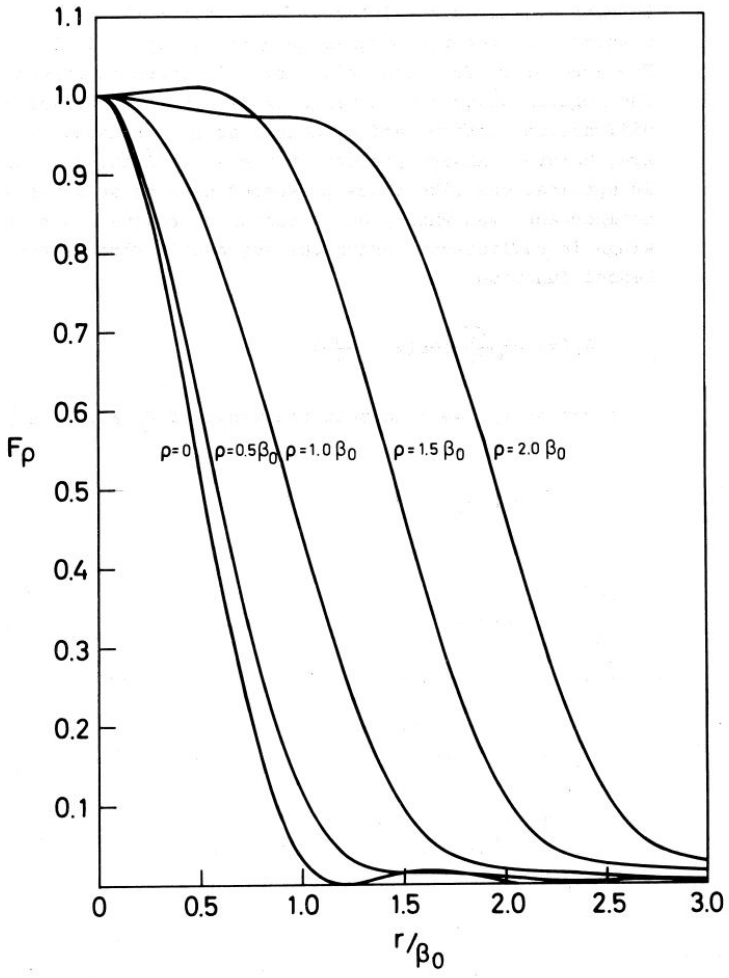



*Figure 2-72. The core of the instrumental profile of the combination of a circular aperture and a scanning pinhole in the focal plane, for a few values of the pinhole radius.*

## 2.5 Scattering

Scattering is a physical process that causes radiation to deviate from a straight trajectory. We saw this in the introductory sections on reflection: If there were microscopic irregularities in the surface we would get diffuse instead of specular reflection. The same goes for radiation passing through a transparent medium: If there are non-uniformities like particles, bubbles, droplets, density fluctuations etc, some of the radiation will deviate from its original trajectory.

In a physical description of the phenomenon, we distinguish between two types of scattering, namely *elastic* and *inelastic*. Elastic scattering involves no (or a very small) loss or gain of energy by the radiation, whereas inelastic scattering does involve some change in the energy of the radiation. If the radiation is substantially or completely extinguished by the interaction (losing a significant proportion of its energy), the process is known as *absorption*.

When radiation is only scattered by one localized scattering center, this is called *single scattering*. Single scattering can usually be treated as a random phenomenon, often described by some probability distribution.

Often, many scattering centers are present, and the radiation may scatter several times, which is known as *multiple scattering*. With multiple scattering, the randomness of the interaction tends to be averaged out by the large number of scattering events, so that the final path of the radiation appears to be a deterministic angular distribution of intensity as the radiation is spread out. Using an imaging device we will not get a sharp, diffraction-limited image of the object, but a more blurred image that is the result of a *convolution* of the image with a point-spread function that includes both diffraction and scattering.

Since the intensity distribution of the object is often unknown, a difficult challenge is the "*inverse scattering problem*", in which the goal is to observe scattered radiation and use that observation to determine either the scattering parameters or the distribution of radiation before scattering. In general, the inverse is not unique, unless the scattering profile can be found by observing the image of some well-known object through the same scattering medium.

Light scattering and absorption are the two major physical processes that contribute to the visible appearance of physical objects. The spectral distribution of absorption determines the color of a surface, while the amount of scattering determines whether the surface is mirror-like or not.

