CONTRIBUTIONS TO THE THEORY OF INTERFERENCE AND DIFFRACTION OF LIGHT AND RELATED INVESTIGATIONS

by

Emil Wolf

A thesis submitted in support of candidature
for the degree of Doctor of Science in
the University of Edinburgh,
September 1955.

PREFACE

The researches described in this thesis are in the main part concerned with Optics and Electromagnetic Theory. They are presented in the form of fifteen papers, published in various scientific journals within the period 1949-1955, and are grouped together under the following headings: (I) The Optical Image, (II) A General Theory of Interference and Diffraction of Light, (III) Studies in Electromagnetic Theory, and (IV) Related Investigations.

In paper 1.1 the light distribution near focus in the three-dimensional aberration-free diffraction image is investigated and formulae are derived for the fraction of the total energy contained within the central core of the diffraction image in any prescribed receiving plane. In paper 1.2 these formulae are used to determine some of the properties of polychromatic star images formed by refracting telescope objectives, taking into account the effect of the secondary spectrum. The next paper is concerned with the effect of the central obstruction of the aperture on the three-dimensional light distribution near focus. An historical and critical survey of Diffraction Theory of Aberrations is presented in paper 1.4, and in paper 1.5 the foundations of the scalar Diffraction Theory of optical image formation are investigated. In the following paper a new aberration function is introduced, which possesses several advantages over aberration functions employed previously, and which may be used in investigations based on Geometrical Optics or Diffraction Theory.

In the usual treatments of Interference and Diffraction, the sources are assumed to be of vanishingly small dimensions (point sources), which emit strictly monochromatic light. Numerous problems encountered, especially in applications to Spectroscopy, Microscopy and Astronomy, make it highly desirable to formulate a theory of Interference and Diffraction on a broader basis, taking into account the finite extension of any physical source, as well as the finite frequency range of radiation which any physical source emits. The need for such a generalization became particularly evident when F. Zernike in 1934 established in his pioneering researches on partial coherence a number of important and unsuspected theorems in this field. In Part II of this thesis, a general theory of Interference and Diffraction of Light is presented, which applies to any stationary field. This theory makes it possible to treat directly problems of Interference and Diffraction with polychromatic light from finite sources, and includes the majority of known results on partial coherence as special cases of much more general theorems. An attractive feature of this theory is that it operates with observable quantities only.

Part III consists of two papers dealing with Electromagnetic Theory. In paper 3.1 a new representation of any Electromagnetic Field in vacuo is described. The field is represented in terms of a single complex scalar wave function, in terms of which the momentum density and the energy density of the field may be defined by means of formulae strictly analogous to the quantum mechanical formulae for the probability current and the probability density.* In paper 3.2 several new theorems are derived which apply to any Electromagnetic Field in which at least one of the field vectors is linearly polarized.

Some related investigations are described in Part IV. In paper 4.1 a systematic derivation is given of the Circle Polynomials of Zernike, which play an important part in some branches of Diffraction Theory. Several new theorems concerning these polynomials are established and a related set of polynomials is investigated. Paper 4.2 is concerned with the X_n and Y_n functions of Hopkins which occur in certain extensions of the analysis of Lommel relating to the three-dimensional light distribution near focus. In paper 4.3 the design of the corrector plate of the Schmidt Camera is discussed

---

* The analysis of this paper has recently been extended by P. Román [Acta Phys. Acad. Sci. Hungaricae, 4 (1954), 209] to regions including currents and charges. Román showed that our method leads to a simplified treatment of the quantization of the Electromagnetic Field.

and a solution is obtained for the design of a plate which leads to an optimum performance over the field taken as a whole, with light covering a given spectral range.

The greater part of the work described in this thesis was carried out whilst I was a Research Assistant to Professor Max Born at Edinburgh University. I wish to acknowledge my sincere gratitude to Professor Born for allowing me to spend much time on work of my own interest and for many stimulating discussions.

In accordance with the regulations I state that the work described in those papers here submitted, which were published under my own authorship, was done entirely by myself, and that none of the work reported in this thesis was carried out under supervision. The papers of Part II are presented as the main contribution.

Manchester University,
August, 1955.

LIST OF CONTENTS

PART III
STUDIES IN ELECTROMAGNETIC THEORY

PART IV
RELATED INVESTIGATIONS

PART I

THE OPTICAL IMAGE

1.1

Reprinted without change of pagination from the
Proceedings of the Royal Society, A, *volume* 204, 1951

Light distribution near focus in an error-free diffraction image

BY E. WOLF, *The Observatory, University of Cambridge*

(*Communicated by R. O. Redman, F.R.S.—Received* 16 *June* 1950)

Expressions are derived for the fraction of the total illumination present within certain regions in receiving planes near focus of spherical waves issuing from a circular aperture. The derivation is based on Lommel's treatment of Fresnel diffraction, and the solution takes the form of rapidly convergent series involving Bessel functions. Numerical results are illustrated by contour lines. The distribution of the illumination in a number of selected planes near focus is studied in greater detail. A comparison with the predictions of geometrical optics is also made. As a special case the fraction of the total illumination present in the geometrical shadow is discussed. The results provide a mathematical basis for a discussion of the imaging properties in optical systems where, as for example in a well-corrected refracting telescope objective, the chromatic variation of focus is the only appreciable aberration.

1. INTRODUCTION

Although a good deal of attention has been paid in recent years to problems concerning the effect of diffraction on monochromatic image formation in the presence of aberrations, a number of important questions dealing with the diffraction of perfectly spherical waves have so far not been treated. Consider, for example, the diffraction of such waves at a circular aperture. With the usual approximations the intensity distribution in the geometrical focal plane is given by the well-known formula due to Airy (1835). The distribution of intensity in other planes was discussed by Lommel (1885) in a classical memoir and more recently by Nijboer (1942) and Zernike & Nijboer (1949). In most practical applications, however, it is desirable to know not only the intensity at points near focus of the waves but also the total illumination in the various rings of the diffraction pattern. In the Fraunhofer case, where the intensity is given by Airy's expression, the fraction L of the total illumination† which is received within a circle of radius r and centred on the axis is given by the formula due to Rayleigh (1881),

$$L = 1 - J_0^2(v) - J_1^2(v). \tag{1·1}$$

Here $v = \dfrac{2\pi R}{\lambda f} r$, R being the aperture radius, f the radius of the spherical wave filling the opening and λ the wave-length. The J's denote Bessel functions of the first kind. The literature contains no corresponding result for the more general Fresnel case. In this case the total illumination could of course be calculated by numerical integrations from Lommel's formulae or from those of Nijboer and Zernike. But such a procedure would involve a prohibitive amount of numerical work because of the form of the expressions and the complicated behaviour of the intensity distribution (see figure 2). In the present paper formulae are derived which permit direct calculation of the total illumination.

Starting from Lommel's expressions we obtain a solution which takes the form of rapidly convergent series involving Bessel functions. The solution is evaluated near focus, and the results are displayed in figure 3a in the form of contour lines. A comparison with results predicted by geometrical optics is also made. It is clearly seen that although in the immediate neighbourhood of the focus the illumination curves predicted by geometrical optics do not resemble those of the physical solution, the two solutions approach each other more and more closely as the distance from the focus becomes large compared with the wave-length. The behaviour of L in several planes near the focus is examined in greater detail.

In the special case when the circle for which the fraction of the total illumination is determined coincides with the geometrical confusion disk, the series can be summed. The solution then reduces to the simple form

$$L = 1 - J_0(v)\cos v - J_1(v)\sin v, \tag{1·2}$$

the expression $J_0(v)\cos v + J_1(v)\sin v$ giving the fraction of light within the geometrical shadow.

The results derived in the present paper supplement those of Lommel, Nijboer and Zernike and, with them, provide the mathematical basis for a quantitative discussion

† Following Rayleigh's usage we define the total illumination in a region D as

$$\iint_D I(x, y)\, dx\, dy,$$

where $I(x, y)$ is the intensity at a typical point (x, y) of D.

of the imaging properties in optical systems where chromatic variation of focus is the only appreciable aberration. For example, they make it practicable to calculate the light scattered by diffraction from stellar images formed by refracting telescope objectives, taking into consideration the presence of the secondary spectrum.

2. Expressions for the total illumination

In what follows, u and v are real variables, n and m are non-negative integers, J_n denotes a Bessel function of the first kind and U_n and V_n are Lommel functions defined by†

$$\left.\begin{aligned} U_n(u,v) &= \sum_{s=0}^{\infty}(-1)^s\left(\frac{u}{v}\right)^{n+2s} J_{n+2s}(v), \\ V_n(u,v) &= \sum_{s=0}^{\infty}(-1)^s\left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v). \end{aligned}\right\} \tag{2·1}$$

Further, Y_n and W_n denote the related functions

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

$$W_n(u,v) = \sum_{s=0}^{\infty}(-1)^s(s+1)\left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v). \tag{2·3}$$

We also define the polynomials

$$P_{n,2m}(v) = \sum_{s=0}^{2m}(-1)^s J_{n+s}(v)\, J_{n+2m-s}(v). \tag{2·4}$$

Finally, we shall find it convenient to set

$$Q_{2m}(v) = P_{0,2m}(v) + P_{1,2m}(v). \tag{2·5}$$

(*a*) We shall be concerned with the light distribution in a receiving plane near focus O, of convergent spherical waves which issue from a circular aperture. Let R be the radius of the opening, C the point of intersection between the spherical wave and the axis joining O to the centre of the aperture, and let $CO = f$. Further, let O' denote the point of intersection of the axis with the receiving plane, both this plane and the plane of the aperture being assumed to be perpendicular to the axis (figure 1).

To describe the effect of the light waves at a point P near the focus, we introduce polar co-ordinates (r, θ) in the receiving plane with origin at O' and set

$$u = \frac{2\pi R^2}{\lambda f^2}\Delta f, \quad v = \frac{2\pi R}{\lambda f} r, \tag{2·6}$$

† Two slightly different definitions of the V_n function appear in the literature. That given by (2·1) is used throughout this paper. The Y_n function has recently been introduced by Hopkins (1949) in a paper concerning waves of non-uniform amplitudes. These functions are related by the equation

$$Y_n(u,v) = \frac{1}{2}\left[\frac{v^2}{u} V_{n-1}(u,v) + u V_{n+1}(u,v)\right]$$

which is implicit in Lommel's memoir and which follows from (2·2) by the application of the identity

$$(n+2s)\, J_{n+2s}(v) = \tfrac{1}{2} v [J_{n+2s-1}(v) + J_{n+2s+1}(v)].$$

where $\Delta f = OO'$ (positive in figure 1) denotes the distance between the receiving plane and the geometrical focal plane $u = 0$; it is assumed throughout this paper that $\Delta f/f$, r/f and R/f are small compared to unity. In this notation the first zero of intensity on the axis is given by $u = 4\pi$, whilst the first dark ring in the geometrical focal plane (i.e. the Airy ring) is given by $v = 1{\cdot}22\pi$. The outline of the geometrical shadow is given by $u = \pm v$.

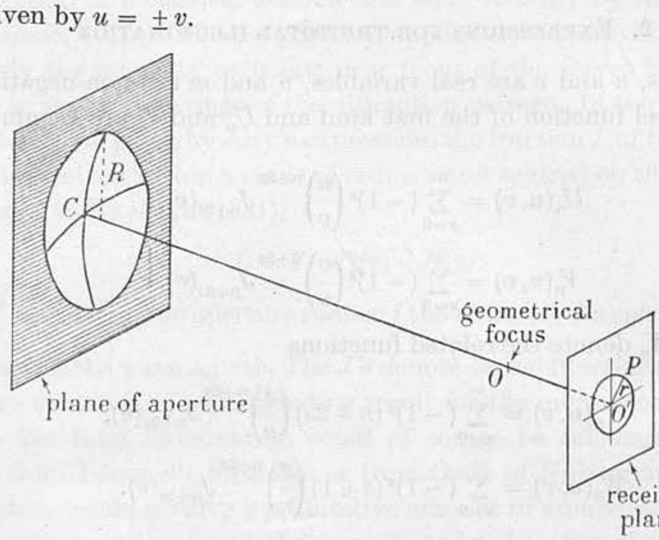


FIGURE 1. $CO = f$; $OO' = \Delta f$; $u = \dfrac{2\pi R^2}{\lambda f^2}\Delta f$; $v = \dfrac{2\pi R}{\lambda f} r$.

The physical significance of the parameters u and v is as follows: If we test in an interferometer the wave in the opening against a spherical wave centred at P, then $u/4\pi$ is the number of fringes of defocusing whilst v/π is the number of fringes of lateral displacement of P from the axis.

Let $\dfrac{A}{f}\sin\dfrac{2\pi}{\lambda}(ct-f)$ be the wave disturbance in the opening. Here A is a constant while c and t denote the velocity of light and the time respectively. The intensity $I(u,v)$ at the point P is then given, in suitable units, by the following formula due to Lommel† (1885):

$$I(u,v) = \frac{4\gamma}{u^2}\,[U_1^2(u,v) + U_2^2(u,v)], \tag{2·7}$$

where

$$\gamma = \left(\frac{A\pi R^2}{\lambda f^2}\right)^2,$$

and the U's are Lommel functions defined by (2·1). When $|u/v| > 1$ the U-series converge too slowly to be useful for computation, and for this case Lommel expressed the solution by means of the related V functions, in the equivalent form

$$I(u,v) = \frac{4\gamma}{u^2}\left[1 + V_0^2(u,v) + V_1^2(u,v) - 2V_0(u,v)\cos\frac{1}{2}\left(u + \frac{v^2}{u}\right) - 2V_1(u,v)\sin\frac{1}{2}\left(u + \frac{v^2}{u}\right)\right]. \tag{2·8}$$

† The solutions (2·7) and (2·8) were also obtained by Struve (1886) by analysis similar to that of Lommel.

Recently Nijboer (1942) and Zernike & Nijboer (1949) derived in the course of a more general investigation another expression for the intensity and gave a figure showing the isophotes near focus (figure 2).

Starting from Lommel's equations we shall now derive expressions for the total illumination received within the area in the receiving plane which is bounded by the circle centred on O' and of radius r_0. We define

$$L(u, v_0) = \frac{1}{B}\int_0^{r_0}\int_0^{2\pi} I(u,v)\, r\, dr\, d\theta, \tag{2·9}$$

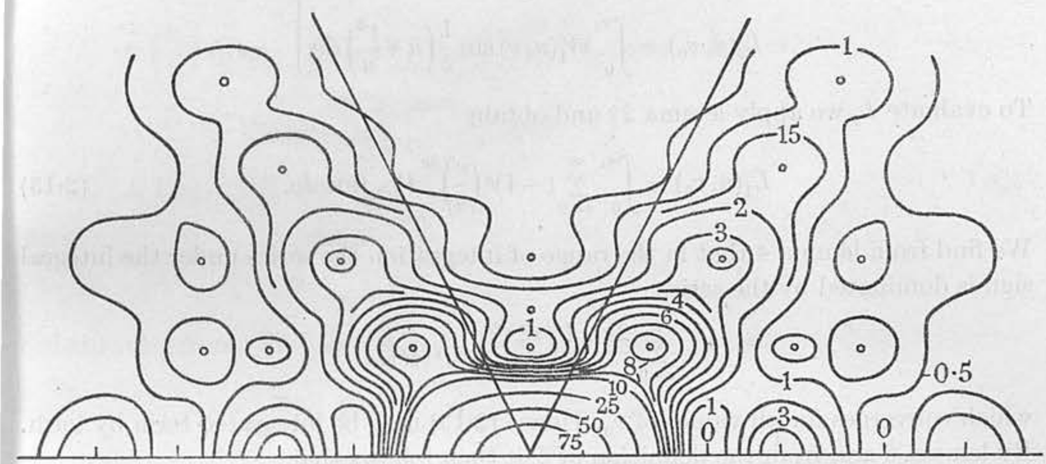


FIGURE 2. Isophotes [contour lines of $I(u, v)$] near focus in a meridional plane. The straight lines indicate the boundary of the geometrical shadow. The numbers give intensity as percentage of the intensity at focus. Axial minima and maxima are indicated by short strokes, others by small circles. The figure covers approximately the range $-35 \leqslant u \leqslant 35$, $0 \leqslant v \leqslant 15$. (After Zernike & Nijboer 1949.)

where v_0 denotes the value of the v parameter for points on the boundary of the circle and

$$B = \left(\frac{A}{f}\right)^2 \pi R^2.$$

To the order of accuracy here in question, $L(u, v_0)$ may be regarded as giving the fraction of the total energy radiated through the aperture which reaches the circle centred at O'. On integrating (2·9) with respect to θ, it follows that

$$L(u, v_0) = \frac{1}{2B}\int_0^{v_0} I(u,v)\, v\, dv. \tag{2·10}$$

The boundary of the circle over which the illumination is to be determined lies inside or outside the geometrical confusion disk according as $|v_0/u| \lesseqgtr 1$, and it coincides with its boundary when $|v_0/u| = 1$. We shall now examine the three cases separately. Since, within the range in which (2·7) and (2·8) are valid, $I(u, v)$ and consequently $L(u, v)$ are even functions of u, we can without loss of generality assume u to be positive. The parameter v is positive by definition.

