Microscopy by reconstructed wave-fronts

By D. Gabor, Dr.-Ing.

*Research Laboratory, British Thomson-Houston Company Ltd., Rugby**

(*Communicated by Sir Lawrence Bragg, F.R.S.—Received* 23 *August* 1948—*Revised* 28 *December* 1948—*Read* 17 *February* 1949)

[Plates 15 to 17]

The subject of this paper is a new two-step method of optical imagery. In a first step the object is illuminated with a coherent monochromatic wave, and the diffraction pattern resulting from the interference of the coherent secondary wave issuing from the object with the strong, coherent background is recorded on a photographic plate. If the photographic plate, suitably processed, is replaced in the original position and illuminated with the coherent background alone, an image of the object will appear behind it, in the original position. It is shown that this process reconstructs the coherent secondary wave, together with an equally strong 'twin wave' which has the same amplitude, but opposite phase shifts relative to the background.

The illuminating wave itself can be used for producing the coherent background. The simplest case is illumination by a point source. In this case the two twin waves are shown to correspond to two 'twin objects', one of which is the original, while the other is its mirror image with respect to the illuminating centre. A physical aperture can be used as a point source, or the image of an aperture produced by a condenser system. If this system has aberrations, such as astigmatism or spherical aberration, the twin image will be no longer sharp but will appear blurred, as if viewed through a system with twice the aberrations of the condenser. In either case the correct image of the object can be effectively isolated from its twin, and separately observed. Three-dimensional objects can be reconstructed, as well as two-dimensional.

The wave used in the reconstruction need not be the original, it can be, for example, a light-optical imitation of the electron wave with which the diffraction diagram was taken. Thus it becomes possible to extend the idea of Sir Lawrence Bragg's 'X-ray microscope' to arbitrary objects, and use the new method for improvements in electron microscopy. The apparatus will consist of two parts, an electronic device in which a diffraction pattern is taken with electrons diverging from a fine focus, and an optical synthetizer, which imitates the essential data of the electronic device on a much enlarged scale.

The theory of the analysis-synthesis cycle is developed, with a discussion of the impurities arising in the reconstruction, and their avoidance. The limitations of the new method are due chiefly to the small intensities which are available in coherent beams, but it appears perfectly feasible to achieve a resolution limit of 1 Å, ultimately perhaps even better.

Introduction

The period of steady progress in the resolving power of electron microscopes which was started in 1931 by Knoll & Ruska came virtually to an end in 1946, when Hillier & Ramberg (1947) eliminated the astigmatism of their objective, and achieved a resolving power only insignificantly different from the theoretical limit. The barrier which stopped progress is of a technical nature, but formidable enough to prevent any really essential improvements along the direct line.

The theoretical limit of conventional electron microscopes is about 5Å. It is determined by a compromise between diffraction and spherical aberration in electron objectives, and at the best compromise it is proportional to the fourth root of the aberration constant. Though several suggestions for correction have been put

* Now at Imperial College, London, Electrical Engineering Department.

forward, they involve such technical difficulties that an improvement by a factor of 2 is the best that can be expected, even optimistically. One can never hope to achieve a resolving power ten times better than the present, which would require a correction of the spherical aberration to about 1 part in 10,000. Such precision can be realized with the technique of the optical workshop, but hardly ever with the means at the disposal of electron optics.

The new method is an attempt to get around the obstacle, instead of across it, by a two-step process, in which the analysis is carried out with electrons, the synthesis by light. The general idea of such a process was first suggested to the author by Sir Lawrence Bragg's 'X-ray microscope' (Bragg 1942; cf. also Boersch 1938). But Bragg's method, in which a lattice is reconstructed by diffraction from an X-ray diffraction pattern, can be applied only to a rather exceptional class of periodic structures. It is customary to explain this by saying that diffraction diagrams contain information on the intensities only, but not on the phases. The formulation is somewhat unlucky, as it suggests at once that since phases are unobservables, this state of affairs must be accepted. In fact, not only that part of the phase which is unobservable drops out of conventional diffraction patterns, but also the part which corresponds to geometrical and optical properties of the object, and which in principle could be determined by comparison with a standard reference wave. It was this consideration which led me finally to the new method.

In order to make the two-step method generally applicable, it had to be combined with a principle apparently not hitherto recognized. If a diffraction diagram of an object is taken with coherent illumination, and a coherent background is added to the diffracted wave, the photograph will contain the full information on the modifications which the illuminating wave has suffered in traversing the object, apart from an ambiguity of sign, which will be discussed later. Moreover, the object can be reconstructed from this diagram without calculation. One has only to remove the object, and to illuminate the photograph by the coherent background alone. The wave emerging from the photograph will contain as a component *a reconstruction of the original wave*, which appears to issue from the object. Conditions can be found in which the remainder can be sufficiently separated from the useful component to allow a true, or very nearly true, reconstruction of the original object.

This principle has been confirmed by numerous experiments. Some of the results are shown in figures 10 to 12 and explained in the last section of this paper.

In light optics a coherent background can be produced in many ways, but electron optics does not possess effective beam-splitting devices; thus the only expedient way is using the illuminating beam itself as the coherent background. This leads us to illumination by a coherent, divergent electron wave, illustrated in figure 1. It will be useful to explain this arrangement first, anticipating the principle of reconstruction which will be proved later.

The apparatus consists of two parts, the electronic analyzer and the optical synthetizer. The analyzer is similar to an electron shadow microscope (Boersch 1939), but with the important difference that it operates with coherent illumination, and under conditions in which the shadow microscope is useless, as the interference diagram has little likeness to the original. An electron gun, combined with a suitable

aperture and electron lens system, produces a coherent illuminating beam, as nearly homocentric as possible. Exactly homocentric illumination is of course impossible, because of the unavoidable spherical aberration of electron lenses, but for simplicity we can talk of the narrow waist of the beam as of a 'point focus'. A small object is arranged some small distance before or behind the point focus, and a photographic plate at a comparatively large distance L. The divergence angle of the beam, γ_m, must be sufficient for the required resolution limit d_A, which is by Abbe's relation

$$d_A = \tfrac{1}{2}\lambda \sin \gamma_m.$$

The factor $\frac{1}{2}$ will be used in this paper to simplify the discussions, except in numerical calculations, where it will be replaced by the more accurate value 0·6.

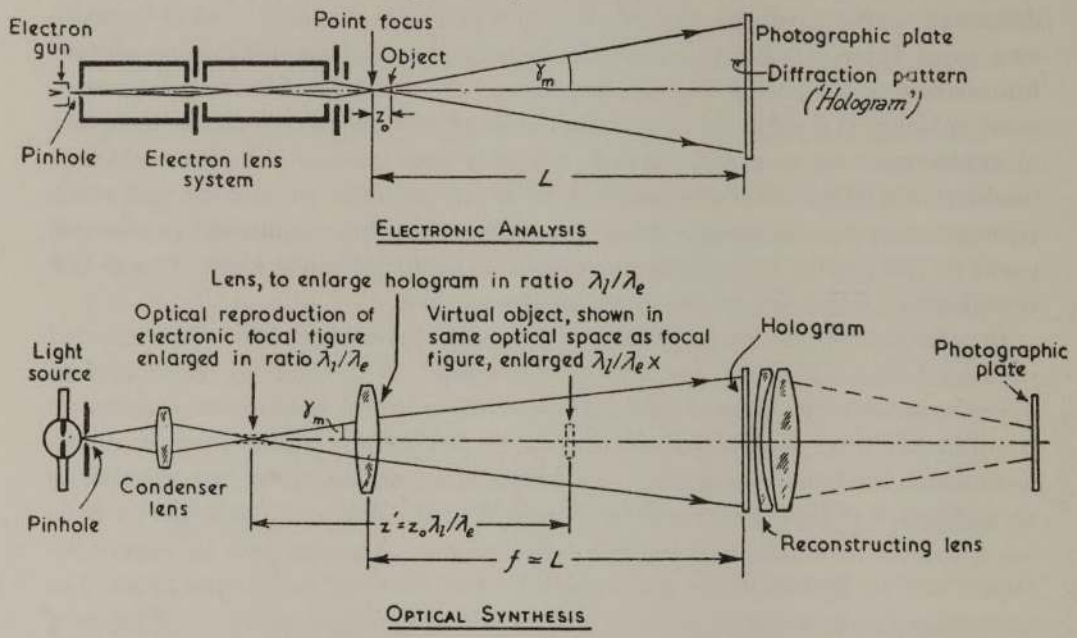


FIGURE 1. Principle of electron microscopy by reconstructed wave-fronts.

As the photograph of a diffraction pattern taken in divergent, coherent illumination will be often used in this paper, it will be useful to introduce a special name for it, to distinguish it from the diffraction pattern itself, which will be considered as a complex function. The name 'hologram' is not unjustified, as the photograph contains the total information required for reconstructing the object, which can be two-dimensional or three-dimensional.

The hologram must be either printed, or developed by reversal, and the positive is transferred to the optical synthetizer, which is a light-optical imitation of the electronic device. All essential dimensions, which determine the shape of the wave, are scaled up in the ratio of light wave-length λ_l to electron wave-length λ_e. As electrons of about 50 keV energy, with a de Broglie wave-length of about 0·05 Å, are the most useful in electron microscopy, this ratio will be of the order 100,000. It may be noted that the focal length of the electron lenses is not an essential dimension, and need not be scaled up.

To avoid scaling up the photographic plate a further lens is provided, which enlarges it in the ratio λ_l/λ_e in the optical space of the enlarged focal figure. This means that the image of the hologram is moved practically to infinity, i.e. it must be in the focal plane of the collimator lens. In the illustration it has been assumed for simplicity that the angles are the same in the analyzer and in the synthetizer, but it will be shown later that the condition $f = L$ is not essential. Nor is it necessary to use a separate condenser lens system. The condenser and the collimator, which have been shown separate in figure 1 to simplify the explanations, form one optical unit, whose function it is to produce an imitation of the original wave-front in the plane of the hologram. The spherical aberration, and the practically unavoidable ellipticity of the electron lenses must be reproduced with great accuracy, with a tolerance of about one fringe for the marginal rays.

Thus in the new method it is no longer necessary to correct the spherical aberration of electron lenses. The aperture can be opened up far beyond the limit of tolerance in ordinary electron microscopy. It is only necessary to *imitate* the aberrations to the same accuracy as they would have to be corrected to achieve a certain resolution. Thus the difficulty is shifted from electron optics to light optics, where refracting surfaces can be figured to any shape, without the limitations imposed in electron optics by the laws of the electromagnetic field. On the electron-optical side we require only a certain moderate constancy, sufficient to avoid readjustment of the optical system at too frequent intervals.

The technical difficulties of the scheme will not be dealt with in this paper. It may be only mentioned that they involve mechanical and electrical stability, operation with objects much smaller than those hitherto dealt with in electron microscopy, and the problem of obtaining the high current densities required under the additional condition of coherence. For the rest the paper will deal mainly with the general wave-theoretical foundations of the new method.

The principle of wave-front reconstruction

Consider a coherent monochromatic wave with a complex amplitude U striking a photographic plate. We write $U = Ae^{i\psi}$, where A and ψ are real. U may be decomposed into a 'background wave' $U_0 = A_0 e^{i\psi_0}$, and a remainder $U_1 = A_1 e^{i\psi_1}$ which is due to the disturbance created by the object and may be called the secondary wave. Thus the complex amplitude at the photographic plate is

$$U = U_0 + U_1 = A_0 e^{i\psi_0} + A_1 e^{i\psi_1} = e^{i\psi_0}(A_0 + A_1 e^{i(\psi_1 - \psi_0)}) \tag{1}$$

and its absolute value $A = [A_0^2 + A_1^2 + 2A_0 A_1 \cos(\psi_1 - \psi_0)]^{\frac{1}{2}}$.

The density of photographic plates, plotted against the logarithm of exposure, is an S-shaped curve, with an approximately straight branch between the two knees. In this region the transmission of intensity is a power $-\Gamma$ of the exposure. The word 'transmission' and the symbol t will be reserved in this paper for the amplitude transmission, which is in general complex; hence the intensity transmission is tt^*, where the asterisk denotes the complex conjugate. For pure absorption, without

phase change, t is a real number, the square root of the intensity transmission. Thus we write for the negative process

$$t_n = (K_n A)^{-\Gamma_n},$$

where K_n is proportional to the time of exposure. In the printing of the negative the exposure is proportional to t_n, hence the transmission of the positive print becomes

$$t_p = [K_p(K_n A)^{-\Gamma_n}]^{-\Gamma_p} = KA^{\Gamma}, \tag{2}$$

where $\Gamma = \Gamma_n \Gamma_p$ is the 'overall gamma' of the negative-positive process. The same type of law applies if reversal development is used.

If now in the reconstruction process we illuminate the positive hologram with the background U_0 alone, a 'substituted wave' U_s will be transmitted, which is, apart from a constant factor

$$U_s = U_0 t_p = A_0 e^{i\psi_0}[A_0^2 + A_1^2 + 2A_0 A_1 \cos(\psi_1 - \psi_0)]^{\frac{1}{2}\Gamma}. \tag{3}$$

The simplest, and as will be seen also the most advantageous choice, is $\Gamma = 2$, which gives

$$\begin{aligned} U_s = U_0 A^2 &= A_0 e^{i\psi_0}[A_0^2 + A_1^2 + 2A_0 A_1 \cos(\psi_1 - \psi_0)] \\ &= A_0^2 e^{i\psi_0}\left[A_0 + \frac{A_1^2}{A_0} + A_1 e^{i(\psi_1 - \psi_0)} + A_1 e^{-i(\psi_1 - \psi_0)}\right]. \end{aligned} \tag{4}$$

Comparing this with (1) one sees that if $A_0 = \text{const.}$, i.e. if the background is uniform, the substituted wave contains a component proportional to the original wave U (the first and third terms). This is not in itself a proof of the principle of reconstruction, as any wave can be split into a given wave and a rest. It remains to be shown that the remainder, i.e. the spurious part of U_s, does not constitute a serious disturbance.

