Geometric Optics

Roland Shack
University of Arizona, Optical Sciences Center, Tucson, Arizona, USA
Phone: 520-621-1356; e-mail: roland.shack@optics.arizona.edu

Abstract
This article reviews the most elementary theory of optics, namely geometric optics, and describes its connection to quantum optics and other high-level theories. The basics of light propagation in the form of rays and wave-fronts are described. The main emphasis is on optical image-forming systems and system aberrations. This article provides basic knowledge for any optical engineer.

Keywords
geometric optics; optical engineering; image formation; aberrations; imaging systems.

1 Introduction

Optics is blessed with several levels of theory. At the top are quantum optics and electromagnetic theory, which account respectively for the microscopic and the macroscopic interaction of light and matter. This is the domain of optical physics, and all optical phenomena are encompassed by them.

However, in applications such as electromagnetic diffraction theory, exact solutions to problems are extremely difficult. To deal practically with such problems there is a simpler approximate theory, scalar wave theory, which provides an adequately accurate result in most applications involving interference and diffraction. It does not, however, account for polarization nor does it predict the quantitative interactions with matter. This is the domain of physical optics.

At a still lower level of approximation, where not even interference and diffraction are accounted for, is geometrical optics. Geometrical optics is the lowest level of optical theory available today. It does not compete with physical optics in explaining optical phenomena, and can only provide a crude approximation to the detailed behavior of light. Why is it still in use?

Its overwhelming virtue is in its simplicity. It can provide useful, if somewhat inexact, answers to many practical problems, where the complexity of a more accurate theory would be prohibitive. It has been applied successfully to problems where one would not expect useful results, for example, in the analysis of waveguides.

Its primary area of application, however, is in optical design. It has proven itself to be indispensable for the efficient development of optical systems. No useful alternative has yet been developed.

Until the development of the computer and its application to optical design, virtually no consideration was given to the physical nature of light. However, in recent years the application of the wave theory of geometrical optics has made it possible to convert from geometrical optics to physical optics in the final image space of an optical system, and to calculate the structure of the image taking diffraction into account. The hard work of tracing the wave front through the optical system is done geometrically.

In this article, we deal with the basic concepts of both the corpuscular theory and the wave theory of geometrical optics. Applications as an approximation to physical optics and to radiometry are touched on, but the major part of this article deals with the geometrical concepts relevant to an understanding of image-forming optics.

2 Basic Concepts

Historically, there have been two different approaches to the theory of geometrical optics. The oldest and most prevalent one is the corpuscular theory in which it is assumed that light consists of particles that transport energy from a source to a receiver. The other is a wave theory, which goes as far back as Christian Huygens, and which is somewhat subtly different from physical wave optics.

2.1 Corpuscular Theory

2.1.1 Rays as Trajectories of Energetic Particles

In the corpuscular theory, light is assumed to consist of energetic particles. Particles of different colors have different energies,

the blue being more energetic than the red. In this respect, they can be thought of as simplified versions of physical photons.

These particles are emitted from a source and travel along trajectories to a receiver. The trajectories are identified as ray paths, or simply rays. In classical geometrical optics, the study of rays and their properties is paramount.

One important characteristic of geometrical particles of light is that they never collide or interfere with each other regardless of the intensity of the light beam. Intense beams can intersect each other with no interaction.

2.1.2 Refractive Index, Color, and Dispersion

In a vacuum, all light particles travel at the well-known speed of light. However, in transparent media they travel more slowly, and particles of different colors travel at different speeds. The ratio of the speed of light in a vacuum to the speed of light in a given medium is known as the refractive index of the medium. Blue light travels more slowly in the medium than red light, and so its refractive index is higher. The variation of refractive index with color is what is known as dispersion.

The designation of color by name results in a rather imprecise determination of dispersion, as indeed it was before Fraunhofer's discovery of spectrum lines. Today the wavelengths of the light producing the spectrum lines are used for a precise designation of color.

2.1.3 Fermat's Principle and Optical Paths

A fundamental property of ray paths is described by Fermat's principle. Given two points along a ray, the time it takes for light to go from one point to the other is given by

$$t_A - t_B = \int_{t_A}^{t_B} \mathrm{d}t = \frac{1}{c}\int_A^B n\,\mathrm{d}s, \qquad (1)$$

which is stationary, usually a minimum, along all possible neighboring paths. If the refractive index of the medium is constant along the path, the path will be a straight line. If the refractive index is not constant, the path will vary in such a way that Fermat's principle is satisfied.

The integral of the product of refractive index and geometrical path length is called the *optical path length*, and is the distance that light would travel in a vacuum in the same time interval. Because the speed of light in a vacuum is constant, Fermat's principle can be expressed in terms of optical path as well as transit time.

2.1.4 Snell's Law

Most optical systems consist of reasonably homogeneous media with constant refractive indices separated by abrupt discontinuities, for example, lenses in air. Given a boundary separating two dissimilar media, a ray will change its direction in going through the boundary unless it is normal to the boundary. This change in direction is governed by Snell's law,

$$n' \sin i' = n \sin i, \qquad (2)$$

where the primed quantities are in the emergent space, i is the angle between the incident ray and the normal to the boundary at the intersection point, and i' is the angle between the emergent ray and the normal.

Snell's law is the basic tool in designing and analyzing optical systems. It is simple and powerful, and can be derived from Fermat's principle.

2.2 Wave Theory

2.2.1 The Point Source

The basic elementary source in geometrical optics is the point source. It is an infinitesimal region of space that emits light. Associated with every point source is a family of rays that trace the trajectories of the light particles emitted. Different point sources and their ray families behave independently of each other.

2.2.2 Geometrical Wave Fronts

Associated with each point source and its family of rays is a family of surfaces. These surfaces are surfaces of constant optical path from the source measured along the rays, and they are called *geometrical wave fronts* (see also WAVE OPTICS). Although there is no characteristic wavelength associated with these surfaces, they are called *wave fronts* because they are generally good approximations to physical wave fronts except in the neighborhood of geometrical shadow boundaries. Christian Huygens was the first to use the idea of geometrical wave fronts in discussing the properties of light.