(1) *The case* $v_0/u < 1$

In this case we have from (2·8) and (2·10)

$$L(u,v_0) = \frac{2}{u^2}[\tfrac{1}{2}v_0^2 + L_1(u,v_0) - 2L_2(u,v_0) - 2L_3(u,v_0)], \tag{2·11}$$

where
$$\left.\begin{aligned} L_1(u,v_0) &= \int_0^{v_0} v[V_0^2(u,v)+V_1^2(u,v)]\,dv, \\ L_2(u,v_0) &= \int_0^{v_0} vV_0(u,v)\cos\frac{1}{2}\left(u+\frac{v^2}{u}\right)dv, \\ L_3(u,v_0) &= \int_0^{v_0} vV_1(u,v)\sin\frac{1}{2}\left(u+\frac{v^2}{u}\right)dv. \end{aligned}\right\} \tag{2·12}$$

To evaluate L_1 we apply lemma 2† and obtain

$$L_1(u,v_0) = \int_0^{v_0} \sum_{s=0}^{\infty} (-1)^s \left(\frac{v}{u}\right)^{2s} P_{0,2s}(v)\,v\,dv. \tag{2·13}$$

We find from lemma 4 that in the range of integration the series under the integral sign is dominated by the series

$$\sum_{s=0}^{\infty} \frac{2s+1}{s\,!}\left(\frac{v_0^2}{2u}\right)^{2s} \exp(\tfrac{1}{2}v_0^2),$$

which converges for all values of v_0. Hence (2·13) may be integrated term by term. By lemma 6 and from the definition of Q it then follows that

$$L_1(u,v_0) = \frac{v_0^2}{2}\sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{v_0}{u}\right)^{2s} Q_{2s}(v_0). \tag{2·14}$$

To evaluate L_2, we write it in the form

$$L_2(u,v_0) = \cos\tfrac{1}{2}u\int_0^{v_0} vV_0(u,v)\cos\frac{v^2}{2u}dv - \sin\tfrac{1}{2}u\int_0^{v_0} vV_0(u,v)\sin\frac{v^2}{2u}dv. \tag{2·15}$$

We have from lemma 9, for $n=0$, on equating real and imaginary parts that

$$\left.\begin{aligned} \int_0^{v_0} vV_0(u,v)\cos\frac{v^2}{2u}dv &= u\left[W_1(u,v_0)\cos\frac{v_0^2}{2u} + W_2(u,v_0)\sin\frac{v_0^2}{2u}\right], \\ \int_0^{v_0} vV_0(u,v)\sin\frac{v^2}{2u}dv &= u\left[W_1(u,v_0)\sin\frac{v_0^2}{2u} - W_2(u,v_0)\cos\frac{v_0^2}{2u}\right]. \end{aligned}\right\} \tag{2·16}$$

(2·15) and (2·16) give

$$L_2(u,v_0) = u\left[W_1(u,v_0)\cos\frac{1}{2}\left(u+\frac{v_0^2}{u}\right) + W_2(u,v_0)\sin\frac{1}{2}\left(u+\frac{v_0^2}{u}\right)\right]. \tag{2·17}$$

In a similar way we find

$$L_3(u,v_0) = u\left[W_2(u,v_0)\sin\frac{1}{2}\left(u+\frac{v_0^2}{u}\right) - W_3(u,v_0)\cos\frac{1}{2}\left(u+\frac{v_0^2}{u}\right)\right]. \tag{2·18}$$

† Lemmas quoted without other reference will be found in the appendix.

Substituting for L_1, L_2 and L_3 into (2·11) we obtain

$$L(u,v_0)=\left(\frac{v_0}{u}\right)^2\left[1+\sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{v_0}{u}\right)^{2s}Q_{2s}(v)\right]$$
$$-\frac{4}{u}\left\{[W_1(u,v_0)-W_3(u,v_0)]\cos\frac{1}{2}\left(u+\frac{v_0^2}{u}\right)+2W_2(u,v_0)\sin\frac{1}{2}\left(u+\frac{v_0^2}{u}\right)\right\}. \quad (2·19)$$

But $W_1(u,v_0)-W_3(u,v_0)$

$$=\sum_{s=0}^{\infty}(-1)^s(s+1)\left(\frac{v_0}{u}\right)^{2s+1}J_{2s+1}(v_0)-\sum_{s=0}^{\infty}(-1)^s(s+1)\left(\frac{v_0}{u}\right)^{2s+3}J_{2s+3}(v_0)$$
$$=\left(\frac{v_0}{u}\right)J_1(v_0)+\sum_{s=1}^{\infty}(-1)^s(s+1)\left(\frac{v_0}{u}\right)^{2s+1}J_{2s+1}(v_0)-\sum_{s=1}^{\infty}(-1)^{s-1}s\left(\frac{v_0}{u}\right)^{2s+1}J_{2s+1}(v_0)$$
$$=\sum_{s=0}^{\infty}(-1)^s(2s+1)\left(\frac{v_0}{u}\right)^{2s+1}J_{2s+1}(v_0)$$
$$=Y_1(u,v_0). \quad (2·20)$$

Also $$W_2(u,v_0)=\sum_{s=0}^{\infty}(-1)^s(s+1)\left(\frac{v_0}{u}\right)^{2s+2}J_{2s+2}(v_0)$$
$$=\tfrac{1}{2}Y_2(u,v_0). \quad (2·21)$$

Substituting from (2·20) and (2·21) into (2·19), we finally obtain for the case $v_0/u\leqslant 1$,

$$L(u,v_0)=\left(\frac{v_0}{u}\right)^2\left[1+\sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{v_0}{u}\right)^{2s}Q_{2s}(v_0)\right]$$
$$-\frac{4}{u}\left[Y_1(u,v_0)\cos\frac{1}{2}\left(u+\frac{v_0^2}{u}\right)+Y_2(u,v_0)\sin\frac{1}{2}\left(u+\frac{v_0^2}{u}\right)\right]. \quad (2·22)$$

(2) *The case $v_0/u=1$*

In this case the boundary of the circle coincides with the edge of the geometrical shadow. The expression for L can then be obtained by putting $v_0=u$ in (2·22). We then find

$$L(u,u)=1+\sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}Q_{2s}(u)-\frac{4}{u}[Y_1(u,u)\cos u+Y_2(u,u)\sin u]. \quad (2·23)$$

Now $$Y_n(u,u)=\sum_{s=0}^{\infty}(-1)^s(n+2s)J_{n+2s}(u), \quad (2·24)$$

and it has been shown by Lommel (1868, (9), p. 40) that for $n\geqslant 1$ the sum of the series on the right-hand side of (2·24) is

$$\frac{u}{2}J_{n-1}(u).$$

Hence for $n\geqslant 1$, $$Y_n(u,u)=\frac{u}{2}J_{n-1}(u). \quad (2·25)$$

We also have, by lemma 7, that

$$\sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}Q_{2s}(u)=J_0(u)\cos u+J_1(u)\sin u. \quad (2·26)$$

Substituting from the last two equations into (2·23) we find that

$$L(u,u)=1-J_0(u)\cos u-J_1(u)\sin u. \quad (2·27)$$

(3) *The case* $v_0/u > 1$

In this case we find it convenient to split the range of integration into the two intervals $0 \leqslant v \leqslant u$ and $u < v \leqslant v_0$. We then have from (2·7), (2·10) and (2·27) that

$$L(u, v_0) = 1 - J_0(u)\cos u - J_1(u)\sin u + \frac{2}{u^2}\int_u^{v_0} v[U_1^2(u,v) + U_2^2(u,v)]\,dv. \qquad (2\cdot 28)$$

From lemma 1,

$$\int_u^{v_0} v[U_1^2(u,v) + U_2^2(u,v)]\,dv = \int_u^{v_0} \sum_{s=0}^{\infty} (-1)^s \left(\frac{u}{v}\right)^{2s+2} P_{1,2s}(v)\,v\,dv.$$

With the help of lemma 4 it can be shown that the series under the integral sign is uniformly convergent in the range of integration and can therefore be integrated term by term. We then find

$$\begin{aligned}
\frac{2}{u^2}\int_u^{v_0} v[U_1^2(u,v) + U_2^2(u,v)]\,dv &= 2\sum_{s=0}^{\infty}(-1)^s u^{2s}\int_u^{v_0} v^{-2s-1} P_{1,2s}(v)\,dv \\
&= \sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}u^{2s}\left[\frac{Q_{2s}(u)}{u^{2s}} - \frac{Q_{2s}(v_0)}{v_0^{2s}}\right] \quad \text{(by lemma 5)} \\
&= \sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}Q_{2s}(u) - \sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{u}{v_0}\right)^{2s} Q_{2s}(v_0) \\
&= J_0(u)\cos u + J_1(u)\sin u - \sum_{s=1}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{u}{v_0}\right)^{2s} Q_{2s}(v_0) \\
&\qquad \text{(by lemma 7).} \qquad (2\cdot 29)
\end{aligned}$$

Substituting from (2·29) into (2·28), we finally obtain for the case $v_0/u > 1$

$$L(u, v_0) = 1 - \sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{u}{v_0}\right)^{2s} Q_{2s}(v_0). \qquad (2\cdot 30)$$

(*b*) Before we discuss the general solution (2·22), (2·27) and (2·30) we shall derive expressions for the total illumination according to geometrical optics.

The rays of light which proceed from the region of the opening between circles of radii h and $h + dh$ centred on the axis cut the receiving plane in points of the annulus formed by the circles of radii r and $r + dr$ centred at O', where, to first order,

$$\frac{r}{\Delta f} = \frac{h}{f} \quad (h \leqslant R). \qquad (2\cdot 31)$$

The intensity I^* at P, according to geometrical optics, is then given by

$$\left.\begin{aligned}
I^*(u, v) &= \frac{2\pi h\,dh}{2\pi r\,dr}\frac{A^2}{f^2} \quad \text{when } r < \frac{R}{f}|\Delta f|, \\
&= 0 \quad \text{when } r > \frac{R}{f}|\Delta f|.
\end{aligned}\right\} \qquad (2\cdot 32)$$

From (2·31) and (2·32) it then follows on eliminating h, that

$$\left.\begin{aligned}
I^*(u, v) &= \left(\frac{A}{\Delta f}\right)^2 = \frac{4\gamma}{u^2} \quad \text{when } \left|\frac{v}{u}\right| < 1, \\
&= 0 \quad \text{when } \left|\frac{v}{u}\right| > 1.
\end{aligned}\right\} \qquad (2\cdot 33)$$

$$\left.\begin{aligned} \text{Hence} \qquad L^*(u, v_0) = \frac{1}{B}\int_0^{r_0}\int_0^{2\pi} I^*(u, v)\, r\, dr\, d\theta = \left(\frac{v_0}{u}\right)^2 &\text{ when } \left|\frac{v_0}{u}\right| < 1, \\ = 1 &\text{ when } \left|\frac{v_0}{u}\right| \geqslant 1. \end{aligned}\right\} \tag{2·34}$$

3. Numerical results and discussion of the solution

The general solution derived in § 2 was evaluated for

$$v = 0, 1, 2, \ldots, 15$$

$$\begin{Bmatrix} A \\ B \end{Bmatrix} = 0, 0{\cdot}05, 0{\cdot}10, \ldots, 1,$$

where $A = u/v$ and $B = v/u$. The use of the parameters A and B in place of u in tabulating Lommel and allied functions was suggested by Hopkins (1949). Although this method possesses some advantages it was not found very satisfactory owing to the resulting uneven spacing of u. Consequently an inadequate picture of the behaviour of L was obtained in some regions, and it was therefore found necessary to calculate a few additional values.

Use was made of the recurrence relation

$$\left.\begin{aligned} P_{0,2m+2}(v) &= 2J_0(v)\,J_{2m+2}(v) - P_{1,2m}(v), \\ \text{so that for } m \geqslant 1 \quad Q_{2m}(v) &= 2J_0(v)\,J_{2m}(v) + P_{1,2m}(v) - P_{1,2m-2}(v). \end{aligned}\right\} \tag{3·1}$$

In table 1 are given the numerical values of some of these polynomials.

Table 1

$2m$	$Q_{2m}(0)$	$Q_{2m}(1)$	$Q_{2m}(2)$	$Q_{2m}(3)$	$Q_{2m}(4)$	$Q_{2m}(5)$	$Q_{2m}(6)$	$Q_{2m}(7)$
0	+1·00000	+0·77917	+0·38274	+0·18259	+0·16209	+0·13885	+0·09925	+0·09007
2	0	−0·01378	−0·15038	−0·39448	−0·48300	−0·36504	−0·27223	−0·27020
4	—	−0·00019	−0·00826	−0·04563	−0·07103	+0·02781	+0·21678	+0·30393
6	—	—	−0·00021	−0·00245	−0·00500	+0·01818	+0·09045	+0·14444
8	—	—	—	−0·00008	−0·00021	+0·00243	+0·01488	+0·02829
10	—	—	—	—	−0·00001	+0·00017	+0·00142	+0·00328
12	—	—	—	—	—	+0·00001	+0·00009	+0·00026
14	—	—	—	—	—	—	—	+0·00002

$2m$	$Q_{2m}(8)$	$Q_{2m}(9)$	$Q_{2m}(10)$	$Q_{2m}(11)$	$Q_{2m}(12)$	$Q_{2m}(13)$	$Q_{2m}(14)$	$Q_{2m}(15)$
0	+0·08452	+0·06834	+0·06237	+0·06056	+0·05220	+0·04776	+0·04706	+0·04227
2	−0·24323	−0·19610	−0·18690	−0·17858	−0·15245	−0·14294	−0·14008	−0·12457
4	+0·26135	+0·24035	+0·26267	+0·24499	+0·21374	+0·21562	+0·20959	+0·18577
6	+0·07522	−0·06864	−0·13713	−0·13011	−0·15245	−0·19062	−0·18476	−0·17476
8	−0·00200	−0·09597	−0·16975	−0·13071	−0·03579	+0·00305	+0·00842	+0·04976
10	−0·00346	−0·03176	−0·06224	−0·03457	+0·06531	+0·14579	+0·13206	+0·08448
12	−0·00067	−0·00579	−0·01295	−0·00325	+0·04182	+0·09133	+0·07432	−0·01573
14	−0·00007	−0·00070	−0·00181	+0·00032	+0·01254	+0·02953	+0·02064	−0·03699
16	—	−0·00006	−0·00019	+0·00014	+0·00241	+0·00630	+0·00346	−0·01810
18	—	—	−0·00001	+0·00002	−0·00033	+0·00098	+0·00035	−0·00515
20	—	—	—	—	+0·00004	+0·00011	0	−0·00103
22	—	—	—	—	—	+0·00003	0	−0·00016
24	—	—	—	—	—	—	—	−0·00002

The series present in the expressions for L were found to converge very rapidly. This is due to the fact that since

$$|J_r(v)| \leqslant 1, \tag{3·2}$$

$$|Q_{2m}(v)| \leqslant \sum_{s=0}^{2m} |J_s(v)\,J_{2m-s}(v)| + \sum_{s=0}^{2m} |J_{s+1}(v)\,J_{2m+1-s}(v)| \leqslant 2(2m+1), \tag{3·3}$$

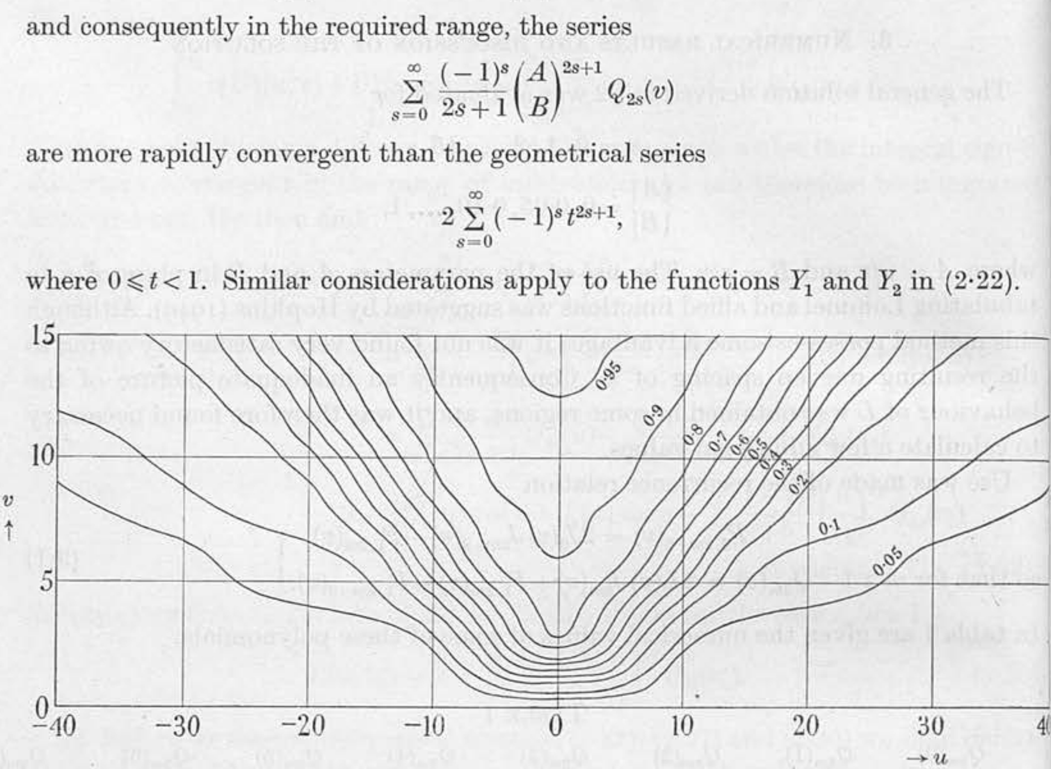
and consequently in the required range, the series

$$\sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1}\left(\frac{A}{B}\right)^{2s+1} Q_{2s}(v)$$

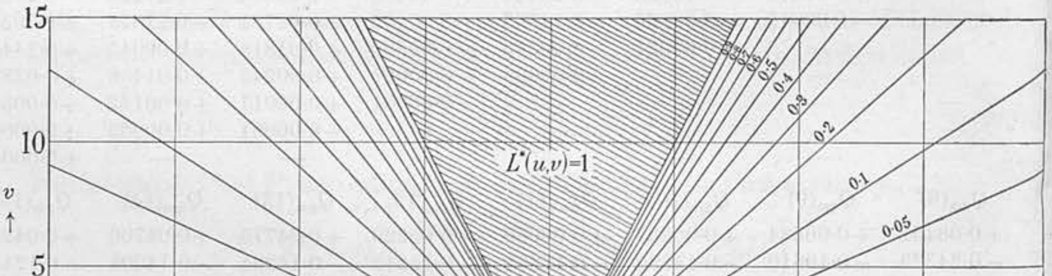
are more rapidly convergent than the geometrical series

$$2\sum_{s=0}^{\infty}(-1)^s t^{2s+1},$$

where $0 \leqslant t < 1$. Similar considerations apply to the functions Y_1 and Y_2 in (2·22).



