CHAPTER 9

Optical Systems

A pile of rocks ceases to be a rock pile when somebody contemplates it with the idea of a cathedral in mind.

—Antoine de Saint-Exupéry, *Flight to Arras*

9.1 INTRODUCTION

An optical system is a collection of sources, lenses, mirrors, detectors, and other stuff that (we hope) does some identifiable useful thing. We've talked about the pieces individually, but until now we haven't spent much time on how they work together. This chapter should help you to think through the behavior of the system from end to end, to see how it ought to behave. That means we need to talk about the behavior of light in systems: practical aberration and diffraction theory, illumination and detection, and how to calculate the actual system output from the behavior of the individual elements.

9.2 WHAT EXACTLY DOES A LENS DO?

In Section 4.11.2, we looked at the Gaussian (i.e., paraxial) imaging properties of lenses. We were able to locate the focus of an optical system, calculate magnification, and generally follow the progress of a general paraxial ray through an optical system by means of multiplication of *ABCD* matrices.

Here, we concentrate on the finer points, such as the aberrations of an optical system, which are its deviations from perfect imaging performance. We will use three pictures: the pure ray optics approach, where the aberrations show up as ray spot diagrams where not all the rays pass through the image point; the pure wave approach, where the aberrations are identified with the coefficients of a polynomial expansion of the crinkled wavefront, derived from exact calculation of the wave propagation through the system; and a hybrid ray/wave picture. (See Figure 9.1.)

The hybrid picture is messy but useful and, in fact, is the basis of most "wave optics" models. It takes advantage of the fact that ray optics does a good job except near focus or in other situations where diffraction is expected to be important. Accordingly, we trace rays to the vicinity of the exit pupil from a single object point, construct a wavefront

Building Electro-Optical Systems, Making It All Work, Second Edition, By Philip C. D. Hobbs

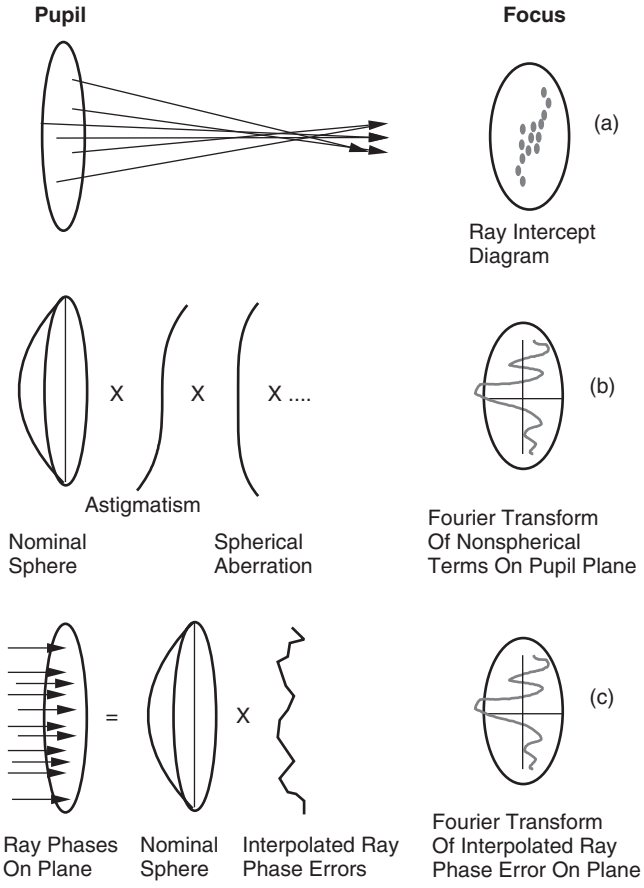

Figure 9.1. Three ways of looking at an imperfect optical system: (a) ray spot diagram, (b) wavefront polynomial expansion, and (c) wave aberration of rays.

whose phase is determined by the calculated propagation phase along the ray paths, and then use wave optics from there to the focus. This is unambiguous unless rays traversing widely different paths get too close to one another. By now you'll recognize this as another case of putting it in by hand, which is such a fruitful approach in all of engineering.

The different pictures contain different information. A ray tracing run is very poor at predicting the shape of the focused spot, but contains lots of information about the performance of the system across the field of view. For example, field curvature and geometric distortion show up clearly in a ray trace, since different field angles are presented at once, but tend to disappear in a pure wave propagation analysis, where only a single field position can readily be considered at a time.

9.2.1 Ray Optics

Ray optics assumes that all surfaces are locally planar, and that all fields behave locally like plane waves. To compute the path of a ray encountering a curved surface, we

notionally expand the ray into a plane wave, and the surface into its tangent plane. We then apply the law of reflection and Snell's law to derive the direction of the reflected and refracted **k** vectors, and these become the directions and amplitudes of the reflected and refracted rays. This is a first-order asymptotic approach, valid in the limit $ka \to \infty$, where a is the typical dimension of the surface (e.g., its radius of curvature, diameter, or whatever is most appropriate). There is a slight additional subtlety. A single ray, being infinitesimally thin, transports no energy; to find out the field amplitudes, we must consider small bundles of rays, or *pencil beams*, occupying an element of cross-sectional area $d\mathbf{A}$, measured in a plane normal to their propagation direction. Conservation of energy requires that the product of the ray intensity $I\, d\mathbf{A}$ be constant along the axis. Curved surfaces and any refraction or diffraction will in general cause $|d\mathbf{A}|$ to change, either by anamorphic magnification or by focusing. Thus in computing the contribution Di of a given ray bundle to the intensity at a given point $\mathbf{x}'$ from that at $\mathbf{x}$, we must multiply by the Jacobian,

$$dI(\mathbf{x}') = dI(\mathbf{x}) \left| \frac{d\mathbf{A}}{d\mathbf{A}'} \right| . \tag{9.1}$$

If the incoming illumination is spatially coherent, we must instead sum the (vector) fields, which transform as the square root of the Jacobian,

$$d\mathbf{E}(\mathbf{x}') = d\mathbf{E}(\mathbf{x}) \sqrt{\left| \frac{d\mathbf{A}}{d\mathbf{A}'} \right|} . \tag{9.2}$$

Going the other way, for example, computing a specular reflection by starting from the obliquely illuminated patch to the propagating beam, we have to put in the reciprocal of the obliquity factor—otherwise energy wouldn't be conserved on reflection from a perfect mirror. (See Section 9.3.5.) The mathematical way of putting this is that the Jacobian of the oblique projection equals the ratio $\cos\theta_2/\cos\theta_1$. We saw this effect in radiation from planar surfaces in Section 1.3.12, and it shows up in wave optics as the obliquity factor (see Sections 9.2.1 and 9.3.4).

9.2.2 Connecting Rays and Waves: Wavefronts

In order to move from one picture to another, we have to have a good idea of their connections. The basic idea is that of a *wavefront* (Figure 9.2). Most people who have had an upper-level undergraduate physical optics class will picture a wavefront as a plane wave that has encountered some object (e.g., a perforated plane screen or a transparency) and has had amplitude and phase variations impressed upon it. While this picture isn't wrong, it also isn't what an optical designer means by a wavefront, and the differences are a frequent source of confusion, especially since the same diffraction integrals are employed and the conceptual differences are seldom made explicit.

A physicist will tell you that a wavefront is a surface of constant phase, which can have crinkles and ripples in it, whereas a lens designer will say it's the deviation from constant phase on a spherical surface centered on the Gaussian image point. In actual fact, whenever people actually calculate imaging with wavefronts (as opposed to waving

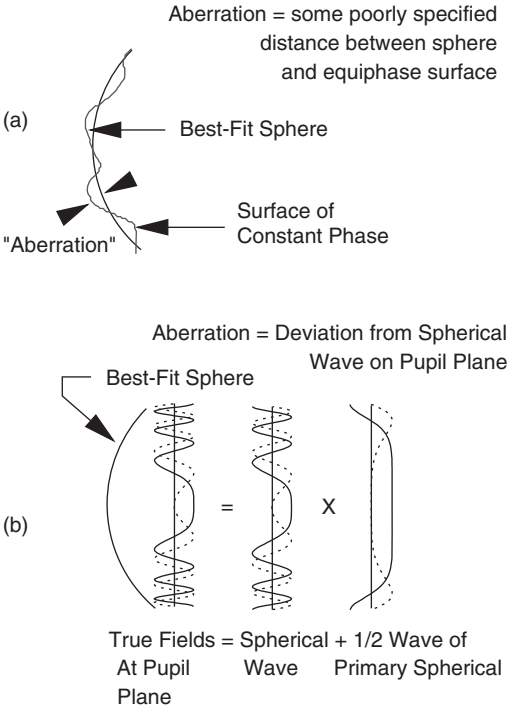

Figure 9.2. Wavefront definitions: (a) surface of constant phase and (b) deviation from a prespecified spherical wave on a plane.

their arms) the real definition is the phase deviation, on a plane, from a spherical wave:

$$W(\mathbf{x}) = K \arg(e^{ik|\mathbf{x}-\mathbf{x}_0|}\psi(\mathbf{x})), \tag{9.3}$$

where K is a constant that expresses the units in use: radians, waves (i.e., cycles), or OPD in meters. As long as the deviations from sphericity are sufficiently small and slow, the two descriptions are equivalent.

Deviations from the perfect spherical wave case are called *aberrations*; aberration theory is in its essence a theory of the way phase propagates in an optical system. Amplitude and polarization information is not given equal consideration, which leads to the total disregard of obliquity factors, among other things. A vernacular translation is, "it gives the wrong answer except with uniform illumination and small apertures, but is close enough to be useful."

Aside: Phase Fronts on the Brain. From what we know already of diffraction theory, concentrating solely on phase like this is obviously fishy; an amplitude-only object such as a zone plate can destroy the nice focusing properties, even while leaving the phase intact initially. Neglect of amplitude effects is our first clue that aberration theory fundamentally ignores diffraction. The software packages sold with some measuring interferometers have serious errors traceable to this insistence on the primacy of phase information. The author's unscientific sample is not encouraging; he has had two such units, from

different manufacturers, manufactured ten years apart. Both were seriously wrong, and in different ways.

9.2.3 Rays and the Eikonal Equation

We have the scalar wave equation, which allows us to predict the fields everywhere in a source-free half-space from an exact knowledge of the full, time-dependent, free-space scalar field on the boundary of the half-space. In the limit of smooth wavefronts and wavelengths short compared to D^2/d (where D is the beam diameter and d the propagation distance), we can neglect the effects of diffraction. In this limit, the gradient of the field is dominated by the $i\mathbf{k}\cdot\mathbf{x}$ term, and each segment of the wavefront propagates locally as though it were its own plane wave, $\psi_{\text{local}}(\mathbf{x}) \approx A\exp(i\mathbf{k}_{\text{local}}\cdot\mathbf{x})$.

Phase is invariant to everything, since it's based on counting; we can look on the phase as being a label attached to a given parcel of fields, so that the propagation of the field is given by the relation between $\mathbf{x}$ and t that keeps ϕ constant. That means that the direction of propagation is parallel to $\mathbf{k}_{\text{local}}$,

$$\mathbf{k}_{\text{local}} \approx \frac{-i\nabla\psi}{\psi}, \tag{9.4}$$

which gives us a natural connection between rays and wavefronts.

If we take a trial solution for the scalar Helmholtz equation,

$$\psi(\mathbf{x}) = A(\mathbf{x})e^{ik_0 S(\mathbf{x})}, \tag{9.5}$$

applying the scalar Helmholtz equation for a medium of index n and taking only leading order terms as $k_0 \to \infty$ suppresses all the differentials of A, which is assumed to vary much more slowly than $\exp(ik_0 S)$, leaving the *eikonal equation*

$$|\nabla S(\mathbf{x})|^2 = n^2(\mathbf{x}), \tag{9.6}$$

where the eikonal $S(\mathbf{x})$ is the optical path length (it has to have length units because $k_0 S$ must be dimensionless). Once we have S, we can get the propagation direction from (9.4). What's more, Born and Wolf show that in a vector version of this, the Poynting vector lies along ∇S, so in both pictures, (9.6) is the natural connection between rays and waves.

This connection is not a 1:1 mapping between rays and waves, however. The eikonal can be used to attach rays to a wavefront, then trace the resulting rays as usual, but that procedure doesn't lead to the same results as propagating the field and then computing the eikonal, because the whole eikonal idea breaks down near foci and caustics, as well as having serious problems near shadow boundaries. The approximation becomes worthless at foci because the phase gradient vectors can never really cross; that would require a field singularity, which is impossible for the wave equation in a source-free region. Another way to say this is that the optical phase can be made a single-valued function of position in any given neighborhood, and therefore its gradient is also a single-valued function of position. So identifying rays purely as gradients of the phase cannot lead to rays that cross. For example, the eikonal equation predicts that a Fresnel amplitude zone plate (see Section 4.13.2) has no effect on wave propagation other than blocking some of the light.

The eikonal approximation shares the wave optics difficulty that there's no simple way to include multiple source points in a single run; the total field has only one gradient at each point.

9.2.4 Geometrical Optics and Electromagnetism

Geometrical optics (GO) is an asymptotic electromagnetic theory, correct in the limit $k \to \infty$. Like most asymptotic theories (e.g., the method of steepest descents), it requires some funny shifts of view. We go back and forth between considering our ray to be infinitesimally narrow (so that all surfaces are planes and all gradients uniform) and infinitely broad (so that the beam steers like a plane wave). The way a GO calculation goes is as follows:

1. Pick some starting rays, for example, at the centers of cells forming a rectangular grid on the source plane. Assign each one an amplitude and phase. (You can do GO calculations assuming incoherent illumination, but you're better off computing the amplitudes and phases of each ray and applying random phasor sums to the results—it's computationally cheap at that stage and avoids blunders.)
2. Assuming that the true optical field behaves locally like a plane wave, apply the law of reflection, Snell's law, and so on, to follow the path of the ray to the observation plane.
3. Add up the field contributions at the observation plane by computing the optical phase (the integral of $\mathbf{k} \cdot d\mathbf{s}$ over the ray path) and amplitude from each ray and summing them up. Remember to apply the Jacobian—if you imagine each ray occupying some small patch of source region, the area of the patch will in general change by the time it gets to the observation plane. This isn't mysterious, it's just like shadows lengthening in the evening. Field amplitudes scale as the reciprocal square root of the area. If the ray directions at the observation plane are similar, you can add up the fields as scalars. Otherwise, you'll need to be more careful about the polarization. Nonplanar rotations of $\mathbf{k}$ will make your polarization go all over the place.

You can trace rays in either direction, so if only a limited observation region is of interest, you can start there and trace backwards to the source. Either way, you have to make sure you have enough rays to represent the field adequately. You have to think of a GO calculation as involving nondiffracting rectangular pencil beams, not just rays; in general, the patches will overlap in some places, and you have to add all the contributions in complex amplitude.

In an inhomogeneous but isotropic medium, the geometric optics laws need to be generalized slightly. The local direction of wave propagation is the gradient of the phase. This leads to the eikonal equation (9.6) or the curvature equation (9.12), which are differential equations giving the change of the ray direction as a function of distance along the ray. Note that the Jacobian has to be carried along as well if you want to get the correct answer for the field amplitudes. If the medium is anisotropic as well as inhomogeneous, life gets a good deal harder, as you have to carry along the polarization state and any beam walkoff as well as the $\mathbf{k}$ vector and Jacobian as you go. If you have sharp edges, caustics, or shadows, geometric optics will give you the wrong answers there—it ignores diffraction, will exhibit square-root divergences at caustics and edges, and will predict zero field in the shadow regions.