2.2.3 Connection with Physical Optics

As stated above, the precise specification of a color in geometrical optics is in terms of the wavelength of spectral lines. This wavelength is of course a physical quantity that in itself has no other significance in geometrical optics. However, it is a unit of length, and geometrical optical path lengths can be measured in wavelength units. This combined with the geometrical wave front gives us a geometrical model of a physical wave field.

2.2.4 Significance in Imaging Systems

Traditionally, optical image-forming systems have been designed and evaluated using ray optics almost exclusively. Before the advent of computers, the stupendous amount of work required to do it any other way was out of the question. However, today it is possible to trace geometrical wave fronts through an optical system (admittedly using rays to do it) to the final image space where one can convert to physical optics in order to account for diffraction in the image-forming process.

3 General Applications

There are a number of areas where geometrical optics is commonly applied. What they have in common is that the geometrical optical model is much easier to deal with than a more accurate higher-level model, and that, for these applications, the assumptions and approximations are adequate for the purpose. Two application areas are discussed briefly in this section, and the remainder of the article deals with the most important area, that of image-forming optics.

3.1 Physical Optics

3.1.1 Wave-front Propagation

The principal general area of physical optics where geometrical optical tools are commonly employed is in the analysis and evaluation of instruments in which wave-front transmission is the principal phenomenon, such as spectrometers and interferometers. Geometrical optics has even been applied successfully to waveguide analysis.

3.1.2 The Use of *Ad Hoc* Properties

A number of physical wave-field properties do not directly occur in the geometrical optical model. Only by assuming them as *ad hoc* properties can they be accounted for. One such property already mentioned is the use of wavelength as the unit for measuring optical paths. Other properties are wave amplitude, reflectance, transmittance, and polarization. When used with care, *ad hoc* properties are a significant enhancement to the geometrical optical model.

3.1.3 Diffracted Rays

One interesting special application is in extending the geometrical model to include diffraction, at least by sharp-edged apertures. It is based on the Rabinowicz model of diffraction in which he decomposes a diffracted field into the coherent superposition of a standard geometrical field truncated by the aperture and an induced field that appears as that produced by a coherent line source on the edge of the aperture. In the geometrical model, additional ''diffracted'' rays are created on the boundary.

3.2 Radiometry

In most cases, a geometrical optical model is adequate for dealing with the radiometry of optical systems. Although point sources are certainly included in radiometric phenomena, extended sources generally play a greater role. In the geometrical model, an extended source is simply a dense array of independent point sources. The independence of the point sources means that in radiometry, extended sources are assumed to be noncoherent.

4 Imaging Systems: First Order

The principal application of geometrical optics is in the area of image-forming systems. Principles and concepts specific to imaging systems are fairly extensive, and the rest of this article deals with these.

No optical system forms images perfectly, and departures from ideal behavior are called *aberrations*. However, to deal with aberrations we must first define what constitutes ideal behavior. Although there are subtle differences between the terms, ideal behavior is variously described as first order, collinear, paraxial, and Gaussian. Each of these will be dealt with in further detail in the following conceptual development.

4.1 Objects and Images: Optical Spaces

In every optical image-forming system, there is at least one object space and at least one image space. In compound imaging systems, each element takes the image space of the preceding element as its object space and reimages that into its own image space, which in turn is the object space of the next element. At its most basic level, the elements of a compound system are the optical surfaces. A single lens is a compound of two surfaces and has a total of three spaces. A three-lens system has a total of seven spaces.

These spaces are called *optical spaces* and are of infinite extent. They are not bounded by the optical surfaces. That part of the optical space that lies between adjacent surfaces is called the *real part*. The rest of the optical space is its virtual part. In a conventional drawing of an optical system showing ray paths from surface to surface, only the real parts of each space are shown.

4.2 Ideal Behavior: Collinear Mapping

An object consists of a three-dimensional array of point sources. Its image consists ideally of a corresponding array of point images. The ideal image is related to the object through a mapping process.

The most popular mapping to represent ideal behavior is a collinear mapping. The basic rule is that, in addition to there being a one-to-one correspondence between object and image points, there is also a one-to-one correspondence between straight lines in the two spaces. As a consequence, there is in addition a one-to-one correspondence between object and image planes. Corresponding elements are called *conjugate elements*.

4.3 Axially Symmetric Systems

In the vast majority of optical systems, every surface and aperture in the system is intended to be rotationally symmetric about a single common mechanical axis. The basic concepts in the geometric theory of image-forming systems were developed in the context of such systems, and most of the following discussion assumes axial symmetry.

4.3.1 Focal Systems and Gaussian Imagery

Each space in an axially symmetric system has its own axis of symmetry. They all coincide with the mechanical axis of symmetry, but each space is distinct in all its properties.

Each point on the object axis has one and only one point on the image axis corresponding to it. If an object point at infinity on the axis has a conjugate at a finite location on the image axis, the conjugate image point is the rear focal point of the system. There is a corresponding front focal point that has as its conjugate a point at infinity on the image axis. Such a system is called a *focal system*.

4.3.1.1 **Cardinal Points** The two focal points of a focal system are two of the six cardinal points defined by Gauss. The other four are defined as follows.

A point off axis in the object space will have a conjugate image point off axis. The distance of such a point from the axis is called the *height of the point*, and the ratio of the image height to the object height is called *the magnification of the pair*.

An object plane perpendicular to the axis and containing the given point has as its conjugate an image plane, also perpendicular to the image axis, that contains the given image point. All conjugate pairs of points in the two planes have the same magnification. Every pair of conjugate planes has a unique magnification connecting them. No two pairs have the same magnification. The pair of planes having a unit positive magnification are called the *principal planes of the system*. Their axial points are called the *principal points*. These are also two of the cardinal points.

The distance from the principal point to the focal point in image space is called the *rear focal length* of the system, and the corresponding distance in object space is called the *front focal length*.

The last pair of cardinal points are the nodal points. These are conjugate axial points located so that a line passing through the front nodal point making some angle with the axis will have as its conjugate a line passing through the rear nodal point making the same angle with the axis. These make it possible to determine the image size in the rear focal

plane of an object at infinity but with a finite angular extent.