This remainder consists of two terms. One of these has the same phase as the background, with an amplitude $(A_1/A_0)^2$ times the amplitude of the background. This term can be made very small if the background is relatively strong, which does not mean that the contrast in the hologram must be poor. Assume for instance $(A_1/A_0)^2 = 0{\cdot}01$, i.e. a secondary intensity which is only 1% of the primary. This gives $A_1/A_0 = 0{\cdot}1$, and the intensity ratio between the maxima and minima of the interference fringes is $(1{\cdot}1/0{\cdot}9)^2 = 1{\cdot}5$. With $\Gamma = 2$ the intensity transmissions will be in the ratio $1{\cdot}5^2 = 2{\cdot}25$, a very strong contrast. The contrast will fall below the observable limit of about 4% only for $(A_1/A_0)^2 \leqslant 0{\cdot}0001$, i.e. if the flux scattered by the object into the area of the diagram is less than 0·01% of the illuminating flux. This remarkable effect of the coherent background has been systematically utilized by Zernike (1948) for the amplified display of weak interference fringes.

The second term of the remainder has the same amplitude $A_1 A_0^2$ as the reconstruction of the original secondary wave, but it has a phase shift of opposite sign relative to the background. It may be called for brevity the 'conjugate-complex' wave. The two twin waves carry the same energy.

The conjugate wave produces a serious disturbance only in rather exceptional arrangements; in most cases the twin waves can be effectively separated. To make this plausible one may think of Fresnel-zone plates. These can be, in fact, considered

as holograms of a point object, produced by a point source at infinity. Zone plates act simultaneously as positive and negative lenses, producing two focal points, one at each side of the plate, at equal distance, which can be separately observed. As will be shown later, homocentric illumination produces always such twin images, only, with the source at finite distance, these will be in mirror-symmetrical position with respect to the point source, not to the hologram. In beams which are only approximately homocentric the second image is no longer sharp, but effective separation can be always achieved if the object is sufficiently small, and if certain positions are avoided.

While the twin wave cannot be avoided, the spurious term which is proportional to $(A_1/A_0)^2$ and the distortion due to an uneven background can both be eliminated, or at least effectively suppressed by a modification of the photographic process. In the case of small objects, at least over a large part of the hologram, the photographic density difference between two neighbouring interference maxima is insignificant. This makes it possible to wash out the interference fringes by taking a slightly defocused print of the positive hologram, and processing this print with $\Gamma = 1$. If this print, which has a transmission inversely proportional to $A_0^2 + A_1^2$, is placed in register with the positive, and illuminated by the background wave U_0, the substituted wave becomes

$$U_s' = A_0 e^{i\psi_0}[A_0^2 + A_1^2 + 2A_0 A_1 \cos(\psi_1 - \psi_0)]/(A_0^2 + A_1^2) = e^{i\psi_0}\left[A_0 + \frac{2A_1 \cos(\psi_1 - \psi_0)}{1 + (A_1/A_0)^2}\right]$$

$$= e^{i\psi_0}\left[A_0 + 2A_1 \cos(\psi_1 - \psi_0) - 2\frac{A_1^3}{A_0^2}\cos(\psi_1 - \psi_0) + \ldots\right], \quad (5)$$

in which the spurious term is of the order $(A_1/A_0)^3$ as compared with the background, and the distortion due to a non-uniform background is eliminated. If one only wants to eliminate the background by itself, one can also use a negative photograph taken in the illuminating beam without the object, processed with $\Gamma = 2$.

To discuss briefly also the case $\Gamma \neq 2$, we put for simplicity $A_0 = 1$ and $A_1/A_0 = a$, and obtain from (3) by binomial expansion

$$U_s = e^{i\psi_0}[1 + \tfrac{1}{2}\Gamma a^2 + \Gamma a \cos(\psi_1 - \psi_0) + \tfrac{1}{2}\Gamma(\Gamma - 2)\, a^2 \cos^2(\psi_1 - \psi_0) + \ldots]. \quad (6)$$

In the reconstructed wave the contrast is enhanced in the ratio $\frac{1}{2}\Gamma$. But, in addition, one obtains twin waves with phase shifts $2(\psi_1 - \psi_0)$, etc., but with smaller amplitudes. This makes it evident that $\Gamma = 2$ is the best choice, except if the original contrast is so weak that it must be enhanced, even at the cost of faithfulness in the reproduction.

Illumination by a homocentric wave

In order to study the reconstruction cycle in more detail, it will be advantageous to start with the simple case of homocentric illumination, which can be approximately realized by a sufficiently small pinhole as light source. It will be convenient to restrict the discussion for a start to two-dimensional objects, occupying a part of some closed surface Σ which encloses the point source O. The object at a point P of Σ may be characterized by an amplitude transmission coefficient $t(P)$, which is the

ratio of the complex amplitudes at the two sides of Σ, in proximity of the point P. t is in general a complex datum, real only for purely absorbing objects. It is, of course, understood that the concept of a transmission coefficient, real or complex, is not applicable to an object which is two-dimensional in the mathematical sense. Of a physical object to which this concept is applicable we must assume that it is at least several wave-lengths in thickness. Moreover, we must assume that laterally, in the surface Σ, the function $t(P)$ does not vary appreciably within a wave-length. These are the conditions for the applicability of the Fresnel-Kirchhoff theory of diffraction. In electron optics, operating with fast electrons of about 0·05 Å wave-length, this condition is always satisfied, as there exists no material object (except nuclei) whose physical properties change significantly in less than about ten times this wave-length.

With these qualifications we can apply the Fresnel-Kirchhoff diffraction formula (cf., for example, Baker & Copson 1939, p. 73). The notations are explained in figure 2. If the monochromatic source at O is of unit strength, the amplitude in the illuminating wave is

$$U_0 = \frac{1}{r_0} e^{ikr_0},$$

where r_0 is the distance measured from O, and $k = 2\pi/\lambda$. The presence of an object in the surface Σ modifies the amplitude at a point Q outside to

$$U(Q) = \frac{1}{2\lambda}\int_{\Sigma} t(P)\, e^{ik(r_0+r_1)-\frac{1}{2}\pi i}(\cos\theta_0 - \cos\theta_1)\frac{dS}{r_0 r_1}. \tag{7}$$

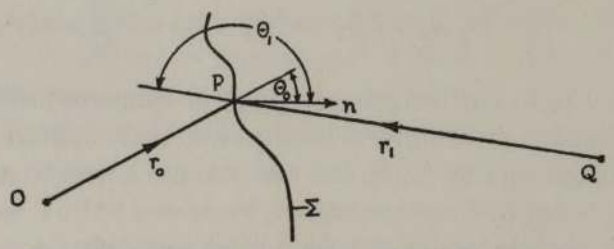


FIGURE 2. Fresnel-Kirchhoff diffraction formula.

We will now apply this formula to calculate the 'physical shadow' of a plane object at infinity. The 'physical shadow' includes the diffraction effects, and is to be distinguished from the 'geometrical shadow' into which it merges at vanishing wave-length.

As the beams to be used in practice will have semi-cone angles of 0·05 or less, we can put $\cos\theta_0 = -\cos\theta = 1$, and consider the factor $1/r_0 r_1$ as constant. We also drop the constant factor $(1/2\lambda)\, e^{-\frac{1}{2}\pi i}$, and use equation (7) in the simplified form

$$U(Q) = \int_{\Sigma} t(P)\, e^{ik(r_0+r_1)}\, dS. \tag{7·1}$$

Using the notations explained in figure 3, the distance r_0 of a point P in the object plane $z = z_0$ is

$$r_0 = (x^2+y^2+z_0^2)^{\frac{1}{2}} = z_0 + \tfrac{1}{2}(x^2+y^2)/z_0 - \tfrac{1}{8}(x^2+y^2)^2/z_0^3 + \ldots.$$

In this section we will use only the first two terms of the expansion.

The observation point Q may be at a distance L in the Z-direction, very large compared with z_0, practically at infinity, so that we can write

$$r_1 = L\sec\gamma - (x\cos\alpha + y\cos\beta).$$

The first terms in the expression for r_0 and r_1 give constant phase factors, independent of x, y, which may be dropped. The remaining essential part may be termed 'the amplitude in the direction ξ, η', and is

$$U(\alpha,\beta) = \iint t(x,y)\exp\{ik[(x^2+y^2)/2z_0 - (x\cos\alpha + y\cos\beta)]\}\,dx\,dy. \tag{8}$$

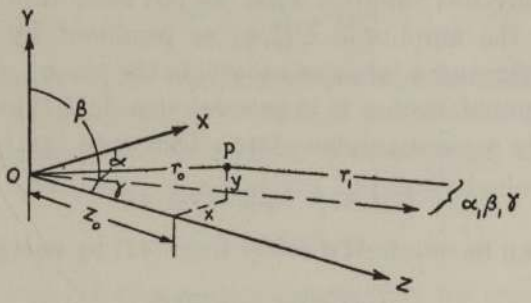


FIGURE 3. Explanation of symbols.

Unless the limits are indicated, integrations in this paper will be always understood to be carried out between infinite limits. As the phase under the integral is valid only for small angles, equation (8) is physically valid only if $t(x, y)$ vanishes rapidly outside a small central area.

It will now be convenient to introduce 'Fourier variables' ξ, η instead of the direction cosines by

$$\xi = \frac{1}{\lambda}\cos\alpha, \quad \eta = \frac{1}{\lambda}\cos\beta. \tag{9}$$

Their connexion with the co-ordinates X, Y in a plane at the large distance $z = L$ is given by

$$X = L\frac{\cos\alpha}{\cos\gamma} = \frac{\lambda L\xi}{[1-\lambda^2(\xi^2+\eta^2)]^{\frac{1}{2}}}, \quad Y = L\frac{\cos\beta}{\cos\gamma} = \frac{\lambda L\eta}{[1-\lambda^2(\xi^2+\eta^2)]^{\frac{1}{2}}}.$$

If the illuminating cone is narrow enough, ξ and η can be taken to represent the co-ordinates in the physical shadow. The geometrical shadow of a point x, y has the Fourier co-ordinates $\xi = x/\lambda z_0, \eta = y/\lambda z_0$. The quantity

$$\mu = \lambda z_0 \tag{10}$$

is the only parameter of the diffraction problem. Its square root can be considered as the characteristic length. Details coarser than $\mu^{\frac{1}{2}}$ will be shown to have shadows more or less similar to themselves, finer details lose all likeness by diffraction.

Using the notations (9) and (10) equation (8) can be written, with the abbreviation $x^2 + y^2 = r^2$,

$$U(\xi,\eta) = \iint [t(x,y)\,e^{\pi i r^2/\mu}]\,e^{-2\pi i(x\xi+y\eta)}\,dx\,dy. \tag{11}$$

Thus the amplitude in the ξ, η direction is the Fourier transform of the function

$$t(x, y)\, e^{i\pi r^2/\mu}$$

in the standard notation of Campbell & Foster (1931). We can at once write down the reciprocal formula

$$t(x, y) = e^{-\pi i r^2/\mu} \iint U(\xi, \eta)\, e^{+2\pi i(x\xi + y\eta)} d\xi\, d\eta. \tag{12}$$

It will be useful to study these reciprocal transformations mathematically, while provisionally disregarding the conditions which must be imposed on the function $t(x, y)$ to give them physical validity. First we put them into a more symmetrical form, by imagining the amplitude $U(\xi, \eta)$ as produced by the passage of the illuminating wave U_0 through a 'shadow object' in the plane ξ, η with a transmission $\tau(\xi, \eta)$. (It may be noted that τ is in general complex; thus the shadow object cannot be replaced by a photographic plate.) That is to say, we put

$$U(\xi, \eta) = U_0(\xi, \eta)\, \tau(\xi, \eta). \tag{13}$$

The background U_0 can be obtained directly from (11) by putting $t = 1$

$$U_0(\xi, \eta) = i\mu\, e^{-\pi i \mu \rho^2}$$

with the abbreviation $\xi^2 + \eta^2 = \rho^2$. We now obtain the symmetrical transformation formulas

$$\tau(\xi, \eta) = \frac{1}{i\mu} e^{\pi i \mu \rho^2} \iint t(x, y)\, e^{-\pi r^2/i\mu}\, e^{-2\pi i(x\xi + y\eta)} dx\, dy, \tag{14}$$

$$t(x, y) = i\mu\, e^{\pi r^2/i\mu} \iint \tau(\xi, \eta)\, e^{-\pi i \mu \rho^2}\, e^{+2\pi i(x\xi + y\eta)} d\xi\, d\eta. \tag{15}$$

These may be called the 'shadow transformations', and $t(x, y)$, $\tau(\xi, \eta)$ a pair of 'shadow transforms'. They are, of course, intimately related to Fourier transforms, though simpler in some respects.

The transformations (14) and (15) can be derived from one another by the rule: Interchange t and τ, x and ξ, y and η, i.e. interchange Latin and Greek symbols, and replace i by $-i$, μ by $1/\mu$. Two transformations in succession restore the original. Physically this means that if instead of a photograph we could produce a 'shadow object' with the absorbing and refractive properties of $\tau(\xi, \eta)$, and illuminated this with the background, we should exactly restore the object $t(x, y)$ in its original position. As a photograph cannot imitate the imaginary part of τ, a certain residual wave arises, to which we will return in the next section. But it will be useful to consider first a few examples of shadow transforms.

As in the case of Fourier integrals, the transforms of exponentials of quadratic forms are particularly simple and instructive. It is convenient to write these in the form

$$t(x, y) = \exp[-\pi(A_1 x^2 + 2B_1 x + A_2 y^2 + 2B_2 y)].$$

This is the product of an x- and a y-factor, and as the transform is again the product of a ξ- and an η-factor, it is sufficient to give the transform of

$$t(x) = e^{-\pi(Ax^2 + 2Bx)}, \tag{16·1}$$

which is $$\tau(\xi) = (1+i\mu A)^{-\frac{1}{2}} \exp\left[-\frac{\pi\mu(\mu A\xi^2 + 2B\xi - iB^2)}{1+i\mu A}\right]. \tag{16·2}$$

Thus the shadow transform of an exponential of a quadratic form is a function of the same type, as in the case of Fourier transforms, but the relation between the parameters is of a different build. For example, if $A = B = 0$, which makes t a constant, τ will be the same constant, while the Fourier transform of a constant is a delta function, which vanishes everywhere except at the argument zero. Moreover, the shadow transform of a harmonic function ($A = 0$)

$$t(x) = \mathrm{e}^{2\pi i x/p} \tag{16·3}$$

is again a harmonic function $$\tau(\xi) = \mathrm{e}^{-i\pi\mu/p^2}\mathrm{e}^{2\pi i\mu\xi/p}. \tag{16·4}$$

The period in the shadow is p/μ, which is *the geometrical shadow of the period* p. The only difference is in the phase factor $\mathrm{e}^{-i\pi\mu/p^2}$. If the period p is long compared with the characteristic length $\mu^{\frac{1}{2}}$, the phase factor tends to unity, which means that if the object contains no details finer than $\mu^{\frac{1}{2}}$ the physical shadow tends towards the geometrical shadow

$$\tau(\xi,\eta) \to t(\mu\xi, \mu\eta).$$

Equations (16·3) and (16·4) contain a simple rule for constructing the shadow transform of an object, by expanding $t(x,y)$ into a Fourier integral with periods p_x, p_y. In the transform the Fourier coefficients will differ from the original only in a phase factor

$$\exp\left[-i\pi\mu\left(\frac{1}{p_x^2}+\frac{1}{p_y^2}\right)\right].$$

As a practical method this may be used with the cautioning remark that infinite trains of periodic functions are not very suitable for the description of small objects, and that the applicability of equations (14) and (15) to the physical process is strictly speaking limited to objects which transmit appreciably only in a region $x/z_0, y/z_0 \ll 1$.