9.2.5 Variational Principles in Ray Optics

Some propagation problems yield to a less brute-force approach: calculus of variations. If the medium is nonuniform, so that n is a function of $\mathbf{x}$, we need to put an integral in the exponent instead,

$$\psi(\mathbf{x}) = A(\mathbf{x}) \exp\left[-i\omega t + ik_0 \int_{\text{path}} n(\mathbf{x}) ds\right], \tag{9.7}$$

as in the eikonal approximation (9.5). Since the ray path depends on n, we don't know exactly what path to do the integral over, so it doesn't look too useful. In fact, we're rescued by *Fermat's principle*, a variational principle that states that

$$\delta S = \delta \int_{\text{path}} n(\mathbf{x}) ds = 0; \tag{9.8}$$

that is, the integral has an extremum on the true ray path P. The path yielding the extremum is said to be an *extremal*. Fermat called it the *principle of least time*, which assumes that the extremal is a global minimum. There's obviously no global maximum—a given path can loop around as much as it wants—so this is a good bet.

We solve variational problems of this sort by imagining we already know the parametrized $\mathbf{P} = \mathbf{x}(u)$, where $\mathbf{x}(0)$ is the starting point $\mathbf{x}_0$, $\mathbf{x}(u_1)$ is the end point $\mathbf{x}_1$, and u is a dummy variable. We demand that a slight variation, $\epsilon Q(u)$ (with $\mathbf{Q} \equiv 0$ at the ends of the interval), shall make a change in S that goes to 0 faster than ϵ [usually $O(\epsilon^2)$] as $\epsilon \to 0$. Since the arc length is all we care about, we can parameterize the curve any way we like so we'll assume that $\mathbf{x}$ is a continuous function of u and that $d\mathbf{x}/du \neq 0$ in $[0, u_1]$. Thus

$$\int_0^{u_1} \left((n(\mathrm{x}) + \epsilon \mathbf{Q} \cdot \nabla n) \left| \dot{\mathbf{x}} + \epsilon \dot{\mathbf{Q}} \right| - n(\mathbf{x}) |\dot{\mathbf{x}}| \right) du = O(\epsilon^2), \tag{9.9}$$

where dotted quantities are derivatives with respect to u. Since it's only the ϵ term we're worried about, we series-expand the squared moduli, cancel the zero-order term, and keep terms of up to order ϵ, which yields

$$\int_0^{u_1} \left(\frac{n\dot{\mathbf{x}} \cdot \dot{\mathbf{Q}}}{|\dot{\mathbf{x}}|} + |\dot{\mathbf{x}}| \mathbf{Q} \cdot \nabla n \right) du = 0, \tag{9.10}$$

which isn't too enlightening until we notice that it's nearly a total derivative, with one term in $\mathbf{Q}$ and one in $\dot{\mathbf{Q}}$. Integrating by parts, and using the fact that $\mathbf{Q} = 0$ at the ends and is continuous but otherwise arbitrary, we get the result

$$\frac{\nabla n - \dot{\mathbf{x}}(\dot{\mathbf{x}} \cdot \nabla n)}{n} = \frac{(\ddot{\mathbf{x}} - \dot{\mathbf{x}}(\dot{\mathbf{x}} \cdot \ddot{\mathbf{x}}))}{|\dot{\mathbf{x}}|^2}, \tag{9.11}$$

which can be written more neatly by changing back to arc length and using the convention that $\nabla_\perp$ is the gradient perpendicular to $\dot{\mathbf{x}}$, yielding the *curvature equation*

$$\frac{\nabla_\perp n}{n} = \left. \frac{d^2\mathbf{x}}{ds^2} \right|_\perp . \tag{9.12}$$

This says that the curvature of the path is equal to the perpendicular gradient of $\log n$, which makes a lot of physical sense, since we don't expect the path to change when we (say) double the refractive index everywhere.[†]

9.2.6 Schlieren Effect

One interesting consequence of the curvature equation (9.12) is that light waves steer like tanks: they turn toward whichever side goes more slowly. Accordingly, a refractive index gradient causes the wave to bend, the *schlieren effect*. Since $dn/dT < 0$ for gases, a temperature gradient in air produces schlieren, which is why there are mirages. On a hot, sunny day, the ground is warmer than the air, so $dn/dz < 0$ and light bends upwards; an image of a patch of sky appears near the ground in the distance, looking like a pool of water. At sea, with the opposite sign of dT/dz, a ship becomes visible before it crosses the geometrical horizon. More complicated gradients can cause multiple images, as in the beautifully named *fata Morgana* (after Morgan Le Fay, King Arthur's nemesis); ghostly shorelines with fantastic mountains can appear in the middle of the ocean. (Who said there was no poetry in optics?)

There are a couple of examples that show up more often in instruments: *thermal lensing*, which is pretty much like a mirage, and gradient-index (GRIN) optics, as we saw in Sections 4.13.1 and 8.3.5. Thermal lensing in nonaqueous liquids can be a big effect—big enough to be a sensitive laser spectroscopy method. A visible probe laser traverses a long path in some solvent, coaxially with an infrared pump beam. An axially symmetric temperature gradient results in progressive defocusing of the probe beam, which can be detected very sensitively with a masked detector. Water is a disappointing solvent for thermal lensing, with a low dn/dT and a high thermal conductivity.

9.2.7 The Geometrical Theory of Diffraction

For objects whose typical dimension a is large compared to a wavelength, the ordinary ray optics laws (the law of reflection and Snell's law) apply with high absolute accuracy except near shadow boundaries and places where rays cross—foci and caustics. (The relative accuracy is of course also very bad inside shadows, where geometric optics predicts zero fields.) For such large objects, it is reasonable to apply a local correction in these situations, the *geometrical theory of diffraction* (GTD), formulated by Keller,[‡] Ufimtsev,[§] and others. Like ray optics, GTD is an asymptotic theory valid in the limit $ka \gg 1$, but it lets us include higher order terms to get better accuracy. The basic idea is that illuminated bodies follow geometrical optics (GTD) or physical optics (PTD) except within a wavelength or two of shadow boundaries and sharp edges. Large objects ($ka \gg 1$) can be looked on as arrangements of flats, curved regions, and edges,

[†]Extremals that minimize some smooth functional such as (9.9) are called *weak extremals*, because the true minimum may not be continuous. If discontinuous and unsmooth functions are considered, the result is called a *strong extremal*. The strong extremal is usually a global minimum, never a maximum, but sometimes just a local minimum. If you don't know any calculus of variations, consider learning it—it isn't difficult, and it's a great help in optical problems. The book by Gelfand and Fomin (see the Appendix) is a good readable introduction.

[‡]J. B. Keller, Geometric theory of diffraction. *J. Opt. Soc. Am*. **52**, 116–130 (1962).

[§]Pyotr Y. Ufimtsev, *Method of Edge Waves in the Physical Theory of Diffraction*. Available at http://handle.dtic.mil/100.2/AD733203.

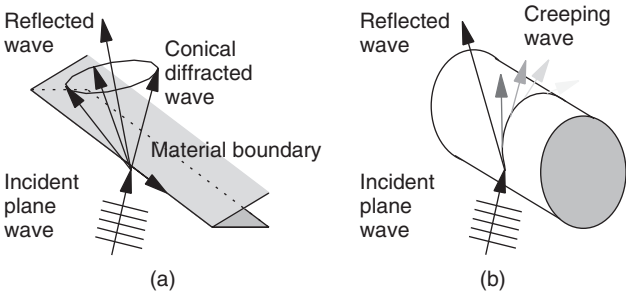

Figure 9.3. GTD and PTD calculations combine geometrical or physical optics calculation with a correction factor due to the vector diffraction from edges and shadow boundaries: (a) edge waves from discontinuities and (b) creeping waves from curves.

and the scattered fields can be decomposed into sums over those individual contributions. Locally these can be described as flat, ellipsoidal, cylindrical, wedge-shaped, or conical—all shapes for which rigorous analytic solutions exist, at least in the far field. The beautiful trick of PTD is to take each of these canonical cases, solve it twice—once rigorously and once by physical optics—*and then subtract the two solutions*, yielding the edge diffraction contribution alone. For sharp edges, this is expressed as a line integral around the edges, turning a 3D problem into a 1D problem. In most calculations the diffracted contributions are further approximated as *diffracted rays*. (Curved surfaces give rise to *creeping rays*, which are harder to deal with—the F-117A stealth fighter is all flats and angles because it was designed using 1970s computers.)

The two kinds of diffracted rays are shown in Figure 9.3: *edge rays*, which emanate from each point of discontinuity, such as a corner or an edge, and *creeping rays*, generally much weaker, which emanate from shadow edges on smooth portions of the surface.

So the way it works is that you do the calculation via geometrical or physical optics, which ignores the edge contributions, and then add in the vector diffraction correction in a comparatively simple and computationally cheap way. Complicated geometries will have important contributions from multiple scattering, leading to a perturbation-like series in which nth order terms correspond to n-times scattered light.

These approximations are usually very complicated, but on the other hand, they contain information about *all* incident and scattered angles, *all* positions, and *all* wavelengths, in one formula. The information density in that formula dwarfs that of any numerical solution, and frequently allows algebraic optimization of shapes and materials, which is very difficult with numerical solutions. This makes GTD and PTD well suited for design problems, especially with computer algebra systems available for checking.

GTD approximations tend to diverge as $x^{-1/2}$ at shadow boundaries, caustics, and foci, which of course are points of great interest. The same idea, local approximation by analytically known results, can be used to get a uniform asymptotic approximation, valid everywhere. For details, see the IEEE collected papers volume[†] and the excellent monographs of Ufimtsev and of Borovikov and Kinber referenced in the Appendix.

[†]Robert C. Hansen, ed., *Geometric Theory of Diffraction*, IEEE Press, New York, 1981.

9.2.8 Pupils

The entrance pupil is an image of the aperture stop, formed by all the elements ahead of the stop, and the exit pupil is the image of the same stop formed by all succeeding ones. Thus they are images of one another, and each point in the entrance pupil has a conjugate point in the exit pupil. Since nobody really knows what a lens does, we rely on this property heavily in wave optics calculations of imaging behavior. The most consistent high-NA approach to the problem is to use the ray optics of thin pencil beams to construct the fields on the pupil plane, and then propagate from there to the image using the Rayleigh–Sommerfeld integral. Remember the ray/wave disconnect: in the wave picture, *pupil* refers to the Fourier transform plane, not to the image of the aperture stop. (Like most ray/wave terms, the two are related but generally not identical.)

Not all imaging optical systems possess proper pupils; for example, a scanning system with the x and y deflections performed by separate mirrors lacks one unless intervening optics are included to image one mirror onto the other. An optical system without a pupil is not a shift-invariant system, so that Fourier imaging theory must be applied with caution.

9.2.9 Invariants

There are a number of parameters of an optical beam which are invariant under magnification. One is the state of focus: if an object point is 1 Rayleigh range from the beam waist, its image will be at 1 Rayleigh range from the waist of the transformed beam (neglecting diffraction). This is because the longitudinal magnification of an image is not M but M^2.

The best known is the *Lagrange invariant*, which we've encountered already as the conservation of étendue. You can get this by putting two rays as the columns of a 2×2 matrix R. No matter what $ABCD$ matrix you hit R with, the determinant of the result is equal to $\mathrm{Det}(R)$: $x_1\theta_2 - x_2\theta_1$ is invariant in the air spaces in any paraxial optical system. If we generalize to the case $n \neq 1$, the $ABCD$ matrix that goes from n_1 into n_2 is

$$\begin{bmatrix} 1 & 0 \\ 0 & n_1/n_2 \end{bmatrix}, \tag{9.13}$$

whose determinant is n_1/n_2, so the generalized Lagrange invariant L is

$$L = n(x_1\theta_2 - x_2\theta_1). \tag{9.14}$$

The more usual form of this is the theorem of Lagrange, where for a single surface between media of indices n_1 and n_2,

$$n_1 x_1 \theta_1 = n_2 x_2 \theta_2. \tag{9.15}$$

Another invariant is the number of resolvable spots, which is the field of view diameter or scan distance measured in spot diameters; if we take the two ends of the scan to be the two rays in the Lagrange invariant, the range goes up as the cone angle goes down, and hence the spot size and scan angle grow together.

9.2.10 The Abbe Sine Condition

The Lagrange invariant holds for paraxial systems, but not for finite apertures. Its most natural generalization is the *Abbe sine condition*,

$$n_1 x_1 \sin\theta_1 = n_2 x_2 \sin\theta_2, \tag{9.16}$$

which we don't get for free. (Optical design programs include *offense against the sine condition* (OSC) in their lists of aberrations.) A system obeying the sine condition is said to be *isoplanatic*, and has little or no coma.[†] Like other aberration nomenclature, this term has a related but not identical meaning in the wave picture: an optical system is said to be isoplanatic if its transfer function does not vary with position in the image. You can see that this is a different usage by considering vignetting; a few missing rays won't make the sine condition false, but they will certainly change the transfer function.

Aside: NA and f-Number. It's possible to get a bit confused on the whole subject of numerical aperture and f-number, because there are two competing definitions of $f\#$ in common use. One, historically coming from photography, is

$$f\# = D/\mathrm{EFL} = 1/(2\tan\theta),$$

where D is the pupil diameter, θ is the half-angle of the illuminated cone, and EFL is the effective focal length (just focal length f to us mortals). There's no clear upper limit to this number—light coming in from a hemisphere effectively has an infinite radius at any nonzero focal length, so $f\# = \infty$.

The other definition, coming from microscopy, is

$$f\# = 1/(2\sin\theta) = 0.5/\mathrm{NA},$$

assuming $n = 1$. Since $\mathrm{NA} \leq 1$ in air, in this definition a hemispherical wave would be coming in at $f/0.5$. The two are equivalent for small NA and distant conjugates, so they're often confused. Photographers care most about image brightness, since that determines exposure, so the quoted $f\#$ on the lens barrel actually applies on the *image* side of the lens, and is nearly constant as long as the object distance $d_o \gg f$. Microscopists care most about resolution, so microscope NA is quoted on the object side, where it's also nearly constant because of the small depth of focus. The two definitions express the same information, but confusion is common when we don't keep them straight. (The author recommends the 0.5/NA definition as being closer to the imaging physics as well as giving a simpler exact formula for image brightness, since $n^2\Omega' = \pi(\mathrm{NA})^2$.)

9.3 DIFFRACTION

There is not enough space in this book to treat diffraction in complete detail. For purposes of measurement systems, diffraction is important in four ways: in imaging; in gratings

[†]Optical system design is full of forbidding terms like that, but don't worry—half an hour's work and you'll be obfuscating with the best of them.

and holograms; in spatial filtering; and in vignetting, the incidental cutting off of parts of the beam by the edges of optical elements, apertures, and baffles. We've already covered the ordinary Huyghens–Fresnel theory in Section 1.3, so this section concentrates on the finite-aperture case.

9.3.1 Plane Wave Representation

Monochromatic solutions to the scalar wave equation in free space can be expressed exactly as sums of plane waves of different **k**. The **k**-space solution is exactly equivalent to the real-space solution; no approximation is involved. Thus if we have a focused beam, and we know its plane wave spectrum exactly, we can calculate its amplitude and phase at any point ($\mathbf{x}$, t) we like. It's important to hold on to this fact in discussing diffraction; once we have specialized to the scalar case, there are no further approximations in the actual wave propagation calculation. The additional approximations of diffraction theory involve how spatial Fourier coefficients on surfaces couple into plane waves, and how an obstruction in a free-space beam modifies its plane wave spectrum.

The easiest case is a plane boundary, because different plane waves are orthogonal on that boundary; thus a Fourier transform of the fields on the surface, appropriately weighted, gives the plane wave spectrum directly. Life gets significantly harder when the boundary is nonplanar. There are a handful of other coordinate systems in which the Laplacian separates, but the only three useful ones are Cartesian, cylindrical, and spherical. (Ellipsoidal coordinates are a special case of spherical for electrostatics, but not for electrodynamics.) Generally, though, unless you're a glutton for punishment, you have to choose among plane interfaces, asymptotically large spheres, and numerical solution.