The cardinal points of a system are used to establish the constants used in the collinear mapping of object space into image space. The term Gaussian imagery is used to describe this application of the cardinal points, but the underlying model is collinear.

4.3.1.2 **Gaussian Imagery** It is important in developing the mapping equations to be consistent in the assumed sign convention. We assume a standard Cartesian coordinate system in each space with the standard Cartesian sign conventions. We also identify the axis with the optical axis in each of the spaces.

The usual mapping equations take the origin of the coordinate system in each space to be located at the principal point. The distances to a pair of conjugate planes are given by

$$\frac{n'}{z'} = \frac{n}{z} + \phi, \tag{3}$$

where

$$\phi = \frac{n'}{f'} = -\frac{n}{f} = \frac{1}{f_e} \tag{4}$$

is the power of the system, f and f' are the front and rear focal lengths respectively, and f_e, the reciprocal of the power, is the effective focal length. The magnification is given by

$$m = \frac{z'/n'}{z/n}, \tag{5}$$

and the transverse coordinates by

$$x' = mx, \quad y' = my. \tag{6}$$

4.3.1.3 **Gaussian Reduction** It is a relatively simple thing to determine the Gaussian properties (the power and the location of the principal points) of a compound system if the Gaussian properties of the components are known. The reduction is carried out for two components at a time. The most elementary component is a single refracting surface, and the Gaussian properties of a single lens can be obtained by this method. Once one has the Gaussian properties of the individual lenses in a lens system, the properties of the system as a whole can be obtained by continuing the process.

If we have two components, where the power of the first is ϕ_1 and the power of the second is ϕ_2, there are three spaces involved. The object space for the first component is the object space for the system, and the image space of the second component is the image space for the system. Let the refractive index of the object space be n_o and the image space n'_o.

The front principal point P_1 of the first component is in the object space, as is the front principal point P of the system. The rear principal point P'_2 of the second component is in the image space, as is the rear principal point P' of the system.

The space between the two components is common to both. It is the image space of the first component and the object space of the second component, and contains respectively the rear principal point P'_1 of the first component and the front principal point P_2 of the second component. Let the distance from P'_1 to P_2 be t and the refractive index of the space be n. Then the power of the system is given by

$$\phi = \phi_1 + \phi_2 - \phi_1\phi_2 t/n, \tag{7}$$

and the principal points of the system are located by

$$\frac{\overline{P_1P}}{n_o} = \left(\frac{\phi_2}{\phi}\right)\frac{t}{n}, \quad \frac{\overline{P'_2P'}}{n'_o} = -\left(\frac{\phi_1}{\phi}\right)\frac{t}{n}, \tag{8}$$

where $\overline{P_1P}$ is the distance in object space from the front principal point of the first component to the front principal point of the system, and $\overline{P'_2P'}$ is the distance in image space from the rear principal point of the second component to the rear principal point of the system. The resultant system can in turn be a component in a larger system.

4.3.2 Afocal Systems

It is also possible for a system to have an image point at infinity corresponding to an object point at infinity. Such a system is called an *afocal system*.

Afocal systems have no cardinal points. Instead, they have a constant characteristic magnification; that is, all pairs of conjugate planes have the same magnification. An arbitrarily selected pair of conjugate planes can be used to establish the origins of the coordinate systems. The mapping equations are then

$$\frac{z'}{n} = \frac{m^2 z}{n}, \quad x' = mx, \quad y' = my. \quad (9)$$

4.3.3 Ray Tracing and Paraxial Optics

The general procedure for tracing a ray through an optical system is sequential. An incident ray at an optical surface is defined by its location on the surface and its direction of propagation. By applying Snell's law the ray is refracted, determining the direction of the emergent ray. The ray is transferred to the next surface, for which it is the incident ray, and the process is repeated. The operations are fairly complex, and very tedious if done by hand.

However, in a region sufficiently close to the axis that angles, sines, and tangents cannot be distinguished from each other, the ray tracing is considerably simplified. Both the refraction and the transfer operations reduce to simple linear equations. This paraxial behavior of the optical system is congruent with the paraxial behavior of the collinear model of the optical system, and because the ray-tracing equations are linear, they can be applied to the extended collinear model of the optical system without restriction to its paraxial region.

If the real ray-tracing equations had been represented by a power-series expansion, the first-order terms would have to be congruent with the paraxial behavior. This is where the idea of first-order optics comes from.

In the collinear model that represents the ideal optical system, every refracting surface is represented by its principal planes, which are coincident at the surface. In the ray tracing, refraction occurs at the principal planes and transfer occurs between adjacent principal planes. The rays should properly be called *collinear rays*, but they are commonly called *paraxial rays*, even though they are not restricted to the paraxial region.

In the refraction process, a ray incident on a surface is determined by its height at the surface and the angle that it makes with the optical axis. The emergent ray is determined by the refraction equation

$$n'u' = nu - y\phi, \quad (10)$$

where $\phi = (n'n)c$ is the power of the surface and c is the curvature of the surface. The transfer to the next surface is governed by

$$y' = y + n'u'\left(\frac{t'}{n'}\right), \quad (11)$$

where t' is the axial distance from the first surface to the second. This ray is now incident on the second surface, and the process can be repeated until the ray trace is complete.

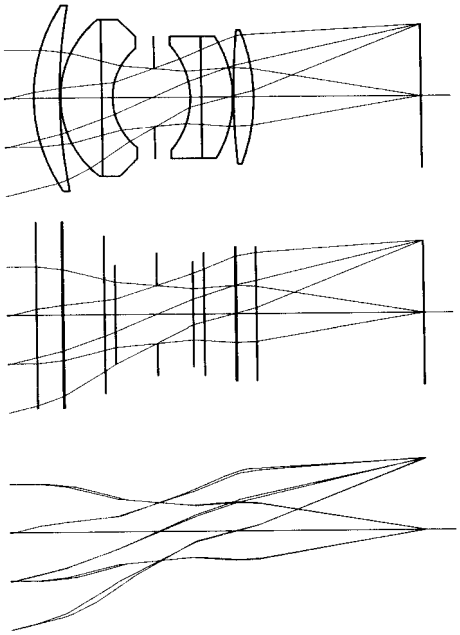

Fig. 1 Collinear (first-order) modeling. The top part represents a real system with real ray paths. The middle part shows the collinear model with collinear (paraxial) rays. The bottom part compares the real ray paths with the collinear ray paths

The cardinal points of the complete system can be determined by tracing just two rays without concern for the cardinal points of the components (see Fig. 1). In general, Gaussian reduction is more efficient if there are no more than three or four surfaces. For larger numbers, the ray trace is much more efficient.