(a). Contour lines of $L(u, v)$



(b). Contour lines of $L^*(u, v)$

FIGURE 3. Contour lines for the fraction of total illumination in circles centred on axis in receiving planes near focus, according to physical optics and according to geometrical optics. Throughout the shaded region $L^*(u, v) = 1$.

From the calculated values the contour lines of L were constructed for the range $-40 \leqslant u \leqslant 40$, $0 \leqslant v \leqslant 15$ by graphical interpolation. These are displayed in figure $3a$; the corresponding curves predicted by geometrical optics are shown in figure $3b$ for comparison. It is seen that in the immediate neighbourhood of the focus geometrical optics does not give a good approximation to the physical solution. With increasing departure from the focus the two solutions approach each other more and more closely, the level curves of L being in good qualitative agreement with those of L^* already at distances a few fringes away from the focus.

The behaviour of L and L^* in the planes $u = 0, 2\pi, 4\pi, 6\pi, 8\pi, 10\pi$ and 12π corresponding to defocusing of $0, \frac{1}{2}, 1, 1\frac{1}{2}, 2, 2\frac{1}{2}$ and 3 fringes respectively is shown in figures $4a$ and $4b$. In the special case when $u = 0$, the receiving plane coincides with the geometrical focal plane and L is then given by (2·30) with $u = 0$, viz. by

$$L(0, v) = 1 - Q_0(v) = 1 - J_0^2(v) - J_1^2(v) \tag{3·4}$$

in agreement with Rayleigh's formula (1·1).

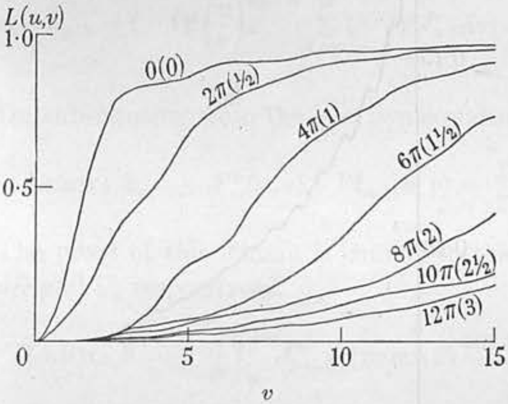


FIGURE $4a$. $L(u, v)$

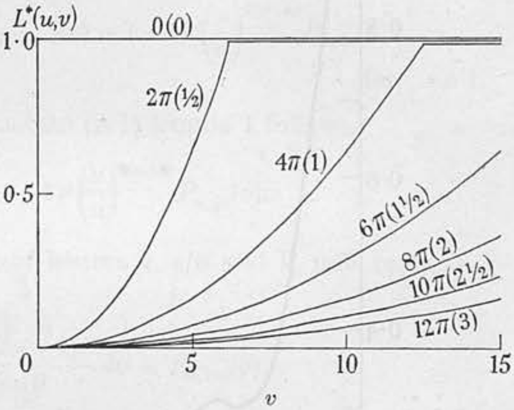


FIGURE $4b$. $L^*(u, v)$

The ratio of total illumination in circles centred on axis in receiving planes near focus, according to physical optics (figure $4a$) and according to geometrical optics (figure $4b$). The numbers along the curves are the values of u while the numbers in brackets denote defocusing in fringes. The horizontal parts of the curves $u = 2\pi$ and $u = 4\pi$ in figure $4b$ correspond to $L^* = 1$.

It is of interest to examine what fraction $\epsilon(u)$ of the total energy is present within the geometrical shadow. From (2·27) it follows that to a sufficient approximation

$$\epsilon(u) = 1 - L(u, u) = J_0(u)\cos u + J_1(u)\sin u, \tag{3·5}$$

where, as before, u specifies the position of the receiving plane. This function is displayed in figure 5. It is not a strictly decreasing function but it has maxima (apart from $u = 0$) when $J_1(u) = 0$ and minima when $\sin u = 0$ $(u \neq 0)$.

We note that for the particular case of diffraction of spherical waves at a circular aperture our solution illustrates the well-known theorem that geometrical optics may be regarded as the limiting case of physical optics as $\lambda \to 0$. For if we keep r and

Δf fixed and let $\lambda \to 0$, then $u \to \infty$ and $v \to \infty$, but u/v remains finite and it can be verified that in each of the equations (2·22), (2·27) and (2·30) all the terms except the first tend to zero. Consequently as $\lambda \to 0$,

$$L(u, v_0) \to \left(\frac{v_0}{u}\right)^2 \text{ when } \left|\frac{v_0}{u}\right| < 1,$$
$$\to 1 \text{ when } \left|\frac{v_0}{u}\right| \geqslant 1,$$

in agreement with the solution (2·34) of geometrical optics.

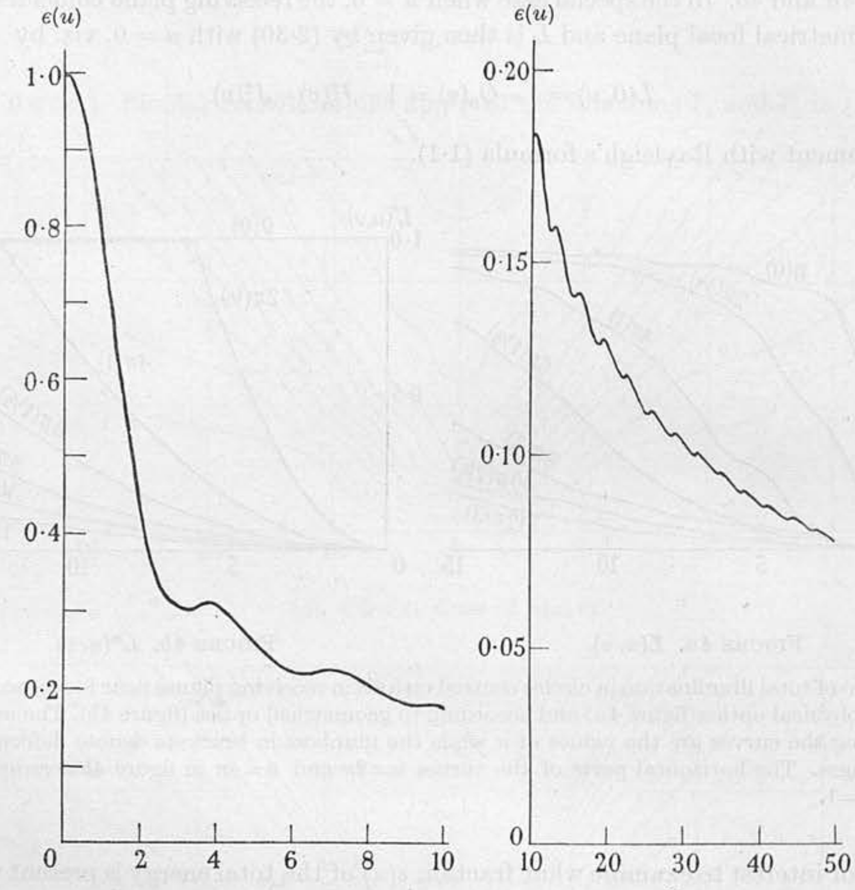


FIGURE 5. The fraction $\epsilon(u)$ of the total energy within the geometrical shadow.

APPENDIX. SOME MATHEMATICAL LEMMAS

The functions U, V, Y, W, P and Q which occur in the lemmas we now derive are defined by equations (2·1) to (2·5)

LEMMA 1. $$U_n^2(u,v) + U_{n+1}^2(u,v) = \sum_{s=0}^{\infty} (-1)^s \left(\frac{u}{v}\right)^{2(n+s)} P_{n,2s}(v).$$

Proof. Since the U series are absolutely convergent for all u and v, a series for $U_n^2+U_{n+1}^2$ can be obtained by performing the various operations term by term.

Let
$$\alpha_s = (-1)^s\left(\frac{u}{v}\right)^{n+2s} J_{n+2s}(v), \quad \beta_s = (-1)^s\left(\frac{u}{v}\right)^{n+1+2s} J_{n+1+2s}(v).$$

Then
$$U_n^2 = \sum_{s=0}^{\infty} a_s, \quad U_{n+1}^2 = \sum_{s=0}^{\infty} b_s,$$

where
$$a_s = \sum_{i=0}^{s} \alpha_i \alpha_{s-i} = (-1)^s\left(\frac{u}{v}\right)^{2(n+s)} \sum_{i=0}^{s} J_{n+2i}(v)\, J_{n+2s-2i}(v),$$

$$b_s = \sum_{i=0}^{s} \beta_i \beta_{s-i} = (-1)^s\left(\frac{u}{v}\right)^{2(n+1+s)} \sum_{i=0}^{s} J_{n+1+2i}(v)\, J_{n+1+2s-2i}(v).$$

Hence
$$U_n^2+U_{n+1}^2 = \sum_{s=0}^{\infty} a_s + \sum_{s=0}^{\infty} b_s = a_0 + \sum_{s=1}^{\infty} (a_s+b_{s-1}). \tag{A 1}$$

Now
$$a_0 = \left(\frac{u}{v}\right)^{2n} J_n^2(v) = \left(\frac{u}{v}\right)^{2n} P_{n,0}(v),$$

$$a_s+b_{s-1} = (-1)^s\left(\frac{u}{v}\right)^{2(n+s)} \sum_{i=0}^{2s} (-1)^i J_{n+i}(v)\, J_{n+2s-i}(v) = (-1)^s\left(\frac{u}{v}\right)^{2(n+s)} P_{n,2s}(v) \quad \text{for} \quad s \geqslant 1.$$

On substituting from the last two equations into (A 1) lemma 1 follows.

LEMMA 2.
$$V_n^2(u,v)+V_{n+1}^2(u,v) = \sum_{s=0}^{\infty} (-1)^s \left(\frac{v}{u}\right)^{2(n+s)} P_{n,2s}(v).$$

The proof of this lemma is similar to that of lemma 1, v/u and V_n now replacing u/v and U_n respectively.

LEMMA 3.
$$\frac{2}{\pi}\int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v\cos\theta)\,\frac{\cos(2m+1)\,\theta}{\cos\theta}\,d\theta = P_{n,2m}(v).$$

Proof. By definition,
$$P_{n,2m}(v) = \sum_{s=0}^{2m} (-1)^s J_{n+s}(v)\, J_{n+2m-s}(v).$$

From Watson (1922, 5·43 (1), p. 150),
$$J_{n+s}(v)\, J_{n+2m-s}(v) = \frac{2}{\pi}\int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v\cos\theta)\cos 2(m-s)\,\theta\, d\theta.$$

Hence
$$P_{n,2m}(v) = \sum_{s=0}^{2m} (-1)^s \frac{2}{\pi}\int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v\cos\theta)\cos 2(m-s)\,\theta\, d\theta$$
$$= \frac{2}{\pi}\int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v\cos\theta)\,\sigma_m(\theta)\,d\theta, \tag{A 2}$$

where
$$\sigma_m(\theta) = \sum_{s=0}^{2m} (-1)^s \cos 2(m-s)\,\theta$$
$$= \frac{\cos(2m+1)\,\theta}{\cos\theta}. \tag{A 3}$$

Lemma 3 now follows from (A 2) and (A 3).

LEMMA 4. $$| P_{n,2m}(v) | \leqslant \frac{2m+1}{(n+m)!} (\tfrac{1}{2}v)^{2(n+m)} \exp(\tfrac{1}{2}v^2).$$

Proof. Since $$P_{n,2m}(v) = \sum_{s=0}^{2m} (-1)^s J_{n+s}(v) J_{n+2m-s}(v),$$

it follows that $$| P_{n,2m}(v) | \leqslant \sum_{s=0}^{2m} | J_{n+s}(v) J_{n+2m-s}(v) |.$$

By Watson (1922, (4), p. 16) we have, for $r \geqslant 0$,
$$| J_r(v) | \leqslant \frac{(\frac{1}{2}v)^r}{r!} \exp(\tfrac{1}{4}v^2), \tag{A4}$$

so that
$$\begin{aligned} | P_{n,2m}(v) | &\leqslant \sum_{s=0}^{2m} \frac{(\frac{1}{2}v)^{n+s}}{(n+s)!} \exp(\tfrac{1}{4}v^2) \frac{(\frac{1}{2}v)^{n+2m-s}}{(n+2m-s)!} \exp(\tfrac{1}{4}v^2) \\ &= \left(\frac{v}{2}\right)^{2(n+m)} \exp(\tfrac{1}{2}v^2) \sum_{s=0}^{2m} \frac{1}{(n+s)!\,(n+2m-s)!} \\ &\leqslant \frac{2m+1}{(n+m)!} \left(\frac{v}{2}\right)^{2(n+m)} \exp(\tfrac{1}{2}v^2). \end{aligned}$$

LEMMA 5.
$$\int^v v^{-(2n+2m+1)} P_{n+1,2m}(v)\, dv = -\frac{1}{2(2n+2m+1)} v^{-2n-2m} [P_{n,2m}(v) + P_{n+1,2m}(v)].$$

Proof. By definition
$$\begin{aligned} \int^v v^{-(2n+2m+1)} P_{n+1,2m}(v)\, dv &= \int^v v^{-(2n+2m+1)} \left[\sum_{s=0}^{2m} (-1)^s J_{n+1+s}(v) J_{n+1+2m-s}(v) \right] dv \\ &= \sum_{s=0}^{2m} (-1)^s \int^v v^{-(2n+2m+1)} J_{n+1+s}(v) J_{n+1+2m-s}(v)\, dv. \end{aligned}$$

The integral under the summation sign can be evaluated by a lemma on cylinder functions [see, for example, Watson (1922, (1), p. 136)]. The above expression then becomes
$$\begin{aligned} &\sum_{s=0}^{2m} \frac{(-1)^{s+1}}{2(2n+2m+1)} v^{-(2n+2m)} [J_{n+s}(v) J_{n+2m-s}(v) + J_{n+s+1}(v) J_{n+1+2m-s}(v)] \\ &\qquad = -\frac{1}{2(2n+2m+1)} v^{-(2n+2m)} [P_{n,2m}(v) + P_{n+1,2m}(v)]. \end{aligned}$$

LEMMA 6.
$$\int^v v^{2n+2m+1} P_{n,2m}(v)\, dv = \frac{1}{2(2n+2m+1)} v^{2n+2m+2} [P_{n,2m}(v) + P_{n+1,2m}(v)].$$

This lemma can be proved by a similar argument as in lemma 5, using Watson (1922, (2), p. 136) in place of Watson (1922, (1), p. 136).

LEMMA 7. $$\sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) = J_0(u) \cos u + J_1(u) \sin u.$$

Proof. By definition of Q_{2s} and from lemma 3 it follows that
$$Q_{2s}(u) = \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} [J_{2s}(2u\cos\theta) + J_{2s+2}(2u\cos\theta)] \frac{\cos(2s+1)\theta}{\cos\theta}\, d\theta,$$

and, by applying the addition formula for Bessel functions, we then find that

$$Q_{2s}(u) = \frac{2}{\pi}\int^{\frac{1}{2}\pi} \frac{2s+1}{u\cos^2\theta} J_{2s+1}(2u\cos\theta)\cos(2s+1)\theta\, d\theta. \tag{A 5}$$

Hence $$\sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) = \frac{2}{\pi}\sum_{s=0}^{\infty}\int_0^{\frac{1}{2}\pi} \frac{(-1)^s}{u\cos^2\theta} J_{2s+1}(2u\cos\theta)\cos(2s+1)\theta\, d\theta. \tag{A 6}$$

Interchanging the summation and integration sign, which is justified since the series is uniformly convergent in $0 \leqslant \theta \leqslant \frac{1}{2}\pi$, and, using the well-known result of Jacobi [see, for example, Watson (1922, (4), p. 22)] that

$$\sum_{s=0}^{\infty} (-1)^s J_{2s+1}(z)\cos(2s+1)\theta = \tfrac{1}{2}\sin(z\cos\theta),$$

(A 6) becomes $$\sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) = \frac{1}{\pi}\int_0^{\frac{1}{2}\pi} \frac{\sin(2u\cos^2\theta)}{u\cos^2\theta} d\theta. \tag{A 7}$$

Let $$F(u) = \frac{1}{\pi}\int_0^{\frac{1}{2}\pi} \frac{\sin(2u\cos^2\theta)}{u\cos^2\theta} d\theta. \tag{A 8}$$

Then $$\frac{d}{du}[uF(u)] = \frac{2}{\pi}\int_0^{\frac{1}{2}\pi} \cos(2u\cos^2\theta)\, d\theta$$

$$= \frac{2\cos u}{\pi}\int_0^{\frac{1}{2}\pi} \cos(u\cos 2\theta)\, d\theta - \frac{2\sin u}{\pi}\int_0^{\frac{1}{2}\pi} \sin(u\cos 2\theta)\, d\theta.$$

Now $$\frac{2}{\pi}\int_0^{\frac{1}{2}\pi} \cos(u\cos 2\theta)\, d\theta = J_0(u),$$

$$\frac{2}{\pi}\int_0^{\frac{1}{2}\pi} \sin(u\cos 2\theta)\, d\theta = 0,$$

so that $$\frac{d}{du}[uF(u)] = J_0(u)\cos u.$$

Integrating several times by parts and using the recurrence relations for Bessel functions or otherwise it can easily be verified that the solution of this differential equation subject to the boundary condition $F(0) = 1$ [which is imposed by (A 8)] is

$$F(u) = J_0(u)\cos u + J_1(u)\sin u. \tag{A 9}$$

Lemma 7 now follows from (A 7), (A 8) and (A 9).

LEMMA 8. $$\sum_{s=0}^{\infty} (-1)^s V_{n+2s}(v) = W_n(u, v).$$

Proof. $$\sum_{s=0}^{\infty} (-1)^s V_{n+2s}(v) = \sum_{s=0}^{\infty} (-1)^s \sum_{p=0}^{\infty} (-1)^p \left(\frac{v}{u}\right)^{n+2s+2p} J_{n+2s+2p}(v)$$

$$= \sum_{s=0}^{\infty}\sum_{p=0}^{\infty} (-1)^{s+p} \left(\frac{v}{u}\right)^{n+2s+2p} J_{n+2s+2p}(v).$$

Collecting terms of the same order in J and arranging the series in ascending order lemma 8 follows. The grouping of terms is justified since the series is absolutely convergent.