9.3.2 Green's Functions and Diffraction

The study of diffraction is based on the idea of the *Green's function*, which is the response of a system consisting of a linear partial differential equation plus boundary conditions to a source term of $\delta(\mathbf{x} - \mathbf{x}')$. There is so much confusion around as to what the origins and limitations of diffraction theory are that it seems worth going through the math here. The following discussion follows Jackson fairly closely, so look there for more detail if this is unfamiliar. We solve the equation for the Green's function, and then we can solve the equation by a superposition integral of the source term $f(\mathbf{x}')$ (neglecting boundary terms),

$$\psi(\mathbf{x}) = \iiint_{\text{all space}} f(\mathbf{x}')G(\mathbf{x}, \mathbf{x}')d^3x'. \tag{9.17}$$

The usual case in diffraction theory is a bit more complicated, in that we actually have the field (rather than a source) specified on some surface, which may or may not be one of the boundaries of the space. Boundary conditions couple to the normal derivative of the Green's function, $\hat{\mathbf{n}} \cdot \nabla G$. (Don't confuse n the refractive index with $\mathbf{n}$ the unit vector normal to the surface.)

We'll specialize to the Helmholtz wave equation, so the defining equation for G is

$$(\nabla^2 + k^2)G(\mathbf{x}, \mathbf{x}') = -\delta^3(\mathbf{x} - \mathbf{x}'). \tag{9.18}$$

The two Green's functions of interest are G_0, the one for free space,

$$G_0(\mathbf{x}, \mathbf{x}') = \frac{\exp\left(ik|\mathbf{x} - \mathbf{x}'|\right)}{4\pi|\mathbf{x} - \mathbf{x}'|} \tag{9.19}$$

and G_+, the one for Dirichlet boundary conditions on the plane $z = 0$,

$$G_+(\mathbf{x}, \mathbf{x}') = \frac{\exp\left(ik|\mathbf{x} - \mathbf{x}'|\right)}{4\pi|\mathbf{x} - \mathbf{x}'|} - \frac{\exp\left(ik|\mathbf{x} - \mathbf{x}''|\right)}{4\pi|\mathbf{x} - \mathbf{x}''|}, \tag{9.20}$$

where $\mathbf{x}''$ is the mirror image of $\mathbf{x}'$.

Green's theorem is a straightforward corollary of the divergence theorem,

$$\iiint_V (\phi\,\nabla^2\psi - \psi\,\nabla^2\phi)d^3x = \oiint_S \left(\phi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\phi}{\partial n}\right)dA, \tag{9.21}$$

where surface S encloses volume V. If we choose $\phi = G_+$, and make S the plane $z = 0$ plus a hemisphere off at infinity, then by applying the wave equation and the definition of G, we get the *Rayleigh–Sommerfeld integral*,

$$\psi(x) = \frac{1}{i\lambda}\iint_{z=0} \frac{\exp\left(ik|\mathbf{x} - \mathbf{x}'|\right)}{|\mathbf{x} - \mathbf{x}'|}\left(1 + \frac{i}{k|\mathbf{x} - \mathbf{x}'|}\right)\frac{\mathbf{n}\cdot(\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}\psi(\mathbf{x}')d^2x'. \tag{9.22}$$

A limiting argument shows that the contribution from the hemisphere goes to 0.

If we choose G_0 instead, we get the *Kirchhoff integral* ,

$$\psi(\mathbf{x}) = -\frac{1}{4\pi}\int_{z=0} \frac{\exp(ikR)}{R}\left[\nabla'\phi + ik\left(1 + \frac{i}{kR}\right)\frac{\mathbf{R}}{R}\psi\right]\cdot\mathbf{n}'\,dA'. \tag{9.23}$$

These scary-looking things actually turn out to be useful—we'll revisit them in Section 9.3.6.

Ideally what we want is to find the exact plane wave spectrum of the light leaving S for a given plane wave coming in, because that makes it easy to do the propagation calculation. Getting the correct plane wave spectrum is easy for a planar screen, since different plane waves are orthogonal on a plane, and because we can use the correct Green's function in the planar case (the Rayleigh–Sommerfeld theory). For more complicated boundaries, life gets very much harder since analytically known Green's functions are rare, and plane waves are not then orthogonal on S, so we can't just Fourier transform our way out of trouble. The surfaces of interest are usually spheres centered on some image point, so we'd need to expand in partial waves, and then find the plane wave spectrum from that. Fortunately, there's an easier way.

Aside: Theory That's Weak in the Knees. One problem for the outsider coming to learn optical systems design is that it's a pretty closed world, and the connections between the scalar optics of lens design and the rest of optics are not clearly brought out in books on the subject, or at least those with which the present author is familiar—it isn't at all obvious how a given ray intercept error influences the signal-to-noise ratio, for example. This is not helped by the uniformly inadequate presentation of the theoretical

underpinnings, which almost always base Fourier optics on the Fresnel approximation and aberration theory on a sloppy use of the Huyghens propagator.

A charitable interpretation of this is that it is an attempt to make the subject accessible to undergraduates who don't know anything about Green's functions. Yet it is unclear how they are aided by such sloppiness as defining the wavefront on a spherical reference surface near the exit pupil, then doing the integrals as though it were a plane.

Some claim that this is the Kirchhoff approximation (it isn't), and others unapologetically toss around the (paraxial) Huyghens integral on the spherical surface, even for large-aperture lenses. The funny thing about this is that, apart from neglect of obliquity, they get the right result, but for the wrong reasons. It matters, too, because the confusion at the root of the way the subject is taught damages our confidence in our results, which makes it harder to calculate system performance with assurance. If you're an optics student, ask lots of rude questions.

9.3.3 The Kirchhoff Approximation

Usually we have no independent way of measuring the actual fields on the boundary and are reduced to making a guess, based on the characteristics of the incoming wave. The *Kirchhoff approximation* says that, on surface S, the fields and their derivatives are the same as the incoming field in the unobstructed areas and 0 in the obstructed ones. This turns out to work reasonably well, except for underestimating the edge diffraction contribution (see Section 9.2.7). The degree of underestimate depends on the choice of propagator (see below); empirically the Kirchhoff propagator does a bit better on the edge diffraction contribution than the Rayleigh–Sommerfeld propagator.

You can find lots more on this in Stamnes, but the net is that these physical optics approximations work pretty well for imaging and for calculating diffraction patterns, but it won't get fine details right, for example, the exact ripple amplitude, and will underestimate the field in the geometric shadow regions. You need GTD or PTD to do that properly.

9.3.4 Plane Wave Spectrum of Diffracted Light

In Section 1.3, we used the Huyghens propagator, which in real space is

$$\Theta(x, y, z) = \frac{-i}{\lambda} \iint_P \Theta(x', y', z') \frac{\exp\left(ik \dfrac{(x - x')^2 + (y - y')^2}{2(z - z')}\right)}{(z - z')} dx' dy', \tag{9.24}$$

where P is the xy plane, and in $\mathbf{k}$-space is

$$\Theta(x, y, z) = \iint_{P'} \Theta(u, v)|_{z=0} e^{i(2\pi/\lambda)(ux+vy)} e^{-i(2\pi z/\lambda)(u^2+v^2)/2} du\, dv, \tag{9.25}$$

where P' is the uv plane.

If a beam gets cut off sharply, it scatters strong fringes out to high angles. Being a paraxial approximation, the Huyghens integral requires very large values of $z - z'$ to be used in that situation. The Rayleigh–Sommerfeld result (9.22) is the rigorously correct scalar solution for a correctly given $\psi(\mathbf{x})$ on the plane $z = 0$, because it is based on the correct Green's function for a half-space above that plane. To get the $\mathbf{k}$-space

representation (angular spectrum), we choose $\mathbf{x}$ to be on the surface of a very large sphere of radius R, and neglect the constant term $-ie^{ikr}/R$, which yields

$$\psi(u, v) = \frac{w \ \mathrm{circ}(1-w)}{\lambda} \iint_P \exp\left(\frac{i2\pi}{\lambda}(ux' + vy')\right) \psi(x', y')dx'dy', \tag{9.26}$$

where u and v are the direction cosines in the x and y directions as before, $w = (1 - u^2 - v^2)^{1/2} = k_z/k$, and $\mathrm{circ}(x)$ is 1 for $0 \leq x < 1$ and 0 otherwise. It is clear from this equation that the $\mathbf{k}$-space solution is the Fourier transform of the fields on the boundary, multiplied by a factor of $-ik_z = 2\pi i w/\lambda = ik \cos\theta$, where θ is the angle of incidence of the outgoing light. A heuristic way of looking at this uses a pencil beam rather than a plane wave. A circular beam coming from a surface at an incidence angle of θ occupies an elliptical patch on the surface, whose area is $\pi a^2 \sec\theta$. On this patch, the field strength is not diminished by spreading out (different places on the long axis of the patch are seeing the same beam at different times), so the *obliquity factor* $w = \cos\theta$ is required to counteract the tendency of the integral to become large as the angle approaches grazing. (We saw this as the Jacobian in Section 9.2.1 and in the drinking-straw test of Section 1.3.12.)

The $\mathbf{k}$-space Kirchhoff integral is similar,

$$\psi(u, v) = \frac{(w_{\mathrm{inc}} + w) \ \mathrm{circ}(1-w)}{2\lambda} \iint_{P'} \exp\left(\frac{i2\pi}{\lambda}(ux' + vy')\right) \psi(x', y')dx'dy', \tag{9.27}$$

which is just the same as the far-field Rayleigh–Sommerfeld integral except for the obliquity factor. The Neumann boundary condition case, where $\hat{\mathbf{n}} \cdot \nabla\psi$ is specified on the boundary, yields the same Fourier transform expression with an obliquity factor of w_{inc}. The three propagators are all exact since they predict the same fields if the source distribution is correct—they differ only when we make an inaccurate guess at $\phi(x, y)$.

9.3.5 Diffraction at High NA

Diffraction from apertures in plane screens can be calculated for all z by assuming that the field is the same as the incident field in the aperture, and zero elsewhere. In imaging problems, the screen has reflection or transmission amplitude and phase that depend on position. If we just take the incident field as our guess, we wind up suppressing the high-angle components by the obliquity factor (see Section 9.2.1), so in fact we have to put the reciprocal of the obliquity factor into the illumination beam in order for energy to be conserved (i.e., multiply by the Jacobian of the inverse transformation). This is physically very reasonable, since the screen could have a transmission coefficient of 1 (i.e., not be there at all), in which case the plane wave components had better propagate unaltered.

If the illumination beam has high NA, then the obliquity factors of the plane wave components of the illumination beam will be different, and that has to be taken into account. If the object has only low spatial frequencies, and doesn't have large phase variations due to topography, then each plane wave will be scattered through only a small angle, so that $\cos\theta$ doesn't change much, and the obliquity factors cancel out. This effect is partly responsible for the seemingly unaccountable success of Fourier optics at high NA.

As we discussed in Section 1.3.9, the simple thin object model used in diffraction imaging theory is a complex reflection coefficient, which depends on $\mathbf{x}$ and not on $\mathbf{k}$. Height differences merely change the phase uniformly across the pupil. This works fine as long as the maximum phase difference across the pupil is smaller than a wave, i.e., we're within the depth of focus, and providing we take a weighted average of the phase shift over the entire pupil, i.e., the phase shift with defocus isn't $k_z z$ anymore (see Example 9.4).

9.3.6 Propagating from a Pupil to an Image

We're now in a position to say what the exact scalar field propagator is between a pupil and an image. Consider the exit pupil plane of an optical system, with a mildly wrinkled wavefront that is basically a spherical wave centered on the nominal image point $\mathbf{x}_0$,

$$\psi(\mathbf{x}') = \tilde{A}(\mathbf{x}')\frac{e^{-ik|\mathbf{x}'-\mathbf{x}_0|}}{|\mathbf{x}'-\mathbf{x}_0|}, \tag{9.28}$$

where the *pupil function* $\tilde{A}$ is a complex envelope that carries the amplitude and phase information we care about. (In a little while it will be apparent that the natural variables for expressing $\tilde{A}$ are the direction cosines u and v, just as in the paraxial theory.) We're interested in the structure of the image, so we use the Rayleigh–Sommerfeld integral to propagate to $\mathbf{x}_1 = \mathbf{x}_0 + \boldsymbol{\zeta}$, where $|\boldsymbol{\zeta}|$ is assumed to be small compared to $|\mathbf{x}' - \mathbf{x}_0|$. We further assume that $1/(k|\mathbf{x}' - \mathbf{x}|) \ll 1$, that is, we're in the limit of large Fresnel number, which allows us to discard that term (which turns out to represent the evanescent fields and pupil edge diffraction), so we write (where P' is the uv plane as before)

$$\psi(\mathrm{x}) = \frac{1}{i\lambda}\iint_{P'} \frac{\exp(ik|\mathbf{x} - \mathbf{x}'|)}{|\mathbf{x} - \mathbf{x}'|}\frac{\exp(-ik|\mathbf{x_0} - \mathbf{x}'|)}{|\mathbf{x_0} - \mathbf{x}'|}\tilde{A}(\mathbf{x}')\frac{\mathbf{n}\cdot(\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}d^2x'. \tag{9.29}$$

Note that we haven't made any assumptions about small angles or slowly varying envelopes—apart from the scalar field and discarding the evanescent contributions, this is an exact result. Providing that ζ is small compared to $|\mathbf{x} - \mathbf{x}'|$, we can ignore it in the denominator, but since it isn't necessarily small compared to $1/k$, we have to keep it in the exponent. Doing a third-order binomial expansion of the exponent, we get

$$|\mathbf{x}_0 + \boldsymbol{\zeta} - \mathbf{x}'| - |\mathbf{x}_0 - \mathbf{x}'| = \boldsymbol{\zeta}\cdot\frac{\mathbf{x}_0 - \mathbf{x}'}{|\mathbf{x}_0 - \mathbf{x}'|} + \boldsymbol{\zeta}\cdot\frac{\left(\boldsymbol{\zeta} - \dfrac{(\mathbf{x}_0 - \mathbf{x}')(\boldsymbol{\zeta}\cdot(\mathbf{x}_0 - \mathbf{x}'))}{|\mathbf{x}_0 - \mathbf{x}'|^2}\right)}{2|\mathbf{x}_0 - \mathbf{x}'|} + O(\zeta^3). \tag{9.30}$$

The first term is the phase along the radial vector, which as usual is going to turn into the kernel of a Fourier transform; the second is the phase due to the length of the vector changing. (Note that there is no radial component of the phase in order ζ^2.) If we were in the paraxial case, we'd just forget about terms like that, or at most say that the focal surface was a sphere centered on $\mathbf{x}'$, but the whole point of this discussion is that $\mathbf{x}' - \mathbf{x}_0$ be allowed fractional variations of order 1, so we can't do that.

What we do need to do is restrict $\boldsymbol{\zeta}$. In order for the ζ^2 term to be small compared to $1/k$, it is sufficient that

$$|\boldsymbol{\zeta}| \ll \sqrt{\frac{\lambda|\mathbf{x}' - \mathbf{x}_0|}{\pi}}. \tag{9.31}$$

Since we imagine that the pupil function has been constructed by some imaging system, the rays have been bent so as to construct the spherical wavefront. For consistency, we must thus put in the inverse of the obliquity factor, and the $\hat{\mathbf{n}} \cdot (\mathbf{x}_0 - \mathbf{x}')$ term then goes away to the same order of approximation as neglecting $\boldsymbol{\zeta}$ in the denominator.[†] We also transform into direction cosines, so that $(dx, dy) = |\mathbf{x}_0 - \mathbf{x}'|(du, dv)$, which leaves a pure inverse Fourier transform,

$$\psi(\mathbf{x}) = \frac{1}{\lambda} \iint_P e^{ik\mathbf{u}\cdot\boldsymbol{\zeta}} \tilde{A}(\mathrm{u}) du dv. \tag{9.32}$$

For a pupil-image distance of 20 mm and a wavelength of 0.5 μm, this Fraunhofer-type approximation is valid in a sphere of at least 100 μm in diameter, *even at NA* $= 1$. In order to cause the image to deviate seriously from the Fourier transform of the pupil function, there would have to be hundreds of waves of aberration across the pupil, so that for all interesting cases in imaging, where the scalar approximation applies, Fourier optics remains valid. This applies locally, in what is called the *isoplanatic patch*, and does not imply that the whole focal plane is the Fourier transform of the whole pupil, as it is in the paraxial case, because that depends on things like the field curvature and distortion of the lens, and the different obliquities at different field positions.