4.3.4 Fields and Pupils

It is customary in describing an optical image-forming system to identify a unique object plane perpendicular to the axis as the object, and its conjugate as the image. The useful areas of these surfaces are also usually limited in extent, either by a physical aperture known as a field stop, or by a virtual limit imposed by the application. The resulting limited surfaces are called the *object and image fields*, and paraxial (collinear) ray tracing is used to find the location and size of the image field if the object field is given.

For an object point on axis, only a limited beam of light can actually get through the optical system to form the image point on axis. The aperture in the system that limits the size of this beam is called the *aperture stop* of the system, and it often is located internally. The object-space conjugate to the aperture stop is called the *entrance pupil* and the image-space conjugate the exit pupil. The beam in object space appears to be limited by the entrance pupil and in image space by the exit pupil.

4.3.5 Marginal and Chief Rays

Because of the axial symmetry assumed for the optical system, the axial beam of light is completely characterized by a single ray, one that goes from the axial object point to the edge of the aperture stop and on to the axial image point. This ray is called the *marginal ray*.

For an object point at the edge of the field, the beam connecting it with its image point is also limited by the aperture stop, and to a first approximation has the same cross section everywhere as the axial beam except for being displaced from the axis. The displacement is entirely characterized by a ray going from the point at the edge of the field through the center of the aperture stop and on to the conjugate image point. This ray is called the *chief ray*.

Only these two rays are necessary to specify the fields and pupils. Wherever the marginal ray crosses the axis we have an image, and its size is given by the chief ray height at that location. Wherever the chief ray crosses the axis we have a pupil, and its size is given by the marginal ray.

The chief ray properties are distinguished from the marginal ray properties by the variables being barred, whereas the marginal ray variables remain unbarred.

4.3.6 The Optical (Lagrange) Invariant

At any plane in the optical system, the marginal ray is characterized by its height on the plane and its angle with respect to the axis. The chief ray is similarly characterized. A very interesting and useful relationship between the two rays is obtained by forming the combination

$$L = n\bar{u}y - nu\bar{y}. \tag{12}$$

This quantity is constant throughout the system and is called the *optical*, or Lagrange, invariant. It is particularly useful at the field and pupil planes, having the same value at all of them.

One simple application involves the Lagrange invariant at the object and image planes. At the object and image planes respectively,

$$L = -nu\bar{y} = -n'y'\bar{y}', \tag{13}$$

and so we can express the magnification connecting the image with the object in terms of the marginal ray angles:

$$m = \frac{y'}{y} = \frac{nu}{n'u'}. \tag{14}$$

4.3.7 Numerical Aperture and the *f* Number

The magnitude of the quantity nu is a paraxial approximation to a quantity known as the numerical aperture, which is strictly given by $n \sin u$ and is usually represented by the symbol NA. The numerical aperture is an important quantity for several reasons. In physical optics, it determines the resolution of the system, but even in geometrical optics the numerical aperture in the image space determines the irradiance on the image plane; the irradiance varies as the square of the numerical aperture.

In the early days of photography, it was recognized that some simple yet effective way of controlling the exposure of the photographic negative was needed. The amount of light getting to the film clearly was proportional to the area of the entrance pupil. Also, the irradiance on the film plane for a given pupil area was inversely proportional to the focal length of the lens. A long–focal-length lens with the same pupil area put less light on the film plane in the same exposure time. Thus, the ratio of the focal length of the lens to its entrance pupil diameter, or f number, was established as a quantity that could be used to control the exposure. Lenses of the same f number give the same exposure in the same exposure time, regardless of the focal length.

When the object is at infinity, the numerical aperture is equal to the reciprocal of twice the f number.

Complications arose when the object was so close to the lens that the image distance was considerably larger than the focal length. The exposure of course went down, and because the definition of the f number involved the focal length, not the image distance, some modification of the exposure factor involving the magnification was in order. Various fixes, some more ungainly than others, have persisted to this day. The fundamental problem is that the numerical aperture is the quantity that actually controls the exposure, and the f number was an empirical creation, adequate for the conditions it was designed for. Today, the numerical aperture is as easy to measure as the f number, and because

it is completely general, one might, for the convenience of those who persist in using the f number, define an effective f number that is the reciprocal of twice the numerical aperture.

4.4 Systems with No Axial Symmetry

Most of the above properties seem to have depended on the presence of an axis of rotational symmetry. If you do not have an axis, how can you have paraxial behavior?

The fact is, we do have a kind of axis. There is some point on the object defined as the center of the object field, and there is some element in the system that limits the beam of light coming from that point. The latter is the aperture stop, and its image in object space is the entrance pupil. A real ray traced from the central object point through the center of the entrance pupil will eventually pass through the center of the aperture stop and on through the system until it reaches the image surface at what is therefore defined as the center of the image field. This real ray is called the *optical-axis ray*, and is the nearest thing we have to an optical axis.

Unfortunately, rays that are paraxial to the optical-axis ray can in general have large angles of incidence at the refracting or reflecting surfaces, and the ray-tracing equations do not simplify the way they do for an axially symmetric system. On the other hand, the collinear mapping process does not require any imposed symmetry conditions. For a focal system, unique front and rear focal planes exist, and object planes parallel to the front focal plane have as conjugates planes parallel to the rear focal plane. Moreover, there is a line normal to the front focal plane that has as its conjugate a line normal to the rear focal plane. These lines can be identified as the axes in their respective spaces.