Reconstruction with homocentric illumination

Stigmatic illumination is a particularly simple and instructive illustration of the principle of reconstruction which was broadly explained in the first section. It may be recalled that if the hologram is replaced in the original position and illuminated by the background alone, one obtains in addition to the illuminating or primary wave two other waves, one of which is proportional to the original secondary wave emitted by the object, and the other differs from this only by having phase shifts of opposite sign relative to the background. The other small spurious terms may be disregarded for the moment.

It will now be convenient to subtract the background, i.e. the primary wave, both in the object plane, and in the plane of the photographic plate, and to consider instead of t and τ the functions

$$t_1 = t - 1 \quad \text{and} \quad \tau_1 = \tau - 1. \tag{17}$$

As $t = 1$ corresponds to $\tau = 1$, the functions $t_1(x,y)$ and $\tau_1(\xi,\eta)$ are connected by the relations (14) and (15), the same which connect t and τ. We will talk of t_1 as 'the object proper' and of τ_1 as its shadow.

By equation (6), substituting a photographic plate for the physical shadow means replacing τ_1 by

$$\tfrac{1}{2}\Gamma(\tau_1+\tau_1^*).$$

Substituting this into the inverse shadow transformation (15), we obtain two terms t_1. The first of these differs from the original object proper only in the factor $\frac{1}{2}\Gamma$. But in the second term, derived from $\frac{1}{2}\Gamma\tau^*$, the sign of i has been reversed, and this results in a spurious figure in the object plane, superimposed on the correct reconstruction of the object.

We can give a simple interpretation to the wave corresponding to τ^* if we observe that in the equation (14) applied to the object proper t_1

$$\tau_1(\xi,\eta)=\frac{1}{i\mu}e^{i\pi\mu\rho^2}\iint t_1(x,y)\,e^{-\pi r^2/i\mu}\,e^{-2\pi i(x\xi+y\eta)}dx\,dy, \tag{14·1}$$

reversing the sign of i is equivalent to reversing the sign of x, y and of $\mu=\lambda z_0$ and replacing $t_1(x,y)$ by a function

$$t_1(x,y)=t\ (-x,-y). \tag{18}$$

The transformation has now a parameter $-\mu$ instead of μ, i.e. it corresponds to an object in the plane $-z_0$ instead of in $+z_0$. By equation (18) this object arises from the original by mirroring it on the Z-axis, and changing phase delays into phase advances. Summing up, *the 'twin' wave τ^* corresponds to an apparent 'twin object', in central symmetrical position with respect to the point focus* 0, *and with opposite phase-shifting properties.*

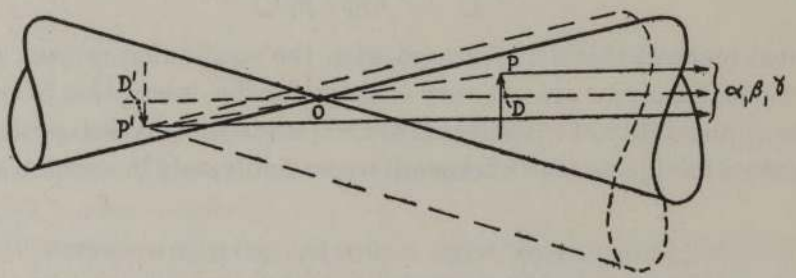


FIGURE 4. The twin images arising in the reconstruction.

Figure 4 is an illustration of the twin objects, from which one can verify this conclusion. The Fresnel-Kirchhoff formula can be interpreted as the sum of elementary spherical waves, originating from the object points P, with amplitudes proportional to $t(P)$. At infinity, in a direction α, β these are plane waves, and their phase difference relative to the background is given by the difference between the ray $\bar{O}\bar{P}$, and its projection on the direct ray, $\bar{O}\bar{D}$, apart from the phase shift which arises in the object. Figure 4 makes it clear that the same phase difference, but with opposite sign, would be produced by an object point P', in central symmetrical position to P, if the sign of the phase shift at P' is also reversed.

The interpretation of the residual wave in the reconstruction as a wave emitted by a twin object makes it at once clear that conditions can be found which allow a fairly effective isolation of the reconstructed object, by making use of the limited focal depth of the viewing system. Separation becomes possible if the distance

between the twin objects, $2z_0$, exceeds the focal depth D_f, which can be defined as the resolution limit d, divided by the total cone angle utilized in the image formation, $2\gamma_m$. Using Abbe's value d_A for d, the criterion of separation is

$$2z_0 > \frac{d}{2\gamma_m} = \frac{\lambda}{4\gamma_m^2}. \tag{19}$$

If the point focus is not produced by a physical aperture, but by the image of an aperture, formed by an optical system, this is equivalent to the condition that the object must be outside the diffraction region, in which the wave cannot be considered as homocentric.

Outside the focal diffraction region separation is possible, but not complete separation. The twin images will always interfere with one another to some extent, and the interference cannot be regulated at will. This follows from the structure of the transformation equations, which contain only one characteristic length $(\lambda z_0)^{\frac{1}{2}}$, and there is no second length with which to form a dimensionless separation factor. Thus the spurious part of the reconstructed image depends only on the object itself, and on the parameter μ. This disturbance will now be investigated in some detail.

The spurious part of the reconstruction in homocentric illumination

The simplicity of the transforms (16·1) and (16·2) suggests building up arbitrary plane objects from 'probability spots'. In the limit these tend to two-dimensional delta functions,which can represent any function $t_1(x, y)$, but it is not necessary, nor is it physically justifiable, to pass to this limit. Optical imagery does not operate with points, but with elementary regions of the size of the resolution limit. Inside such a small area the values of $t_r(x, y)$, which describe the reconstituted object, are not independent of one another.

First we carry out the reconstruction cycle for a single probability spot. Assume the transmission in the object plane in the form

$$t(x, y) = 1 - A \exp\left\{-\frac{\pi}{a^2}[(x-x_0)^2 + (y-y_0)^2]\right\} = 1 - A\mathrm{e}^{-\pi(r'/a)^2}, \tag{20}$$

where the abbreviation r' has been used for the distance measured from the centre x_0, y_0 of the spot. $(1-A)$ is the amplitude transmitted at the centre of the spot, at unit background. If the object is a pure absorber, A is real, positive, and less than unity. If the object has pure-phase contrast $|1-A| = 1$, and $|A|$ is in the limits 0 to 2.

Equations (16·1) and (16·2) give for the physical shadow of (20)

$$\tau(\xi, \eta) = 1 + \frac{i\epsilon A}{1-i\epsilon} \exp\left(\frac{\pi i \mu \rho'^2}{1-i\epsilon}\right) \tag{21}$$

with the abbreviations $\epsilon = a^2/\mu = a^2/\lambda z_0$

and $\rho'^2 = (\xi - x_0/\mu)^2 + (\eta - y_0/\mu)^2 = (\xi - \xi_0)^2 + (\eta - \eta_0)^2,$

where ξ_0, η_0 is the geometrical shadow of x_0, y_0. The diffraction figure (21) centres around this point. Its character is determined by the dimensionless parameter ϵ. If ϵ is large τ approaches the geometrical shadow $t(\mu\xi, \mu\eta)$. The more important case is $\epsilon \ll 1$, which allows simplifying (21) to

$$\tau(\xi, \eta) = 1 + i\epsilon A\, \mathrm{e}^{-\pi(a\rho')^2 + \pi i \mu \rho'^2}. \tag{21·1}$$

The smaller the original spot, the larger its physical shadow.

The photograph substitutes for the complex physical shadow (21·1) the real transmission function

$$\tau_s(\xi, \eta) = |\tau|^{\Gamma} \doteqdot 1 + \tfrac{1}{2}\Gamma i\epsilon A\, \mathrm{e}^{-\pi(a\rho')^2} \mathrm{e}^{\pi i \mu \rho'^2} - \tfrac{1}{2}\Gamma i\epsilon A^*\, \mathrm{e}^{-\pi i \mu \rho'^2}. \tag{22}$$

This approximation is valid if $\epsilon^2 \ll 1$.

The inverse transformation (15), applied to the first two terms of (22), restores the original object (20), but the contrast is $\frac{1}{2}\Gamma$ times the original. The same transformation applied to the last term of (22) gives the spurious or error term

$$t_e(x, y) = -\tfrac{1}{4}\Gamma i\epsilon A^*\, \mathrm{e}^{-\pi(ar'/2\mu)^2} \mathrm{e}^{-\pi i (r'^2/2\mu)}. \tag{23}$$

This is the amplitude (at unit background) which the twin image produces in the original object plane. The spurious image centres on x_0, y_0, but it has a character quite different from the original. The amplitude t_e falls off only slowly with the distance r' from the centre, the slower the smaller the original spot radius a, while the phase changes rapidly, according to the last factor in (23), in a manner independent of the spot size. Thus the spurious image will manifest itself in a system of fine and weak interference fringes, superimposed on the true reconstruction.

The exact value of the reconstructed transmission function t_r in the case $\Gamma = 2$ may be also given for reference:

$$t_r(x, y) = 1 - A \exp\left[-\pi\left(\frac{r'}{a}\right)^2\right] - \frac{i\epsilon A^*}{2 + i\epsilon} \exp\left[-\frac{\pi(\epsilon + 2i)}{\mu(4 + \epsilon^2)} r'^2\right] + \frac{\epsilon^2 A A^*}{1 + \epsilon^2 - 2i\epsilon} \exp\left[-\frac{2\pi\epsilon(1 + \epsilon^2 + 2i\epsilon)}{\mu[(1 + \epsilon^2)^2 + 4\epsilon^2]} r'^2\right]. \tag{22·1}$$

The first two terms stand for the exact reconstruction, the last two for the spurious amplitude. They differ from (23) only in terms of the order ϵ^2 or higher.

The reconstruction cycle in the case $\Gamma = 2$ is illustrated in figure 5 for a probability spot with a black centre. One must be careful not to go beyond $\Gamma = 2$ if there are sharp contrasts in the object. As shown in figure 6, a lighter centre will appear inside a black ring, and black lines will appear doubled.

Up to this point we have assumed unlimited apertures; consequently there was no lower limit to the spot size a which could be correctly reproduced. The effects of limited resolution can be very simply discussed by assuming a mask used in the taking of the hologram, with an amplitude transmission

$$\mathrm{e}^{-\pi(c\rho)^2}.$$

Such graded masks are preferable to sharp apertures, not only from the point of view of mathematical simplicity, but also because they reduce the 'false detail' resulting

from the cut-off to a minimum. Their use is well known in structure analysis (Bunn 1945, p. 350).

As the mask is used twice, in the taking of the photograph and in the reconstruction, its total effect is

$$e^{-\pi(\Gamma+1)(c\rho)^2}.$$

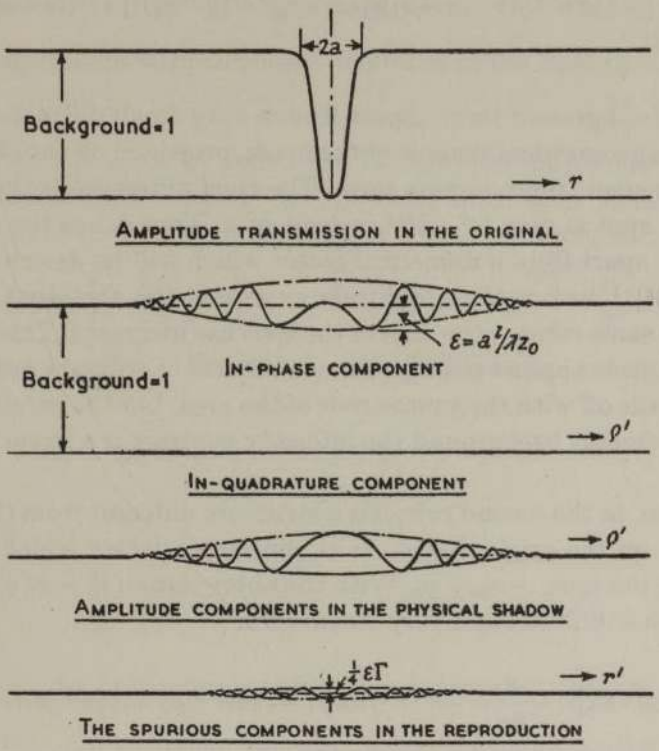


FIGURE 5. Reproduction cycle of a 'probability spot'.

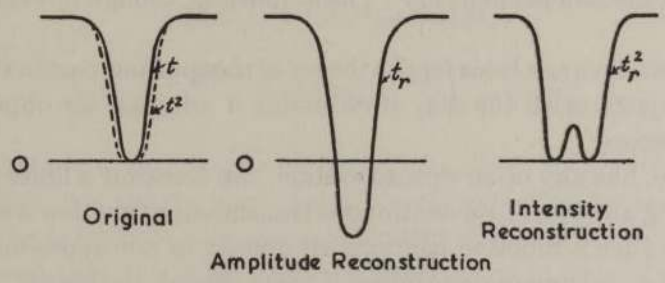


FIGURE 6. Distortion by exaggerated contrast.