This analysis applies backwards as well, in going from a small-diameter object to the entrance pupil of the lens, although if the object is a material surface and not itself an aerial image, the usual thin-object cautions apply. The combination of the two shows that an object point is imaged to an image point, and that the point spread function of the system is the Fourier transform of the pupil function $\tilde{A}$, at least in the large Fresnel number limit.

This is really the main point: the imaging of an object point into an image point via a pupil is controlled by Fourier optics in all cases, and for an imaging system faithful enough to deserve the name, the amplitude PSF of the imaging operation really is the Fourier transform of the pupil function $\tilde{A}$, regardless of NA.

***Example 9.1: High-NA Fourier Optics—Metal Lines on Silicon at* NA $= 0.95$.** Figures 9.4 and 9.5 show the Fourier optics result versus experiment for a 90 nm tall line of gold on silicon. Even though the scalar Fourier optics approximation to high-NA imaging is a fairly sleazy one, it nevertheless works extremely well in practice.

Example 9.2: When Is the Paraxial Approximation Valid? The special case of a perturbed spherical wave is very important in applications, but the usual Fourier optics result is more general; the far-field pattern is the Fourier transform of the pupil function. What is the range of validity of that approximation?

[†]For a system of unit magnification, this cancellation is exact when both the object-to-pupil and pupil-to-image transforms are computed; when the magnification is not 1, the pupil function $\tilde{A}$ will need some patching up, but that's not a fundamental objection at this point.

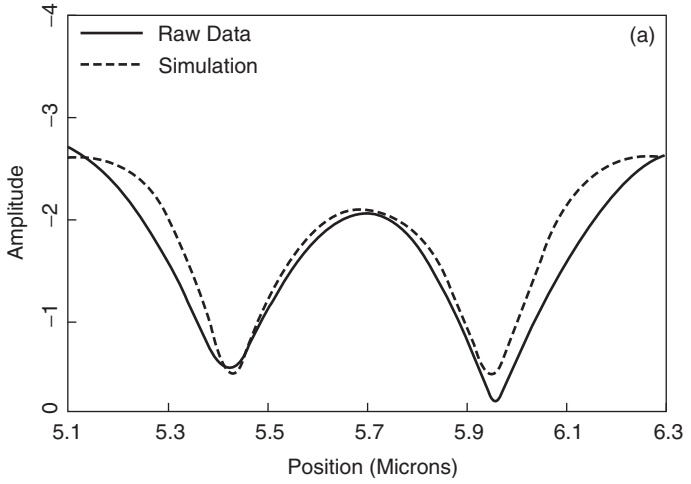

Figure 9.4. Heterodyne microscope image of Au lines on Si, 515 nm, 0.90 NA: amplitude.

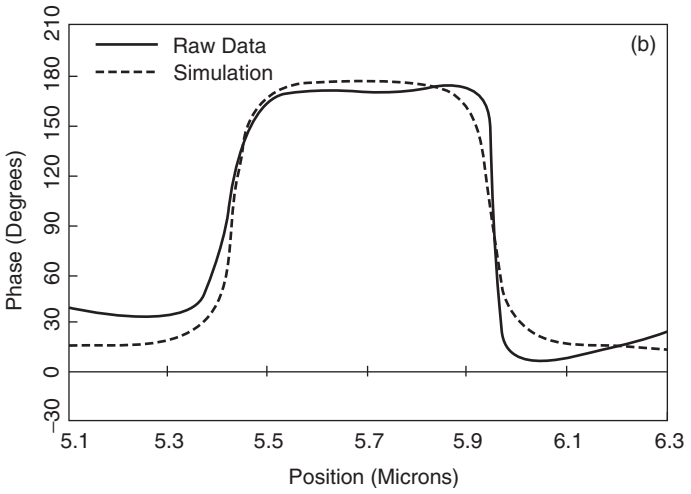

Figure 9.5. Au lines on Si: phase.

Comparison of the Huyghens integral with the Kirchhoff and Rayleigh–Sommerfeld ones shows two differences: the Huyghens integral omits the obliquity factor, and for a plane wave component $\exp(i2\pi(ux + vy)/\lambda)$, the Huyghens integral replaces the true phase shift $kz(w - 1)$ by the first term in its binomial expansion, apparently limiting its use to applications where this causes a phase error of much less than 1 (it is not enough that it be much less than kz because it appears in an exponent). If we require that the next term in the binomial expansion be much less than 1, we find that this requires that

$$|z| \gg \sqrt[3]{\frac{x^4}{\lambda}}. \tag{9.33}$$

This restriction is necessary for general fields, but *not* for paraxial ones. The slowly varying envelope equation is

$$\frac{d^2\Theta}{dx^2} + \frac{d^2\Theta}{dy^2} + 2ik\frac{d\Theta}{dz} = 0. \tag{9.34}$$

Its validity depends solely on the initial conditions; a sufficiently slowly varying envelope will be accurately described by this equation for all z. For slowly varying Θ and small $z - z'$, the error in the phase term does indeed become large, but a stationary phase analysis shows that the large-x contribution to the integral goes strongly to zero as $z \to z'$, due to the rapidly varying phase factor, so that the integral remains valid for all z, and the Huyghens integral is not limited to far-field applications. This is perhaps easier to see in the spatial frequency representation.

If we take $\Theta_\alpha(\mathbf{x}) = e^{i\alpha x}e^{i\gamma z}$ in (9.34), requiring the phase error to be small compared to 1 leads to (9.33) for fixed α of order k, and an absolute phase error that grows secularly with z, as one would expect. This is not a deadly error, as it amounts only to a deviation of the field curvature from spherical to parabolic; if we take as our reference surface a parabola instead of a sphere, it goes away; it may make the calculated optical path incorrect, and in applications where that matters, it should be checked by comparison with the Rayleigh–Sommerfeld result.

For fixed z, the restriction can be applied to α instead:

$$|\alpha| \ll \sqrt[4]{\frac{k^3}{z}}. \tag{9.35}$$

This is easily satisfied for small z as well as large.

9.3.7 Telecentricity

As Figure 9.6 illustrates, a telecentric optical system is one in which the principal ray is parallel to the optical axis. This means that, roughly speaking, the axis of the cone of light arriving at the image or leaving the object is not tilted, and is equivalent to saying that the pupil is at infinity. An optical system can be telecentric in the object space, the image space, or both.

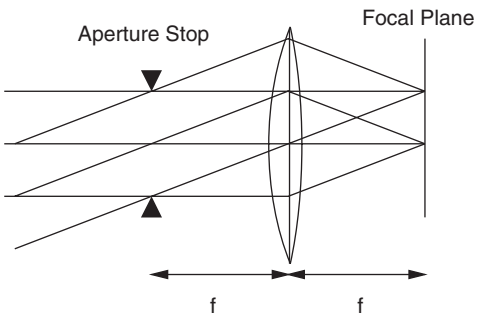

Figure 9.6. A telecentric optical system.

This property is of more practical interest than it may sound. In a telecentric system, tilting the sample or moving it in and out for focusing does not change the magnification—the image is an orthographic projection, like an engineering drawing. Light reflected from a plane sample, such as a microscope slide or a flat mirror, retraces its path. Both of these properties are very useful for scanning or imaging interferometers such as shearing interference microscopes.

In a telecentric imaging system with telecentric illumination, the illumination diagram is independent of position; all points in the field of view are imaged with light covering the same range of angles. Because both the illumination and collection NAs are constant, the range of spatial frequencies received is the same everywhere too. These two properties together give telecentric systems nearly space-invariant point spread functions. This is of great benefit when interpreting or postprocessing images, for example, in automatic inspection systems. Obviously a telecentric system can have a field of view no larger than its objective (the last optical element on the outside), and usually it's significantly smaller.

9.3.8 Stereoscopy

Stereoscopic vision requires the ability to look at a scene from two different directions and synthesize the resulting images. This is different from merely binocular vision. A binocular microscope presents the same image to each eye, whereas a properly stereoscopic microscope splits the pupil into two halves, presenting one half to each eye. Since pupil position corresponds to viewing angle, this reproduces the stereo effect. Splitting the pupil reduces the resolution, but the gain in intuitive understanding is well worth it.

9.3.9 The Importance of the Pupil Function

Pupil functions don't get the respect they deserve. The point spread function $h(\mathbf{x})$ of an optical system is the Fourier transform of the pupil function $\tilde{A}(\mathbf{u})$, and the point spread function ranks with the étendue as one of the two most important attributes of an imaging system. The pupil function is the filter that is applied to the spatial frequency spectrum of the sample to generate the image.

In signal processing, we choose our filter functions very carefully, so as to get the best measurement, but this is less often done in optics, which is odd since optics are much more expensive. One reason for it is confusion of two quite different objects, both called *transfer functions*, and both giving rise to *point spread functions* (PSFs). The confusion has arisen because, for historical reasons, the one less clearly connected to the electromagnetic field quantities $\mathbf{E}$ and $\mathbf{B}$ has staked out the high ground.

9.3.10 Coherent Transfer Functions

When we describe the actions of an optical system in terms of the plane wave decomposition of the scalar optical field E, and apply Fourier transform theory to describe how a sinusoidal component of the field distribution at a sample plane propagates to the image plane, we are using the *coherent transfer function* (CTF) of the system. The CTF is the convolution of the illumination and detection pupil functions, because the amplitude PSF of the measurement is the product of the illumination and detection PSFs. Most of the

time, one of the two is far more selective than the other, so the CTF and the broader of the two pupil functions are often interchangeable.

The CTF is the right description of a translationally invariant phase-sensitive optical system; this class includes holography setups, scanning heterodyne microscopes, and phase shifting imaging interferometers, as well as any system producing an aerial image, such as binoculars. To determine the output of such a system, multiply the Fourier transform of the sample's complex reflection coefficient, as a function of position, by the instrument's CTF, and take the inverse transform. This is of course mathematically equivalent to convolving the sample function with the instrument's 2D amplitude point spread function. Since the optical phase information is preserved, digital postprocessing can be used to transform the complex fields in a great variety of ways.

The net effect is that provided you measure both phase and amplitude, on a sufficiently fine grid and at sufficiently high SNR, you can do with postprocessing anything you could do on an aerial image with optical elements; this is a remarkable and powerful result.

Example 9.3: Heterodyne Microscope. A heterodyne microscope is basically a heterodyne Michelson interferometer, using an AO deflector as its beamsplitter, and with a microscope objective in one or both arms. Some versions use separate lenses, and some send both beams down to the sample through the same lens, as shown in Figure 9.7. A uniform pupil has an illumination pupil function $L = \mathrm{circ}((u^2 + v^2)/\mathrm{NA}^2)$, which transforms to an illumination PSF of

$$l(\chi) = \frac{J_1(\pi\chi)}{\chi}, \tag{9.36}$$

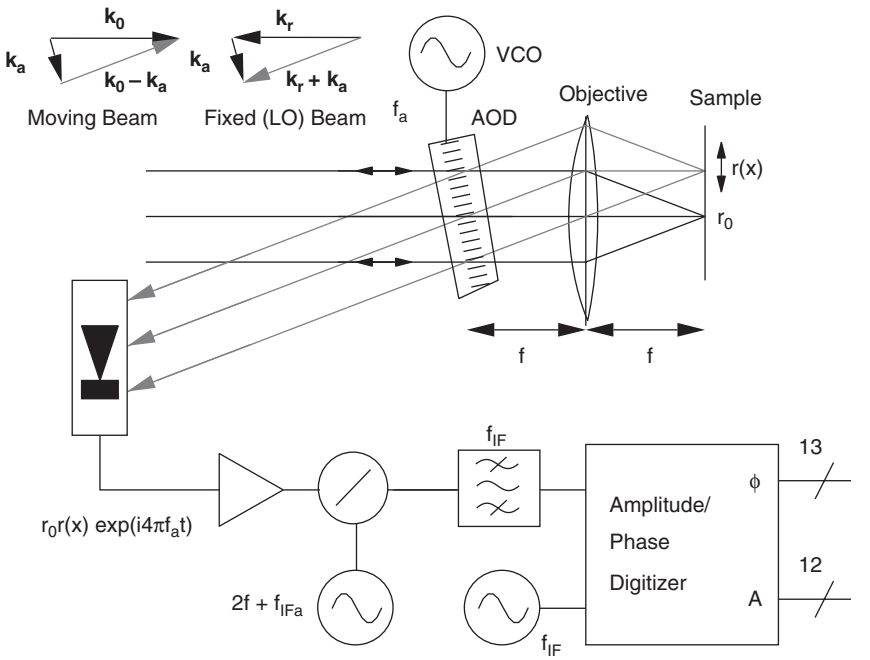

Figure 9.7. Heterodyne confocal microscope.

where $\chi = r\mathrm{NA}/\lambda$. The coherent detector uses interference with a nominally identical beam $s(\chi)$ to produce the AC photocurrent

$$\begin{aligned} i_{\mathrm{AC}} &= 2\mathcal{R}\ \mathrm{Re}\left\{\iint_{\mathrm{det}} d^2x\,\psi_{\mathrm{LO}}\psi_s^*\right\} \\ &= 2\mathcal{R}\ \mathrm{Re}\left\{\exp(-i\,\Delta\omega t)\iint_{\mathrm{det}} |\psi_{\mathrm{LO}}(\mathbf{x})||\psi_s(\mathbf{x})|\exp(i\,\Delta\phi(\mathbf{x},t))\,dA\right\}, \end{aligned} \tag{9.37}$$

as in Section 1.5. By the power theorem, this dot product can be computed in the pupil or the image, or anywhere in between. For our purposes, it is easiest to see what happens if we compute it at the sample surface. There, the two jinc functions are superimposed and multiplied by the local complex reflection coefficient $\tilde{r}$ of the sample S. Thus the total complex AC photocurrent is

$$\tilde{i}_{\mathrm{AC}} = \mathcal{R}\iint_{\mathrm{sample}} l(\mathbf{x})\tilde{r}(\mathbf{x})s^*(\mathbf{x})d^2x, \tag{9.38}$$

which if both beams are unaberrated and focused on $\mathbf{x}$ is

$$\tilde{i}_{\mathrm{AC}}(\mathbf{x}) = \mathcal{R}\lambda\iint_S \tilde{r}(\mathbf{x}')\left[\frac{J_1((\pi\mathrm{NA}/\lambda)|\mathbf{x}-\mathbf{x}'|)}{|\mathbf{x}-\mathbf{x}'|\mathrm{NA}}\right]^2 d^2x', \tag{9.39}$$

so by construction, the amplitude PSF of such a microscope is

$$g(\mathbf{x}) = \left[\frac{J_1(\pi\chi)}{\chi}\right]^2. \tag{9.40}$$

The CTF of this microscope is the *Chinese hat function*,

$$H(\omega) = \frac{2}{\pi}\left(\cos^{-1}(\omega) - \omega\sqrt{1-\omega^2}\right), \tag{9.41}$$

whose name makes perfect sense if you look at Figure 9.8 and remember that it's cylindrically symmetric. This function has a cusp at 0 and extends out to $\omega = 2\mathrm{NA}$. In talking about the CTF, we're subtly sliding into the thin-object Fourier optics approximation, where a spatial frequency component at $\nu = 2\mathrm{NA}/\lambda$ scatters light coming in at u all the way to $-u$, which can still just make it back through the pupil.