The mapping between any conjugate pair of the planes parallel to the focal planes is affine, that is, the image is at most anamorphic. This means that there are a pair of orthogonal directions in the image plane where the magnification is constant, although it may be different in the two directions. We can select the orientation of the x and y axes to correspond with these directions. The result is a set of mapping equations that are very nearly as simple as those for an axially symmetric system. The only added feature is the anamorphism between object and image planes.

The tough part of the problem is how to obtain the collinear mapping parameters from the system properties. A general solution to this problem has not yet been obtained.

5 Imaging Systems: Aberrations

5.1 Aberrations as Departures from Ideal Behavior

The collinear model represents ideal behavior in an optical image-forming system. Real systems depart from this behavior, and these departures are called *aberrations*. (For a more mathematical presentation, see OPTICAL ABERRATIONS.)

From a ray point of view, all real rays from any object point should pass through the ideal collinear image point. If a real ray misses the ideal image point, we say that the ray is aberrated. The transverse displacement of the ray intersection with image plane from the ideal image point is the usual measure of a ray aberration.

From a wave-front point of view, the wave front emerging from the exit pupil

should be spherical and centered on the ideal image point. If it is not spherical, or if it is not centered at the ideal image point, we say that the wave front is aberrated. The ideal wave front is usually called the *reference sphere* from which the wave aberration is measured. At any given point on the reference sphere, the wave aberration is the optical path distance from the reference sphere to the aberrated wave front along the ray passing through that point. The wave aberration function is the wave aberration as a function of the position on the reference sphere. The latter is the proper surface of reference for the pupil.

5.2 Wave and Ray Aberrations

Rays are everywhere normal to geometrical wave fronts, and so ray aberrations and wave aberrations must be connected. As a function of pupil coordinates, the wave aberration function is a scalar function, whereas the transverse ray aberrations are vector quantities. If the pupil coordinates are chosen to be the transverse coordinates on the reference sphere, it can be shown that, to an excellent approximation, the transverse ray aberrations are proportional to the gradient of the wave aberration function.

Between the two types of aberration, the wave aberration is a simpler function of the pupil coordinates, and the ray aberrations are simply derived from it. The reverse is not true.

Another significant advantage of the wave aberration is that the geometrical wave front is a good approximation to a physical wave front, especially in the exit pupil where Fresnel diffraction effects are minimal. The wave aberration function is easily and simply converted into a phase variation over the reference sphere, the latter being the proper surface of integration for a diffraction integral.

5.3 Monochromatic Aberrations

In general, different wavelengths of light result in different aberration functions, and these variations with wavelength are identified as chromatic variations. Before dealing with them, we first must understand the aberrations at a given wavelength, the monochromatic aberrations.

5.3.1 Expansions and Classification

The wave aberration function is a function of pupil coordinates and is also parametrically a function of field coordinates. That is, for every point in the field there is a corresponding wave aberration function that may be different for different field points. Taking into account the axial symmetry of the optical system, it is possible to represent the wave aberration function by a two-dimensional power-series expansion in the pupil coordinates, the terms of which are the elementary wave aberrations.

The principal classification of the terms is by order determined dimensionally (see Fig. 2).

5.3.2 Elementary (Third-order) Aberrations

The lowest nontrivial order for monochromatic aberrations is called *third order*, although the terms are dimensionally fourth degree. The labeling was historically determined by a ray classification, and the ray aberrations are determined by the gradient of the wave aberration function. There are five significant third-order aberrations.

5.3.2.1 **Spherical Aberration** Spherical aberration is the only aberration that is

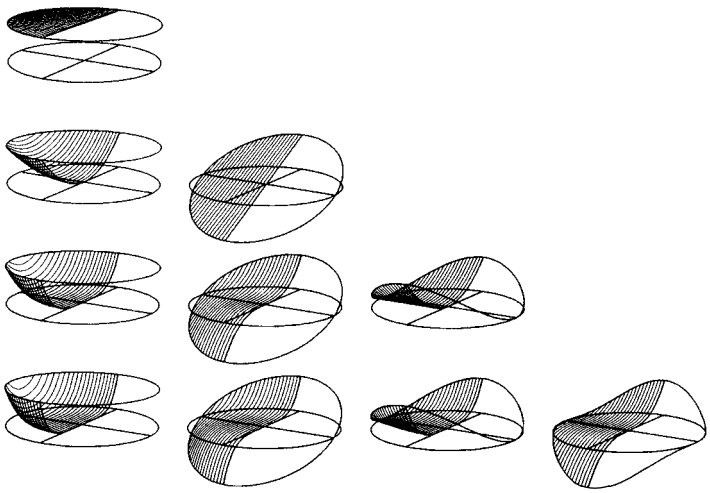

Fig. 2 Wave aberration functions. Each row represents a different order of aberration. The third row down shows the third-order aberration functions for spherical aberration, coma, and astigmatism

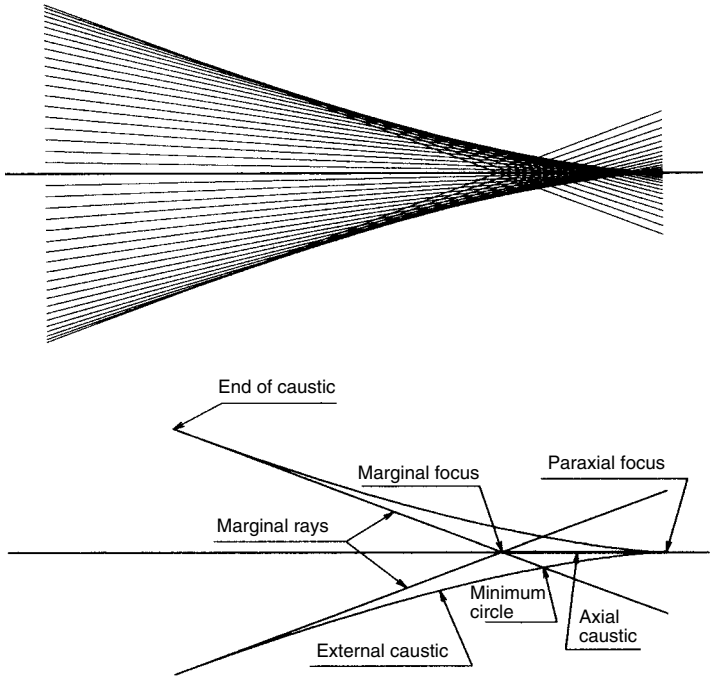

Fig. 3 The caustic produced by spherical aberration. The top part shows the ray paths; the bottom part, at the same scale, shows the structure of the caustic

independent of the field. It is therefore the only monochromatic aberration to affect the axial image. Although its effect is felt over the entire field, it is simpler to describe as an axial aberration.