We now assume $\Gamma = 2$, so as to obtain correct contrast reconstruction, and we put $3c^2 = b^2$. We have now to reconstruct the object, the probability spot (20), from the physical shadow

$$\tau_s = e^{-\pi(b\rho)^2}\{1 + i\epsilon A\, e^{-\pi(a^2-i\mu)\rho'^2} - i\epsilon A^* e^{-\pi(a^2+i\mu)\rho'^2}\}, \qquad (24)$$

which differs from (22) only in the masking factor. Introducing the small dimensionless parameter

$$\sigma = b^2/\mu = b^2/\lambda z_0,$$

one obtains by the transformation (15), neglecting powers of ϵ and σ higher than the first,

$$t_r(x,y) = (1+i\sigma)\exp\left[-\frac{\pi\sigma}{\mu}(1+i\sigma)r^2\right] - \frac{A}{1+(b/a)^2}\exp\left\{-\frac{\pi}{a^2+b^2}[r'^2+\epsilon\sigma(x_0^2+y_0^2)]\right\}$$
$$-\tfrac{1}{2}i\epsilon A^*\exp\{-(\tfrac{1}{2}\pi i r'^2/\mu)-(\pi/4\mu)[\epsilon[(x-x_0)^2+(y-y_0)^2]+\sigma[(x+x_0)^2+(y+y_0)^2]]\}. \quad (25)$$

The first is the background term. Apart from a very small diffraction effect, of the order σ^2, it is the geometrical shadow of the mask, projected on the object plane. The second is the 'correct' reproduction term. The chief difference is that the spread of the reproduced spot is now $(a^2+b^2)^{\frac{1}{2}}$ instead of a. Thus b has the meaning of the *resolution limit*, apart from a numerical factor which will be determined later. The factor $[1+(b/a)^2]^{-1}$ before the amplitude expresses the fact that the amplitude decreases in the same ratio as the area of the spot has increased. This loss of contrast for very small objects appears stronger than in the case of ordinary microscopy, where the amplitude falls off with the square root of the area, but the result is the same, as with a strong coherent background the intensity contrast is a linear function of the amplitude.

The error term, in the second row, has a structure different from (23); it no longer centres exactly on the original spot, as it contains a factor which centres on the mirror image of the spot, $-x_0, -y_0$. With the abbreviation $r_0^2 = x_0^2+y_0^2$ we can write the error term in a different and very useful form

$$t_e = -\tfrac{1}{2}i\epsilon A^*\exp\left[-\frac{\pi}{4\mu}(\epsilon-\sigma)r'^2\right]\exp\left[-(\tfrac{1}{2}\pi i r'^2/\mu)-(\tfrac{1}{2}\pi\sigma/\mu)(r^2+r_0^2)\right]. \quad (25{\cdot}1)$$

This is particularly useful in the case $\epsilon = \sigma$, i.e. $a = b$, as in this case the fringe system t_e has an amplitude independent of r'. The amplitude, though not the phase, centres on $x, y = 0$.

This result can serve as a basis for the theory of the spurious part in the reproduction of arbitrary objects, with the aim of obtaining a criterion for objects suitable for two-step microscopy.

A microscope, like any other optical system, can transmit a finite number of data only. Describing an object by a continuous transmission function is an objectionable idealization, as such a function contains an infinity of non-reproducible detail. We come nearer to an adequate description if we divide up the object by a network of lines into cells of the size of the resolution limit, associate a complex datum with each cell, and investigate the transmission of these data through the optical system.

Equation (25·1) suggests that particularly simple results will be obtained if we represent the object by a two-dimensional lattice of probability spots with a spread $a = b$. As illustrated in figure 7, we arrange these spots in a hexagonal lattice, with a distance d between adjacent centres. d is the resolution limit, which we define in a way slightly different from the usual, by postulating that three (instead of two) equal probability spots with spreads $a = b$ can just be resolved if their centres are at distances d from one another, i.e. the minimum in the centre just vanishes. By

equation (25) the amplitude in the correct reproduction term follows a law $e^{-\frac{1}{2}\pi(r'/b)^2}$ if $a = b$. In the middle between three centres $r' = d/\sqrt{3}$; thus the condition for d is

$$\exp\left[-\frac{\pi}{2}\left(\frac{d}{\sqrt{3}b}\right)^2\right] = \frac{1}{3},$$

which gives $$d = 1{\cdot}45b.$$

This is in good agreement with the usual definition of the resolution limit,

$$d = 0{\cdot}6\lambda/\sin\gamma_m,$$

if we define γ_m as the angle at which the background amplitude has dropped to $1/\sqrt{3}$, i.e. the background intensity to 1/3 of its maximum value. Denoting the corresponding radius in the object plane by $R = z\sin\gamma_m$, we have

$$\exp\left[-\frac{2\pi\sigma}{\mu}R^2\right] = \exp\left[-2\pi\left(\frac{b}{\lambda}\sin\gamma_m\right)^2\right] = \frac{1}{3},$$

which gives $$b = 0{\cdot}42\lambda/\sin\gamma_m, \quad d = 0{\cdot}61\lambda/\sin\gamma_m.$$

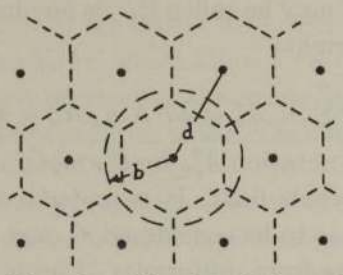


FIGURE 7. Dividing the object into independent elements.

Call N the number of independent elements inside the illuminated field, i.e. number of cells contained in the disk of radius R. Each cell occupies an area

$$\tfrac{1}{4}\sqrt{3}d^2 = 0{\cdot}433d^2 = 0{\cdot}91b^2;$$

thus the number N is

$$N = \frac{\pi}{0{\cdot}91}\frac{R^2}{b} = 3{\cdot}45(0{\cdot}42)^2\left(\frac{\lambda z_0}{b^2}\right)^2 = 0{\cdot}61\,\frac{1}{\sigma^2}. \tag{26}$$

N can be easily made a very large number, of the order 10^6 to 10^8. This suggests a *statistical evaluation* of the spurious part of the reproduction, by assuming random distribution of the amplitude over the independent elements of the object. It is, of course, understood that this might lead to gross errors in special cases, but it is certainly an acceptable assumption if a great number and variety of objects are considered.

Number the elements from 1 to N. The spurious amplitude in the reconstruction at a point x, y results from the superposition of the error terms of the form (25·1), one for each cell with centres x_n, y_n. The distance of x, y measured from x_n, y_n may be called r'_n. With the simplification resulting from $\sigma = \epsilon$ the resulting error amplitude is

$$t_e(x,y) = \tfrac{1}{2}i\sigma\, e^{-\pi\sigma r^2/2\mu}\sum^{N} A_n^* e^{-\frac{1}{2}\pi\sigma r_n^2/\mu} e^{-\frac{1}{2}\pi i r_n^2/\mu}, \tag{27}$$

where we have written r_n for the distance of x_n, y_n from 0, 0.

The relative distance r'_n between x, y and x_n, y_n occurs here only in the phase factor $e^{-\frac{1}{2}\pi i r'_n/\mu}$. The two probability decay factors fall off slowly. The first of these, before the sum, is the square root of the background attenuation, i.e. it falls off at half the rate of the background amplitude. The second factor centres on 0, 0 and falls off at the same rate as the first factor. Thus it is admissible to put both equal to unity as an approximation, and simplify equation (27) to

$$t_e(x, y) = \tfrac{1}{2}i\sigma \sum^{N} A^*_n e^{-\frac{1}{2}\pi i r'^2_n/\sigma}, \tag{27·1}$$

that is to say, in order to obtain the error amplitude at x, y we have to superimpose at this point a large number N of *undamped* waves, with wave-lengths $2\mu/r'$, emanating from all image points x_n, y_n. This wave-length is always longer than the resolution limit d. Its smallest value is at $r' = 2R$ and is μ/R, while the resolution limit is $0{\cdot}61\mu/R$.

Introduce now the hypothesis that there is no correlation between the phases of these waves. With this assumption the mean square of the component of t_e in phase with the background, which may be called $t^2_{\text{eff.}}$, is one-half of the sum of the absolute squares of the terms at the right:

$$t^2_{\text{eff.}} = \tfrac{1}{8}\sigma^2 \sum^{N} A_n A^*_n = \tfrac{1}{8}\sigma^2 N \overline{A_n A^*_n} = \tfrac{1}{8}\sigma^2 N A^2_{\text{eff.}}. \tag{28}$$

Here we have introduced the notation $A^2_{\text{eff.}}$ for the mean square secondary amplitudes, $\overline{A_n A^*_n}$, averaged over the whole field. It is understood that the average level of transmission of the object has to be considered as part of the background, and $A_{\text{eff.}}$ is a measure of the departure from uniformity. Combining (28) with equation (26), $N\sigma^2 = 0{\cdot}61$, we obtain

$$t_{\text{eff.}} = 0{\cdot}28 A_{\text{eff.}}. \tag{29}$$

Equation (29) enables us to formulate a criterion for suitable objects. A background can be considered as practically even if the intensity contrast does not exceed about 5 %, i.e. if the amplitude contrast is less than 2·5 %. This means that for suitable objects we must have

$$A_{\text{eff.}} \leqslant 0{\cdot}1 \tag{30}$$

averaged over the whole field. As an example consider a black-and-white object in which the black part, where $A = 1$, covers a fraction κ of the illuminated field, while for the rest $A = 0$. In this case $A_{\text{eff.}} = \sqrt{\kappa}$, and we obtain the simple rule that not more than about 1 % of the illuminated field should be covered with black dots or lines. If, for instance, the object is a disk, half black and half white, its diameter should not exceed one-seventh of the field diameter.

As a second example consider an object with pure phase contrast, but with random distribution of phase delays. We must qualify this by the condition which precedes every application of the Fresnel-Kirchhoff theory; the phase must not vary appreciably between points spaced at less than a wave-length. In other words, the object must appear even and transparent if it is sharply focused. In ordinary microscopy a crinkled sheet of celluloid, or even reticulated gelatine, will satisfy this condition, but not an opal glass with colloidal dispersion. With this qualification in mind we can apply equation (29) and it can be shown that the value of $A_{\text{eff.}}$ is again unity. In the case of pure phase contrast the complex transmission vector $t = 1 - A$ moves on the

unit circle, all orientations of t are equally probable. Hence $\bar{t} = 0$, which makes $\bar{A} = 1$, and

$$A^2_{\text{eff.}} = \overline{|A-\bar{A}|^2} = \overline{|A-1|^2} = |t|^2 = 1.$$

This means that an object of this type, if it covers the whole field, produces $t_{\text{eff.}} = 0{\cdot}28$, a very serious disturbance. This result is of interest, because it shows that an irregular transparent support for the object, even if it would be invisible in ordinary microscopy, will make all but the most contrasty or regular features of the object invisible. As it is rather doubtful whether an 'optically flat' or at least acceptably regular supporting membrane can be found in electron microscopy, it appears preferable to use supporting membranes only in a small fraction of the field, or to dispense with them altogether.

Improving the separation by masking and other means

These results lead to the conclusion that high-grade purity in the reproduction cannot easily be achieved even with very small objects, as the spurious intensity is proportional to the square root of the object area. But in the case of small objects special techniques become available, which allow a very effective elimination of the spurious amplitudes. The first of these is the masking of the geometrical shadow in the hologram. The second technique is the masking of the background in the reconstruction process.

The spurious amplitude is objectionable only in the area occupied by the true image. Thus we need eliminate only those rays issuing from the twin object which pass through the object area. As may be seen from figure 4, if the object is small these rays will have substantially the same direction as the primary rays which illuminate the object. This means that we can substantially reduce the spurious amplitude if we mask out the geometrical shadow in the hologram.

This masking process, however, will introduce two new disturbances. First, the mask itself will produce a system of interference fringes. This effect can be reduced to a very low level if a 'probability mask' is used. Secondly, the mask will eliminate some of the data required for a complete reconstruction. Evidently the coarser detail will suffer most, as this is contained in or near to the geometrical shadow area in the hologram, while the finer detail is spread over a larger area outside. But if the object is of the order of the characteristic length $\mu^{\frac{1}{2}}$ or smaller, the suppressed detail becomes insignificant. Thus masking of the shadow is a very effective method for improving the reproduction of very small objects.

In the second method the background, i.e. the primary wave, is suppressed *after* it has traversed the hologram. This can be done by producing a real image of the point source by means of the reconstructing lens in figure 1, and arranging a small black mask at this point, preferably a probability mask. This arrangement is similar to the well-known 'schlieren' method. The result is, that instead of an amplitude in the object plane

$$1-t_c-t_e,$$

where c stands for 'correct' and e for 'error', we now obtain

$$-t_c-t_e,$$

neglecting the diffraction effects at the mask. Hence an absorbing object will now appear bright on a dark background, as in 'dark-field illumination'. While in the ordinary or 'bright field' method the intensity is approximately

$$1-2t_c-2t_e,$$

the 'dark field' intensity is $t_c^2+2t_et_c+t_e^2$.

One can consider t_e^2 as the spurious background, while $2t_et_c$ is the interference product of the two images. The spurious background is now the square of its previous value, proportional to the coverage fraction instead of to its square root, and becomes negligible for objects which cover only a few percent of the illuminated field. There remains, however, the interference product $2t_et_c$. This contributes nothing to the background, as it is zero everywhere outside the object, where $t_c = 0$. In the object area it represents merely a small modulation of the correct density values. In the case of black-and-white objects this effect is negligible, as the outlines remain unchanged. How far it can distort graded objects is a matter for further investigation.

A combination of the two methods, i.e. masking the geometrical shadow *and* the primary wave, appears to be particularly promising in the case of small objects.

A third, somewhat laborious method for improving the separation is taking a series of reconstructions, with different values of μ. While the true image always remains the same, the spurious image varies, and can thus be discriminated. A fourth method will be discussed later, in connexion with non-homocentric illumination.

Illuminating waves with astigmatism and spherical aberration

Following a method first introduced by Debye, we build up a general coherent illuminating wave of plane wavelets, normal to the direction α, β, γ, with an amplitude $Ad\Omega$ in the infinitesimal solid angle $d\Omega$:

$$A(\alpha,\beta)\exp\{ik[x\cos\alpha+y\cos\beta+z\cos\gamma-p(\alpha,\beta)]\}\,d\Omega. \tag{31}$$

The amplitude A is assumed as real, the phase factor e^{-ikp} expressing the advance of the phase compared with the direct ray through the origin O. Assuming that O coincides with the 'mean paraxial focus' of the beam, let the phase function p be

$$p(\alpha,\beta) = \tfrac{1}{2}A_s(\cos^2\alpha-\cos^2\beta)+\tfrac{1}{4}(C_x\cos^4\alpha+2C_{xy}\cos^2\alpha\cos^2\beta+C_y\cos^4\beta). \tag{32}$$

The first term is the phase advance due to astigmatism, the second is 'elliptical' spherical aberration. It has been assumed for simplicity that the elliptical errors of second and fourth order have the same principal axes x, y.