The size of a scattering particle is defined by the ratio of its characteristic dimension and the wavelength of the scattered light:

$$x = \frac{2\pi r}{\lambda}$$

The wavelength dependence of scattering is first and foremost determined by this ratio of the size of the scattering particles to the wavelength of the light:

- Scatter diameters much less than the wavelength results in Rayleigh scattering.
- Larger diameters result in Mie scattering.

**Rayleigh scattering** occurs when light travels in transparent solids and liquids, but is most prominently seen in gases. The amount of Rayleigh scattering that occurs to a beam of light is dependent upon the size of the particles and the wavelength of the light; in particular, the scattering coefficient, and hence the intensity of the scattered light, varies for small size particles inversely with the fourth power of the wavelength. This wavelength dependence ($\sim\lambda^{-4}$) means that blue light is scattered much more than red light. In the atmosphere, the result is that blue light is scattered much more than light at longer wavelengths, and so one sees blue light coming from all directions of the sky. At higher altitudes, high up in the mountain or in an airplane, we can observe that the sky is much darker because the amount of scattering particles is much lower.

When the Sun is low on the horizon the sunlight must pass through a much greater air mass to reach an observer on the ground. This causes much more scattering of blue light, but relatively little scattering of red light, and results in a pronounced red-hued sky in the direction towards the sun.

Surfaces described as *white* owe their appearance almost completely to the scattering of light by the surface of the object. Absence of surface scattering leads to a shiny or glossy appearance. Light scattering can also give color to some objects, usually shades of blue (as with the sky, the human iris, and the feathers of some birds).

Scattering by spheres larger than the Rayleigh range is usually known as **Mie scattering**. In the Mie regime, the shape of the scattering center becomes much more significant and the theory only applies well to spheres and, with some modification, spheroids and ellipsoids. Closed-form solutions for scattering by certain other simple shapes exist, but no general closed-form solution is known for arbitrary shapes.

The wavelength dependence of Mie scattering is approximately described by $1/\lambda$.

Both Mie and Rayleigh scattering are considered elastic scattering processes, in which the energy (and thus wavelength and frequency) of the light is not substantially changed. However, electromagnetic radiation scattered by moving scattering centers does undergo a Doppler shift, which can be detected and used to measure the velocity of the scattering centers in forms of techniques such as LIDAR and radar. This shift involves a slight change in energy.

At values of the ratio of particle diameter to wavelength more than about 10, the laws of geometric optics are mostly sufficient to describe the interaction of light with the particle, and at this point the interaction is not usually described as scattering.

### 2.5.1 Some effects of scattering

If we perform passive imaging of a simple object that is illuminated by e.g. sunlight, a detector element in the focal plane of the imaging system will in general receive part of the radiation that is reflected off a specific part of the surface of the object. The incident radiation will consist of several components:

1. Specular reflection, governed by the law of reflection (Section 2.1.3).
2. Diffuse reflection, governed by Lambertian reflection (Section 2.1.2).
3. In addition the detector element will receive some diffuse radiation from other parts of the object.
4. The detector will also receive some radiation that has been scattered in the air:
   a. Radiation that was scattered before it reached the object, and through one or more scatterings ended up on the detector element in question.
   b. Both specular and diffuse reflection from other parts of the object that is not directed towards our detector element, but that is scattered onto it.

Even if we shield the detector so that it can only receive radiation from a small patch of the object, components 4-a and 4-b will be present in addition to components 1 and 2. Correction for these effects may be of some importance in passive remote sensing applications, as the radiation is passing twice through the Earth's atmosphere.

Scattering in air is usually a minor effect compared to when the density of scatterers is increased, as in liquids, tissues, and other translucent materials. We know for instance that an object that is seen through mist or fog will look blurred. And at some distance the object will disappear into the background fog.

The same effect may be observed as *turbidity* in water. Here, particles or minute organisms act as scatterers, causing a haziness that is an indication of the water quality. Turbidity in lakes, reservoirs, and the ocean can be measured using a *Secchi disk*. This black and white disk is lowered into the water until it can no longer be seen; the depth (Secchi depth) is then recorded as a measure of the transparency of the water (inversely related to turbidity).

All non-metallic materials are translucent to some degree. This means that light penetrates the surface and scatters inside the material before being either absorbed or leaving the material at a different location. This phenomenon is called subsurface scattering. Even solid materials like marble display sub-surface scattering. The effect is a "softer" image than a metallic surface would give.