LEMMA 9.

$$\int_0^{v_0} vV_n(u,v)\exp(-iv^2/2u)\,dv = u[W_{n+1}(u,v_0)+iW_{n+2}(u,v_0)]\exp(-iv_0^2/2u).$$

Proof.

$$\int_0^{v_0} vV_n(u,v)\exp(-iv^2/2u)\,dv = \int_0^{v_0}\sum_{s=0}^{\infty}(-1)^s v\left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v)\exp(-iv^2/2u)\,dv$$

$$= \sum_{s=0}^{\infty}\int_0^{v_0}\frac{(-1)^s}{u^{n+2s}}v^{n+2s+1}J_{n+2s}(v)\exp(-iv^2/2u)\,dv. \quad \text{(A 10)}$$

The interchanging of the summation and integration sign above is justified, since by (A 4) the series under the integral sign is dominated by the series

$$\sum_{s=0}^{\infty}\frac{v_0\exp\frac{1}{4}v_0^2}{(n+2s)!}\left(\frac{v_0^2}{2u}\right)^{n+2s},$$

and is therefore uniformly convergent in the range of integration. The integrals in (A 10) can be evaluated with the help of a result due to Lommel (see, for example, Walker (1904 (10), p. 131) which states, that for $\nu>0$, $l>0$,

$$\int_0^r (lx)^\nu J_{\nu-1}(lx)\exp(-ikx^2/2)\,dx = \frac{l^{2\nu-1}}{k^\nu}[U_\nu(kr^2,lr)+iU_{\nu+1}(kr^2,lr)]\exp(-ikr^2/2)$$

Setting $l=1$, $x=v$, $r=v_0$, $\nu=n+2s+1$, $k=1/u$ and multiplying by $(-1)^s/u^{n+2s}$, we find

$$\frac{(-1)^s}{u^{n+2s}}\int_0^{v_0} v^{n+2s+1}J_{n+2s}(v)\exp(-iv^2/2u)\,dv$$

$$= (-1)^s u[V_{n+2s+1}(u,v_0)+iV_{n+2s+2}(u,v_0)]\exp(-iv_0^2/2u). \quad \text{(A 11)}$$

Lemma 9 now follows on substituting from (A 11) into (A 10) and using Lemma 8.

In conclusion, it is a pleasure to thank Dr E. H. Linfoot for much encouragement and for helpful discussions in the course of this work. I also wish to thank Dr H. H. Hopkins, Mr F. Ursell and Mr P. A. Wayman for useful suggestions. Finally, I wish to acknowledge my indebtedness to the Cambridge University Mathematical Laboratory for assistance with some of the computations and to Miss C. M. Munford for the skill and patience with which she carried them out.

REFERENCES

Airy, G. B. 1835 *Trans. Camb. Phil. Soc.* **5**, 283.
Hopkins, H. H. 1949 *Proc. Phys. Soc.* **62** B, 22.
Lommel, E. 1868 *Studien über die Besselschen Functionen*. Leipzig: Teubner.
Lommel, W. 1885 *Abh. Bayer. Akad. Wiss.* **15**, 229. A more accessible, though condensed account of Lommel's treatment is given in: Gray, A., Mathews, G. B. & MacRobert, T. M. 1931 *A treatise on Bessel functions*, 2nd ed. London: Macmillan.
Nijboer, B. R. A. 1942 Thesis, Groningen.
Rayleigh, Lord 1881 *Phil. Mag.* **11**, 214. Also *Collected papers*, **1**, 513; **3**, 91.
Struve, H. 1886 *Mém. Acad. Sci. St-Pétersb.* (7), **34**, no. 5, 1–15.
Walker, J. 1904 *The analytic theory of light*. Cambridge University Press.
Watson, G. N. 1922 *A treatise on the theory of Bessel functions*. Cambridge University Press.
Zernike, F. & Nijboer, B. R. A. 1949 *La théorie des images optiques*. Éditions de la Revue d'Optique, p. 227.

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, CAMBRIDGE
(BROOKE CRUTCHLEY, UNIVERSITY PRINTER)

CONTRIBUTIONS FROM THE CAMBRIDGE OBSERVATORIES

No. 10

ON TELESCOPIC STAR IMAGES

by

E. H. LINFOOT *and* E. WOLF

PRINTED IN ENGLAND BY TAYLOR & FRANCIS, LTD., LONDON

Reprinted from the MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY,
Vol. 112, *No.* 4, 1952

ON TELESCOPIC STAR IMAGES

E. H. Linfoot and E. Wolf

(Received 1951 December 5)

Summary

Diffraction theory is applied to discuss the properties of the polychromatic, three-dimensional star images formed by refracting doublet objectives. The effects of the secondary spectrum on the light distribution in the diffraction image and on the position of best focus are examined quantitatively in two special cases: (A) an $f/15$ refractor of $24\frac{7}{8}$ inches aperture, (B) a similar refractor of one-third the linear dimensions. The image formed by an $f/15$ reflector working without central obstruction (Case R) is analysed for comparison purposes.

---

1. *Introduction*

The first explanation of the main features of the stellar images seen in a good telescope was given by Airy (1835) (**1**). He showed that, as a direct consequence of the undulatory theory of light, a stellar image should exhibit the well-known appearance of a bright central dot (the "star disc") surrounded by faint concentric rings. Airy assumed the telescope to be of circular aperture and free from aberrations or central obstruction, the effects of air tremor to be negligible, and the light to be monochromatic. Thus his analysis, of fundamental importance as the starting point for later work, was far from providing an adequate treatment of the formation of star images in the large astronomical telescopes which have been developed since his day.

Rayleigh (**2**) in 1880 carried the theory forward by investigating the case where small amounts of primary spherical aberration are present; like Airy, he took the light to be monochromatic. He did not discuss the image in detail, but showed that, with amounts of spherical aberration not exceeding one-quarter of a wave-length, the intensity at the centre of the star disk at paraxial focus was diminished by not more than 20 per cent.

Lommel (1885) (**3**)*, in a paper of great power and scope, discussed in detail the properties of the out-of-focus, aberration-free monochromatic image of a point source by a circular objective and confirmed the predicted appearances experimentally, using a monochromatic light source. The numerical part of his work was only carried through for images fairly close to focus. A more convenient method of calculating the images far from focus was developed thirteen years later by K. Schwarzschild (**4**), who used asymptotic expressions for the intensities.

Almost at the same time as Lommel, H. Struve (1886) (**5**) published an independent and very similar analysis; he did not work out the numerical consequences in such detail, but he gave valuable approximate expressions for the intensity near the edge of the geometrical shadow, where the Lommel–Struve expansions are rather slowly convergent.

* The date of this paper is sometimes given as 1884 or 1886. A later paper (Lommel, *ibid.*, 531–664, 1886) deals with diffraction by rectangular apertures, obstacles and slits.

These investigations still left a rather wide gap between the theory of telescopic star images and their observed appearance. Not only did they leave out of account the effects of atmospheric tremor, but they did not discuss the powerful effects of the secondary spectrum on star images in large refracting telescopes. However, by 1886, Lommel's and Struve's work had provided a means of predicting the three-dimensional light distribution near the focus of an error-free monochromatic image.

No actual diagram of this distribution was published until Berek's paper (**6**) appeared in 1926. A new and more accurate diagram, based on a different expansion of the Huygens–Kirchhoff diffraction integral from those obtained by Lommel and Struve, was recently worked out by Zernike and Nijboer (**7**); it is reproduced as Fig. 2 below.

The delay in working out the light-distribution diagram was the more surprising since Dennis Taylor, in his stimulating and entertaining paper (**8**) on the secondary colour aberrations of visual telescopes, had in 1894 made abundantly clear the importance of a knowledge of this distribution to the makers and users of astronomical refractors.

It was not until 1919 that the subject was carried further by A. E. Conrady (**9**), who worked out in some detail the light distribution in slightly defocused images, and in images possessing small amounts of spherical aberration, the receiving plane in the latter case being placed at paraxial focus, at marginal focus and midway between the two. So far as defocusing was concerned, his results did not go beyond Lommel's. His computations were later extended by A. Buxton (**10**) and L. C. Martin (**11**).

The next substantial contributions to the theory were made in 1925 by Picht (**12**) and by Steward (**13**). Both considered the effects of Seidel aberrations on the light distribution in monochromatic diffraction images, and this subject has since been fairly extensively discussed during the past 25 years. However, apart from the computed figure of Zernike and Nijboer already mentioned, the only work published during this period which has an immediate bearing on the topic of star images in visual telescopes is Steward's investigation, in the paper just referred to, of the effect on the images of a central obstruction of the telescope aperture.

In the present paper we take up once more the old question of star images in visual refracting telescopes and develop the theory to the point where a quantitative estimate can be obtained of the light lost from the central part of the image by the combined effects of diffraction and of the secondary spectrum. Some comparison results for reflectors are also obtained. A preliminary survey of the problem showed that to carry out this programme on the basis of the existing mathematical literature would involve a very large amount of computational labour, and the attempt was therefore postponed until an extension of Lommel's and Struve's analysis by one of us (Wolf (**14**)) had made the work easier.

The part played by atmospheric tremor is very different in the visual and in the photographic use of telescopes. In the former, the rare moments of best "seeing" are seized upon by the eye, and eye and memory work together to build up a picture from what has been glimpsed. In the latter, the photographic plate records indiscriminately whatever reaches it during the exposure time and consequently it is the tremor disk—or rather its average over the long exposure time—which is recorded on the plate.

It follows that a study of the effect of atmospheric tremor is an essential part of any adequate theory of the photographic star images formed by large astronomical telescopes, while a theory of visual star images *as glimpsed during moments of nearly perfect seeing* may, and indeed should, be developed without taking account of the effects of tremor. Accordingly, we have not attempted to incorporate a theory of tremor disks into the present analysis, although we do include some results which could be used to obtain a rough estimate of the space-penetrating power of large refractors and reflectors under conditions of average seeing.

2. *Mathematical Preliminaries*

We begin with a brief account of some known results (Lommel (**3**), H. Struve (**5**), Wolf (**14**)) on monochromatic light distribution in space near the focus of converging spherical waves issuing from a circular aperture. The formulae derived in (**3**) and (**5**) concern intensity distribution; those in (**14**) the total illumination within an arbitrarily given circle about the origin in the receiving plane. The intensity formulae form the basis of the later work and because the handling of the approximations and error terms by Lommel and Struve was not quite precise enough for the purposes of the present investigation, they are rederived here by a more accurate discussion.

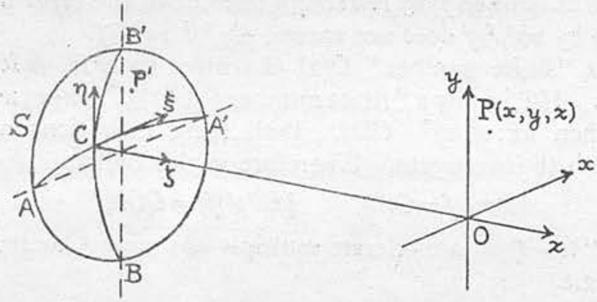


FIG. 1.

In Fig. 1, ABA′B′ represents a circular aperture through which issues a train of converging spherical waves of wave-length λ. $2a = \text{AA}'$ is the diameter of this aperture, C the pole of the wave surface S which momentarily fills it, O the centre of curvature of S. We call CO the axis of the wave train and set $\text{CO} = f$.

Ox, Oy, Oz are axes of Cartesian coordinates (x, y, z) in the space near O; Cξ, Cη, Cζ axes of Cartesian coordinates (ξ, η, ζ) in the space near C. It is assumed throughout that $a/f \ll 1$.

By Huygens' principle, the complex displacement at $\text{P}(x, y, z)$, near O which results from waves of unit amplitude on S is

$$D_\lambda(\text{P}) = \frac{i}{\lambda} e^{ikf} \iint_S \frac{e^{-iks}}{s}\, dS, \tag{2.1}$$

where s denotes the distance of P from the element dS located at $\text{P}'(\xi, \eta, \zeta)$, on the wave front filling the aperture and $k = 2\pi/\lambda$. Write

$$\left.\begin{aligned} \xi &= a\rho\cos\phi, & x &= r\cos\psi, \\ \eta &= a\rho\sin\phi, & y &= r\sin\psi. \end{aligned}\right\} \tag{2.2}$$

Then

$$dS = a^2 \rho\, d\rho\, d\phi, \tag{2.3}$$

with an error which in the case of an $f/15$ telescope nowhere exceeds one part in 1000, and hence

$$D_\lambda(\mathrm{P}) = \frac{ia^2}{\lambda} e^{ikf} \int_0^1 \int_0^{2\pi} \frac{e^{-iks}}{s} \rho \, d\rho \, d\phi. \tag{2.4}$$

Now on the surface S

$$\begin{aligned} \zeta &= f - \sqrt{(f^2 - a^2\rho^2)} = f - f\left\{1 - \frac{1}{2}\frac{a^2\rho^2}{f^2} - \frac{1}{8}\frac{a^4\rho^4}{f^4} - \dots\right\} \\ &= \frac{a^2\rho^2}{2f} + \frac{a^4\rho^4}{8f^3} + \dots\dots \end{aligned} \tag{2.5}$$

Therefore

$$\begin{aligned} s^2 &= \mathrm{PP}'^2 = (x-\xi)^2 + (y-\eta)^2 + (z-\zeta+f)^2 \\ &= (x^2+y^2+z^2) + (\xi^2+\eta^2+\zeta^2) - 2(x\xi + y\eta + z\zeta) + 2f(z-\zeta) + f^2 \\ &= [(f+z)^2 + r^2] + [(f-\zeta)^2 + a^2\rho^2] - f^2 - 2ar\rho\cos(\phi-\psi) - 2z\zeta \\ &= \mathrm{CP}^2 + \mathrm{OP}'^2 - f^2 - 2ar\rho\cos(\phi-\psi) - 2z\zeta \\ &= R'^2 - 2ar\rho\cos(\phi-\psi) - 2z\zeta \end{aligned} \tag{2.6}$$

(where R' is written for CP)

$$= R'^2 - 2ar\rho\cos(\phi-\psi) - z\left(\frac{a^2\rho^2}{f} + \frac{a^4\rho^4}{4f^3} + \dots\right). \tag{2.7}$$

When $a/f = \frac{1}{30}$ (as it is in an $f/15$ refracting telescope), the error in replacing the last term of (2.7) by $za^2\rho^2/f$ does not exceed $\frac{1}{3600}$ of $za^2\rho^2/f$.

If P is only a "finite number" $O(1)$ of fringes away from focus laterally*, then $ar\rho/f = O(\lambda)$. If P is only a "finite number" $O(1)$ of fringes away from focus longitudinally, then $az \,.\, a\rho/2f^2 = O(\lambda)$. Both these conditions are satisfied in the region we wish to investigate. Therefore in this region

$$2ar\rho/f = O(\lambda), \qquad \tfrac{1}{2}za^2\rho^2/f^2 = O(\lambda), \tag{2.8}$$

where O means "less than a moderate multiple of", say 5 or 10.

From (2.7), (2.8)

$$\begin{aligned} s &= R'\left[1 - \frac{2ar\rho}{R'^2}\cos(\phi-\psi) - \frac{z}{R'^2}\left(\frac{a^2\rho^2}{f} + \frac{a^4\rho^4}{4f^3} + \dots\right)\right]^{1/2} \\ &= R' - \frac{ar\rho}{R'}\cos(\phi-\psi) - \frac{za^2\rho^2}{2fR'} + O\left(\lambda\frac{a^2}{f^2}\right) + O\left(\frac{\lambda^2}{f}\right). \end{aligned} \tag{2.9}$$

The term $O(\lambda^2/f)$ is negligible since $\lambda \simeq 2 \times 10^{-5}$ inch and f is not small. The term $O(\lambda a^2/f^2)$ is $O(\lambda/900)$ in an $f/15$ pencil and so is negligible in a refracting telescope when P is only $O(1)$ fringes away from focus.