The wave aberration function is a rotationally symmetric, fourth-degree departure from the reference sphere. All the rays in a given pupil zone come to a common focus on the axis, but the axial location of these zonal foci varies with the radius of the zone. In addition to this zonal focal shift, the rays produce a trumpet-shaped external caustic, seen as the envelope of the rays, that is characteristic of spherical aberration (see Fig. 3).

The geometrically predicted image is represented by tracing a uniformly dense mesh of rays in the pupil and observing the array of intersection points with an observation plane in the neighborhood of the ideal image. This array of points is called a *spot diagram*. In the case of spherical aberration, if we move the observation plane through focus (see Fig. 4), we find that there is no symmetry with respect to focus.

5.3.2.2 **Coma** Coma is an aberration that varies linearly with field height and is therefore absent on axis. Its wave aberration function has a cubic cross section, and the ray aberrations are all on one side of the ideal image, either toward the center of the field or away from it, depending on the sign of the aberration coefficient. Moreover, at best focus the rays are contained within a 60° angle with the ideal image at its vertex.

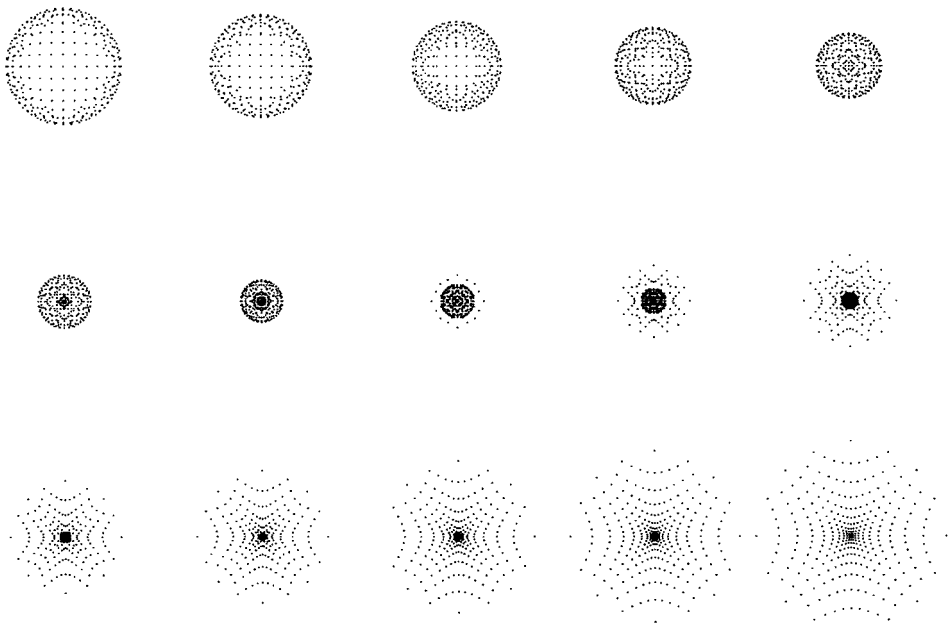

Fig. 4 A through-focus sequence of spot diagrams for spherical aberration. In this figure, and in Figs. 5 and 6, the sequence of fifteen focal positions is from left to right along each row and from top row to bottom row. Note the asymmetry through focus

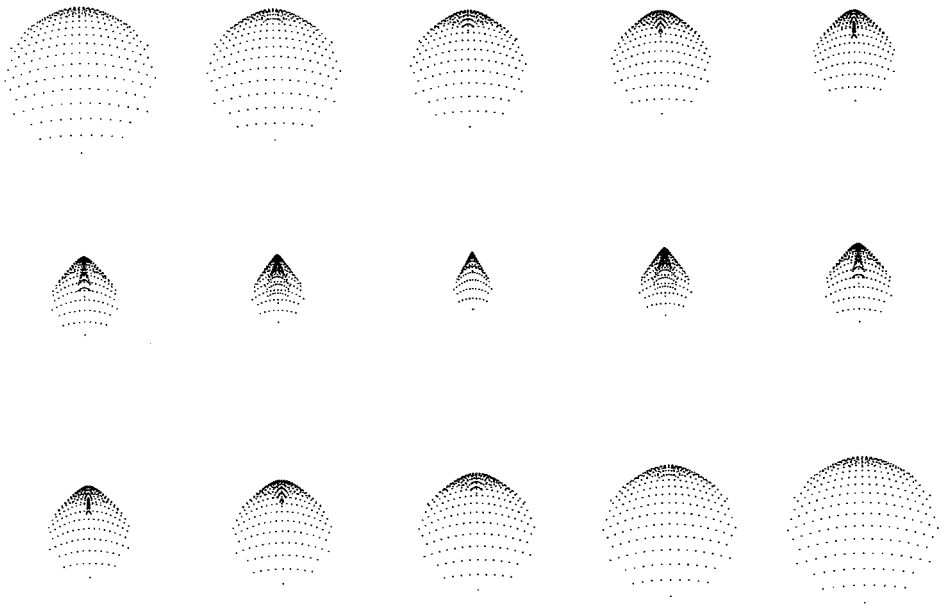

Fig. 5 A through-focus sequence of spot diagrams for coma. The images are symmetrical through focus, but they are asymmetrical in the vertical direction

Coma is symmetrical through focus, as shown in Fig. 5.