With the polar angles γ, θ, connected with α, β by

$$\cos\alpha = \sin\gamma\cos\theta, \quad \cos\beta = \sin\gamma\sin\theta, \tag{33}$$

p can be put into the form

$$p(\gamma,\theta) = \tfrac{1}{2}A_s\sin^2\gamma\cos 2\theta$$
$$+\tfrac{1}{4}\sin^4\gamma[\tfrac{3}{8}(C_x+C_y)+\tfrac{1}{4}C_{xy}+\tfrac{1}{2}(C_x-C_y)\cos 2\theta-\tfrac{1}{4}[C_{xy}-\tfrac{1}{2}(C_x+C_y)]\cos 4\theta]. \tag{34}$$

The fourth-order term now appears as the sum of spherical aberration and two astigmatism terms, one elliptical, the other with fourfold periodicity. If the lens is round

$$C_x = C_y = C_{xy} = C_s, \tag{35}$$

and the fourth-order astigmatic terms vanish. C_s is the constant of spherical aberration. Its meaning is illustrated in figure 8, which shows the ray structure of a beam. The geometric-optical approximation is well justified in the most important practical applications of the present theory, as it is proposed to use beams with apertures about ten times larger than in ordinary electron microscopy, where the diffraction disk is of the same order as the geometrical aberrations. As the minimum cross-section of the beam increases with the third power of the aperture, and the diffraction effect is inversely proportional to the first power, it will represent a small correction only, of the order of 10^{-4} of the geometrical dimensions.

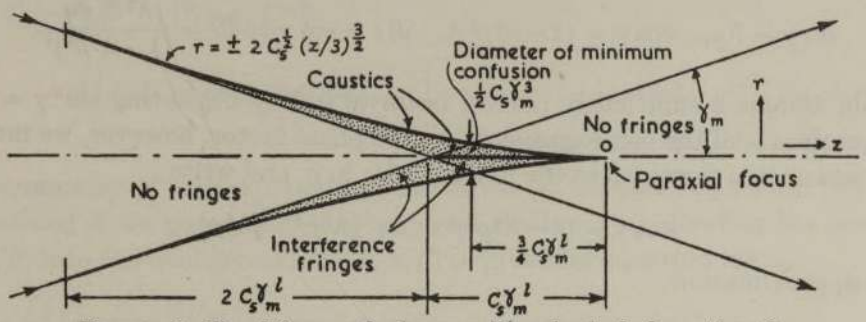


FIGURE 8. Focal figure of a beam with spherical aberration C_s.

If the aperture angle is γ_m, all rays cross the axis in the axial caustic, a line of length $C_s\gamma_m^2$ behind the paraxial focus O. The diameter of the beam in the Gaussian plane $z = 0$ is $2C_s\gamma_m^3$, but at the cross-section of minimum confusion, at $z = -\frac{3}{4}C_s\gamma_m^2$, it is four times less. The minimum cross-section is the intersection of the envelope or outer caustic, a rotational surface with an equation $r = \pm 2C_s^{\frac{1}{4}}(z/3)^{\frac{3}{4}}$, with the cone of maximum divergence, $r = \pm(z + C_s\gamma_m^2)\gamma_m$. This cone and the outer caustic divide up the beam into four regions of different character, of which two are dotted, to indicate that they contain interference fringes. The first of these is inside the envelope but outside the cone. The rays cross in every point of this region. The second is the region surrounding the axial caustic, limited by the envelope and by the cone of maximum divergence, which has three rays crossing in every point. The interference fringes in both regions are so sharp and contrasty as to make objects placed into them almost invisible; thus the whole dotted volume must be ruled out as a possible location for objects. In the remaining two regions, at the right and left, there is only one ray to every point, and they represent smoothly graded backgrounds, suitable for microscopic objects. In the region at the left the illumination density is largest near the edge; in the second, at the right, the density has a maximum on the axis.

If, in addition, as is always the case in electron optics, the beam is also astigmatic, figure 8 can still serve as an illustration, but only for the principal sections of the beam, and these must be imagined as displaced longitudinally by $\pm A_s$. Thus O will be now in the middle between the two focal lines, at right angles to one another and to the beam axis, and a distance $2A_s$ apart.

Returning to the wave-optical representation, summing the wavelets (31) gives for the complex amplitude at the point x, y, z

$$U_0(x,y,z) = \iint A(\gamma,\theta)\exp\{ik[(x\cos\theta + y\sin\theta)\sin\gamma + z\cos\gamma - \tfrac{1}{2}A_s\sin^2\gamma\cos 2\theta - \tfrac{1}{4}C_s\sin^4\gamma]\}\sin\gamma\,d\gamma\,d\theta. \quad (36)$$

We have used here the simplifications arising from equation (35), and these will be assumed also in the following formulae to simplify the discussion, but the results will be of such a nature as to permit their extension without difficulty also to the more general case expressed by equations (32) and (34).

Introduce under the integral sign in (36) the Fourier variables $\xi = (\cos\alpha)/\lambda$, $\eta = (\cos\beta)/\lambda$ and $\rho = (\xi^2+\eta^2)^{\frac{1}{2}}$. The exact transformation equations are

$$\sin\gamma = \lambda\rho, \quad \cos\gamma = (1-\lambda^2\rho^2)^{\frac{1}{2}}, \quad d\Omega = \sin\gamma\,d\gamma\,d\theta = \frac{\lambda^2\,d\xi\,d\eta}{(1-\lambda^2\rho^2)^{\frac{1}{2}}}.$$

We again assume a sufficiently narrow beam to justify neglecting $\sin^2\gamma = \lambda^2\rho^2$ in the denominator of the last expression. In the phase factor, however, we must take into consideration terms up to the fourth order in ρ, and write

$$\cos\gamma = (1-\lambda^2\rho^2)^{\frac{1}{2}} = 1 - \tfrac{1}{2}\lambda^2\rho^2 - \tfrac{1}{8}\lambda^4\rho^4.$$

In this approximation

$$U_0(x,y,z) = \lambda^2\,e^{ikz}\iint A(\xi,\eta)\exp\{2\pi i[x\xi + y\eta - \tfrac{1}{2}z\lambda\rho^2 - \tfrac{1}{2}A_s\lambda(\xi^2-\eta^2) - \tfrac{1}{8}(z+2C_s)\lambda^3\rho^4]\}\,d\xi\,d\eta. \quad (36{\cdot}1)$$

This integral, like the exact expression (36), can be easily evaluated at large distances R from the origin, in a direction α, β. One obtains

$$U_0(R,\alpha,\beta) = -i\frac{\lambda}{R}A(\gamma,\theta)\,e^{ik(R-p)}, \quad (36{\cdot}2)$$

where p is given by equations (32) or (34). The factor $-i$, which expresses an advance of the wave-front by $\frac{1}{4}\lambda$ as compared with the components (31) arises in the transition from plane to spherical waves, and is familiar in diffraction theory. Equation (36·2) supplies the background to the physical shadow of an object, which we are now going to calculate.

The object, in a plane $z = z_0$, may be characterized, as before, by the complex transmission function $t(x, y)$. Using the fundamental premissa of the Fresnel-Kirchhoff diffraction theory, we assume that the amplitude immediately before the object is that of the undisturbed illuminating wave, $U_0(x, y, z_0)$, and $t(x, y)$ times this immediately behind it. We must now give the angular variables in the illuminating wave by suffixes '0' ('original') to distinguish them from the variables of the outgoing wave, without suffixes.

The problem is building up the outgoing wave from the diffraction products of the plane wavelets which compose the original wave. The Fresnel-Kirchhoff formula in the simplified form (7·1) can be again applied, but with the modification that the

wave $r_0^{-1}\,e^{ikr_0}$ must now be replaced by the sum of the wavelets (34). There is no change in the meaning of r_1, the distance of the observation point $Q(R,\alpha,\beta)$ from the object point P. Thus the Fresnel-Kirchhoff formula now assumes the form

$$U(Q) = U(R,\alpha,\beta)$$

$$= \frac{1}{i\lambda R}\iiiint t(x,y)\,A(\alpha_0,\beta_0)\exp\{ik[r_0(\alpha_0,\beta_0)+r_1(\alpha,\beta)]\}\,dx\,dy\,d\alpha_0\,d\beta_0, \quad (37)$$

where

$$r_0(\alpha_0,\beta_0) = x\cos\alpha_0 + y\cos\beta_0 + z_0\cos\gamma_0 - p(\alpha_0,\gamma_0),$$

$$r_1(\alpha,\beta) = R - x\cos\alpha - y\cos\beta - z_0\cos\gamma.$$

Expressing the angles by the Fourier variables ξ,η and ξ_0,η_0 we obtain, with the same approximations as in (36·1),

$$U(R,\xi,\eta) = \frac{\lambda}{iR}e^{ikR}\iint t(x,y)\,dx\,dy$$

$$\iint A(\xi_0,\eta_0)\exp\{2\pi i[x(\xi_0-\xi)+y(\eta_0-\eta)-\tfrac{1}{2}z_0\lambda(\rho_0^2-\rho^2)-\tfrac{1}{8}z_0\lambda^3(\rho_0^4-\rho^4)$$
$$-\tfrac{1}{2}A_s\lambda(\xi_0^2-\eta_0^2)-\tfrac{1}{4}C_s\lambda^3\rho_0^4]\}\,d\xi_0\,d\eta_0. \quad (38)$$

The symmetry of this expression is disturbed by the last two terms, but it is at once restored if we go over to the 'physical shadow', by dividing the amplitude $U(R,\alpha,\beta)$ into the background $U_0(R,\alpha,\beta)$ as given by equation (36·2):

$$\tau(\xi,\eta) = \iiiint t(x,y)\,\frac{A(\xi_0,\eta_0)}{A(\xi,\eta)}\exp[2\pi i\{x(\xi_0-\xi)+y(\eta_0-\eta)-\tfrac{1}{2}z_0\lambda(\rho_0^2-\rho^2)$$
$$-\tfrac{1}{2}A_s\lambda[(\xi_0^2-\xi^2)-(\eta_0^2-\eta^2)]-\tfrac{1}{8}\lambda^3(z_0+2C_s)(\rho_0^4-\rho^4)\}]\,dx\,dy\,d\xi_0\,d\eta_0. \quad (39)$$

This is the formula for the physical shadow at infinity of an object at $z = z_0$, illuminated by a beam with fourth-order aberrations, but which can be evidently extended to aberrations of any order. It is the equivalent of the transformation formula (14) for homocentric illumination, but it cannot be put into the form of an integral over the object plane, as the integration over the angular variables cannot be carried out in terms of the transcendentals recognized in analysis. On the other hand, it can be immediately reduced to a double integral over the angular variables by means of the Fourier transform $T(\xi,\eta)$ of $t(x,y)$ which is

$$T(\xi,\eta) = \iint t(x,y)\,e^{-2\pi i(x\xi+y\eta)}\,dx\,dy,$$

which converts equation (42) into

$$\tau(\xi,\eta) = \iint T(\xi-\xi_0,\eta-\eta_0)\,\frac{A(\xi_0,\eta_0)}{A(\xi,\eta)}\exp[\pi i\{z_0\lambda(\rho^2-\rho_0^2)$$
$$+A_s\lambda[(\xi^2-\xi_0^2)-(\eta^2-\eta_0^2)]+\tfrac{1}{4}\lambda^3(z_0+2C_s)(\rho^4-\rho_0^4)\}]\,d\xi_0\,d\eta_0. \quad (40)$$

This transformation may be illustrated by a few simple examples. If $t = 1$, i.e. if there is no object, T is a delta function

$$T(\xi-\xi_0,\eta-\eta_0) = \delta(\xi-\xi_0,\eta-\eta_0),$$

which means that the integral (40) is the value of the integrand for $\xi_0 = \xi$, $\eta_0 = \eta$, which is unity, as before.

If $t(x, y)$ is a harmonic function of x, y with periods $1/a, 1/b$

$$t(x, y) = e^{2\pi i(ax+by)}. \tag{41·1}$$

T is again a delta function, but shifted to the point a, b

$$T(\xi-\xi_0, \eta-\eta_0) = \delta(\xi-\xi_0-a, \eta-\eta_0-b),$$

and we have again to take the value of the integrand, but this time at $\xi_0 = \xi-a$, $\eta_0 = \eta-b$. The physical shadow is

$$\tau(\xi, \eta) = \frac{A(\xi-a, \eta-b)}{A(\xi, \eta)} \exp\{\pi i a(2\xi-a)[\lambda(z_0+A_s)+\tfrac{1}{4}\lambda^3(z_0+2C_s)(2\xi^2-2\xi a+a^2)]\}$$
$$\exp\{\pi i b(2\eta-b)[\lambda(z_0-A_s)+\tfrac{1}{4}\lambda^3(z_0+2C_s)(2\eta^2-2\eta b+b^2)]\}. \tag{41·2}$$

We have met the first factors under the exponential in the shadow transformation with homocentric illumination. But the period in the shadow is no longer a constant, that is to say, the shadow of a sinusoidal grid is not of the same type as the original. If, for example, $b = 0$, i.e. the grid is parallel to y, the spacing between two maxima is

$$1/(\lambda z_0 a)[1+A_s/z_0+\tfrac{1}{4}\lambda^2(1+2C_s/z_0)(2\xi^2-2\xi a+a^2)].$$

The first factor is the geometrical shadow of the period $1/a$, the second is the correction arising from astigmatism and spherical aberration, and also from the fourth-order term which expresses the departure of a spherical wave-front from a paraboloid. In all practical applications z_0 will be of the order $C_s\gamma_m^2$, and z_0 can be neglected against $2C_s$. Thus the astigmatism and spherical aberration of a beam can be determined from two holograms of a sinusoidal grid, taken in two positions, at right angles to one another. But the method is not very sensitive. Near the edge of the field where $\xi \gg a, \eta \gg b$, the spacing of two neighbouring maxima will be a fraction

$$1\Big/\left(1+\frac{C_s}{z_0}\gamma_m^2\right)$$

of the geometrical spacing. But as z_0 will be of the order $C_s\gamma_m^2$ if good photographs are to be obtained, this fraction will be of the order unity. This shows that a sinusoidal grid, even if it were available, would not be a very suitable test object. Spherical aberration can be much better determined from the physical shadow of a thin wire, but the discussion of this case cannot be carried out in elementary terms, and may be omitted.