Images of human skin, salmon fillets, or other tissue samples will not only arise from reflected light from the sample surface. Some of the radiation incident on the sample will pass through the surface, and be subject to subsurface scattering. Some of this subsurface scattering will emerge from the surface and contribute to the image. This subsurface scattering may depend on the wavelength of the light used, the directional inhomogeneities in the tissue, etc., but may also depend on the condition of the tissue. Thus, measuring subsurface scattering may be useful for quality inspection of e.g., fish and meat.

### 2.5.2 Doppler-shifted strayligh

Straylight will cause a redistribution of intensity within the image of an extended object, usually by blurring the image. However, straylight will also redistribute spectral information.

As an example of a practical - yet unusual - effect of straylight, let us look at the problem of Doppler measurement of the line-of-sight velocity of a rotating extended object, specifically the rotational velocity of the Sun, and high-resolution observations of global solar oscillations that provide a seismic modeling of the solar interior.

If we merge all angular redistribution effects (instrumental diffraction, blurring, scattering) into a single circular symmetric straylight function, Ψ(r), we may express the relation between the observed intensity distribution I(p) and the true continuum intensity distribution $\Phi_c(p')$ by the integral equation

$$I(p) = \int_{\oplus} \Phi_C(p')\,\Psi(r)\,d\omega$$

Here p and p' are directions in the sky and r is the angle between them. The integration is performed over the solid angle of the Sun. We make the simplification that the observed point, p, is situated on the solar equator, and that the rotation axis is perpendicular to the line of sight. The straylight calculation may then be performed on one side of the equator.

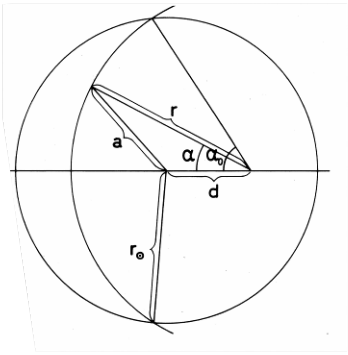



*Figure 2-73. Geometry of integration over half the solar disc.*

Using the symbols from figure 2-73 we rewrite the equation above in polar coordinates:

$$I(p) = 2\int_{\rho_0}^{\rho_1}\int_{0}^{\alpha_0} \Phi_C(a)\,\Psi(\rho)\,d\rho\,d\alpha$$

The limits of the outer integral are given by

$$\rho_0 = 0, d \le r_\oplus,$$

$$\rho_0 = d - r_\oplus, d \ge r_\oplus,$$

$$\rho_1 = d + r_\oplus.$$

The upper limit of the inner integral is given by

$$\alpha_0 = \cos^{-1}\left[\left(\rho^2 + d^2 - r_\oplus^2\right)/2\rho d\right], \quad \rho + d > r$$

$$\alpha_0 = \pi, \quad \rho + d \le r$$

The important point to realize so far is that the observed intensity in a given direction may be seen as a weighted sum of two functions: the circular symmetric straylight function and the intensity distribution of the object. As we have seen, the effects of instrumental diffraction, blurring and sampling may be combined into an "instrumental profile", often expressed as a weighted sum of Gaussians. Long range scattering is often parameterized by a dispersion function. The continuum intensity distribution $\Phi_c$ is often described by a set of polynomials, giving a limb darkening that varies markedly with wavelength. References to the parameterization of the straylight function and the true continuum intensity distribution $\Phi_c$ across the solar disc can be found elsewhere [3].

The line-of-sight velocity component, $V'_{ls}$, is deduced from observations by an equal area fit to the lowest 5% of a given spectral absorption line profile in the radiation from the solar atmosphere. However, the observed intensity at a given point p will contain components of straylight with line-of-sight velocities different from that at p. So, the line profile of the straylight from p' that is added to the true profile at p must therefore be Doppler-shifted by an amount $\Delta\lambda = \lambda\Delta V/c$. Here $\lambda$ is the wavelength, $\Delta\lambda$ is the wavelength shift, c the velocity of light, and $\Delta V$ the difference in the line-of-sight velocity component between p and p'.

The surface layers of the Sun rotate faster at the heliographic equator than at higher heliographic latitudes (differential rotation). The line-of-sight velocity component $V_{ls}$ is often modeled by

$$V_{ls}(q,l) = \left(a_r + b_r \sin^2 l \left(1 + \sin^2 l\right)\right) \sin q + V_{\lim}$$

where $a_r$ and $b_r$ are coefficients of solar rotation, l is the heliographic latitude, and q is the angular distance from the central meridian. The last component describes the line-of-sight component of the granular velocity field, known as the limb effect. This has to be determined separately for each spectral line.