Therefore in investigating the value of (2.4) in the present problem we may set

$$s = R' - \frac{ar\rho}{R'}\cos(\phi-\psi) - \frac{za^2\rho^2}{2fR'},$$

$$ks = kR' - f/R'[v\rho\cos(\phi-\psi) + \tfrac{1}{2}u\rho^2],$$

where the new variables u, v are defined by the equations

$$u = \frac{ka^2z}{f^2}, \qquad v = \frac{kar}{f}. \tag{2.10}$$

In physical terms, $u/4\pi$ is the number of fringes of defocusing and v/π the number of fringes of lateral displacement of P relative to O. We note that $|v/u| \lesssim 1$

* That is to say, if the displacement of P from O would correspond to the appearance of only a "finite number" $O(1)$ of fringes on the surface S seen under test in an interferometer. $ar/f = \lambda/2$ gives "one fringe of lateral displacement"; $za^2/f^2 = 2\lambda$ gives "one fringe of defocusing".

according as P lies in the geometrical cone of rays or in the geometrical shadows. From (2.4), on substituting for ks and noting that $R' = f + O(\lambda)$ in the region where $u/4\pi$, v/π are both $O(1)$, we now obtain the approximate formula

$$D_\lambda(\mathrm{P}) = \frac{ia^2}{\lambda f} \exp[ik(f-R')] \int_0^1 \int_0^{2\pi} \exp\{i[\tfrac{1}{2}u\rho^2 + v\rho\cos(\phi-\psi)]\}\rho\, d\rho\, d\phi$$
$$= \frac{2\pi i a^2}{\lambda f} \exp[ik(f-R')] \int_0^1 \exp(\tfrac{1}{2}iu\rho^2) J_0(v\rho)\rho\, d\rho. \tag{2.11}$$

The integral on the right of (2.11) can be evaluated in terms of the functions

$$U_n(u, v) = \sum_{m=0}^{\infty} (-1)^m \left(\frac{u}{v}\right)^{n+2m} J_{n+2m}(v) \tag{2.12}$$

introduced by Lommel for this purpose; in fact (see Watson, "Bessel Functions", p. 541)

$$2\int_0^1 J_0(v\rho)\exp(\tfrac{1}{2}iu\rho^2)\rho\, d\rho = C(u, v) + iS(u, v), \tag{2.13}$$

where

$$\left.\begin{aligned} C(u, v) &= \frac{\cos\frac{1}{2}u}{\frac{1}{2}u} U_1(u, v) + \frac{\sin\frac{1}{2}u}{\frac{1}{2}u} U_2(u, v), \\ S(u, v) &= \frac{\sin\frac{1}{2}u}{\frac{1}{2}u} U_1(u, v) - \frac{\cos\frac{1}{2}u}{\frac{1}{2}u} U_2(u, v). \end{aligned}\right\} \tag{2.14}$$

(2.11) therefore gives

$$D_\lambda(\mathrm{P}) = \frac{\pi a^2}{\lambda f} \exp\left\{ik\left[(f-R') + \chi(u, v) + \frac{\pi}{2}\right]\right\} \sqrt{(C^2+S^2)}, \tag{2.15}$$

where C, S are written for $C(u, v)$, $S(u, v)$ respectively,

$$\cos\chi = \frac{C}{\sqrt{(C^2+S^2)}}, \qquad \sin\chi = \frac{S}{\sqrt{(C^2+S^2)}} \tag{2.16}$$

and

$$\sqrt{(C^2+S^2)} = \frac{2}{u}\sqrt{\{U_1^2(u, v) + U_2^2(u, v)\}}. \tag{2.17}$$

The intensity $I_\lambda(z, r) = I(u, v)$ at P is then given by the equation

$$\begin{aligned} I_\lambda(z, r) &= |D_\lambda(\mathrm{P})|^2 \\ &= \frac{4\pi^2 a^4}{\lambda^2 f^2}(C^2+S^2) \\ &= \frac{4\pi^2 a^4}{\lambda^2 f^2}\frac{1}{u^2}[U_1^2(u, v) + U_2^2(u, v)]. \end{aligned} \tag{2.18}$$

(2.18) is valid, subject to the limitations already imposed by our approximations, for all u, v; but it is only convenient for computation when $|v/u| > 1$. When $|v/u| < 1$ it may, as Lommel showed, be replaced with advantage by the equivalent formula*

$$I_\lambda(z, r) = \frac{4\pi^2 a^4}{\lambda^2 f^2}\frac{1}{u^2}\left[1 + V_0^2(u, v) + V_1^2(u, v) - 2V_0(u, v)\cos\left\{\frac{1}{2}\left(u + \frac{v^2}{u}\right)\right\} - 2V_1(u, v)\sin\left\{\frac{1}{2}\left(u + \frac{v^2}{u}\right)\right\}\right], \tag{2.19}$$

* When $|v/u| = 1$, (2.18) and (2.19) reduce to

$$I_\lambda(z, r) = \frac{\pi^2 a^4}{\lambda^2 f^2}\frac{1}{u^2}[J_0^2(u) - 2J_0(u)\cos u + 1].$$

where

$$V_n(u,v)=\sum_{m=0}^{\infty}(-1)^m\left(\frac{v}{u}\right)^{n+2m}J_{n+2m}(v). \tag{2.20}$$

Lommel used in his argument slightly different approximations from those made in the above modernized version; for example, his parameters specifying the position of P are not strictly identical with ours. But the final formulae (2.18), (2.19) are the same.

In the geometrical focal plane, $u=0$ and (2.18) reduces to

$$I_\lambda(0,r)=\frac{4\pi^2a^4}{\lambda^2f^2}\left(\frac{2J_1(v)}{v}\right)^2, \tag{2.21}$$

in agreement with Airy (**1**). On the axis, $v=0$ and (2.19) gives

$$I_\lambda(z,0)=\frac{4\pi^2a^4}{\lambda^2f^2}\left(\frac{\sin\frac{1}{4}u}{\frac{1}{4}u}\right)^2. \tag{2.22}$$

Lommel's equations (2.18), (2.19) for the intensity distribution in space near focus formed the starting point for the derivation by one of us (Wolf (**14**)) of expressions, needed below, for the fraction of the total illumination L which falls inside a given small circle about the (x,y)-origin in the receiving plane $z=$ constant. We define

$$L_\lambda(z,r_0)=\frac{1}{\pi a^2}\int_0^{r_0}\int_0^{2\pi}I_\lambda(z,r)r\,dr\,d\phi. \tag{2.23}$$

To the order of accuracy here in question, $L_\lambda(z,r_0)$ measures the fraction of the total energy issuing from the aperture which reaches the circle $r\leq r_0$ in the given receiving plane. Then

$$\begin{aligned}L_\lambda(z,r)&=\left(\frac{v}{u}\right)^2\left[1+\sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{v}{u}\right)^{2s}Q_{2s}(v)\right]\\&\quad-\frac{4}{u}\left[Y_1(u,v)\cos\left\{\frac{1}{2}\left(u+\frac{v^2}{u}\right)\right\}+Y_2(u,v)\sin\left\{\frac{1}{2}\left(u+\frac{v^2}{u}\right)\right\}\right] \qquad (2.24)\\&\qquad\qquad(|v/u|\leq 1)\\&=1-\sum_{s=0}^{\infty}\frac{(-1)^s}{2s+1}\left(\frac{u}{v}\right)^{2s}Q_{2s}(v),\qquad(|v/u|\geq 1), \qquad (2.25)\end{aligned}$$

where

$$Y_n(u,v)=\sum_{s=0}^{\infty}(-1)^s(n+2s)\left(\frac{v}{u}\right)^{n+2s}J_{n+2s}(v), \tag{2.26}$$

$$Q_{2s}(v)=\sum_{i=0}^{2s}(-1)^i[J_i(v)J_{2s-i}(v)+J_{i+1}(v)J_{2s+1-i}(v)]. \tag{2.27}$$

When $u=0$, (2.25) reduces to Rayleigh's formula

$$L_\lambda(0,r)=1-Q_0(v)=1-J_0^2(v)-J_1^2(v). \tag{2.28}$$

When $|v/u|=1$, (2.24) and (2.25) reduce to $L_\lambda(z,r)=1-J_0(u)\cos u-J_1(u)\sin u$. Figs. 2 and 3 show the distribution of $I_\lambda(z,r)$ and $L_\lambda(z,r)$ near focus in each meridional plane. In Fig. 3 the curves $L_\lambda(z,r)=$ const., or $L(u,v)=$ const. can be regarded as analogues in a certain sense of the rays of the geometrical theory. Their form near the geometrical focus $(u,v)=(0,0)$ agrees well with that postulated by Dennis Taylor (**8**) on experimental grounds for the "cone" of light near focus. The comparison is rather rough and ready, since Taylor's observations were made in polychromatic light. Nevertheless, his value of just below $\pm 0{\cdot}2$ mm for the permissible focal tolerance of an $f/15$ pencil is in good accordance with Figs. 2 and 3.

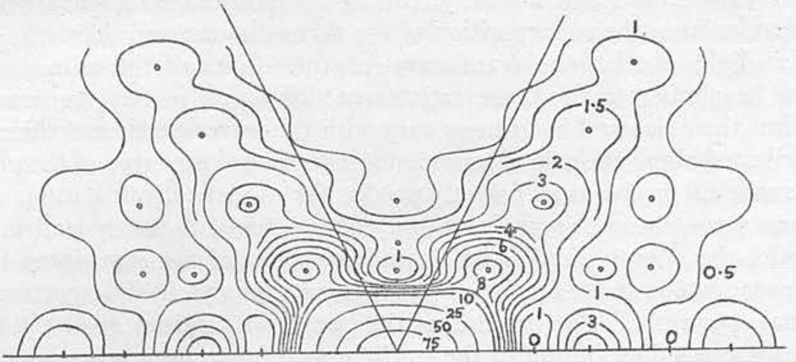


FIG. 2.—*Isophotes of an aberration-free pencil near focus.* (*Contour lines of* $I(u, v) = I_\lambda(z, r)$.) *The figure covers approximately the region* $-35 \leq u \leq 35$, $0 \leq v \leq 15$. *The straight lines show the boundary of the geometrical shadow. The numbers give intensities as a percentage of the intensity at focus. Axial maxima and minima are indicated by short strokes; others by small circles.* (*After Zernike and Nijboer.*)

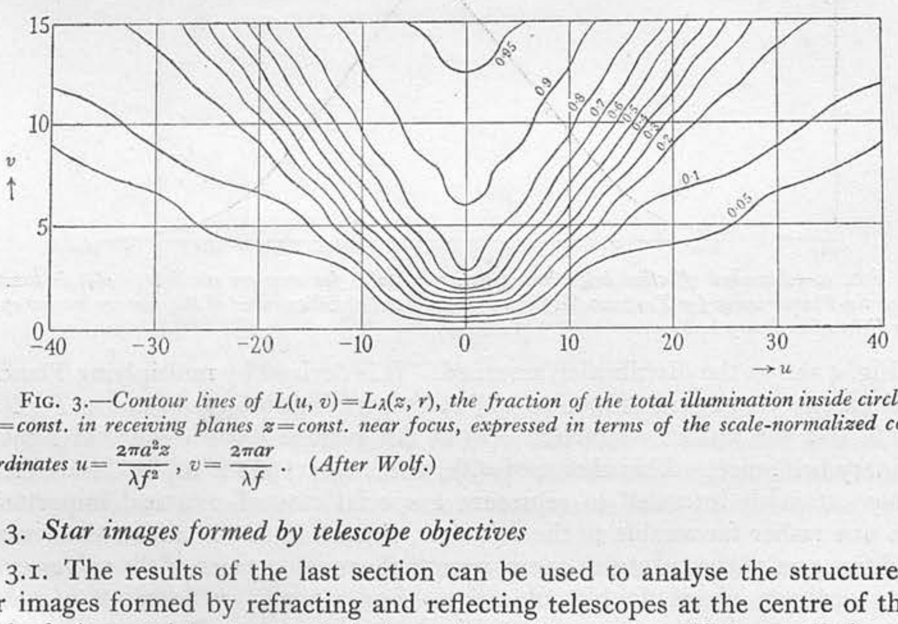


FIG. 3.—*Contour lines of* $L(u, v) = L_\lambda(z, r)$, *the fraction of the total illumination inside circles* r = *const. in receiving planes* z = *const. near focus, expressed in terms of the scale-normalized co-ordinates* $u = \dfrac{2\pi a^2 z}{\lambda f^2}$, $v = \dfrac{2\pi a r}{\lambda f}$. (*After Wolf.*)

3. *Star images formed by telescope objectives*

3.1. The results of the last section can be used to analyse the structure of star images formed by refracting and reflecting telescopes at the centre of their field of view and to draw conclusions about the relative efficiency of the two types of instruments under conditions of perfect "seeing".

In a refracting telescope with a well-figured doublet objective achromatized for visual use, the only appreciable optical error affecting the axial images is the secondary spectrum, the effect of which is to bring the light of each separate wave-length to a separate focus.

This focal spread leads to a rather serious loss of light in the central diffraction disks of star images formed by large refractors; Dennis Taylor (**8**) estimated the loss as high as 42 per cent in a 24-inch refractor of 30 feet focal length. But this estimate rested on too primitive a theoretical foundation to be quantitatively reliable, and an ingenious attempt two years later by the same author

(15) to investigate the problem experimentally likewise failed to give more than a very rough estimate of the light loss.

To obtain a quantitative description of the image, we need to analyse, in terms of wave-length and intensity, the light distribution in space near focus which results from the superposition of the three-dimensional diffraction images formed in light of all the relevant wave-lengths. Each of these images can be taken to be similar to the "aberration-free diffraction image" represented in Fig. 2, but their size and brightness vary with the wave-length and their centres are distributed along the axis in accordance with the colour curve of the objective.

To make the problem sufficiently precise for numerical calculation, we must introduce some assumptions about the colour curve of the objective and its focal ratio, the energy distribution in the spectrum of the star whose image is being investigated, and the relative sensitivity of the eye in different regions of the visual spectrum. The product of the last two quantities gives the effective visual brightness distribution of the starlight as a function of wave-length.

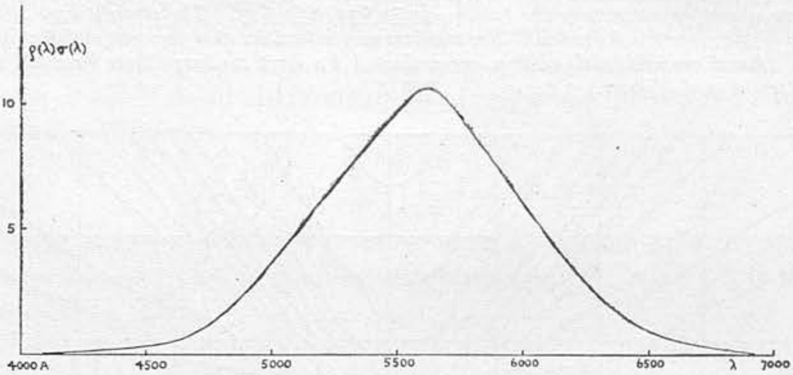


FIG. 4.—*Assumed effective brightness function $\rho(\lambda)\sigma(\lambda)$ for average starlight. $\rho(\lambda)$ is taken from the Planck curve for T=*6000 deg. K; *$\sigma(\lambda)$ from the visibility curve of the average human eye for light of ordinary brilliance.*

Fig. 4 shows the distribution assumed. It is derived by multiplying Planck's ideal energy distribution function $\rho(\lambda)$ corresponding to a temperature $T = 6000$ deg. K and the visibility function $\sigma(\lambda)$ of the average human eye* for light of ordinary brilliance. This choice of $\rho(\lambda)$, which corresponds to the "smoothed" G-type stars, is intended to represent a special case of practical importance. It is one rather favourable to the refractor, since it is one in which the visually brightest part of the spectrum comes near to the turning point of the colour curve of an ordinary visual doublet objective. For a considerable range of colour temperatures it is the σ-function, with its maximum at about 5500 A, which dominates the results and in this range our numerical findings continue to give a fairly accurate picture of the situation. But when used on a blue or on a red star, a visual refractor of normal design would not perform as well as on the yellow stars considered in the remainder of this paper.

We make two different assumptions about the colour curve of the objective, corresponding respectively to the cases of a large visual refractor of about 24 inches aperture and 30 feet focal length and of a medium-sized visual refractor of the same design but of one-third the linear dimensions. These cases seem useful ones to examine, since experience shows that, for refractors working at about the

* Taken from Strong, *Procedures in Experimental Physics*, New York, p. 449, 1936.

usual focal ratio $f/15$, the colour correction is satisfactory to the eye in an 8-inch objective and far from satisfactory in a 24-inch, even though the angular colour aberrations in the two systems may be identical.

More precisely, we consider the following cases of practical interest:

Case A. (The Newall telescope at Cambridge.) A refractor of $24\frac{7}{8}$ inches aperture and 29 feet focal length, with a colour curve shown in Fig. 5. The minimum focus is in light of wave-length $\lambda_m = 5660$ A. We take this as a typical large visual refractor.

Case B. A refractor of one-third the linear dimensions of the Newall telescope; that is, one of aperture $8\frac{7}{24}$ inches and focal length $9\frac{2}{3}$ feet, whose colour curve is obtained from Fig. 5 on reducing the ordinates in the ratio 1:3. This is near to the dimension and colour curve of the Thorrowgood refractor at Cambridge and of many other excellent 8-inch refractors made by Messrs Cooke and Sons in the early part of this century.

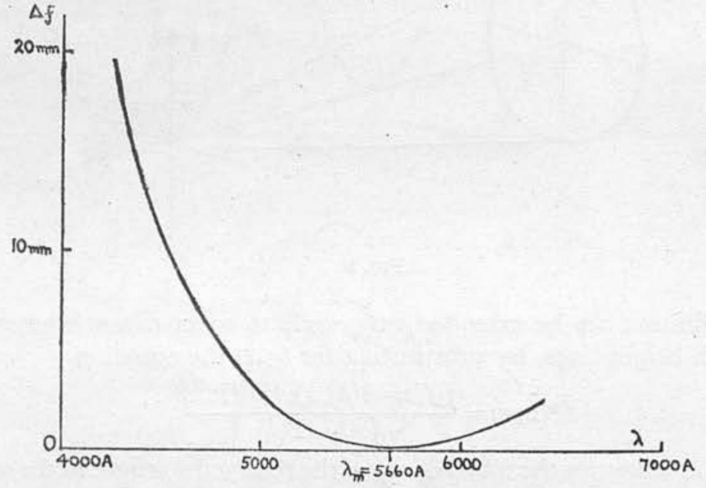


FIG. 5.—*Colour curve of the Newall telescope.*

We also consider, for comparison purposes, the hypothetical Case R of a reflecting telescope of the same focal ratio as the Newall telescope and as the refractor in Case B. (The image formed by a reflector working in polychromatic light depends, in all respects except brightness, on its focal ratio and not on its linear dimensions.)

Case R probably does not correspond to any existing telescope, since one must go back to Herschel's time to find examples of reflectors working, without central obstruction, at focal ratios as long as $f/15$. But from Case R may be derived at once the corresponding results for a Herschelian reflector working at any other focal ratio, since the polychromatic diffraction images at different focal ratios differ only by a simple linear transformation.

3.2. *The "summed visual intensity" and the "summed visual illumination".*—Fig. 6 represents a pencil of light of wave-length λ issuing from the circular exit pupil of a telescope objective and converging towards the corresponding geometrical focal point F_λ on the axis of the telescope. Let λ_m denote the value of λ for which the geometrical focal length of the objective is a minimum. We take the corresponding focal point F_{λ_m} as the origin O of Cartesian coordinates (x, y, z), the axis Oz being along the principal ray of the pencil. In these coordinates

F_λ is the point $(0, 0, \delta(\lambda))$, where the value of $\delta(\lambda) = F_{\lambda_m} F_\lambda$ can be read off from the colour curve of the objective. In the selected special Cases A and B, $\lambda_m = 5660$ A; in Case A, $\delta(\lambda)$ is the quantity Δf of Fig. 5; in Case B, $\delta(\lambda)$ is one-third of this quantity. Case R is covered by setting $\delta(\lambda) = 0$ for all values of λ.