Reconstruction in the presence of spherical aberration and astigmatism

Assume that a photograph has been taken of the physical shadow of an object, according to equations (39) or (40). We have seen that, if the background is relatively strong, this is equivalent to substituting for τ_1 its real part, $\frac{1}{2}(\tau_1+\tau_1^*)$, where, as before, τ_1 relates to the 'object proper' without the background. In order to find the spurious term in the reconstructed object, we must apply to τ_1^* the transformation inverse to (39). But this is rather complicated, while an interpretation in terms of 'twin images' is easy, and leads to much simpler and clearer results.

An expression for τ_1^*, the complex conjugate of the physical shadow τ_1 is obtained from (39) by reversing the sign of i. Assume now, as before, in the plane $z = z_0$ a twin object with a transmission function

$$t_1'(x, y) = t_1^*(-x, -y).$$

Renaming the integration variables $-x, -y$ instead of x, y, one obtains for τ_1^* the expression

$$\tau_1^*(\xi, \eta) = \iiiint t_1'(x, y) \frac{A(\xi_0, \eta_0)}{A(\xi, \eta)} \exp\{2\pi i[x(\xi_0 - \xi) + y(\eta_0 - \eta) + \tfrac{1}{2}z_0\lambda(\rho_0^2 - \rho^2) + \tfrac{1}{2}A_s\lambda[(\xi_0^2 - \xi^2) - (\eta_0^2 - \eta^2)] + \tfrac{1}{8}\lambda^3(z_0 + 2C_s)(\rho_0^4 - \rho^4)]\} dx\,dy\,d\xi_0\,d\eta_0. \quad (42)$$

This is the physical shadow of an object t_1' in the plane $-z_0$, according to equation (39), but with the important difference that the sign of A_s and C_s has been also reversed. The physical significance of this becomes clearer if instead of τ_1^* we consider the complementary wave U_1' which arises in the reconstruction, and which is obtained from (42) by multiplying it with the background (36·2). The result can be written

$$U_1'(R, \xi, \eta) = \frac{\lambda}{iR} e^{ikR} \iint t_1'(x, y)\, dx\,dy \iint A(\xi_0, \eta_0) \exp\{2\pi i[x(\xi_0 - \xi) + y(\eta_0 - \eta) + \tfrac{1}{2}z_0\lambda(\rho_0^2 - \rho^2) + \tfrac{1}{8}z_0\lambda^3(\rho_0^4 - \rho^4) + \tfrac{1}{2}A_s\lambda(\xi_0^2 - \eta_0^2) + \tfrac{1}{4}C_s\lambda^3\rho_0^4]\} \times \exp\{-2\pi i[A_s\lambda(\xi^2 - \eta^2) + \tfrac{1}{2}C_s\lambda^3\rho^4]\} d\xi_0\,d\eta_0. \quad (43)$$

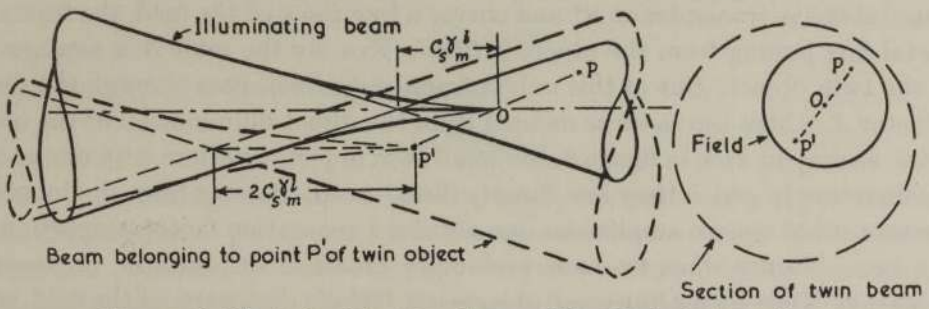


FIGURE 9. The twin object in a beam with spherical aberration.

Comparing this with equation (41), it can be seen that the first two lines represent the emission of an object t_1' in the plane $-z_0$, but *illuminated by a wave in which the signs of the astigmatism A_s and of the spherical aberration C_s are reversed.* This assures complete symmetry in the illumination of the object and its twin. But the emitted wave is modified by the phase factor in the last line. This means that the wavelet issuing from any element $t_1'(x, y)\,dx\,dy$ of the twin object has astigmatism $2A_s$ and spherical aberration $2C_s$. Thus *in the presence of astigmatism or spherical aberration the twin object which appears in the reconstruction will be no longer sharp, but will appear as if viewed through a system with twice the aberrations of the condenser system.* One could, of course, view the twin object instead of the original by means of a viewing system with aberrations of the opposite sign, but not both simultaneously.

This result is illustrated in figure 9, which allows also an elementary verification. The illuminating beam envelope is shown in continuous lines, the beam appearing to issue from a point P' of the twin object in interrupted lines. The axial caustic of this beam is always twice the caustic of the illuminating beam. This can be immediately

understood if one imagines the axial caustic as the locus of the centres of homocentric beams, each emitting rays only in a certain cone. For each of these partial beams there exists a sharp twin point to P, on the line joining P with its centre. Equation (43) proves that this geometric-optical reasoning is in fact justified.

Figure 9 shows also that the beam associated with any point of the twin object intersects the object plane in an area four times larger than the field. From this we can infer at once that if the illumination were even, the spurious amplitude in the object plane would bear the same relation to the correct amplitude as in the case of homocentric illumination, i.e. equation (29) would apply again. In fact the illumination is very uneven in a beam with spherical aberration in cross-sections not very far from the caustic, and on this is based a fourth method of improving the separation, in addition to the others which have been discussed in a previous section. Masking is not very efficient in the presence of spherical aberration, as the geometrical shadow of a point object is a radial line, the projection of the axial caustic. This becomes small only if the object is in the axis, but in electron optics it is not possible to fix small objects by means of a transparent support in the middle of the field.

This fourth method for improving the separation is to place the object in a position where it receives less than the average of illumination density. To explain this briefly, define as 'coefficient of illumination', J, the ratio of the mean intensity over a small object area to the mean over the *whole* illuminated field. If the object has the average intensity transmission tt^* and covers a fraction κ of the field, the fraction of the total flux issuing from the object is $tt^*\kappa J$. Exactly the same flux emerges also from the twin object. But of this only a fraction $\frac{1}{4}\kappa J$ will pass through the object. The factor J is here the same as defined from the direct illumination of the object, because, as may be seen in figure 9, the small twin objects interfere with one another in the direction in which they are directly illuminated. Passing from the intensities to the root mean square amplitudes one obtains a separation factor proportional to $\sqrt{(\kappa J)}$, i.e. $\sqrt{J}$ times what we have previously obtained for uniform, homocentric illumination. Thus by placing small objects in *relatively dark* parts of the field, where $J<1$, one improves the separation, by reducing the spurious background in the object area. Correspondingly more light is sent by the twin object to other regions of the field, but the spurious amplitude is of course harmless if it falls well outside the reconstructed object.

It may be noted that relatively weak illumination does not affect the contrast in the reconstructed object, so long as it is not submerged by ghosts, scattered light, and impurities arising from uneven development of the photograph.

Coherence criteria

Up to this point we have assumed an absolutely coherent monochromatic illuminating wave, originating from a point source, but distorted by passing through a lens system. Absolute coherence means interference fringes of any order, but it means of course zero intensity. In practice we must strike a compromise between these two conflicting claims. The best compromise is obtained if the degree of coherence is just sufficient to produce an interference pattern from which the object can be reconstructed with the required resolution limit.

A necessary criterion of coherence can be immediately formulated, without any regard to details of the hologram. Imagine that an absolutely coherent illuminating beam is moved during the exposure parallel to itself, so that a representative point of it, e.g. the mean paraxial focus, fills a circular disk with diameter d_c. But this is equivalent to moving the object within a disk of the same diameter, as only the relative position of beam and object matters for the physical shadow at infinity, and from such a 'wobbled' hologram we could at best reconstruct an image with a resolution limit d_c. Thus we obtain the necessary condition that the Gaussian or nominal diameter of the illuminating disk, d_c, must not exceed the Abbe limit d_A

$$d_c \leqslant d_A = \tfrac{1}{2}\lambda/\sin\gamma_m. \tag{44}$$

But we can show that this necessary condition is also sufficient, because it will produce holograms practically indistinguishable from one taken with an absolutely coherent beam, within a plate radius corresponding to the maximum angle γ_m. Express in equation (47) the wave-length by de Broglie's relation as

$$\lambda = h/p,$$

where p is the momentum associated with the wave. This relation is valid for photons as well as for electrons or any other particles. Interpreting $p\sin\gamma_m$ as the maximum transversal momentum p_t of the particles in the beam, we write (47) in the form

$$d_c 2p_t \leqslant h. \tag{44·1}$$

Confront this with Heisenberg's inequality

$$d_c' 2p_t' \geqslant h, \tag{45}$$

where d_c' is the maximum transversal uncertainty of position of particles in the beam in the Gaussian focal plane, and $2p_t'$ is the maximum uncertainty in the transversal momentum. Consider first the case that the beam is limited by a physical aperture in the plane considered, i.e. $d_c = d_c'$. Heisenberg's principle states that if the particles composing the beam are specified to the limit (45), they are indistinguishable, that is to say, they produce effects, such as interference fringes, which cannot be distinguished from one another by observation within the cone-angle corresponding to that value of p_c' which changes the inequality into an equality. Comparing (44·1) and (45) we see that if $d_c = d_c'$, we must have $p_t < p_t'$, thus the interference fringes inside the cone γ_m are *a fortiori* the same for all beam particles.

But if d_c is not a physical aperture, but the Gaussian image of one, formed by an optical system, the criterion still holds, because $d_c \sin\gamma_m$ is an invariant in Gaussian optics. If the criterion (44) were not sufficient, it would be possible to break through Heisenberg's principle by placing a suitable lens system in front of the physical aperture to produce observable differences in the fringe system, which would make the particles to some extent distinguishable.

These very general considerations are of course uncertain to a factor of the order unity. In order to obtain a more quantitative idea of the changes which are produced in the hologram by departure from absolute coherence, consider the simple case of illumination through a physical aperture of diameter d, and investigate its effect on

the fringe system produced by a point object on the axis, at a distance z_0 from the aperture. Each point of the illuminating aperture produces a fringe system concentric with the axis which connects this point with the point object. These fringe systems are incoherent with one another, hence their intensities must be summed. At the edge of the hologram the angular spacing of two fringes is $\lambda/z_0 \sin\gamma_m$. Two fringe systems will just wipe out one another if they are displaced by half this amount. This will be the case if the spacing of the two point sources is $\frac{1}{2}\lambda/\sin\gamma_m$, which is just the Abbe limit d_A.

With Zernike (1948), we define the 'degree of coherence' D_c, as the range of intensity difference between maxima and minima in the fringe system at the marginal angle γ_m, divided by the corresponding quantity if the same light flux issues from a point source at the centre of the aperture. Assuming that the intensity variation in the fringe system is sinusoidal, one obtains

$$D_c = \iint \cos(\pi x/d_A)\,dx\,dy \Big/ \iint dx\,dy, \tag{46}$$

where the integration has to be carried out over the area of the illuminating aperture of diameter d. The integrand $\cos(\pi x/d_A)$ expresses the fact that two points spaced in the X-direction by d_A just oppose one another. The integration gives

$$D_c = J_0\left(\tfrac{1}{2}\pi\frac{d}{d_A}\right) + J_2\left(\tfrac{1}{2}\pi\frac{d}{d_A}\right), \tag{47}$$

where J_0 and J_2 are the Bessel functions of zero and second order. Some values are

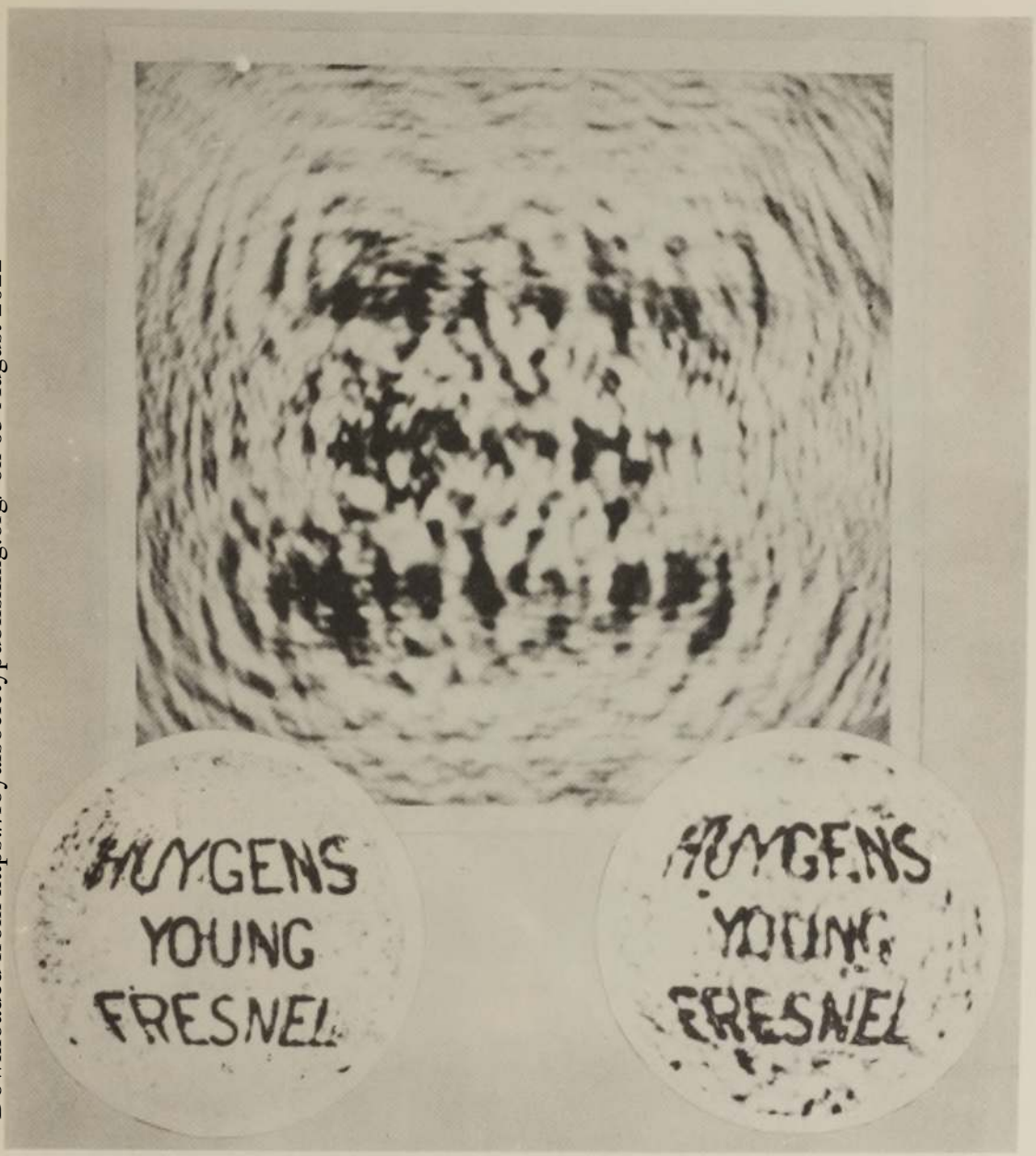
d/d_A	0	0·5	0·75	1·0	1·25	1·5	1·75	2·0
D_c	1	0·925	0·837	0·723	0·590	0·448	0·312	0·181

This justifies the expectation that the fringes system at the edge of the hologram will be rapidly effaced if the diameter of the light source appreciably exceeds the Abbe limit.