If we assume that the intensity is observed at a given wavelength within a spectral line, we must now modify the first two equations of this section that described the observed continuum intensity including straylight. If we assume that the true continuum intensity distribution $\Phi_c$ across the solar disc is known, and that a Gaussian absorption line profile with Doppler width w and a central intensity $I_c$ is being Doppler-shifted as described above and weighted by a circular symmetric straylight function, we will observe the following intensity at a given wavelength $\lambda$ within a spectral line and at a given distance d from the centre of the solar disc:

$$I(d,\lambda) = 2\int_{\rho_0}^{\rho_1}\int_{0}^{\alpha_0} \Phi_C(a)\left(1 - I_C(a)\right)\exp\left[-\left(\lambda - \Delta\lambda\right)^2 / w^2(d)\right]\Psi(\rho)\,d\rho\,d\alpha$$

The equatorial rotation velocity is often given as 2016 m/s ± 13 m/s, obtained during periods of very low straylight. Figure 2-74 shows that for a typical set of straylight parameters, the measured velocity will generally be different from the true one.

[3] F. Albregtsen and B.N. Andersen, Solar Physics 95, 239-249, 1985.

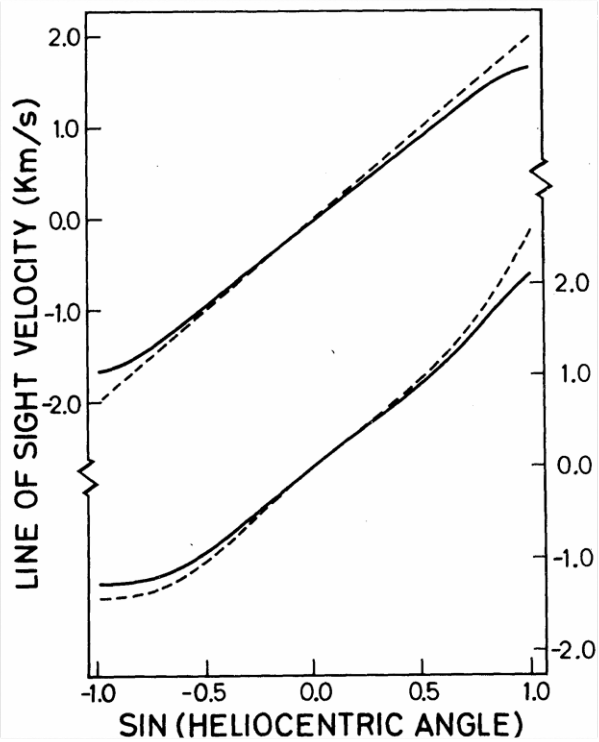



*Figure 2-74. The true (- - -) and the straylight-influenced (---) line-of-sight velocity along the solar equator for a spectral line without limb effect (upper panel) and for a line with limb effect (lower panel). The straylight induced error is exaggerated by a factor of 5 in this figure.*

The main contribution to the error signal is caused by the long range scattering component of the straylight. Typical straylight intensities observed at e.g. Stanford will introduce errors in the velocity signal of 0.1 – 1.0 m/s, while the observed global solar oscillations with periods from 5 to 160 minutes all have velocity amplitudes of the order of 0.1 m/s. Thus, the error signal is at least of the same magnitude as the velocity oscillation signal, and quasiperiodic variations in the straylight may therefore influence the modeling of the solar interior.

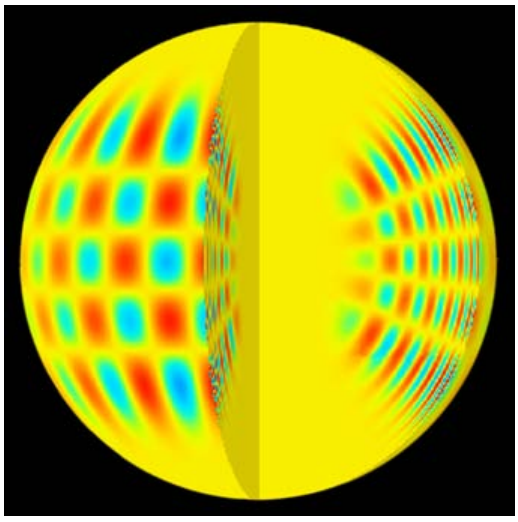


*Figure 2-75. Global solar oscillations*
*(from [http://en.wikipedia.org/wiki/Helioseismology](http://en.wikipedia.org/wiki/Helioseismology) ).*