Denoting $\sqrt{(x^2+y^2)}$ by r, we can then define the "summed visual intensity" $I^*(z, r)$ at the point $P(x, y, z)$ of the polychromatic diffraction image by the equation

$$I^*(z,r) = \frac{\int I_\lambda(z-\delta(\lambda), r)\sigma(\lambda)\rho(\lambda)\,d\lambda}{\int \sigma(\lambda)\rho(\lambda)\,d\lambda}; \tag{3.1}$$

here $\rho(\lambda)$ measures the energy distribution in the starlight as a function of wave-length and $\sigma(\lambda)$ the colour sensitivity of the eye, while I_λ is the intensity function of (2.18) and (2.19).

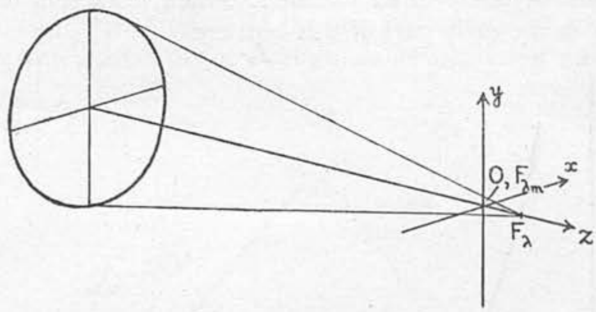


FIG. 6.

The definition can be extended very easily to cover monochromatic light, or spectra with bright lines, by substituting for (3.1) the equation

$$I^*(z,r) = \frac{\int_\lambda I(z-\delta(\lambda), r)\sigma(\lambda)\,dP(\lambda)}{\int \sigma(\lambda)\,dP(\lambda)}, \tag{3.2}$$

in which $P(\lambda)$ measures the total energy of the received starlight in the wave-length range $(0, \lambda)$.

We can also define the " summed visual illumination" $L^*(z, r_0)$ contained in a circle of radius r_0, centred on Oz, in the receiving plane specified by the parameter z, by the equation

$$L^*(z,r_0) = \frac{\int L_\lambda(z-\delta(\lambda), r_0)\sigma(\lambda)\,dP(\lambda)}{\int \sigma(\lambda)\,dP(\lambda)}, \tag{3.3}$$

which reduces to the form

$$L^*(z,r_0) = \frac{\int L_\lambda(z-\delta(\lambda), r_0)\sigma(\lambda)\rho(\lambda)\,d\lambda}{\int \sigma(\lambda)\rho(\lambda)\,d\lambda} \tag{3.4}$$

in the case of a continuous spectrum. It follows at once from (2.23), (3.2) and (3.4) that

$$\begin{aligned} L^*(z,r_0) &= \frac{1}{\pi a^2}\int_0^{r_0}\int_0^{2\pi} I^*(z-\delta(\lambda), r) r\,dr\,d\theta \\ &= \frac{2}{a^2}\int_0^{r_0} I^*(z-\delta(\lambda), r) r\,dr. \end{aligned} \tag{3.5}$$

3.3. *Choice of receiving plane.*—In attempting a definition of best focus which shall correspond reasonably well with practical requirements, two obvious procedures suggest themselves. One is to define the best focal setting as that position of the receiving plane for which the visual brightness at the centre of the image is greatest. The other is to define it as the position for which as much visual illumination as possible is contained in a circle of suitable radius lying in the receiving plane and centred on the principal ray.

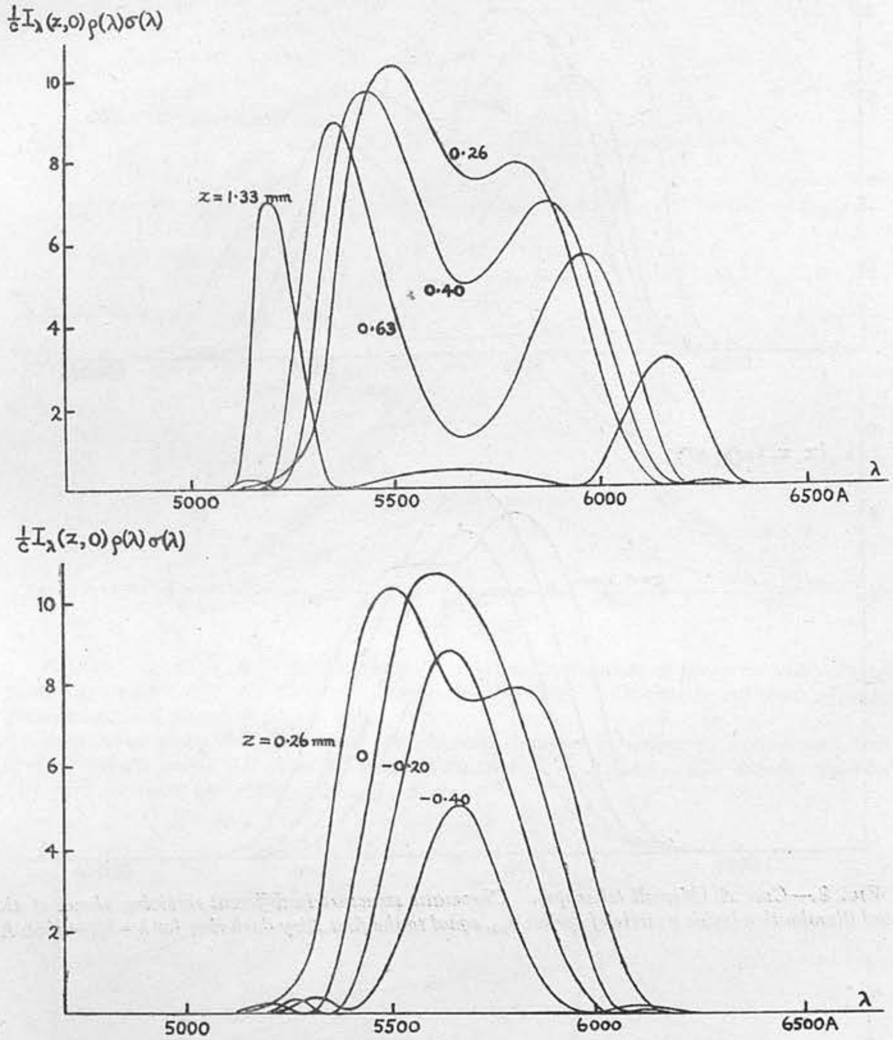


FIG. 7.—*Case A (Newall telescope). Chromatic structure of axial intensity in different receiving planes. The " condensation factor "* $c=\dfrac{4\pi^2a^4}{\lambda_m^2f^2}=1\cdot56\times10^{10}$ *measures the ratio of the intensity of the light of wave-length* λ_m *at geometrical focus to its intensity in the entering beam.*

In Case R (or in the case of any aberration-free image formed by a reflector) the best focal setting according to either of these definitions agrees with that according to geometrical optics. In Cases A and B the determination of best focus on either definition involves a fairly detailed examination of the chromatic structure of the image.

Figs. 7 and 12 give the required information about the visual intensities along the axis of the image in Cases A and B respectively. Figs. 8 and 13 give the corresponding information about the total visual illumination in a small circle drawn about the centre of the image in each of a selected set of receiving

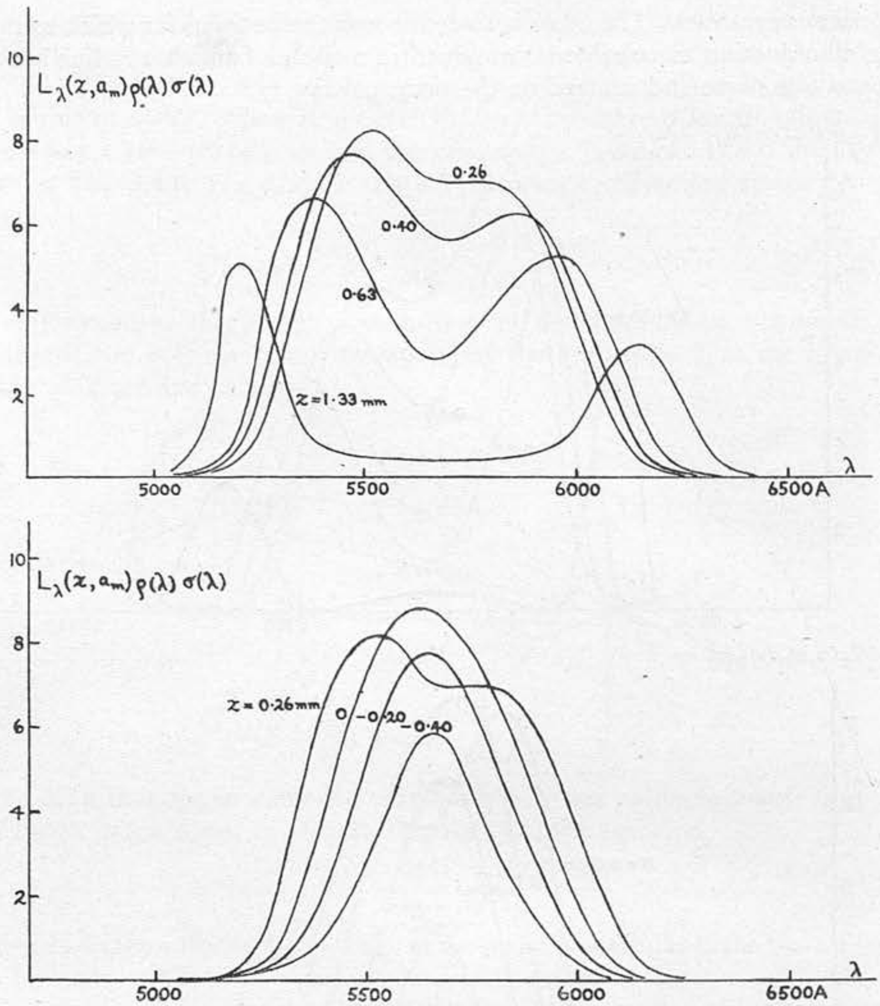


FIG. 8.—*Case A (Newall telescope). Chromatic structure in different receiving planes of the visual illumination inside a circle of radius a_m, equal to the first Airy dark ring for $\lambda = \lambda_m = 5660$ A.*

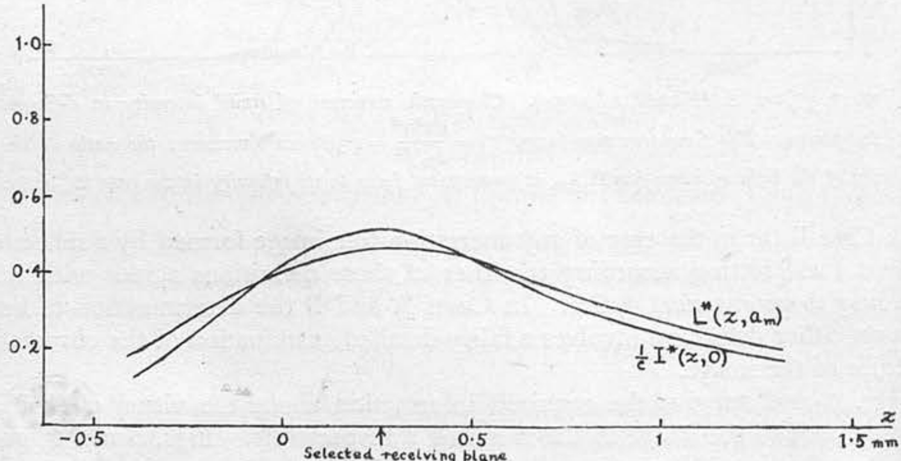


FIG. 9.—*Case A (Newall telescope). Choice of receiving plane.*

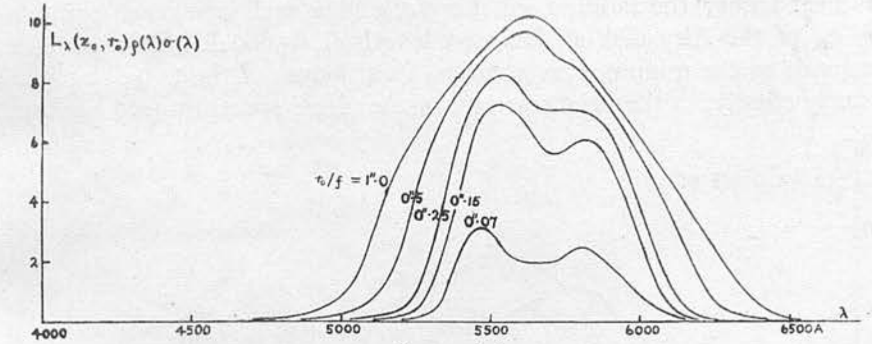


FIG. 10 (*a*).

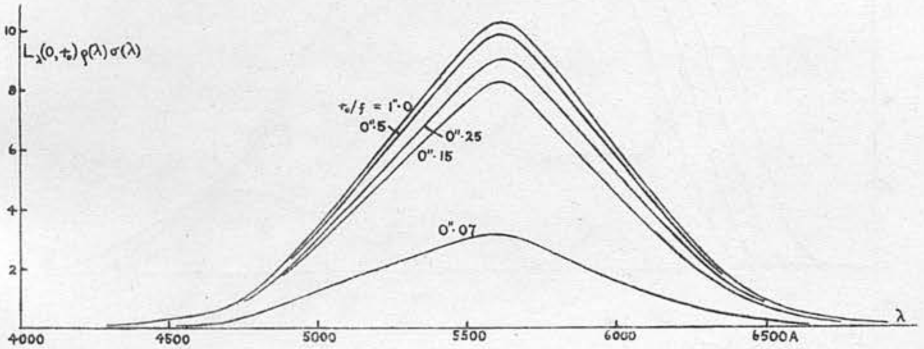


FIG. 10 (*b*).

FIG. 10.—(*a*) *Case A (Newall telescope.) Chromatic structure of image in selected receiving plane* ($z_S = 0\cdot 26$ mm); (*b*) *Case R (Comparison reflector). Chromatic structure of image in geometrical focal plane.*

Each curve shows the total visual illumination, classified according to wave-length, inside a circle of angular radius r_0/f about the origin in the selected focal plane. The angular radius of the first Airy dark ring ($\lambda = 5660$ A) *is* $a_m/f = 0''\cdot 23$.

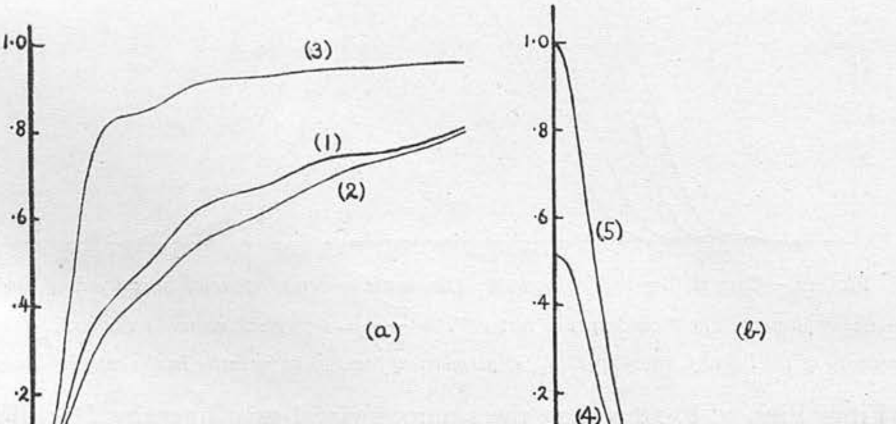


FIG. 11.—(*a*) *Summed visual illumination inside circles of prescribed angular radii r_0/f. Curve 1: Case A (Newall telescope) at selected focus; Curve 2: Newall telescope at minimum geometrical focus; Curve 3: Comparison reflector at geometrical focus.* (*b*) *Summed visual intensity at prescribed angular distances r/f from the centre of the image. Curve 4: Newall telescope at selected focus; Curve 5: Comparison reflector at geometrical focus.*

planes near focus; the radius r_0 of this circle is in each case taken equal to the radius a_m of the Airy disk at the wave-length $\lambda_m = 5660$ A. This wave-length corresponds to the minimum geometrical focal length of the two objectives and also, fairly closely, to the brightest part of the visual spectrum (see Fig. 4).