The coherence condition (47) represents a severe limitation of the available intensities, and it is the chief reason why the applications of the method of reconstructed wave-fronts will be probably restricted to light, with wave-lengths not very far from the visible, and to electrons. X-rays, protons and other particles will have to be excluded, as no sufficiently intense sources are available. Even in the case of electrons rather long exposures will be necessary, unless the present-day technique is improved.

The optical reconstruction

So far we have assumed in the formulae, for simplicity, that the reconstruction is carried out with the same wave-length as used in the production of the diffraction pattern. Let us now distinguish the first wave-length by λ', the second by λ'', and use the primes $'$ and $''$ also for distinguishing the data A_s, C_s in the analyzer and in the synthetizer. The same formal distinction will be used also for z_0' and z_0'', but here a word of explanation is required. z_0' is a datum of the analysis; it is the actual distance of the object from the mean paraxial focus of the illuminating beam. But there is no

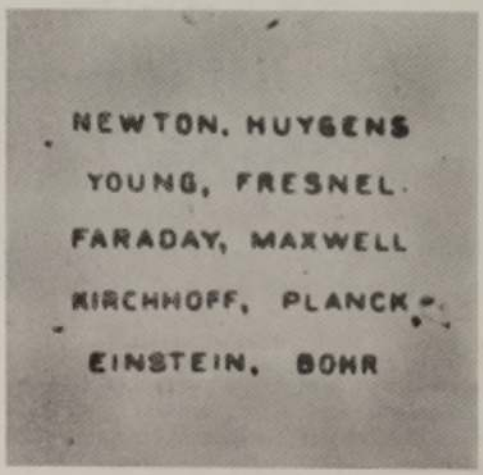


original hologram reconstruction

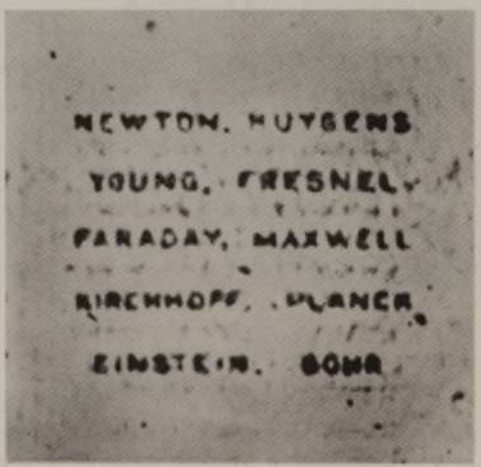
FIGURE 10. Optical reconstruction cycle. The original was a microphotograph of 1·5 mm. diam. Illuminated with $\lambda = 4358$ Å through pinhole 0·2 mm. diam., reduced by a microscope objective to 5μ nominal diameter, at 50 mm. from object. Geometrical magnification 12. Effective aperture of lens used in reconstruction 0·025. Noisy background chiefly due to imperfections of illuminating objective.

(*Facing p.* 480)

hologram

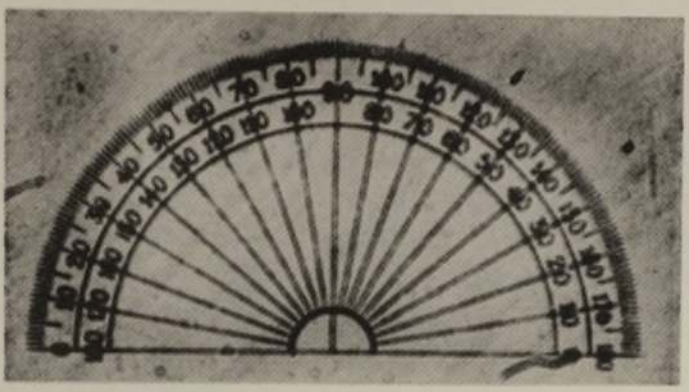


original

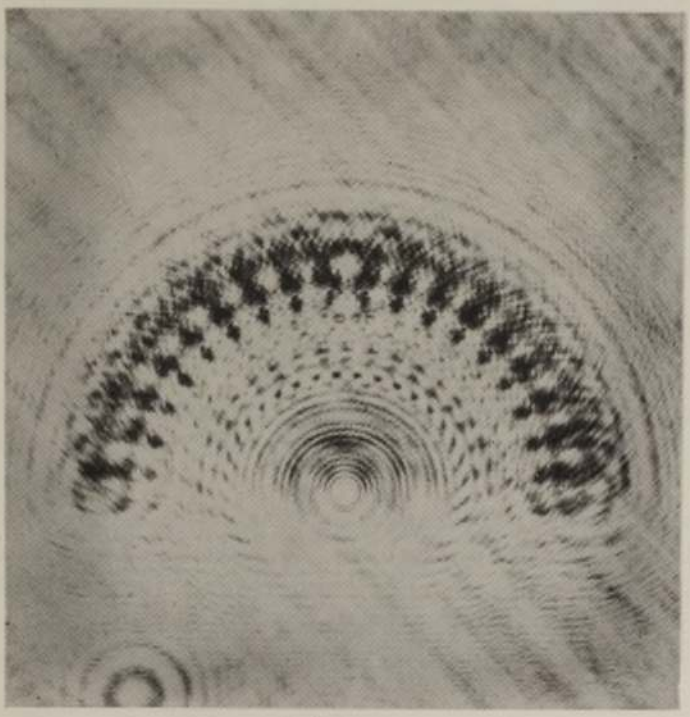


reconstruction

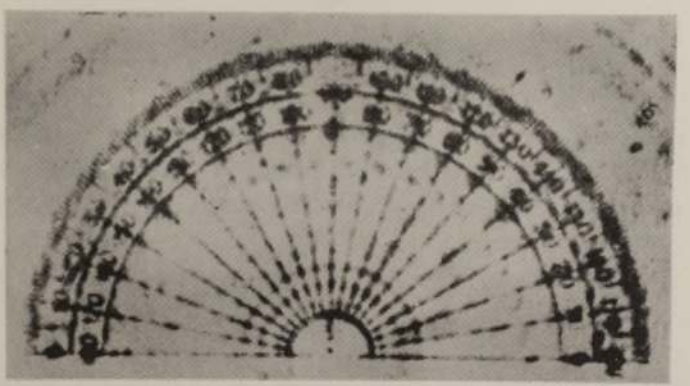
FIGURE 11. Reconstruction cycle with pinhole illumination. The letters in the original were inscribed in a rectangle 0·65 × 0·5 mm. Illumination with $\lambda = 4358$ Å through pinhole of 5μ diam. at 18 mm. from object. Geometrical magnification 10. Effective aperture used in reconstruction 0·075.



original



hologram



reconstruction

Figure 12. Reconstruction with pinhole illumination and wave-length change. The original was a micro-protractor of 1 mm. diam. Same conditions as in figure 11, but the wavelength used in the reconstruction was $\lambda = 5461$ Å.

physical object in the synthetizer, and z_0'' means merely the plane on which the viewing system must be focused in order to obtain a true, or at least the truest possible image of the original object.

The result of the analysis, the physical shadow, now to be called τ', is described by equation (42). We write down this equation again, but replace the Fourier variables ξ, η by the angles α, β, γ. For reasons of symmetry it will be convenient to attach the prime $'$ not only to the data of the analyzer, but also to the co-ordinates $x, y, \alpha, \beta, \gamma$ and $\alpha_0, \beta_0, \gamma_0$ used in the analysis. We write

$$\tau'(\alpha', \beta') = \iiint \frac{A'(\alpha_0', \beta_0')}{A'(\alpha', \beta')} t'(x', y') \times \exp\{2\pi i[Q(\alpha_0', \beta_0') - Q(\alpha', \beta')]\} \frac{d(\cos\alpha_0')\, d(\cos\beta_0')}{\cos\gamma_0'} dx'\, dy', \quad (48)$$

where the phase function Q is

$$Q(\alpha, \beta') = \frac{1}{\lambda'}[x' \cos\alpha' + y' \cos\beta' - \tfrac{1}{2} z_0' \sin^2\gamma_0' - \tfrac{1}{2} A_s'(\cos^2\alpha' - \cos^2\beta') - \tfrac{1}{8}(z_0' + 2C_s')\sin^4\gamma']. \quad (49)$$

The same equation applies to the synthesis, i.e. to the reconstruction of an object t'', with all primes $'$ changed into $''$. The fact that the hologram obtained in the analysis is used in the reconstruction is expressed by

$$\tau''(\alpha'', \beta'', \gamma'') = \tau'(\alpha', \beta', \gamma'), \quad (50)$$

where the angles α', β', γ' and $\alpha'', \beta'', \gamma''$ belong to corresponding points of the hologram. The relation between them is given by the geometries of the analyzer and of the synthetizer.

Consider first the simple case, illustrated in figure 1, in which the focal length f of the collimator lens in the synthetizer, which moves the hologram optically to infinity, is the same as the throw L in the analyzer. In this case the angles α', β' and α'', β'' are the same, and their primes can be disregarded. It can be seen by inspection of equation (48) that it is transformed into the corresponding equation for $\tau'' = \tau'$ if we put

$$x' = \frac{\lambda'}{\lambda''} x'', \quad y' = \frac{\lambda'}{\lambda''} y'', \quad A_s' = \frac{\lambda'}{\lambda''} A_s'', \quad C_s' = \frac{\lambda'}{\lambda''} C_s'', \quad z_0' = \frac{\lambda'}{\lambda''} z_0'', \quad (51)$$

and

$$\tau''(x'', y'') = \tau'\left(\frac{\lambda'}{\lambda''} x'', \frac{\lambda'}{\lambda''} y''\right). \quad (52)$$

The transformation of the integration variables is purely formal. The next two equations postulate the scaling up of the aberrations A_s', C_s' in the synthetizer, and the last of the conditions (51) states that one must focus on the plane z_0'' in order to see the object t'' given by equation (52).

Consider now the more general case

$$f = kL, \quad (53)$$

i.e. we use a collimator lens of focal length k times the throw in the analyzer, always assuming of course that the hologram is in the focal plane of the lens. (This covers

also the case in which the hologram used in the synthesis is an m times enlarged replica of the original; in this case the parameter k which figures in the following equations has the value f/mL.) The angles $\alpha' \ldots$ and $\alpha'' \ldots$ are now connected by the relations

$$\frac{\cos\alpha'}{\cos\gamma'} = k\frac{\cos\alpha''}{\cos\gamma''}, \quad \frac{\cos\beta'}{\cos\gamma'} = k\frac{\cos\beta''}{\cos\gamma''}. \tag{54}$$

The solution of these equations can be written in the form

$$\cos\alpha' = k\cos\alpha''[1 - \tfrac{1}{2}(k^2-1)\sin^2\gamma'' - \tfrac{3}{8}(k^2-1)^2\sin^4\gamma'' - \ldots]. \tag{54·1}$$

Only the first two terms of the expansion will be required. Introduce these into equation (48), where for simplicity we put $\lambda' = \lambda''$, to separate the change of geometry from the change of wave-length. The essential properties of the transformation can be deduced from the phase function Q, equation (49), which now assumes the form

$$\begin{aligned}\lambda Q = {} & k(x'\cos\alpha'' + y'\cos\beta'') - \tfrac{1}{2}k^2 z_0'\sin^2\gamma'' - \tfrac{1}{2}k^2 A_s'(\cos^2\alpha'' - \cos^2\beta'') \\ & - \tfrac{1}{8}k^4\left(\frac{z_0'}{k^2} + 2C_s'\right)\sin^4\gamma'' - \tfrac{1}{2}k(k^2-1)(x'\cos\alpha'' + y'\cos\beta'')\sin^2\gamma'' \\ & + \tfrac{3}{8}z_0'k^2(k^2-1)\sin^4\gamma'' + \tfrac{1}{2}A_s'k^2(k^2-1)\sin^2\gamma''(\cos^2\alpha'' - \cos^2\beta'').\end{aligned} \tag{55}$$

The terms in the first row and the first term in the second correspond to an exact reproduction, the others represent errors which arise only if $k^2 \neq 1$. Considering the first four terms only, equation (48) transforms into an identical equation for τ'' instead of τ' by putting

$$kx' = x'', \quad ky' = y'', \quad k^2A_s' = A_s'', \quad k^4C_s = C_s'', \quad k^2z_0' = z_0'', \tag{56}$$

and

$$t''(x'', y'') = t'\left(\frac{x''}{k}, \frac{y''}{k}\right). \tag{57}$$

This means that in order to see an image which is a k times enlarged replica of the original we must scale up the astigmatism k^2 times, the spherical aberration k^4 times, and focus the viewing system on a plane $z_0'' = k^2z_0'$.