5.3.2.3 **Astigmatism** Astigmatism is an aberration that varies as the square of the field height and is absent on axis. The wave aberration function is a quadratic cylinder, constant in a direction perpendicular to the meridional plane containing the ideal field point. (A meridional plane is a plane that contains the axis of rotational symmetry.)

The distinguishing feature of astigmatism is the presence of two linear foci, one in the meridional plane and the other perpendicular to it (see Fig. 6). The first is called the *sagittal focus* and the second the tangential focus. The two foci are separated along the chief ray by a distance that is proportional to the amount of astigmatism. In the absence of other aberrations the sagittal field is flat, but because the focus shift to the tangential image varies as the square of the field height, the surface containing the tangential images is curved.

5.3.2.4 **Field Curvature** Field curvature is an aberration in which the wave front is spherical and all the rays for a given image pass through a single point, but this point is shifted along the chief ray by an amount that increases quadratically with field height. As a consequence, in the absence of other aberrations, the images are locally very good, but they lie on a curved surface.

5.3.2.5 **Distortion** Distortion is also an aberration in which the wave front is spherical and all the rays pass through a single point, but in this case the points remain in the image plane. However, they are displaced toward or away from the axis by an amount that varies as the cube of the field height. The consequence of this is

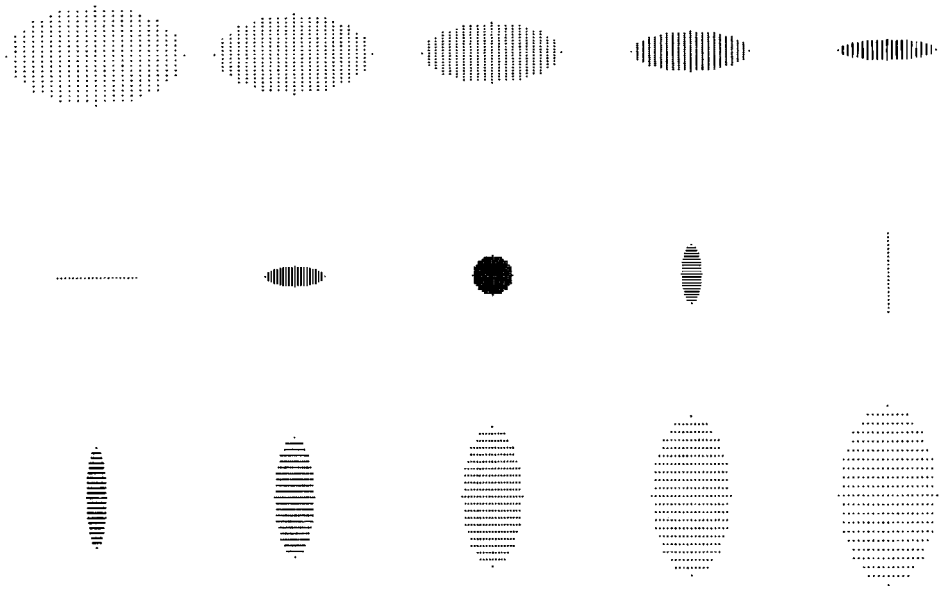

Fig. 6 A through-focus sequence of spot diagrams for astigmatism. The images are symmetrical about the vertical and horizontal axes, but there is a twist symmetry through focus

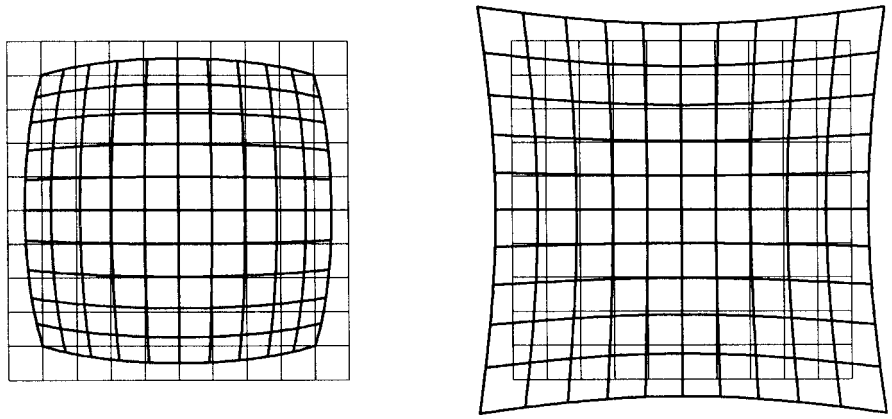

Fig. 7 Distortion. The part on the left shows barrel distortion; the part on the right shows pincushion distortion

that images are radially distorted. A square centered on the axis is imaged with curved sides. If the sides are concave the result is called *pincushion distortion*, whereas if the sides are convex it is called *barrel distortion* (see Fig. 7).

5.4 Chromatic Aberrations

Chromatic aberrations arise because the optical properties of a lens system depend on the refractive indices of the

components, and the refractive indices vary with wavelength.

5.4.1 Elementary (Primary) Chromatic Aberrations

Even the first-order properties of a system can change with wavelength. The resulting aberrations are the primary chromatic aberrations. There are only two, longitudinal chromatic aberration and transverse chromatic aberration. The first is a change in focal position with wavelength and the second is a change in image size, or magnification, with wavelength.

Mirror systems do not suffer from chromatic aberrations because they do not contain dispersive elements.

5.4.2 Chromatic Variation of Monochromatic Aberrations

In addition to the primary chromatic aberrations, the aberration types that have been identified as monochromatic can also vary with wavelength. In some cases, the chromatic variation of these aberrations is common enough for them to be given their own names; for example, the chromatic variation of spherical aberration is called *spherochromatism*.

5.5 The Control of Aberrations in Optical Design

Some of the aberrations of a single lens can be modified by bending the lens (changing the shape of the lens without changing its power) and by locating the aperture stop away from the lens. Others – for example, longitudinal chromatic aberration – are insensitive to these variables. At least two lenses of different glasses are required to correct chromatic aberrations.

All of the third-order aberrations and the primary chromatic aberrations are correctable with only three lenses, but the performance of this triplet, although fairly good, is limited by higher-order aberrations. The complexity of a high-speed camera lens is due entirely to the need for a high degree of aberration correction.