The line spread function is

$$\begin{aligned} l_2(x) &= \frac{8\pi\mathrm{NA}}{\xi^2}\mathbf{H}_1(\xi) \\ &= \frac{16\mathrm{NA}}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m\xi^{2m}}{(2m+1)!!(2m+3)!!} \end{aligned} \tag{9.42}$$

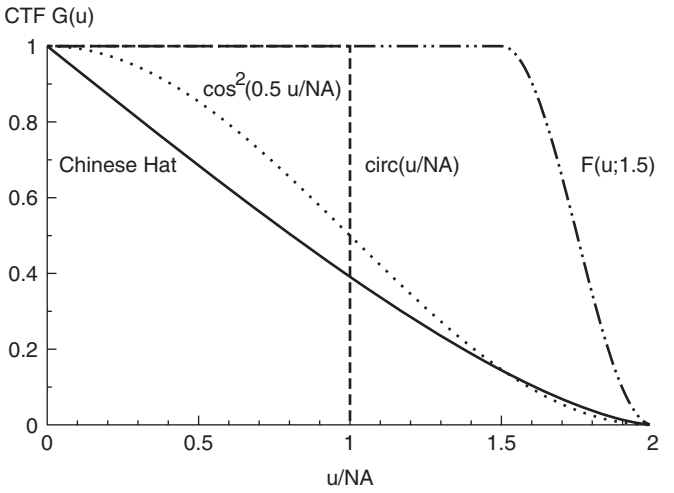

Figure 9.8. CTFs of a heterodyne interference microscope before and after Fourier postprocessing.

where $\xi = 2kx\text{NA}$, and $\mathbf{H}_1(x)$ is the Struve function of order 1 (see Abramowitz and Stegun). This function has an asymptotic series for large x,

$$l_2(x) \sim \frac{16\text{NA}}{\pi}\xi^{-2} - \frac{8\text{NA}(\cos\xi + \sin\xi)}{\sqrt{\pi\xi^5}} + O(\xi^{-4}) \text{ monotonic} + O(\xi^{-4.5}) \text{ oscillatory}, \tag{9.43}$$

which is a distressingly slow falloff. The slowness is due to the cusp at the origin and the higher order nondifferentiability at the outer edges. Because the heterodyne system preserves phase information, this cusp can be removed by digital filtering in a postprocessing step (see Section 17.7.1). Even a very gentle filter can make a big difference to the settling behavior; for example, $F(u) = \cos^2(u/\text{NA})/G(u)$, which turns the Chinese hat function into a von Hann raised cosine (see Section 17.4.9). This filter removes the cusp and makes the edges go to 0 quadratically, and as Figures 9.8–9.10 show, the step response settles at its final value when the uncorrected LSF is still 5% away. It does this with essentially 0 dB noise gain, so there's no penalty whatever.

A different filter, which boosts the wings of the transfer function further, can yield results that approach those expected from a microscope operating at half the wavelength, provided the noise is sufficiently low[†]; the 10–90% edge rise going over a $\lambda/6$ phase step can be reduced from 0.45λ to 0.19λ, that is, 90 nm for $\lambda = 514$ nm. (Phase edges are a bit sharper than pure amplitude ones, since the amplitude dip in the middle sharpens the edge; the pictures in Figures 9.4 and 9.5 were preternaturally sharp because the step happened to be close to $\lambda/4$ tall.) Since an optical image is one of the few real examples of a band-limited function (because waves with spatial frequency higher than $1/\lambda$ cannot propagate to the detector), this is as much as can be achieved in a model-independent fashion.

[†]P. C. D. Hobbs and G. S. Kino, Generalizing the confocal microscope via heterodyne interferometry and digital filtering. *J. Microsc*. **160**(3) 245–264 (December 1990).

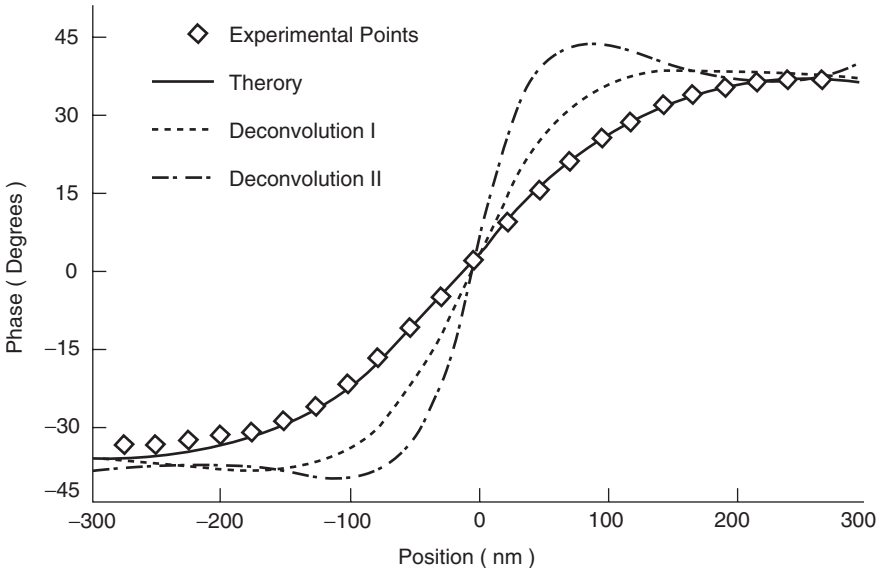

Figure 9.9. Experimental and theoretical phase plots for a heterodyne confocal microscope looking at an aluminum-on-aluminum step, 80 nm tall, before and after deconvolution.

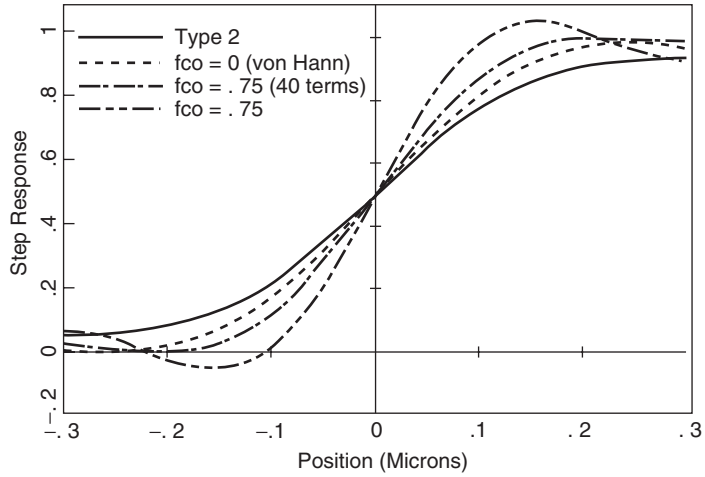

Figure 9.10. Theoretical step response of a heterodyne confocal microscope to an amplitude step, and several deconvolutions.

Example 9.4: Modeling Defocus. Another way to illustrate the difference between coherent and incoherent optical systems, and the value of coherent postprocessing, is the compensation of defocus. In an incoherent system, it is impossible to distinguish between positive and negative defocus, because (apart from aberrations) the difference is only the sign of the phase shift, which gives rise to no intensity change. Although there are minor differences in the behavior of lenses depending on the direction of the defocus, this does not change the basic point.

In a coherent system, on the other hand (provided that we measure the amplitude and phase independently, with adequate spatial resolution and signal-to-noise ratio), we can numerically refocus an image, or focus on any given depth. This can be done as follows: decompose the beam into plane waves, multiply by $\exp(-ikz\sqrt{1-u^2-v^2})$, where u and v are the x and y direction cosines as usual, then put it back together. This is a kind of convolution filter.

The author's former colleagues, Paul Reinholdtsen and Pierre Khuri-Yakub, used this idea with a confocal acoustic microscope to remove blurring caused by out-of-focus structures, by numerically defocusing an in-focus image of the interfering top surface. Looking at a quarter, they were able to read QUARTER DOLLAR on the back, right through the coin, by defocusing George Washington and subtracting him out.

Performing the convolution and taking the real part, we can get the (complex) vertical response of a confocal reflection microscope (where the phase shift is doubled):

$$\tilde{i}(z) = 2\pi \int_0^{\mathrm{NA}} d\omega \exp\left(-i2kz\sqrt{1-\omega^2}\right). \tag{9.44}$$

Here we've assumed that the pupil function is uniform, so that the obliquity factors in transmit and receive cancel out exactly, and that the medium is air. With a change of variable from $\omega = \sin\theta$ to $r = \cos\theta$, this becomes

$$\tilde{i}(z) = 2\pi \int_r^1 \exp(-i2kzr')r'dr'. \tag{9.45}$$

This is easily done by partial integration, but the result is a mess. We can get a good memorable result accurate to about 0.2% up to $\mathrm{NA} = 0.5$ by setting the factor of r' outside the exponent to 1 and computing the envelope and carrier:

$$\tilde{i}(z) = 2\pi(1-r)\mathrm{sinc}(rz/\lambda)\exp(-i2krz), \tag{9.46}$$

that is, the amplitude response is a sinc function and the phase shift is not $2kz$ but is reduced by a factor $[1-(\mathrm{NA})^2]^{1/2}$. The exact result shows that the phase slope reduction reaches a factor of 2 at $\mathrm{NA} = 1$.

9.3.11 Optical Transfer Functions

The CTF is not the most commonly encountered transfer function in the literature. The more usual *optical transfer function* (*OTF*) is another beast altogether, and it's important to keep the distinction crystal clear. The OTF predicts the *intensity* distribution of the image based on that of the sample, with certain assumptions about the spatial coherence of the illuminator, i.e., the statistical phase relationships between the various Fourier components. There is no 1:1 correspondence to the propagation of plane wave components through the optical system. As we'll see, the OTF isn't a proper transfer function.

The intensity[†] of the beam is $\psi\psi^* \cos\theta$. Since the propagation of $\tilde{A}$ to the image plane is governed by the CTF H, the autocorrelation theorem gives us the OTF O:

$$O(u, v) = H(u, v) \star H(u, v). \tag{9.47}$$

The OTF for an ideal system whose pupil function is $\text{circ}[(u^2 + v^2)/(\text{NA})^2]$ is our old friend the Chinese hat; the circ function is real and symmetric, so its transform is real and symmetric, and therefore its self-convolution equals its autocorrelation. (This is only true in focus, of course.)

The OTF and CTF each presuppose a concept of spatial frequency, but it must be understood that these two concepts do not map into each other in a simple way. Intensity is related to the squared modulus of the field variables; this nonlinearity results in the field amplitude spatial frequencies of the CTF undergoing large-scale intermodulation and distortion in the process of becoming the optical intensity spatial frequencies of the OTF. In particular, the width of the OTF is twice that of the CTF, but that does not imply the ability to resolve objects half the size. In discussing the OTF, we still use the variable names u and v, but do be aware that they no longer correspond directly to pupil plane coordinates, nor to the direction cosines of the plane wave components of ψ. (This is another example of a problem that's endemic in optics: reusing nomenclature in a confusing way.)

Being autocorrelations, optical transfer functions always droop at high spatial frequencies, and since intensity is nonnegative, OTFs must always have a maximum at zero. Interestingly, the OTF can go negative at intermediate values of spatial frequency, leading to contrast inversion for objects with periodicities falling in that region, an effect called *spurious resolution*. The OTF is purely real for symmetric optical systems but can exhibit phase shifts in systems lacking an axis of symmetry.

The justification for centering on the OTF is that, with thermal light, the phases of image points separated by more than a couple of spot diameters are uncorrelated, so there is no utility in keeping the phase information. This is of course fine if an in-focus image is being projected on film or an intensity detector, which is intrinsically insensitive to optical phase, but is inadequate for an aerial image or a phase-preserving system like a phase shifting interferometer or a laser heterodyne system, where the phase information still exists and can be very important, not least in fixing the imperfections of the image.

Perhaps the most intuitive way of capturing the distinction is that the OTF is not changed by putting a ground-glass screen at the image, whereas the CTF has its phase scrambled. Classical lens and optical systems designers use the OTF as one of their primary tools, which explains some of the difficulty encountered by workers in the different fields when they talk to each other.

Aside: Nonuniqueness of the Intensity Pattern. Since the relative phases of the plane waves in the CTF are lost when going to the OTF, any two patterns whose fields differ only in phase will produce the same intensity pattern, for example, positive and negative defocus.

[†]Well, irradiance, to be exact—the author resists the Humpty-Dumpty approach to radiometric nomenclature that goes around redefining preexisting terms to mean something else, in this case *intensity* being used for total power per steradian through some surface.

9.3.12 Shortcomings of the OTF Concept

The classical formulation of the optical transfer function is not a good analogue to transfer functions as used in circuit theory, ordinary differential equations, and so forth, although it might superficially look like it.

The behavior of fields is much more intuitive than that of image irradiance, because the fields exist throughout the optical system, whereas image irradiance doesn't. There are other ways in which the OTF isn't really a transfer function, the most important one being that you can't compute the OTF of two systems in cascade by simply multiplying the individual OTFs.

For example, consider a 1:1 relay system consisting of two lenses of focal length f, spaced $4f$ apart, as in Figure 12.1a. With an object at $-2f$ from the first lens, there will be a good image at the center of the system and another one at $2f$ past the second lens. If we choose the reference plane for the individual OTFs to be the center, everything works reasonably well. On the other hand, if we choose it to be off center, the image at that plane will be out of focus, leading to an ugly OTF, falling off very rapidly from zero spatial frequency. The second lens will also be defocused, leading to another ugly OTF, so their product will be ugly squared. This is exactly the right answer, *provided we put a diffuser at the reference plane*.

In real life, of course, an odd choice of reference plane doesn't affect the system operation at all—the defocus of the first half is undone by the defocus of the second, leading to a good image. The OTF gets this wrong, but the CTF gets it right—the phase curvatures of the two CTFs compensate correctly, and you get the right answer.

Lest anyone say that this is just silly, that nobody would set up a calculation that way, let's go a bit deeper into the problem. A symmetric optical system such as this 1:1 relay has no odd-order wave aberrations, because the second half's aberrations cancel out the first half's. (The even orders add.) Computing the overall OTF by multiplying the two half-OTFs will get this wrong, because the phase information is lost, so all the aberrations add in RMS instead of directly. Odd-order contributions will be overestimated, and even-order ones underestimated. Yet this weird OTF thing is called "the transfer function" and tossed about as though it had physical meaning. Beware.

9.3.13 Modulation Transfer Function

The modulation transfer function (MTF) is the magnitude of the OTF, normalized to unity at zero spatial frequency, and is most commonly used to describe the resolution performance of lenses, while not considering their photon efficiency.

9.3.14 Cascading Optical Systems

Under appropriate assumptions, when two optical systems are cascaded, their transfer functions are multiplied to get the transfer function of the cascade. If there is a diffuser, image intensifier, television system, or other phase-randomizing device between the two, use the OTF or MTF. Otherwise, use the CTF.

9.3.15 Which Transfer Function Should I Use?

This depends on the properties of the illuminator, and to a lesser degree on those of the detector. The assumptions leading to the derivation of the OTF are: an illuminator

with very low spatial coherence, and a detector that is sensitive only to intensity, such as a television camera or photodiode, with no phase reference (as in an interferometer). The resulting near-total loss of phase information severely limits the opportunities to gain from postprocessing, although the work of Fienup and others has demonstrated that some phase information can often be retrieved.