But this image will appear with certain aberrations, which are indicated by the new terms in (55). The second term in the second row represents a *coma*. The first term in the last row is an addition to the spherical aberration, which can be incorporated in C_s''. The last term shows that the astigmatism A_s' of second order in the analyzer has produced astigmatism of the fourth order in the analyzer, i.e. a spherical aberration of the elliptical type.

All these error terms can be kept very small unless $k^2 \ll 1$. It can be shown that the best positions of the object are near $z_0' = -C_s'\sin^2\gamma_m'$, hence x', y' will be of the order $C_s'\sin^3\gamma_m'$, even if the object is in a marginal position. Hence the coma term in (55) will be of the order

$$k(k^2-1)C_s'\sin^3\gamma_m'\sin^3\gamma_m'' \simeq \frac{k^2-1}{k^2}C_s'\sin^6\gamma_m',$$

i.e. unless $k^2 \ll 1$ this will be a very small term, except in extreme cases when the spherical aberration $C_s'\sin^4\gamma_m'$ is of the order of several hundred fringes. In such cases the coma might amount to a few fringes, and coma compensation in the viewing system may become necessary.

The last term in (55) is of the order

$$\frac{k^2-1}{k^2} A'_s \sin^4 \gamma'_m,$$

which is again very small unless $k^2 \ll 1$. A'_s in good electron lenses is 10^{-4} or even less of C'_s; thus even if the spherical aberration is of the order of a thousand fringes, this term will represent a fraction of a fringe only.

Thus it is admissible to make the length of the optical synthetizer appreciably different from the throw in the electronic analyzer. It may be particularly advantageous to make $k < 1$, that is to say, not to make use of the full magnification λ''/λ' which is about 100,000, but only of a part of it. The rest can be supplied by the viewing system. This has the advantage that one can work with smaller lenses, though with proportionately larger numerical aperture. Assuming, for instance, $C'_s = 1$ cm. and $\sin \gamma'_m = 0{\cdot}05$, the minimum diameter of the electron beam is $0{\cdot}625\mu$, and if one makes $k = 1$ one requires an optical system capable of handling a light beam with 6·25 cm. minimum diameter. It will be advantageous to reduce this to one-half, or even to one-quarter, as optical systems with numerical apertures of 0·1 to 0·2 present no difficulties if the lenses need not be large.

To sum up, if in the optical synthetizer the data of the electronic condenser system are scaled up according to

$$A''_s = k^2 \frac{\lambda''}{\lambda'} A'_s, \quad C''_s = k^4 \frac{\lambda''}{\lambda'} C'_s, \tag{58}$$

the transversal dimensions of the object will appear scaled up in a ratio $k\lambda''/\lambda'$ and the longitudinal dimensions in the ratio $k^2\lambda''/\lambda'$. Thus *the geometrical or k-part of the transformation is of the type as produced by optical instruments, with a longitudinal magnification equal to the square of the transversal, while the λ-part is a uniform scaling-up, not realizable by ordinary optical imagery.*

The accuracy with which the conditions (58) have to be fulfilled can be best stated in terms of fringes. The maximum admissible deviation of a wave-front from the spherical shape without loss of resolving power has been estimated by Glaser (1943) as 0·4 of a wave-length, by Bruck (1947) as one wave-length. The second can be considered as the more reliable estimate. Thus the condition (58) for C''_s must be observed to an accuracy of one fringe. Assuming again $C'_s = 1$ cm. and a resolution limit of 1 Å, one requires by Abbe's rule an aperture $\sin \gamma'_m = 0{\cdot}025$, and with the more accurate numerical factor 0·6, $\sin \gamma'_m = 0{\cdot}030$. This gives 200 or 400 fringes at the edge of the field, according to which numerical factor one adopts. Thus the spherical aberration in the optical model must imitate C'_s to about one fringe in 200 or in 400.

The astigmatism tolerance at the edge of the field is about a quarter fringe. In carefully manufactured electron objectives A_s is of the order of a few microns, and it can be reduced by the compensation methods introduced by Hillier & Ramberg (1947) by at least one order of magnitude. This is necessary for realizing the full resolving power of present-day electron microscopes. In terms of fringes, the astigmatism in carefully manufactured but not compensated electron lenses amounts to a few fringes at apertures of 0·003, and if this is opened up ten times, to realize

a ten times improved resolving power, the distortion will be of the order of a few hundred fringes. Thus A'_s must be also imitated in the optical synthetizer to an accuracy of one part in a few hundred.

One could think of imitating the data of the electron-optical system by first carefully measuring A'_s and C'_s and computing an optical system with these data. But this is hardly a practicable method. Apart from the difficulties of measuring to the required accuracy, by the time the computation is finished and the optical replica is made the data of the electron-optical system are likely to have changed by far more than the error tolerance. It will be much preferable to make the astigmatism and the spherical aberration of the synthetizer variable, and adjust them until certain known parts of the object, such as the support, or certain standard test objects appear with maximum sharpness. The spherical aberration can be made variable by shifting a fourth-order plate, the astigmatism by crossed cylindrical lenses or by tilting lenses. Expert opticians will be doubtlessly able to work out a schedule to carry out the three adjustments of focus, astigmatism and spherical aberration in a systematic way. Thus only a moderate degree of constancy is required of the electron-optical system, sufficient at least for a series of reconstructions, without too frequent readjustments.

Experimental tests

Experiments were started almost as soon as the idea of reconstruction first emerged. They confirmed the soundness of the basic principle, but pointed to the necessity of elaborating and modifying the original, somewhat primitive views on the mechanism of reconstruction, which have been described elsewhere (Gabor 1948). The experiments were later continued in order to test the conclusions from the quantitative theory described in this paper.

In these tests analysis and synthesis were both carried out with visible light, though not always with the same wave-length. The arrangement for taking holograms was substantially as shown in the upper part of figure 1, but with optical instead of with electron lenses. A condenser threw an image of a high-pressure mercury arc (of the 'compact' type, with tungsten electrodes) through a colour filter on an aperture of about 0·2 mm. diameter. The lines used were 4358Å (violet), and 5461Å (green), isolated by Wratten light filters nos. 47 and 61. In the earlier tests a microscope objective was used to produce an image of this aperture, about 40 times reduced, i.e. with a nominal diameter of about 5μ, which formed the 'point source'. The objects were mostly microphotographs, sandwiched with immersion oil between two polished glass plates. In the earlier experiments the distance between the point source and the object was about 50 mm., the distance from the object to the photographic plate 550 mm., thus the geometrical magnification was about 12.

The photographic plate was held in position against three locating pins. Originally it was planned to develop the holograms by reversal, to make sure of exactly identical positions in the analysis and in the synthesis. In the negative-positive process the printing was carried out on the same locating pins. These precautions proved unnecessary in those experiments in which not only the Gaussian but also the

physical diameter of the source was of the order of the resolution limit, which proves that in these cases the theory of homocentric illuminating beams is a satisfactory approximation. But they were required later, in experiments with very strong spherical aberration in the illuminating beam. Reversal development, however, was found unnecessary, and the far more flexible negative-positive photographic process was used throughout. The negative hologram was usually processed with $\Gamma = 1{\cdot}2$ to $1{\cdot}6$, and the positive with $\Gamma = 0{\cdot}7$ to $1{\cdot}6$, so that a wide range of overall gammas could be tested. When it was confirmed that an overall gamma of 2 gave the best results, this was realized as closely as possible.

In the reconstruction the positive hologram was sandwiched with immersion oil between polished glass plates, which had to be carefully selected. It was backed by a viewing lens, which was an achromatic doublet, cemented and bloomed, with a focal length of 175 mm. and a linear aperture of 47 mm. The spherical aberration was 3 fringes at infinite conjugates. The diameter which satisfies the quarter-wave tolerance can be estimated at 27 mm., and the numerical aperture figures given below are based on this 'effective diameter'. The reconstructed image was viewed in a microscope, and photographed on plates introduced into the eyepiece.

Figure 10, plate 15, is a record of one of these earlier experiments. The figure at the left is a direct photograph of the original, which was a microphotograph of the names of the three founders of the wave theory of light. It was taken through the viewing system, with the same optics as used for the reconstruction. The top figure is the central part of the hologram, and the one at the right, is the reconstruction. All three were taken with the violet mercury line 4358 Å. The effective numerical aperture was $0{\cdot}025$, thus the resolution limit $0{\cdot}6 \times 0{\cdot}436/0{\cdot}025 = 10\mu$. This is $\frac{1}{150}$ of the diameter of the reproduced part of the microphotographs, and corresponds about to the gap between the 'Y' and the 'G' in 'HUYGENS'.

Though in its best parts the reconstruction almost attains the resolution of the direct photograph, the picture is very 'noisy'. This is due only to a smaller part to the essential disturbance created by the twin image, to a greater part it is due to specks of dust, and inhomogeneities in the two microscope objectives. It may be noted that these very troublesome effects, unwelcome concomitants of the great phase-discriminating power of the methods using a coherent background, cannot be expected to appear in an electronic analyzer. However imperfect an electron lens may be from the point of view of theoretical optics, it can contain neither dust nor 'schlieren', as the electromagnetic field smoothes itself out automatically, and in this respect any electron lens is superior to all but the best optical lenses.

In order to avoid these inessential disturbances, in some later experiments the optical surfaces were reduced to a minimum. In the experiments of which figures 11 and 12, plates 16 and 17, are records, the source was a pinhole of 3μ diameter, pierced into tinfoil with a very fine needle. Thus no glass surfaces other than those of the microphotographs were involved in the taking of the hologram. In the reconstruction the optics was also reduced to a minimum by cutting out the second microscope. The spacing between the object and the viewing lens was reduced to 180 mm., the distance between the lens and the plate increased to 700 mm., so that a fourfold enlargement of the object was produced by the viewing lens, sufficient for

direct photography on not unduly slow plates. Further enlargement was obtained in some cases by taking the hologram with the violet line, but reconstructing it with the green line.

The effective numerical aperture in this experimental series was 0·075, and the theoretical resolution limit 3·5μ. This is about $\frac{1}{350}$ of the diameter of the part of the microphotograph which is reproduced in figure 11 and which contains ten great names in the theory of light. The resolution is just about sufficient to resolve the hole in an 'A'. The theoretical resolution of the reconstruction is less, because the pinhole source of 3μ, used both in the analysis and in the synthesis, is of the same order. It can be estimated at about 5·5μ, by the thumb rule of orthogonal composition of errors. This resolution has been in fact very nearly achieved in the case of figure 11 and also in figure 12. It can be also seen that the background is very much more even than in figure 10. The residual disturbance is mostly essential, and due to the twin object. In these experiments the twin object could be separately focused, and as regards sharpness could not be distinguished from the 'true' image.

Experiments for testing the theory in the case of illuminating beams with large spherical aberration are in progress, but they have already confirmed its main results.

Conclusion

The new principle can be applied in all cases where coherent monochromatic radiation of sufficient intensity is available to produce a divergent diffraction pattern, with a relatively strong coherent background. While the application to electron microscopy promises the direct resolution of structures which are outside the range of ordinary electron microscopes, probably the most interesting feature of the new method for light-optical applications is the possibility of recording in one photograph the data of three-dimensional objects. In the reconstruction one plane after the other can be focused, as if the object were in position, though the disturbing effect of the parts of the object outside the sharply focused plane is stronger in coherent light than in incoherent illumination. But it is very likely that in light optics, where beam splitters are available, methods can be found for providing the coherent background which will allow better separation of object planes, and more effective elimination of the effects of the 'twin wave' than the simple arrangements which have been investigated.

I thank Mr L. J. Davies, Director of Research of the British Thomson-Houston Company, for permission to publish this paper and Mr J. Williams for assistance in the experimental work.

References

Baker, B. B. & Copson, E. T. 1939 *The mathematical theory of Huygens' principle.* Oxford: Clarendon Press.
Boersch, H. 1938 *Z. techn. Phys.* **19**, 337.
Boersch, H. 1939 *Z. techn. Phys.* **20**, 346.
Bragg, W. L. 1942 *Nature*, **149**, 470.
Bruck, H. 1947 *C.R. Acad. Sci., Paris*, **224**, 1553.

Bunn, W. 1945 *Chemical crystallography*. Oxford University Press.
Campbell, G. A. & Foster, R. M. 1931 *Fourier integrals for practical applications*. Bell Telephone System Monograph, B. 584. New York.
Gabor, D. 1948 *Nature*, **161**, 777.
Glaser, W. 1943 *Z. Phys.* **121**, 647.
Hillier, J. & Ramberg, E. G. 1947 *J. Appl. Phys.* **18**, 48.
Zernike, F. 1948 *Proc. Phys. Soc.* **61**, 147.

The thermal equilibrium at the tropopause and the temperature of the lower stratosphere

By R. M. Goody, *St John's College, University of Cambridge*

(*Communicated by D. Brunt, F.R.S.—Received* 16 *August* 1948—*Revised* 14 *February* 1949)

Further considerations along the lines of Emden's methods lead to continuity of temperature at the tropopause as a condition for a stable transition from a state of convective to radiative equilibrium. This explains the characteristic appearance of the temperature distribution near the tropopause. Application of this condition leads to a simple explanation of the latitude variation of stratosphere temperature, mainly in terms of the effects of water vapour and carbon dioxide. The variation of stratosphere temperature with ozone concentration may be calculated, which confirms Dobson's hypothesis that anomalous seasonal variations in stratosphere temperature are due to seasonal variations of ozone concentration. Reasons for the approximately isothermal character of the lower stratosphere are also discussed.

1. Introduction

The facts relating to the temperature of the lower stratosphere which require explanation are now well established. They are:

(i) that the transition of temperature from the troposphere to the stratosphere is smooth;

(ii) that the temperature is lower over the tropics than over the arctic;

(iii) that temperature gradients are small relative to those occurring in the troposphere, and often positive; and

(iv) that the seasonal variation of temperature differs from that of the troposphere immediately below the tropopause.

The lower stratosphere is taken to mean the approximately isothermal region between the tropopause and a height of approximately 30 km., above which temperature begins to increase rapidly with height.

Any theory of heat interchange in the lower stratosphere must attempt to explain these facts. Such a theory, in the present state of knowledge, must necessarily be highly simplified, for the existence of winds at high altitudes, and of moving air masses with distinct boundaries, make an exact treatment almost impossible. In problems of similar complication, however, it is sometimes possible to isolate one