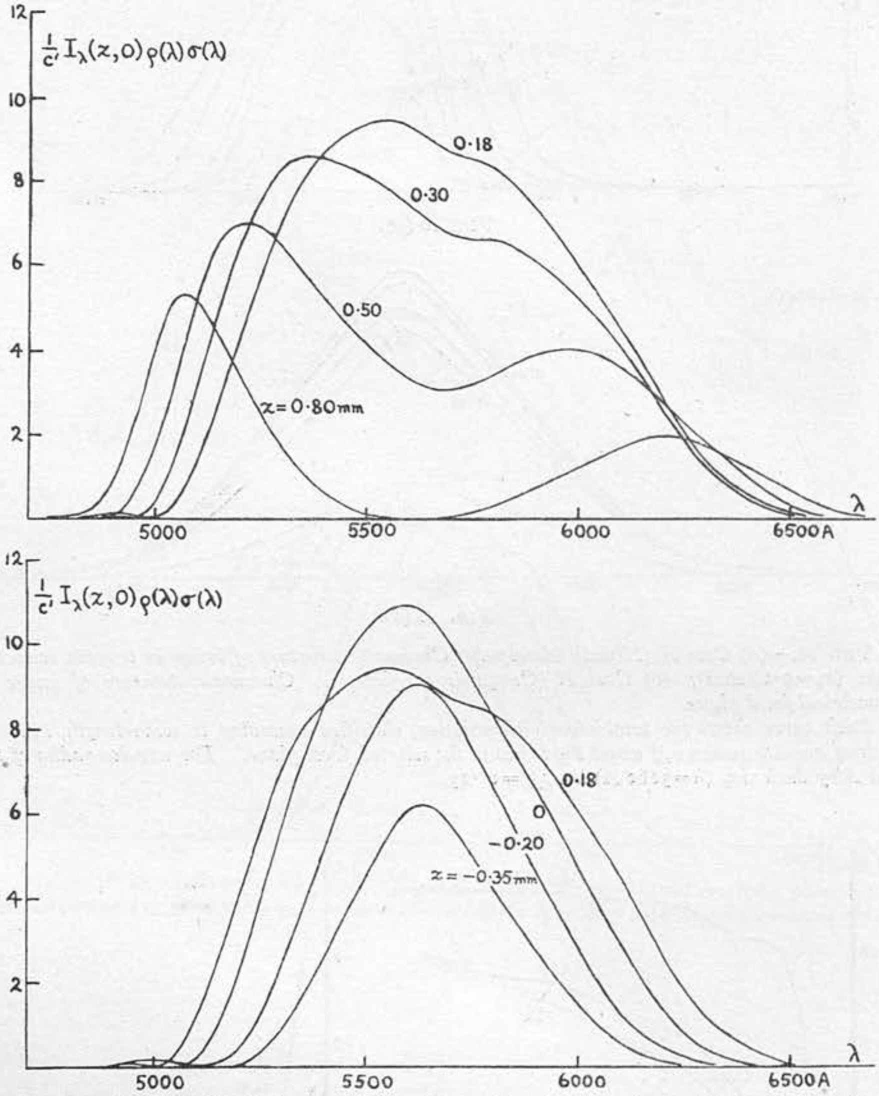


FIG. 12.—*Case B (8·3-inch refractor). Chromatic structure of axial intensity in different receiving planes. The " condensation factor "* $c' = \dfrac{4\pi^2 a^4}{\lambda_m{}^2 f^2} = 1\cdot 73 \times 10^9$ *measures the ratio of the intensity of the light of wave-length* λ_m *at geometrical focus to its intensity in the entering beam.*

From Figs. 7, 8, 12 and 13 the summed visual axial intensity $I^*(z, 0)$ and the summed visual illumination $L^*(z, r_0)$ in Cases A and B can be obtained with sufficient accuracy with a planimeter. They are represented, with the appropriate normalizing coefficients, in Fig. 9 (Case A) and Fig. 14 (Case B). The point of greatest axial intensity $I^*(z, 0)$ corresponds to the best focal setting according to the first definition; that of greatest visual illumination $L^*(z, a_m)$ to the best focal setting according to the second.

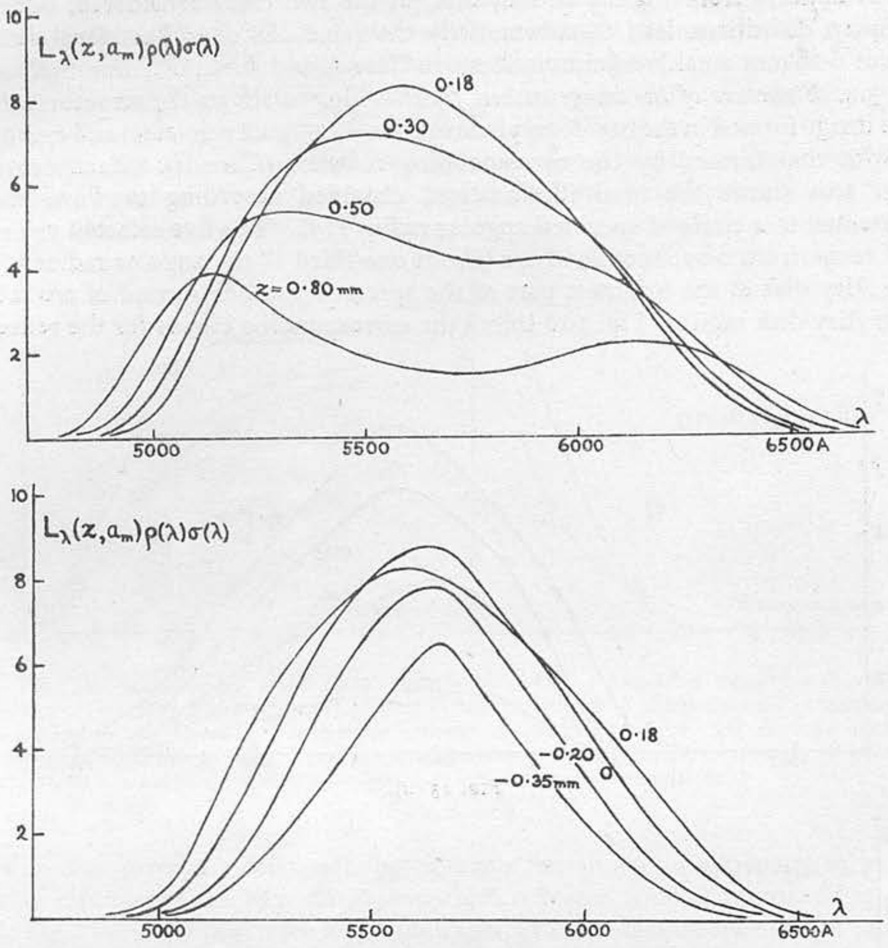


FIG. 13.—*Case B (8·3-inch refractor). Chromatic structure in different receiving planes of the visual illumination inside a circle of radius a_m, equal to the first Airy dark ring for $\lambda = \lambda_m = 5660$ A.*

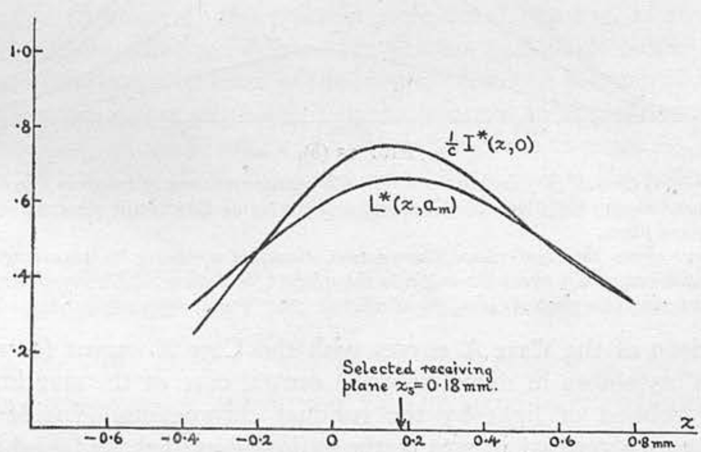


FIG. 14.—*Case B (8·3-inch refractor). Choice of receiving plane.*

It appears from Figs. 9 and 14 that, in the two cases considered, both the adopted definitions lead to substantially the same choice of best focal setting, about 0·26 mm outside minimum focus in Case A and about 0·18 mm in Case B.

3.4. *Structure of the image at best focus.*—Fig. 10 shows the structure of the star image formed in the best focal plane in Case A (25-inch refractor) and compares it with that formed by the corresponding reflector (Case R). Each curve in Fig. 10 *a* shows the total illumination, classified according to wave-length, contained in a circle of specified angular radius r_0/f. The five selected values of r_0/f range from 0·07 seconds of arc (about one-third of the angular radius of the full Airy disk at the brightest part of the spectrum) to one second of arc (about four Airy disk radii). Fig. 10 *b* shows the corresponding curves for the reflector.

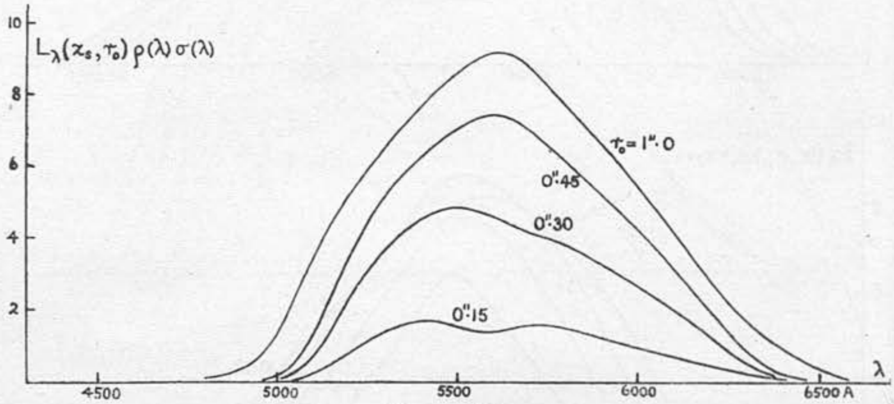


FIG. 15 (*a*).

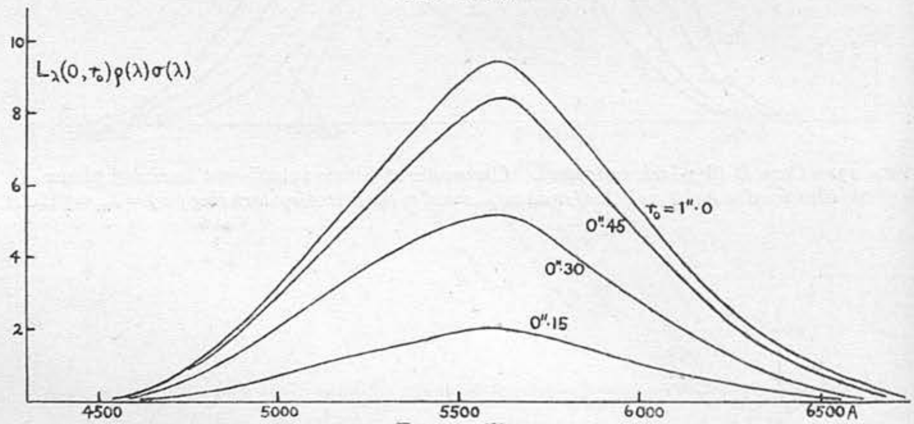


FIG. 15 (*b*).

FIG. 15.— (*a*) *Case B* (8·3-*inch refractor*). *Chromatic structure of image in selected receiving plane* (z_s=0·18 *mm*); (*b*) *Case R* (*Comparison reflector*). *Chromatic structure of image in geometrical focal plane.*

Each curve shows the total visual illumination, classified according to wave-length, inside a circle of angular radius r_0/f about the origin in the selected focal plane. The angular radius of the first Airy dark ring (λ=5660 *A*) *is* $a_m/f = 0''\cdot 68$.

Comparison of the Case A curves with the Case R curves for $r_0/f = 0''\cdot 07$, $0''\cdot 15$ and $0''\cdot 25$ shows in detail how the central core of the star image in the refractor is robbed of light by the residual chromatism. Inside the circle $r_0/f = 0''\cdot 25$, roughly equal in size to the yellow Airy disk, is found hardly any light outside the wave-length range 5000–6200 A. Light outside this range, that is to say from both ends of the visual spectrum, therefore makes no appreciable contribution to the penetrating power of the telescope; it provides the violet or purple halo which is such a disagreeable accompaniment to the star images formed by every large visual doublet objective. Fig. 4 shows that this purple light contains an appreciable fraction of the whole visual illumination, and comparison of the curves for $r_0/f = 1''\cdot 0$ in Figs. 10 *a* and 10 *b*, shows that the colour error sends only a small part of it outside a circle two seconds of arc in diameter.

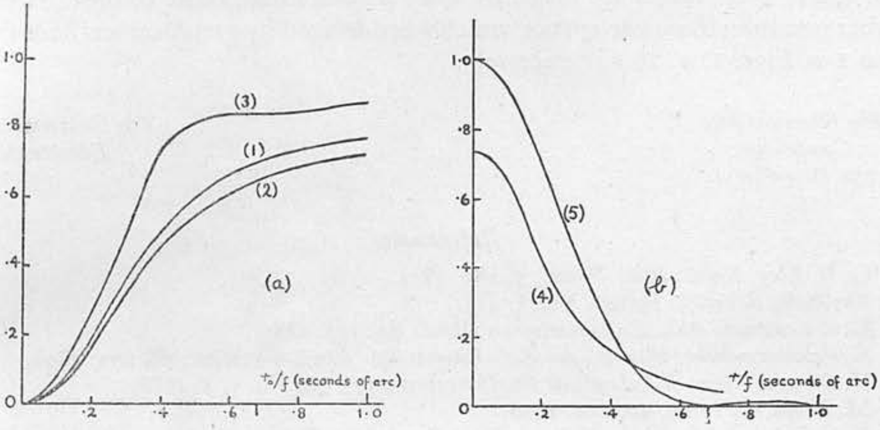


FIG. 16.—(*a*) *Summed visual illumination inside circles of prescribed angular radius* r_0/f. *Curve* 1 : *Case B* (8·3-*inch refractor*) *at selected focus; Curve* 2 : 8·3-*inch refractor at minimum geometrical focus; Curve* 3 : *comparison reflector at geometrical focus.* (*b*) *Summed visual intensity at prescribed angular distances* r/f *from the centre of the image. Curve* 4 : 8·3-*inch refractor at selected focus; Curve* 5 : *comparison reflector at geometrical focus.*

Marked abscissae: a_m/f.

Fig. 15 *a* gives the corresponding information about image structure in the smaller refractor (Case B) and its comparison reflector. Comparison of Fig 15 *a* with Fig. 10 *a* allows us to see how much smaller is the disturbance of the central core of the diffraction image by residual chromatism in Case B than in Case A. For example, in Case B the chromatism reduces the total visual illumination inside a circle of half the diameter of the (yellow) Airy disk* by about 30 per cent; in Case A by about 50 per cent.

Figs. 11 and 16 follow up this point in more detail. In Fig. 11 *a* is shown the "summed visual illumination" (defined in Section 3.2) inside circles of different angular radii r_0/f about the centre of the image. Curve 1 refers to Case A in the selected focal plane, curve 2 to Case A in the minimal focal plane and curve 3 to the monochromatic image ($\lambda = \lambda_m = 5660$ A) in the plane through F_{λ_m}. Curve 3 may also stand for Case R in the geometrical focal plane, since the difference between the curves for Case R and for the monochromatic case turns out to be inappreciable.

(The last result explains why the images in a reflector appear free from spurious colour, although they are formed by the superposition of coloured diffraction images of different sizes.)

Fig. 16 *a* gives the corresponding curves with Case B in place of Case A. For the same reason as before, curve 3 may be taken as referring either to the

* M. Verdet (16) estimates the apparent diameter of the visible bright nucleus of a star image under nearly perfect seeing conditions at about half the diameter of the first Airy dark ring.

/.3

REPRINTED FROM THE
PROCEEDINGS OF THE PHYSICAL SOCIETY, B, VOL. LXVI, p. 145, 1953

Diffraction Images in Systems with an Annular Aperture

BY E. H. LINFOOT AND E. WOLF*
The Observatories, University of Cambridge

MS. received 23rd September 1952

ABSTRACT. The paper is concerned with the effects of central obstruction of the aperture on the three-dimensional light distribution near the focus of an aberration-free optical system. Diagrams are given of the isophotes (lines of equal light intensity) in two selected special cases and are compared with the corresponding diagram for an unobstructed aperture. It appears that when the central obstruction is large the bright central nucleus of the three-dimensional image becomes longer and narrower, so that focal depth and resolving power are both increased. A central obstruction of ratio 0·25, on the other hand, is found to have practically no effect on the size and shape of the bright nucleus.

§ 1. INTRODUCTION

THE intensity distribution in space near the geometrical focus of an error-free pencil of monochromatic light bounded by a circular aperture has been given in a diagram by Zernike and Nijboer (1949). The diagram exhibits in a striking way a peculiarity of the diffraction image already noted by Dennis Taylor (1893), namely the quasi-tubular structure of the luminous core of the image. It is this tubular elongation of the bright central nucleus of the diffraction image, exceeding what we could expect on the basis of more elementary considerations, which explains the excellent performance of 6-in. or 8-in. refracting telescopes in spite of their considerable secondary spectrum.

In reflecting telescopes there is no secondary spectrum. Colour effects arise from the fact that the diffraction patterns surrounding the geometrical focus are of different sizes in different wavelengths, but these are so inconspicuous that their existence is usually ignored even in the literature of diffraction theory †. However, in most reflecting telescopes there is a different complication which affects the diffraction images, namely the central obstruction (usually circular in form) of the aperture by the diagonal flat or by the secondary mirror.

It seems unlikely that anyone who has interested himself in this subject should not feel a strong desire to know what is the effect of a central obstruction of the aperture on the three-dimensional light distribution near focus. In the present note this question is discussed by a straightforward application of Lommel's classical formulae (Lommel 1885)‡, and diagrams of the light distribution are

* Now at the Department of Applied Mathematics, University of Edinburgh.

† They are considered by Mecke (1920) and Picht (1931). A more detailed discussion, on somewhat different lines, is contained in our paper (Linfoot and Wolf 1953) on telescopic star images.

‡ Some extensions of Lommel's tables of U_1 and U_2 were needed for figs. 4 and 5, and these were kindly prepared for us by the Cambridge University Mathematical Laboratory. Since this paper was written, A. Boivin, 1952, *J. Opt. Soc. Amer.*, **42,** 60, has published interesting new expansions for the Lommel functions with arguments $u'=\epsilon^2 u$, $v'=\epsilon v$ (see eqn. (4)) and applied them to the discussion of diffraction by arrays of ring-shaped apertures.

given, in figs. 4 and 5 respectively, for two selected values of the linear obstruction ratio ϵ, namely, $\epsilon = 0{\cdot}25$ and $\epsilon = 0{\cdot}707$. The corresponding diagram for the unobstructed aperture is given for comparison in fig. 3; it is substantially identical with Zernike and Nijboer's, which was obtained by a different method. An earlier diagram by Berek (1926), based, like ours, on Lommel's results, contained inaccuracies.