Example 9.5: OTF of an Ideal CCD Camera. As an example of the use of the OTF, consider a CCD camera with square pixels of pitch δ, a 100% fill factor, QE = 1 everywhere, and negligible bleed of one pixel into its neighbor. This is a spatial analogue of the sampled-data systems we'll encounter in Section 17.4.3, so although the detector is not shift invariant, we lose no information about the true OTF as long as the pixel pitch obeys the Nyquist criterion, and it is still sensible to talk about the OTF and MTF of such a system. The detector sensitivity pattern is $\text{rect}(x/\delta)\ \text{rect}(y/\delta)$, which is unaltered by squaring. Since u and x/λ are the conjugate variables, the detector CTF is the product of x and y sinc functions scaled by δ/λ, and its OTF is the same, so the OTF of the lens/CCD system is

$$\text{OTF}_{\text{tot}}(u, v) = \text{OTF}_{\text{lens}}(u, v) \left(\frac{\delta}{\lambda}\right)^2 \text{sinc}\left(\frac{u\delta}{\lambda}\right) \text{sinc}\left(\frac{v\delta}{\lambda}\right). \tag{9.48}$$

(We couldn't use this detector coherently without an LO beam, of course, so we can think of the spatial filtering action corresponding to this CTF as occurring on a Gaussian surface just above the detector.)

9.4 ABERRATIONS

A lens is often thought of as imaging an object plane on an image plane, but really it images a volume into another volume. A perfect imaging system would image every point in its object space to a corresponding point in its image space. The fidelity of the image would be limited only by diffraction, and in the transverse direction, it would be perfectly faithful geometrically as well. Squares would come out square (no distortion or anamorphic errors), and a flat object would give rise to a flat image (no field curvature), but unless the magnification was unity, the longitudinal magnification would nonetheless differ from the transverse magnification. Paraxial theory, whether the ray model (as in *ABCD* matrices) or the field model, always predicts perfect imaging, apart from defocus. We therefore expect the aberrations to turn up in the higher order terms.

Unfortunately, the algebra gets ugly in a hurry when we're dealing with exact ray tracing or scalar wave propagation; there are lots of square roots. Good quality optical systems ought to have phase aberrations that are small compared to the total phase delay through the system, so we anticipate that a power series expansion will yield useful simplifications. This power series is really not that well defined, because higher orders yield higher spatial frequency information that will eventually be corrupted by edge diffraction and vignetting, so that the aberration series is really a high-order polynomial plus some hard-to-treat residual, which we will assume is small.[†]

[†]This is rather like the distortion polynomial of Section 13.5.

Nobody uses high-order analytical aberration theory anymore. Lens designers use the low-order terms as conveniences, but rely on computer ray tracing and (manually guided) numerical optimization of an intelligently chosen starting configuration. For the system designer, high-order aberrations are of peripheral concern as well.

Aberrations are most obtrusive in wide field, high-NA optics, such as lithographic lenses and fast telescopes. Lots of instruments use lenses like that, but they are seldom fully custom designs, because the engineering and tooling costs would be astronomical. Thus the heroic lens design is someone else's problem—the rest of us mostly live in the low-NA, narrow field region behind those fancy lenses. Instrument designers need to know how aberrations propagate, what produces them, and how to avoid doing it. For this use, the lowest order aberrations are generally enough. For the same reason, we'll ignore the pure ray picture entirely and center on phase shifts of the plane wave components of a focused spot at an arbitrary field position.

9.4.1 Aberration Nomenclature

Aberration theory is somewhat separate from the rest of optics, because it is primarily used by lens designers, who have been doing much the same sort of work for 100 years, in sharp distinction to workers in most of the rest of optics. This is not to disparage the great strides lens design has made in that time, but it remains true that the formalism of classical aberration theory is not clearly related to the rest of the optical world, and a number of the terms used have different meanings than elsewhere. For example, to most practical optics folk, *defocus* means that the focal plane is offset from where it should be. In the paraxial approximation, a defocus d translates to a quadratic phase (time delay) across the pupil,

$$\Delta t_{\text{defocus}} \approx \frac{nd}{c}\left(1 - \frac{u^2}{2}\right), \tag{9.49}$$

whereas the real phase delay is $k_z z$, which in time delay terms is

$$\Delta t_{\text{defocus}} = \frac{nd}{c}\cos\theta = \frac{nd}{c}\sqrt{1-u^2}, \tag{9.50}$$

which of course contains all even orders in u.

In wave aberration theory, *defocus* means the quadratic expression (9.49), *even at large NA*. A pure focus shift in a large-NA system thus comes out as aberrations of all even orders, including spherical aberration and so on, even though a twist of the focus knob will restore the image completely. Those of us who use physical intuition heavily must guard against being led astray by this sort of thing.

Aberrations with the same name in the two different pictures do not correspond uniquely to one another; we've already seen the problem with defocus, but it also exists in other places. The names of aberrations have only mnemonic value—once again, if you expect everything to make sense together, you'll wind up chasing your tail.

As a way of connecting aberration theory with ordinary experience, let's calculate the effects of introducing a plane-parallel slab of dielectric into a perfect, converging spherical wave of limited NA.

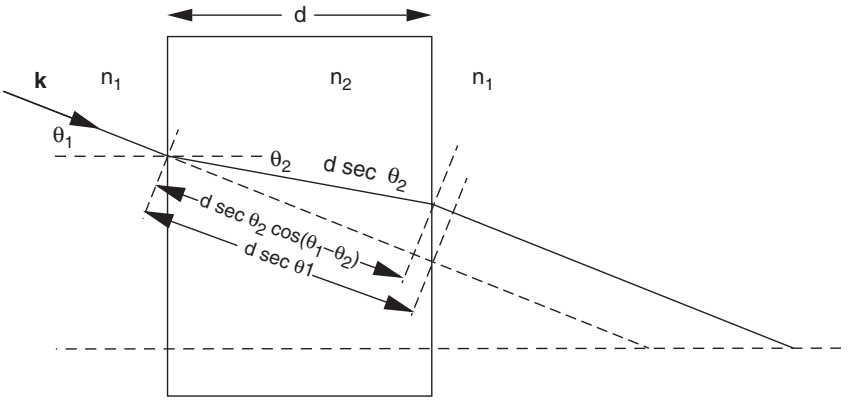

Figure 9.11. A plane-parallel slab of dielectric introduced into a plane wave.

9.4.2 Aberrations of Windows

Figure 9.11 shows the **k** vector of a plane wave incident on a plane-parallel slab of dielectric constant n_2. The refracted wave travels farther, and through a different index material. In the wave picture, this one is easy; the phase shift is just $(k_{z2} - k_{z1})d$. Let's use the hybrid picture, where we calculate the phase difference along the ray paths. Inspection of the figure shows that the change in the propagation time due to the presence of the slab is

$$\Delta t = \frac{d}{c} \sec\theta_2 (n_2 - n_1 \cos(\theta_1 - \theta_2)), \tag{9.51}$$

since a translation perpendicular to **k** has no effect on a plane wave. (If this isn't obvious, it's worth spending a bit of time on. This is a move that the hybrid picture relies on a good deal.) Without loss of generality, if the slab's faces are parallel to the (x, y) plane, and the incident plane wave has direction cosines $(u_1, 0)$, then Snell's law requires that $u_2 = n_1 u_1 / n_2$ (we've used $u_1 = \sin\theta_1$). Writing (9.51) in terms of u_1, we get

$$\Delta t = \frac{d}{c}\left[n_2 \sqrt{1 - \left(\frac{n_1 u_1}{n_2}\right)^2} - n_1 \sqrt{1 - u_1^2} \right], \tag{9.52}$$

which (comfortingly enough) is the same as the wave result. This obviously has terms of all even orders in u_1. Let's look at the first three orders, Δt_0 to Δt_4:

$$\Delta t_0 = \frac{d}{c}(n_2 - n_1),\ \Delta t_2 = \frac{d}{c}\frac{u_1^2}{2}(n_1 - n_1^2/n_2),\ \Delta t_4 = \frac{d}{c}\frac{u_1^4}{8}(n_1 - n_1^4/n_2^3) \tag{9.53}$$

The higher order terms in the series expansion are equally simple, and all have the same form. Specializing to a 6 mm thick plate and BK7 glass ($n_d = 1.517$), we get the result shown in Figure 9.12. The bulk of this is obviously caused by the zero-order time delay and the focus shift, but considering that one cycle of green light takes only 1.8 fs even the residuals may be very large (the right-hand axis goes up to 12,000 waves). We

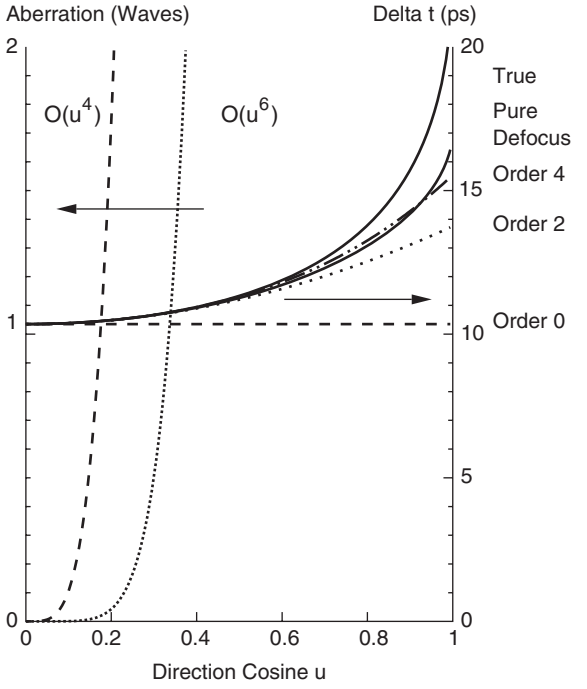

Figure 9.12. Differential time delay t suffered by a plane wave on passing through the dielectric plate of Figure 9.10, together with approximations of orders 0, 2, and 4. The curves at left are aberration residuals up to orders 2 and 4.

find out the effects of the aberrations on our nice focused beam by using the delay as a function of u and v to construct a phase factor to multiply our pupil function $\tilde{A}$, and then using (9.32) to get the new point spread function.

As we've already discussed, the u_1^2 term is usually called simply "defocus," though Figure 9.12 shows up the clear distinction between this and real defocus; to keep things clear, we'll call the quadratic term *paraxial defocus*. The term in u_1^4 is called *primary spherical aberration*. Spherical aberration is independent of field position, and so it occurs even for an on-axis beam.

The curves on the left show the aberrations of fourth order and greater, and of sixth order and greater, in units of waves (i.e., cycles) for 600 THz (500 nm) light, assuming that the pupil function of the beam is symmetric around $u = 0$. If that isn't true, for example, a low-NA beam coming in at a big field angle, the true aberration is v times the spread of time delays across the pupil function, minus the best-fit defocus.

9.4.3 Broken Symmetry and Oblique Aberrations

Other aberrations, such as astigmatism and coma, show up as soon as we run a finite field angle, that is, move the center of the pupil function distribution away from (0, 0). In an axisymmetric system like this plate or most lenses, these *oblique aberrations* are purely an effect of a shift of origin. (If the plate had some wedge angle, that would no longer be true.)

A residual effect of this broken symmetry is that if we move the origin to $(u_0, 0)$ (which loses us no generality), the pupil function is an even function of v. Thus the aberrations of a really symmetric system depend only on even powers of v, and by appropriate rearrangement of terms, that means they depend only on the cosine of the azimuthal angle θ ($u = \rho\cos\theta$, $v = \rho\sin\theta$). Manufacturing and assembly errors are in general asymmetrical and are frequently of the same order as the design residuals, so don't make too much of it.

If we move the center of the pupil function to $(u_0, 0)$, we're calculating the fields at a point $\mathbf{x} = (u_0 f/(1-u_0^2)^{1/2}, 0, f)$, where L is the z distance from the pupil to the focus. For simplicity, we'll call this x coordinate the height h. The aberration polynomial coefficients get a tiny bit more complicated,

$$\Delta t_0 = \frac{d}{c}(a-b), \tag{9.54}$$

$$\Delta t_1 = -\frac{d}{c}\eta(\alpha a - \gamma b), \tag{9.55}$$

$$\Delta t_2 = -\frac{d}{2c}(\eta^2[(\alpha\beta+\alpha^2)a - (\gamma\delta+\gamma^2)b] + v^2[\alpha\beta a - \gamma\delta b]), \tag{9.56}$$

$$\Delta t_3 = -\frac{d}{16c}\left(\eta^3[(8\alpha\beta+\alpha^3)a - (8\gamma\delta+\gamma^3)b] + \eta v^2[8\alpha\beta a - 8\gamma\delta b]\right), \tag{9.57}$$

and so on, where $u = u_0 + \eta$, $\beta = 1/(n_2^2/n_1^2 - u_0^2)$, $\alpha = u_0\beta$, $\delta = 1/(1-u_0^2)$, $\gamma = u_0\delta$, $a = n_1/\beta^{1/2}$, and $b = 1/\delta^{1/2}$. The coefficients of $\eta^i v^j$ are the aberration amplitudes.

9.4.4 Stop Position Dependence

One good way of reducing the effect of badly aberrated edge rays is to block them with a strategically placed stop. This may seem wasteful of light, but those rays weren't doing our measurement any good anyway, so they're no loss. This is one example of a case where the stops may be fairly far from the Fourier transform plane.

9.5 REPRESENTING ABERRATIONS

The standard method of representing the aberration coefficients of a wavefront is the *wave aberration polynomial*,[†]

$$W = \sum_{l,m,n=0}^{\infty} W_{2l+n,2m+n,n} h^{2l+n}\rho^{2m+n}\cos^n\phi, \tag{9.58}$$

where W is the optical path difference in meters (converted to $n = 1$ as usual). Think of kW as the (unwrapped) phase of $\tilde{A}$. Apart from the fact that the practical upper limit of this summation is very finite, it's moderately useful, although more mysterious looking in this form. The coefficients all have names up to order 6 or so (the order of a term

[†]Warren J. Smith, Optical design, Chap. 2 in J.S. Accetta and D.L. Shumaker, *The Infrared and Electro-Optical Systems Handbook*, Vol. 3.

TABLE 9.1. Seidel Aberrations

Piston	W_{000}
Tilt	$W_{111}h\rho\cos\theta$
(Paraxial) defocus	$W_{020}\rho^2$
Spherical	$W_{040}\rho^4$
Coma	$W_{131}\rho^3\cos\theta$
Astigmatism	$W_{222}h^2\rho^2\cos^2\theta$
Field curvature	$W_{220}h^2\rho^2$
Distortion	$W_{311}h^3\rho\cos\theta$

is the sum of the exponents of ρ and h, not $\cos\theta$), which are listed up to order 4 in Table 9.1. Of the ones we haven't talked about, piston is just an overall phase shift, which we often don't care about, and tilt corresponds to a shift in the focal position.

The ray model came first. Ray aberrations are quoted as position errors in the focal plane; because the ray travels along ∇S, the same term shows up in one lower order in the ray model—astigmatism is a fourth-order wave aberration but a third-order ray aberration, which can cause confusion sometimes. We saw in Section 9.2.3 that the local direction of propagation is parallel to $\nabla\Phi$. The ray intercept error is

$$\Delta\mathbf{x} = -\frac{L}{n}\nabla(\text{OPL}). \tag{9.59}$$

The most common way to quote aberration contributions is in peak-to-peak waves over the full diameter of the pupil.

9.5.1 Seidel Aberrations

Wave aberrations up to order 4 are known as Seidel aberrations; their pupil functions are shown in Figure 9.13 and their functional forms in Table 9.1. Looking at the spherical aberration and astigmatism profiles, it is clear that the RMS wavefront error could be significantly reduced by *compensation*, that is, adding a bit of tilt to the coma and a bit of defocus to the spherical aberration so as to minimize $\langle\Delta\phi\rangle$. Compensated waveforms are shown in the bottom row of Figure 9.13.