Even in a well-corrected lens, the state of correction is the result of balancing sometimes surprisingly large aberrations from one or more elements with correspondingly large aberrations of opposite sign from other elements. As a result, manufacturing tolerances are usually very tight.

In the case of mirror systems, there are no chromatic aberrations to be corrected. However, a mirror cannot be ''bent'' like a lens to modify its performance. Aberrations are primarily controlled by making the mirror surfaces aspheric.

Lens systems can also use aspheric surfaces, but until recently this has been avoided because of the expense involved in producing them. Aspheric mirrors are expensive too, but there is no other recourse.

Glossary

Aberration: A departure from ideal behavior.

Afocal Systems: Systems that do not contain focal planes. An object plane at infinity has its conjugate image plane at infinity. Rays parallel to the axis in object space have conjugate rays parallel to the axis in image space.

Aperture Stop: That aperture that limits the size of a beam going from the axial object point to the axial image point.

Cardinal Points: The Focal Points, principal Points, and Nodal Points of a system taken as a set.

Caustic: The envelope of all rays emanating from a point and reflected or refracted by a curved surface.

Chief Ray: A ray that goes from the edge of the object field to the edge of the image field, passing through the center of the aperture stop on the way.

Collinear Mapping: A mapping of an object space into an image space in which there is a one-to-one correspondence between points, between straight lines, and between planes in the two spaces.

Conjugate Elements: Object and image elements (points, lines, or planes) that are in one-to-one correspondence with each other.

Dispersion: The variation in refractive index of a medium.

Effective f Number: The reciprocal of twice the numerical aperture. It is a property of the space (object or image) of the system.

Entrance Pupil: The image of the aperture stop in the object space of the system.

Exit Pupil: The image of the aperture stop in the image space of the system.

f Number: The ratio of the effective focal length of a system to its entrance pupil diameter. It is a property of the system.

Focal Length (Effective): The reciprocal of the power of a focal system.

Focal Lengths (Front and Rear): The distances from the principal points to the focal points.

Focal Systems: Systems that contain focal planes. An object plane at infinity has a conjugate plane at a finite distance from the optical system in the image space.

Front Focal Plane: Object plane that is conjugate to an image plane at infinity.

Front Focal Point: The axial point in the front focal plane.

Height: The distance from the axis of an object or image point.

Ideal Behavior: A Collinear Mapping.

Image Space: The optical space that contains the image of an optical element.

Magnification: The ratio of the image height to the object height for an off-axis object point.

Marginal Ray: A ray that goes from the axial object point to the axial image point, touching the edge of the aperture stop on its way.

Meridional Plane: A plane in an optical system that contains the optical axis.

Nodal Points (Front and Rear): Conjugate axial points where any object ray passing through the front nodal point and its conjugate ray passing through the rear nodal point make equal angles with the optical axis.

Numerical Aperture: The sine of the angle between the real marginal ray in the object

or image space of a system multiplied by the refractive index of the space. It is a property of the space, and not of the system as a whole.

Object Space: The optical space that contains the object of an optical element.

Optical Angles: Ray angles multiplied by the refractive index of the space that contains them.

Optical Path Length: The product of the refractive index and the geometrical path length along a ray.

Optical Space: The object or image space of an optical element.

Paraxial: Close to the axis.

Power: The refractive strength of a system. It is equal to the reciprocal equivalent length of a focal system. It is zero for afocal systems.

Principal Planes (Front and Rear): The conjugate object and image planes that have a unit positive magnification connecting them.

Principal Points (Front and Rear): The axial points of the principal planes.

Ray: (1) The trajectory along which light particles travel from a source to a receiver. (2) The normal to a wave front.

Ray Aberration (Transverse): The departure of the ray intersection with the image plane from its ideal position.

Real Image: An image formed in the real part of an image space.

Real Part of a Space: That part of an optical space that is normally seen in a conventional drawing of an optical system.

Rear Focal Plane: The image plane that is conjugate to an object plane at infinity.

Rear Focal Point: The axial point in the rear focal plane.

Reduced Distances: Axial distances divided by the refractive index of the space in which they exist.

Refractive Index: The ratio of the speed of light in vacuum to the speed of light in a medium.

Virtual Part of a Space: That part of an optical space that is outside the real part.

Wave Aberration: The departure of an aberrated wave front in the exit pupil from the ideal wave front, measured along the ray.

Wave Front: A surface of constant optical path from the source.

Further Reading

Optical Design

Kidger, M. J. (2002), *Fundamental Optical Design*. Bellingham, WA: SPIE Press.

Kingslake, R. (1978), *Lens Design Fundamentals*. New York: Academic Press.

Kingslake, R. (1983), *Optical System Design*. New York: Academic Press.

Malacara, D., Thompson, B. J. (2001), *Handbook of Optical Engineering*. New York: Marcel-Dekker.

O'Shea, D. C. (1985), *Elements of Modern Optical Design*. New York: Wiley-Interscience.

Shannon, R. R. (1997), *The art and science of optical design*. Cambridge: University Press.

Smith, W. J. (1990), *Modern Optical Engineering*. New York: McGraw-Hill.

Theoretical

Buchdahl, H. A. (1993), *An Introduction to Hamiltonian Optics*. New York: Dover Publications.

Hansen, R. C., Ed. (1981), *Geometric Theory of Diffraction*. New York: IEEE Press.

Luneburg, R. K. (1966), *Mathematical Theory of Optics*. Berkeley, CA: University of California Press.

Mahajan, V. N. (1998), *Optical Imaging and Aberrations: Part I. Ray Geometrical Optics*. Bellingham, WA: SPIE Press.

Mahajan, V. N. (2001), *Optical Imaging and Aberrations: Part II. Wave Diffraction Optics*. Bellingham, WA: SPIE Press.

Sommerfeld, A. (1964), *Optics*. New York: Academic Press.

Stavroudis, O. N. (1972), *The Optics of Rays, Wavefronts, and Caustics*. New York: Academic Press.

Welford, W. T. (1986), *Aberrations of Optical Systems*. London: Adam Hilger.