9.5.2 Aberrations of Beams

Frequently in instruments we want to talk about the aberrations of a fixed laser beam, so it doesn't make much sense to talk about dependence on field angle or image height. In that case, the only relevant terms up to fourth order are paraxial defocus ρ^2, astigmatism $\rho^2\cos^2(\theta-\theta_0)$, spherical aberration ρ^4, and coma $\rho^3\cos(\theta-\theta_1)$. Since in general no symmetry constraint applies, the θ_i can be anything.

Aside: Zernike Circle Polynomials and Measurements. The Zernike polynomials are an orthogonal basis set for representing the optical phase *in a circular pupil*. This sounds like a great way of expressing measurement results—decomposing a wavefront into orthogonal polynomials is computationally cheap and well conditioned, and all. Unfortunately, their practical utility is zilch. Due to vignetting and beam nonuniformity, our pupils are almost never exactly circular or uniformly illuminated, and errors in the

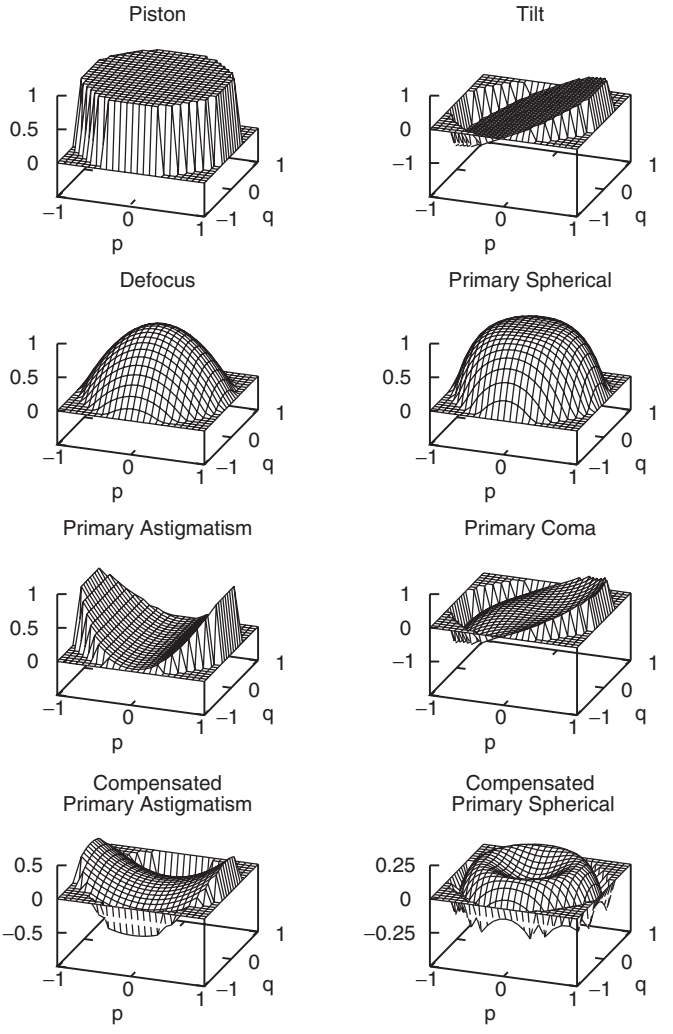

Figure 9.13. Seidel aberrations.

boundary conditions destroy the orthogonality. Defining the "true" Zernike coefficients is especially problematical when our measuring interferometer is intentionally using an elliptical clipping boundary, and perhaps choosing a different ellipse for each run. Even if the pupil stays circular, Zernikes are only obliquely connected to the beam quality measures we care about (e.g., defocus in diopters).

The power series coefficients stay reasonably still even as the bounding ellipse changes, and are pretty well connected to what we care about, so use that up to fourth order in the wavefront. If that isn't accurate enough, do the calculation numerically.

9.5.3 Chromatic Aberrations

Different wavelengths see different values of n, and a time delay Δt produces different phase shifts. Thus all the coefficients of the aberration polynomial are wavelength

dependent. Changes in focal length with λ are *longitudinal chromatic aberration*, changes in magnification are *lateral chromatic aberration* or *lateral color*, and changes in the primary spherical term are *spherochromatism*. Few of these things are under our control as system designers, which isn't to say they aren't important.

9.5.4 Strehl Ratio

The Strehl ratio is the ratio of the central intensity of a focused spot to what it would be with the same amplitude distribution but zero phase error,

$$R = \left| \frac{\iint \tilde{A}(u, v) du\, dv}{\iint \left|\tilde{A}\right| du\, dv} \right|^2 . \tag{9.60}$$

The Schwarz inequality guarantees that this ratio achieves 1 only when the phase error is indeed zero, and never exceeds that value. A Strehl ratio of 0.8 corresponds to Rayleigh's λ/4 criterion for a system that is diffraction limited.† When building focused-beam instruments, we frequently find that the electrical signal power goes as the square of the Strehl ratio, which is a convenient way of including aberration tolerances in our photon budgets. A useful approximation for the Strehl ratio is Marechal's formula,

$$R \approx \exp \frac{-\langle \Delta\phi^2 \rangle}{2\pi^2}, \tag{9.61}$$

where $\langle \Delta\phi^2 \rangle$ is the mean square phase error in rad^2 (remember to weight the mean square calculation by the intensity and area, and normalize correctly). If you prefer to use the phase p in waves (i.e., cycles), it's $\exp(-p^2/2)$, which is pretty easy to remember.

The Strehl ratio suffers from the same problem as equivalent width; if the peak is shifted from the origin, the ratio may come out very small, even for a good beam. Thus we often want to calculate the Strehl ratio after the tilt contribution (which moves the focus sideways) has been removed. Also, it can give some odd results with multiple transverse mode beams—because the different spatial modes have different frequencies, DC intensity measurements like Strehl ratio miss the rapidly moving fringe pattern in the beam, and so underestimate the far-field divergence. The Strehl ratio is an excellent quality measure for good quality beams, where the intensity profile has one main peak; for uglier ones, consider using Siegman's M^2 instead.

Example 9.6: ISICL Signal-to-Noise Calculation. The Strehl ratio shows up by other names in other fields. In antenna theory, it is called the *phase efficiency*, which is the ratio of the on-axis detected signal to that of a perfectly figured antenna, neglecting loss. Strehl ratio thus shows up as a multiplicative factor on the detected signal from focused-beam instruments. The ISICL sensor of Example 1.12 uses coherent detection with a Gaussian beam. Referring to Figure 1.15, a particle with differential scattering cross section $d\sigma/d\Omega$ crossing near the focus produces a scattered photon flux per steradian

$$\frac{dn_s}{d\Omega} = \frac{2\pi P_T (\mathrm{NA}_T)^2}{\lambda hc} \frac{d\sigma}{d\Omega} R_T, \tag{9.62}$$

†Strictly speaking, the 0.8 corresponds to 0.25 wave of uncompensated spherical aberration.

where R_T and NA_T are the transmit beam's Strehl ratio and numerical aperture. When coherently detected by a similar LO beam, the detection NA is $\pi(\mathrm{NA}_R)^2 R_R$ (the factor of R_R accounts for the dephased (aberrated) part of the LO power, which goes into the shot noise but not into the coherently detected signal). Doppler shift makes this an AC measurement, so from Section 1.5.2, the 1 Hz SNR is equal to the number of detected signal photoelectrons per second, which is

$$\mathrm{SNR} = n = \frac{2\pi^2 \eta P_T (\mathrm{NA}_T)^2 (\mathrm{NA}_R)^2}{\lambda h c} \frac{d\sigma}{d\Omega} R_T R_R, \tag{9.63}$$

where η is the detector's quantum efficiency.

The detected SNR goes as the product of the Strehl ratios of the transmit and LO beams. This provides a natural connection between aberrations and signal level (and hence SNR), which is why the Strehl ratio is so useful for instrument designers.

9.6 OPTICAL DESIGN ADVICE

Many books have been written on lens design, and lots of software exists to help. That's not what we're talking about now, and is in fact beyond our scope. Optical design (in the restricted sense used here) is concerned with sticking lenses and other parts together to get the desired result. An analogy from electronics is IC design versus application circuit design, with the lens being like the IC: most of the time you use standard ones, but occasionally you have to do a custom one; it will be a better match, but will cost something to design and will usually be more expensive in production unless you have gigantic volumes.

Computerized exact ray tracing is not usually necessary in instrument design, although if you have easy-to-use software for it, you might as well—it doesn't make anything worse, after all. On the other hand, we often don't have the full optical prescription for the lenses, so exactness doesn't buy us a lot. Thick-lens paraxial ray optics is better than good enough for layout, and our hybrid wave optics model plus some simple aberration theory does the job of calculating image or beam quality, signal levels, and SNR.

If necessary, we can use ray tracing or numerical wavefront simulation to dial in the design once we've arrived at a close approximation, but it is needed surprisingly seldom since the fancy stuff is usually done at very low NA, where life is a lot easier. Apart from etalon fringes, using mostly collimated beams (or parallel light in imaging systems) makes it possible to add and remove components freely, with only minor effects on optical quality.

9.6.1 Keep Your Eye on the Final Output

In Section 1.7.1, we used an expression for the detected photocurrent as the optical system output, and that remains the right thing to do when aberrations and finite aperture systems are considered—you just use the real pupil function instead of the paraxial one, and pay attention to obliquity factors and the Strehl ratio. It's an excellent way to know just when our treatment of system aberrations is losing contact with instrument-building reality. As we saw in Example 9.6, it leads to a natural interest in the Strehl ratio, which appears in the photocurrent and SNR calculations. There are other things that matter

besides the photocurrent (e.g., geometrical distortion), but if you don't see how to relate the figure of merit under discussion to the actual instrument performance, find another way of describing it until you can. Lens designers produce lenses for a living, and we build electro-optical systems.

9.6.2 Combining Aberration Contributions

An optical system is made up of a series of imperfect elements, each contributing its own aberrations. In order to combine them, we note that all the pupils in the system are images of each other, and so the pupil functions multiply together. The exit pupil of each element is imaged at the output of the system by the (ideal) imaging action of all subsequent elements, and the resulting pupil functions multiplied together to get the total pupil function of the system. Watch out for magnification differences—those pupils won't all be the same size, and the smallest one wins.

9.7 PRACTICAL APPLICATIONS

9.7.1 Spatial Filtering — How and Why

Spatial filtering is the deliberate modification or removal of some plane wave components from a beam, and is completely analogous to the ordinary electronic filtering performed in a radio. It is normally done by using a lens to Fourier transform the beam, inserting at the focal plane a mask that is transparent in some places and opaque in others (such as a slit or a pinhole), and then transforming back with a second lens. It is widely used for cleaning up beams and for removing artifacts due to periodic structures in a sample (e.g., IC lines).

Spatial filtering using a pinhole can make a uniform beam from a Gaussian one but will waste around 75% of the light doing it. The smallness of the pinhole required for a good result with a given NA may be surprising—the first Airy null is *way* too big (see Example 9.9).

Spatial filters are not as widely used as one might expect, based on the analogy to electrical filters. They require complex mechanical parts, which is a partial explanation. The real trouble is tweakiness: they are difficult to align, easy to misalign, and sensitive to mechanical and thermal drifts; if ordinary translation stages are used, an expensive and very labor-intensive device results. These problems can be reduced by clever design, for example, by using a laser or an in situ photographic process (with a fixed optical system) to build the mask right inside the filter. If you need a pinhole spatial filter, use a big pinhole ($>20\ \mu$m diameter in the visible) and correspondingly low NA to get the best stability. One other problem is that they nearly all have sharp edges, which isn't usually optimal.

9.7.2 How to Clean Up Beams

Laser beams are frequently rotten. Most gas lasers produce beams of reasonable quality, but these are often surrounded by multiple reflections, scattered light, and spontaneous emission. Diode lasers are worse; their beams suffer from astigmatism and are highly asymmetric. We may need to free these beams of their bad associations in order to make them useful, or to change their spatial distributions to something more convenient.

This is done through spatial filtering and the use of apertures (and occasionally special elements such as anamorphic prism pairs). Since both of these operations are performed by passing the beam through holes of various sizes, the distinction is somewhat artificial but is nonetheless useful: apertures are used on the beam before focusing, and spatial filters in the Fourier transform plane. A seat-of-the-pants test is that if it requires a fine screw to adjust, it's a spatial filter.

Putting an aperture some way out in the wings of the beam (say, four times the $1/e^2$ diameter) has little effect on its propagation characteristics, so use them freely to reduce artifacts. If the artifacts are close to the beam axis, it may be helpful to let the beam propagate for some distance before applying the aperture; a lens may be helpful in reducing the optical path length this might otherwise require (here is where it shades into spatial filtering). Appropriately placed apertures can turn a uniform beam into a good Gaussian beam, or chop off the heavily aberrated wings of an uncorrected diode laser beam.

Example 9.7: How Small Do I Have to Make My Aperture? Slits and pinholes are frequently used to render an instrument insensitive to the direction from which incident light originates. A monochromator (see Example 7.1) uses a slit to ensure that the light incident on its grating arrives from one direction only; this maximizes its spectral selectivity. There's obviously a trade-off between selectivity and optical throughput; because different positions across the slit translate to different incidence angles on the grating, a wider slit translates directly into lower spectral resolution.

More subtly, with a wide slit a change in the position or direction of the incoming light can cause apparent spectral shifts. This is because the slit doesn't obliterate the spatial pattern of the incoming light, it just clips it at the slit edges, and so broadens its angular spectrum. If the slit is made small enough, it can't shift far laterally, and we expect that the width of the diffraction pattern will swamp any likely angular change from the incoming light. On the other hand, using a slit that narrow can make things pretty dim. Exactly how narrow does it have to be?

As we saw in Section 5.7.9, scatterers tend to produce patterns that are aligned with the incident light, but smeared out in angle. The same is true of slits and pinholes; in the Fourier optics approximation, the (amplitude) angular spectrum of the incident light is convolved with the Fourier transform of the aperture's transmission coefficient.

If we illuminate the slit at an incidence angle θ, the main lobe of the sinc function is aligned with the $\mathbf{k}$ vector of the incoming light, and the intensity along the normal to the slit will decrease as θ increases. Since x/λ and u are conjugate variables, if we require that a change of $\pm\delta$ radians cause a fractional decrease of less than ϵ in the central intensity of the diffraction pattern, the slit width w must obey

$$w \sin \delta < \sqrt{\frac{6\epsilon}{\pi}}, \tag{9.64}$$

so that for a relatively modest requirement, for example, a shift of less than 5% from a $\pm 10^\circ$ rotation, $w < 2.25\lambda$. It is difficult to use slits this small, but not impossible. Improving on this, such as requiring a 1% shift from a $\pm 30^\circ$ rotation, is impractical, as it requires a slit very small compared with a wavelength.

Similar considerations govern the sensitivity to lateral motion of a pinhole in a focused beam. The moral of this story is that although spatial filters can reduce the sensitivity of a measurement to angular shifts of illumination, a complete cure is not to be

TABLE 9.2. Effects on a Gaussian Beam of a Circular Aperture of Radius r Placed at the Beam Waist

r/w	ΔI(nom)	w_{best}	ΔI(best)
0.5	0.66	2.72	0.027
1.0	0.30	1.47	0.022
1.5	0.099	1.13	0.014
2.0	0.022	1.0245	0.0055
2.5	0.0028	1.003	0.0014
3.0	0.0002	1.0002	0.0001

found here. Single-mode fibers are dramatically better, because they control the direction from which the light hits the pinhole (see Section 8.2.2). Unfortunately, they'll often make your light source dramatically noisier as well, by turning FM noise into AM; see Section 8.5.13.

Example 9.8: How Small Can I Make My Aperture? On the other hand, if we have a beam that has artifacts out in its wings, such as most gas laser beams, we would like to make the aperture as small as possible without causing objectionable diffraction rings. If the beam is Gaussian, the ripple amplitude is very sensitive to slight vignetting. Table 9.2 gives the maximum absolute deviation ΔI from both the nominal and best-fit Gaussian beam, in the transform plane, due to apertures of different radius placed at the beam waist. Values have been normalized to a central intensity of 1.0.

Example 9.9: Using an Aperture to Make a Nearly Gaussian Beam. Let's consider the use of an aperture in the pupil to make a uniform beam into a reasonable approximation to a Gaussian beam. Although a uniform beam exhibits an extensive set of diffraction rings, an appropriately placed aperture can greatly reduce their amplitude. A uniform beam of amplitude E_0 and radius a at a wavelength λ gives rise to a far-field Airy pattern whose first null is at $\theta = 0.61\lambda/a$. If the beam is Fourier transformed by a lens of focal length f, so that the numerical aperture $\text{NA} = a/f$, the pattern is given by

$$E(r) = E_0 k\ \text{NA}^2 \text{jinc}(kr\text{NA}). \tag{9.65}$$

Figure 9.14 shows the result of truncating this pattern with a circular aperture at the first Airy null, and recollimating with a second lens. Here the radius of the original uniform beam was 100 μm and the wavelength was 530 nm. The Gaussian shown has a $1/e^2$ (intensity) radius of 95 μm, and the sidelobes are below 1 part in 10^3 of the peak intensity. Only about 15% of the light is lost. Note that the peak intensity is twice that of the original uniform beam. Contrary to appearances, the total beam power has gone down—the high peak intensity covers a very small area since it's at the center. A graded neutral density filter, which is the competing technique, cannot increase the central value, so that it is limited to at most half this efficiency, and far less if we expect the Gaussian to drop to nearly zero before the edge of the beam is reached; on the other hand, it requires no lenses. (See Figure 9.15.)

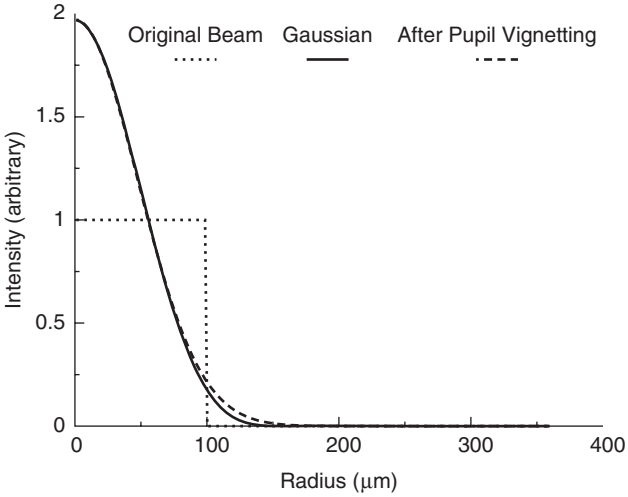

Figure 9.14. Turning a uniform beam into a nearly Gaussian one with a pinhole of the same radius as the first Airy null.

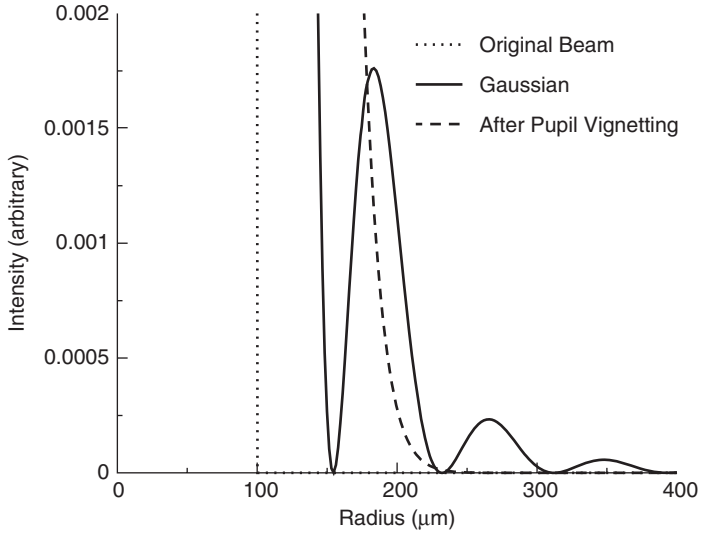

Figure 9.15. The data of Figure 9.14 on an expanded scale.

9.7.3 Dust Doughnuts

In general, a bit of dust on a lens is no big deal. The exception is when the dusty surface is near a focus, in which case dust is extremely objectionable, leading to strong shadows called *dust doughnuts*. (The shadows are annular in shape in a Cassegrain telescope, hence the name.) How far out of focus does the dust have to be?

Assuming the dust is less than 100 μm in diameter, and that a 1% intensity error is acceptable, a focused spot has to be at least 1 mm in diameter at the dirty surface, so

the required defocus is

$$|\delta Z_{\text{defocus}}| \gtrsim \frac{1 \text{ mm}}{2 \cdot \text{NA}}. \tag{9.66}$$

9.8 ILLUMINATORS

This discussion is indebted to an SPIE short course, "Illumination for Optical Inspection," by Douglas S. Goodman.

9.8.1 Flying-Spot Systems

A scanning laser microscope is an example of a *flying-spot* system, as opposed to a *full-field* or *staring* system, in which a larger area is imaged at once. The first flying-spot optical systems in the 1950s used an illuminated point on a cathode ray tube as the light source, and a PMT as the detector because the spot was so dim. That at least had the advantage of speed and no moving parts. Flying-spot systems are simple to analyze, because their PSFs are the product of the illumination and detection spots (whether amplitude or intensity is the appropriate description), and there is no problem with speckle or scattered light smearing out the image.

9.8.2 Direction Cosine Space

Full-field systems require a bit more work to specify accurately. We're usually using thermal light from a bulb of some sort, so the illumination is coming from some reasonably wide range in (u,v). Figure 9.16 shows a sample region illuminated by a cone of light, which is plotted on a large sphere (much bigger than the sample region of interest) and then projected down into the (x,y) plane. Since the exact radius R is of no significance, we normalize it out and plot the illumination in terms of u and v. Due to the curvature of the spherical surface, illuminated patches lying near the horizon are strongly foreshortened in area (by a factor of $\cos\theta$, where θ is the polar angle). This is quite useful actually, since the flux passing into the surface is reduced by the same factor due to the spreading out of the obliquely illuminated patch (equivalently, the apparent source area seen by any one point is reduced by the cosine). Figure 9.17 shows bright- and dark-field systems, with oblique and concentric illumination.

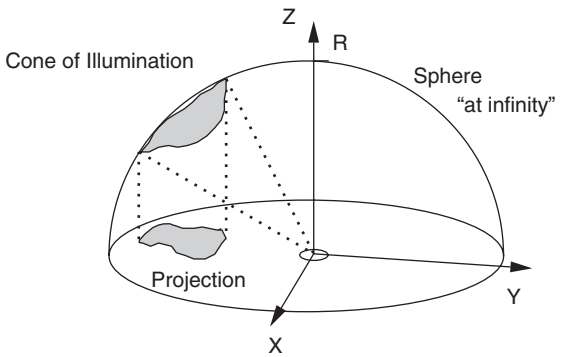

Figure 9.16. Direction cosine space.

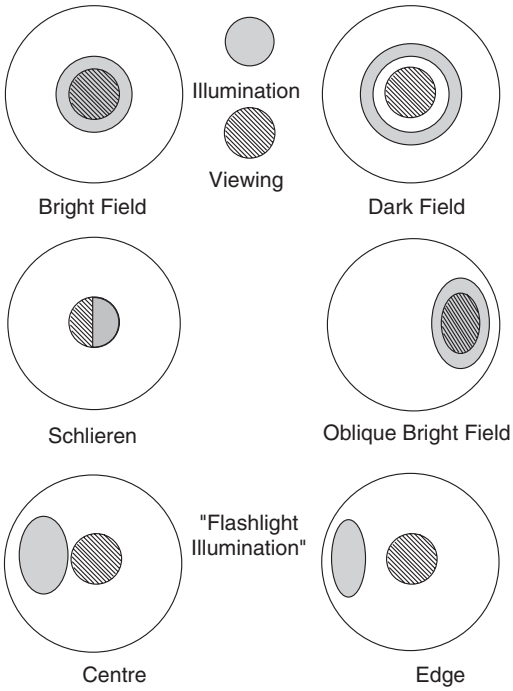

Figure 9.17. The general type of illumination is determined by the overlap of the illumination and detection patterns in u,v space.

9.8.3 Bright and Dark Fields

Illuminators are divided into bright- and dark-field types, depending on whether the image of a featureless sample (e.g., a mirror in reflection or a glass plate in transmission) is bright or dark; this is equivalent to whether the illuminator is directly visible in the image. In direction cosine space, a bright-field system has illumination and collection patterns that substantially overlap, whereas they miss each other completely in a dark-field setup. It is possible to work in intermediate states, generically called dim field.

Dark-field images consist entirely of scattered light, so that highly scattering areas such as dust particles appear bright on a dark background. Bright-field ones see darker areas where light has been absorbed or scattered so far as to miss the collection lens. This means that increasing the NA tends to wash out the contrast in bright field—all that scattered light is collected and reassembled in the image, so only the absorption contrast is left. Dark-field illumination shows up phase objects and weak scatterers better than bright field, but the image intensity is quadratic in the phase shift ϕ, so it is not especially sensitive at small ϕ. In Sections 1.5 and 10.3.5 we discuss ways to get higher sensitivity for small signals.

9.8.4 Flashlight Illumination

The first crack at an illumination system is often to take a fiber bundle illuminator and point it at the sample, keeping it away from the rest of the optical system. This produces a kind of oblique dark-field illumination that varies in angular spectrum and intensity

across the sample. If the sample is diffuse (e.g., white paper), or quantitative results are unnecessary, this may work fine—and it's certainly quick.

9.8.5 Critical Illumination

Of more thoughtful illumination strategies, the simplest is critical illumination: just image the source down onto the sample. Any variations in the source appear directly in the image, so the source must be very uniform. A more subtle problem is that critical illumination is generally nonstationary; that is, the illumination diagram is different at different points in the field of view. This is because the image of the bulb radiates in all directions, and hence its angular spectrum will be vignetted in the pupil of the imaging lens, at least for points sufficiently far off-axis. This vignetting can be so severe that the edges of the field change from bright to dark field, as the specular reflection misses the collection lens completely.

9.8.6 Köhler Illumination

Köhler illumination overcomes the source nonuniformity problem by putting the sample at the Fourier transform plane of the condenser, or equivalently by imaging the source on the entrance pupil of the collecting lens. This strategy makes the illumination conditions stationary across the field (i.e., it doesn't change from bright to dark field the way critical illumination can), because all the unscattered light makes it through the pupil of the imaging lens without being vignetted.[†]

Köhler illumination tends to keep the spatial frequency bandwidth more nearly constant; the center of the light cone from each point goes through the center of the pupil. Vignetting will reduce the spatial frequency bandwidth at the edges but won't make the illumination change from bright to dark field across the sample, as can happen with critical illumination.

9.8.7 Testing Illuminators

Douglas Goodman suggests using the *ball-bearing test*. A sphere reflects a collimated beam into 4π steradians with equal power per steradian in all directions, and this of course works backwards too; you can see the whole illumination pattern in direction cosine space by looking at its image in a sphere. For looking at microscope illumination, mount the ball bearing on a flat black plate in reflection, or on a microscope slide with a tiny glue patch in transmission. You can see the microscope lens in the ball too, so you can align the two for concentricity that way. The direct readout in direction cosine space allows you to see what your illumination function looks like very easily, though you may have to stop down your collection NA to get a clear view of it.

In other situations, the interior of a ping-pong ball cut in half makes a very handy screen for projecting the illumination on. It's especially good as a scatterometer with diffracting structures and laser illumination; shine the laser through a hole in the top of the hemisphere and look at the diffraction spots. (Ping-pong balls can also be used as integrating spheres—see Section 5.7.8.)

[†]Recall that the pupil is usually at or near the transform plane.

9.8.8 Image Radiance Uniformity

We saw in Section 2.4.1 that a planar source with no preferred direction is Lambertian. Whenever such a diffuse planar object is imaged on a planar detector over a wide field, the edges of the image become darker. The darkening gets worse rapidly as the field angle θ increases; the image radiance goes as $\cos^4\theta$, as we can verify by counting powers of the cosine. The object–pupil distance goes as $\sec\theta$, which gives us two cosine factors since the flux at the pupil goes as $1/r^2$. The projected pupil area goes as the cosine, which makes three, and Lambertian angular dependence of the object radiance makes four factors of the cosine altogether. Vignetting at places other than the pupil and the increased Fresnel losses at high angles make real systems somewhat worse in general, though there are ways of increasing the loss near the middle to flatten the curve a bit. The $\cos^4$ falloff is also why it's easier to get uniform illumination by spacing many sources at close intervals, rather than trying to use one source plus diffusers.

9.8.9 Contrast and Illumination

Imaging systems stand or fall by their contrast. A featureless image contains no information; one where the contrast comes from an unexpected mechanism (e.g., stray fringes) may contain even less—you're more confused and misled after the measurement than before.

The word *contrast* is used in two different senses. Contrast is defined as the ratio of p-p intensity change to mean intensity (flux density, irradiance) in the image, which is 0 for no contrast and 1 at perfect contrast:

$$C = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}. \tag{9.67}$$

The other sense denotes the source of the contrast: imaging can be done in phase contrast, interference contrast, and so on. Contrast is a normalized measure; dialing down the illumination power doesn't change the contrast, but of course the signal level and SNR will deteriorate.

The image contrast is a complicated function of the illumination type, collection, detection strategy, and sample. The dependence is not weak, either; different illumination will make bright areas look darker than dark areas. For example, smooth aluminum lines on rough, (nearly black) polycrystalline silicon carbide will look very bright when illuminated in bright field, but will look dark when illuminated obliquely, because their specular reflection causes the returned light to miss the objective lens, whereas the rough surrounding surface will scatter some light into the lens—a black surface can look lighter than a mirror. Other examples are everywhere. The scattering geometry is important too; linear features such as scratches or smears scatter very efficiently perpendicular to the line, but only weakly at other angles; if the illumination doesn't have any component perpendicular to the lines, the scattered light will be weak. (This phenomenon is familiar to us all—think of a smeared car windshield at night.)

The most important thing in choosing an illumination strategy for your measurement is to mess around with a whole lot of alternatives before making a final choice. Automatic inspection, machine vision, lithography, microscopy, and trace defect detection live and die by the quality of the illuminator and by how well it's matched to the characteristics of the problem. Poorly chosen illumination can make the software job many times harder

(or even impossible). Don't just buy a fiber illuminator, shove it in, and expect it to work well.

In the highest resolution optical microscopy, resolution is improved by using diffuse illumination and high-NA microscope objectives, because high spatial frequency components can take a plane wave component near grazing and send it back the way it came, so that the spatial frequency bandwidth is doubled. Unfortunately, most samples show almost no contrast when examined this way, so experimentation is needed even here.

9.8.10 Retroreflectors and Illumination

Any time you're trying to inspect a transparent or specularly reflecting object that's bigger than your lens, there's a problem with illumination uniformity and efficiency, as well as a serious case of changing illumination conditions with field position. One very useful way of solving this is to use a big chunk of retroreflecting material, with the illuminator and sensor optically superposed. Light from the source bounces off the specular surface, hits the retroreflecting material, and retraces its path to the collecting lens. This is a great way of keeping the illumination conditions the same even with a small light source—it's sort of a poor man's telecentric system. The angular spread of the returned beam is a degree or so, so there is a certain amount of ghosting and other artifacts, but for jobs like inspecting plastic film for defects, or looking at semiconductor wafers, it's a very useful trick sometimes (see Section 7.8).