§ 2. ANALYTICAL FORMULAE

We use substantially the same notation and approximations as in Linfoot and Wolf (1953). For convergent spherical waves of unit amplitude, issuing from a circular aperture of radius R and having radius of curvature $f = \mathrm{CO}$ (fig. 1)

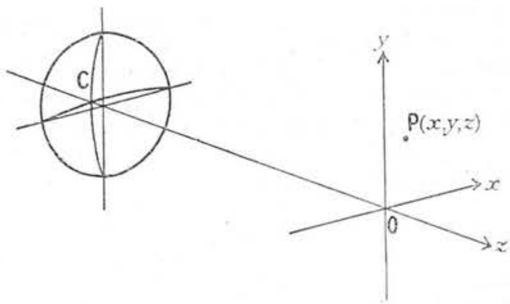


Fig. 1.

at the moment of emergence, the complex displacement at a point $\mathrm{P}(x, y, z)$ in space near the geometrical focus 0 is given by the equation

$$u_\lambda^{(R)}(\mathrm{P}) = \frac{ikR^2}{f}[\exp\{ik(f-\mathrm{CP})\}] \int_0^1 [\exp(\tfrac{1}{2}iu\rho^2)] J_0(v\rho)\,\rho\,d\rho, \quad \ldots\ldots(1)$$

where $$u = kR^2z/f^2, \quad v = kRr/f, \quad k = 2\pi/\lambda, \quad r = +(x^2+y^2)^{1/2}. \quad \ldots\ldots(2)$$

To allow for the effect of a central obstruction of radius $R' = \epsilon R$, we subtract from $u_\lambda^{(R)}(\mathrm{P})$ the complex quantity

$$u_\lambda^{(R')}(\mathrm{P}) = \frac{ikR'^2}{f}[\exp\{ik(f-\mathrm{CP})\}] \int_0^1 [\exp(\tfrac{1}{2}iu'\rho^2)] J_0(v'\rho)\,\rho\,d\rho, \quad \ldots\ldots(3)$$

in which $$u' = kR'^2z/f^2 = \epsilon^2 u, \quad v' = kR'r/f = \epsilon v. \quad \ldots\ldots(4)$$

The intensity at P is then the squared modulus of the quantity

$$\begin{aligned} u_\lambda(\mathrm{P}) &= u_\lambda^{(R)}(\mathrm{P}) - u_\lambda^{(R')}(\mathrm{P}) \\ &= \frac{ikR^2}{f}[\exp\{ik(f-\mathrm{CP})\}] \left[\int_0^1 [\exp(\tfrac{1}{2}iu\rho^2)]\, J_0(v\rho)\,\rho\,d\rho \right. \\ &\qquad \left. - \epsilon^2 \int_0^1 [\exp(\tfrac{1}{2}iu'\rho^2)]\, J_0(v'\rho)\,\rho\,d\rho \right]. \quad \ldots\ldots(5) \end{aligned}$$

Now $$\int_0^1 [\exp(\tfrac{1}{2}iu\rho^2)] J_0(v\rho)\,\rho\,d\rho = \frac{1}{u}[\exp(\tfrac{1}{2}iu)][U_1(u,v) - iU_2(u,v)]$$

$$\int_0^1 [\exp(\tfrac{1}{2}iu'\rho^2)] J_0(v'\rho)\,\rho\,d\rho = \frac{1}{u'}[\exp(\tfrac{1}{2}iu')][U_1(u',v') - iU_2(u',v')],$$

where $U_1(u,v)$, $U_2(u,v)$ are Lommel functions. Thus (5) can be written

$$u_\lambda(\mathrm{P}) = \frac{ikR^2}{f}[\exp\{ik(f-\mathrm{CP})\}] \left[\frac{1}{u}[\exp(\tfrac{1}{2}iu)](U_1 - iU_2) - \frac{1}{u}[\exp(\tfrac{1}{2}i\epsilon^2 u)] \times (U_1' - iU_2') \right], \quad \ldots\ldots(6)$$

where, U_1, U_2, U_1', U_2' are written for $U_1(u,v)$, $U_2(u,v)$, $U_1(u',v')$, $U_2(u',v')$ respectively, and the intensity $I_\lambda(\mathrm{P}) = |u_\lambda(\mathrm{P})|^2$ is given by the equation

$$I_\lambda(\mathrm{P}) = \frac{k^2 R^4}{f^2 u^2} \left| [\exp(\tfrac{1}{2}iu)](U_1 - i\,U_2) - [\exp(\tfrac{1}{2}i\epsilon^2 u)](U_1' - iU_2') \right|^2$$
$$= \frac{\pi^2 R^4}{\lambda^2 f^2} \{M^2(u,v) - 2\epsilon^2 N(u,v;\,u',v') + \epsilon^4 M^2(u',v')\}, \qquad \ldots\ldots(7)$$

where

$$\left.\begin{aligned} M^2(u,v) &= \left(\frac{2}{u}\right)^2 (U_1^2 + U_2^2),\; M^2(u',v') = \left(\frac{2}{u'}\right)^2 (U_1'^2 + U_2'^2) \\ N(u,v;\,u',v') &= \frac{4}{uu'}[(U_1 U_1' + U_2 U_2')\cos\tfrac{1}{2}(1-\epsilon^2)u \\ &\qquad + (U_2 U_1' - U_1 U_2')\sin\tfrac{1}{2}(1-\epsilon^2)u]. \end{aligned}\right\} \qquad \ldots\ldots(8)$$

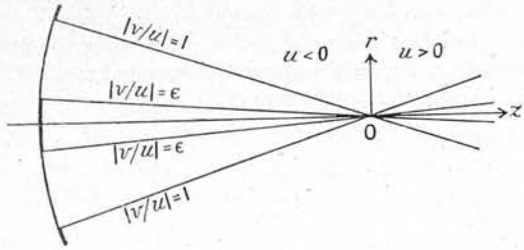


Fig. 2.

On the axis of the converging pencil, $v=0$ and (6) gives for the intensity the expression

$$I_\lambda(z,0) = \frac{4\pi^2 R^4}{\lambda^2 f^2}\left(\frac{\sin\frac{1}{4}u(1-\epsilon^2)}{\frac{1}{4}u}\right)^2. \qquad \ldots\ldots(9)$$

Thus the central obstruction increases the distance between the zeros along the axis of the system by a factor $1/(1-\epsilon^2)$ but they remain equally spaced—at least in the range where our approximations are valid, namely the part of the axis where $u/4\pi$ does not become large compared with unity.

In the geometrical focal plane, $u=0$ and the expression (7) for the intensity reduces to

$$I_\lambda(0,r) = \frac{4\pi^2 R^4}{\lambda^2 f^2}\left[\frac{2J_1(v)}{v} - \epsilon^2\frac{2J_1(\epsilon v)}{\epsilon v}\right]^2. \qquad \ldots\ldots(10)$$

The zeros of this function give, in v-units, the radii of the 'Airy dark rings' corresponding to a centrally obstructed aperture. The expressions (9) and (10) were given (with a different normalization) many years ago by Steward (1925).

§ 3. COMPUTATIONAL RESULTS

To interpret figs. 3, 4 and 5 we note that each of them represents one quarter of a bisymmetrical pattern obtained by reflecting it in the u- and v-axes. This bisymmetrical pattern shows the isophotes (lines of equal light intensity) near focus in any meridional section of the pencil. The diagrams apply to pencils of all sufficiently long focal ratios (to those, in fact, for which Lommel's formulae are valid approximations in the sense of our paper (Linfoot and Wolf 1953)) and the use of (u, v) coordinates in the diagrams is equivalent to scale-normalizing the cylindrical cartesian coordinates (r, z) in accordance with equations (2). The intensity at the geometrical focal point $(0, 0)$ is normalized to unity in each figure.

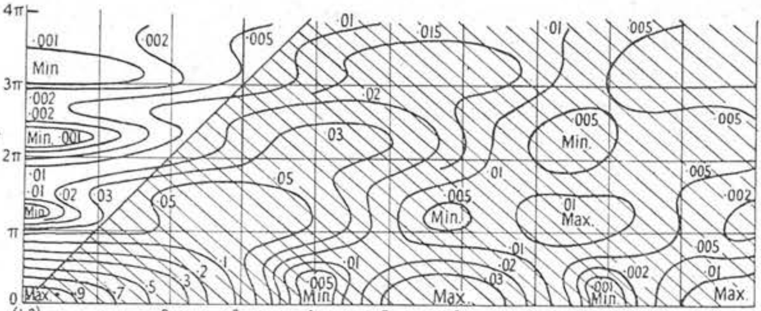


Fig. 3. Isophotes near focus of an aberration-free pencil without central obstruction. The intensity at the focus (0, 0) is normalized to unity. The scale-normalized coordinates (u, v) possess physical interpretations; $u/4\pi$ is the number of fringes of defocusing, v/π the number of fringes of lateral displacement of the point (z, r) from the geometrical focus 0. The bisymmetrical diagram obtained by reflecting the figure in both u- and v-axes shows the light-distribution in any meridional section of the pencil; the u-axis is along the principal ray. The shaded area shows the region of the geometrical cone of rays

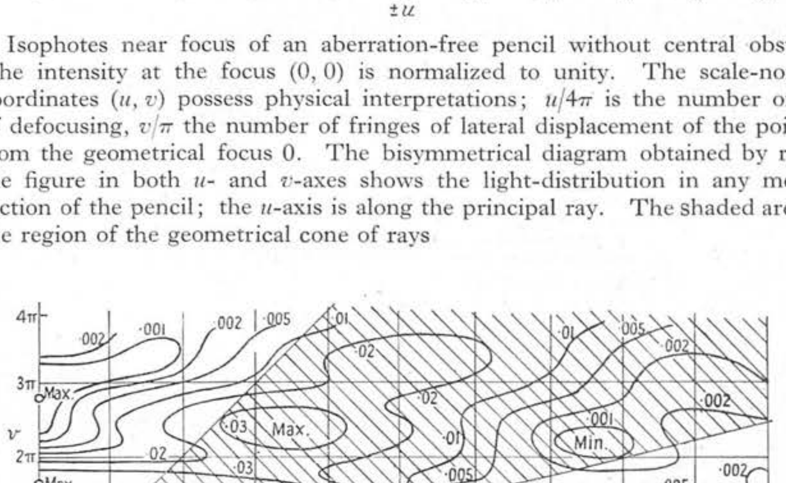


Fig. 4. Isophotes near focus of an aberration-free pencil with central obstruction ratio $\epsilon = 0{\cdot}25$. The intensity at the geometrical focus is normalized to unity. Reflection of the figure in both coordinate axes gives a diagram of the isophotes in any meridional section. The shaded area gives the position of the hollow cone of rays.

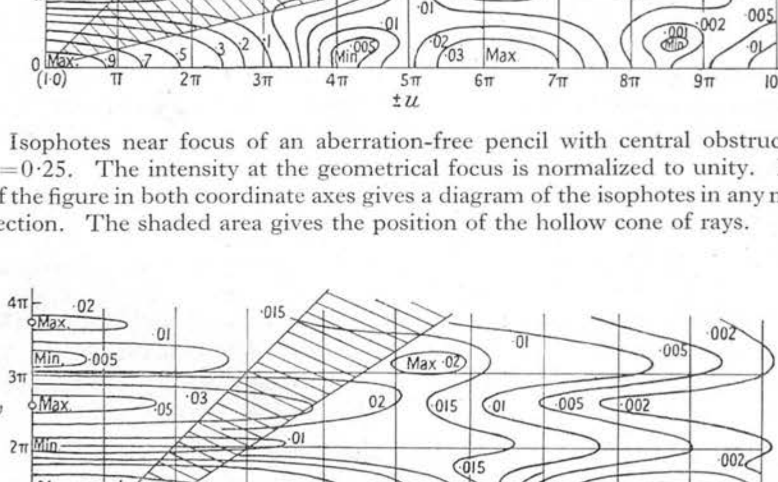


Fig. 5. Isophotes near focus of an aberration-free pencil with central obstruction ratio $\epsilon = 0{\cdot}707$. The intensity at the geometrical focus is normalized to unity. Reflection of the figure in both coordinate axes gives a diagram of the isophotes in any meridional section. The shaded area gives the position of the hollow cone of rays.

The shaded areas in the bisymmetrical patterns show where the geometrical light-cones meet the meridional (u, v)-plane in each case. In the unobstructed case (fig. 3) the cone of rays is a solid cone, with axis lying along the u-axis and the lines $|v/u|=1$ lie in its surface. In the centrally obstructed cases the cone is hollow; its axis lies along the u-axis and its 'body' lies between the two conical surfaces traced out when the lines $|v/u|=\epsilon$ and $|v/u|=1$ are rotated about the u-axis (see fig 2). In fig. 4, $\epsilon=0{\cdot}25$, in fig. 5, $\epsilon=0{\cdot}707$.

The value $\epsilon=0{\cdot}25$ was selected for computation because it corresponds to the greatest central obstruction which is regarded as tolerable by users of visual reflecting telescopes. A comparison of figs. 3 and 4 shows how small is the effect of this obstruction on the relative intensities in different parts of the image. In particular, the size and shape of the bright central nucleus is almost unaffected. There is, however, an increase in the intensity of the first Airy bright ring, the effects of which might occasionally be visible to a keen observer.

The value $\epsilon=0{\cdot}707$, corresponding to an obstruction of half the area of the aperture, was selected with the following idea in mind. When the central obstruction is fairly large, for example when it has the above value, the pencil of rays passing through the geometrical focus has the form of a hollow circular cone (shown in section in fig. 5) and at a sufficient distance from the geometrical focal plane almost all the light is to be found between the walls of this cone, that is to say, in the notation of § 2, in the region $0{\cdot}707<|v/u|<1$. It is natural to ask whether anything at all resembling this type of light distribution is to be found in the near neighbourhood of the geometrical focus. Figure 5 provides an answer to the question: close to focus, the light distribution has the same general character as in the case of the unobstructed aperture and is strikingly different from that predicted by geometrical optics. The main effect, in this region, of the obstruction is to draw out the central nucleus of the image along the axis of the system to approximately twice its former length, its cross section being correspondingly reduced.

These results may be of some practical value in connection with the design of lens–mirror systems, in which the use of mirrors generally involves central obstruction of the beam, while the refractive elements usually introduce some chromatic variation of focus. The existence and dimensions of a tubular core to the diffraction image have here, just as in the case of telescope doublets, an important bearing on the amount of chromatic variation of focus which can be tolerated in the system (Conrady 1923, Linfoot and Wolf 1953).

More academic, but perhaps not entirely without interest, is the point that a large central stop on the objective of a refracting telescope not only increases resolving power by decreasing the lateral diameter of the bright central nucleus of the image but also, by elongating the nucleus in the axial direction, reduces the disturbing effects of chromatism on its colour-composition at best focus.

REFERENCES

Berek, M., 1926, *Z. Phys.*, **40**, 421.

Conrady, A. E., 1923, *Dictionary of Applied Physics*, Vol. IV (London: Macmillan), p. 222.

Linfoot, E. H., and Wolf, E., 1953, *Mon. Not. R. Astr. Soc.*, **112**, 452.

von Lommel, E., 1885, *Abh. der Bayerischen Akad.*, **53**, 233.

Mecke, R., 1920, *Ann. Phys., Lpz.*, (4) **61**, 471; **62**, 623.

Picht, J., 1931, *Optische Abbilding* (Braunschweig: Vieweg), p. 102.

Steward, G. C., 1925, *Phil. Trans. Roy. Soc.* A, **225**, 131.

Taylor, H. Dennis, 1893, *Mem. Roy. Astr. Soc.*, **54**, 77.

Zernike, F., and Nijboer, B. R. A., 1949, Contribution to *Théorie des images optiques* (Paris: Éditions de la Revue d'Optique), p. 227.

1.4

THE PHYSICAL SOCIETY

REPRINTED FROM
REPORTS ON PROGRESS IN PHYSICS, VOL. XIV, p. 95, 1951

PRINTED IN GREAT BRITAIN

THE DIFFRACTION THEORY OF ABERRATIONS

BY

E. WOLF

THE PHYSICAL SOCIETY

Reprinted from

REPORTS ON PROGRESS IN PHYSICS, Vol. XIV, p. 95, 1951

PRINTED IN GREAT BRITAIN

THE DIFFRACTION THEORY OF ABERRATIONS

By E. WOLF*

The Observatory, Cambridge University

CONTENTS

ABSTRACT. A historical and critical survey is given of investigations concerned with image formation in optical instruments in the presence of aberrations. The development of the subject is traced from the early researches of Airy on an aberration-free image. The advances made in recent years are discussed in greater detail; these mainly concern the effects of small aberrations, tolerance criteria, and the asymptotic treatment of diffraction problems. A section is also included on investigations into the effects of waves of non-uniform amplitude. The detailed light distribution in typical images is illustrated by isophote-diagrams and photographs.

§ 1. INTRODUCTION

In a perfect optical system the light waves which proceed from each element of the object emerge in the image space as convergent waves spherical in form. Such an instrument represents an idealization which cannot be realized in practice; in the image space of an actual instrument, the wave surfaces will, as a rule, be of a more complicated form. The deviations from the ideal spherical form of these surfaces may range in practice from a fraction of a wavelength in a well corrected telescope or microscope objective, to several dozen wavelengths in instruments required for less precise work. In many cases the resulting image bears little resemblance either to the Airy diffraction pattern or to the confusion figure predicted by geometrical optics.

The diffraction theory of aberrations is concerned with the study of images formed by actual optical instruments. Recent years have seen important advances in this branch of optics and many valuable results have been obtained. In particular, in the domain of very small aberrations (amounting only to a fraction of a wavelength) greatly simplified series expansions for the light distribution in diffraction patterns have been given, supplemented by a detailed study of a number of typical cases. The effects of aberrations of intermediate size (about one to ten wavelengths) have been studied with the help of a specially constructed mechanical integrator. Attempts have also been made to examine the effects of large aberrations. Further, a new and systematic investigation of the maximum amount of aberrations which may be tolerated in optical instruments has been carried out. The possibility has also been examined of improving the quality of

* Now at the Department of Mathematical Physics, University of Edinburgh.

the image by deliberately introducing non-uniformity of amplitude and sometimes, in addition, varying the phase of the disturbance over the exit pupil.

As no article on the diffraction theory of aberrations has been published in the preceding volumes of these Reports we have included in the present account a historical review of the subject. The period before 1940 is, however, treated only briefly; additional information can be found in the articles of von Laue (1928), König (1929), Martin (1946) and in a thesis by Nijboer (1942). A survey of experimental investigations on diffraction images was recently given in a thesis by Nienhuis (1948). We begin with a short preliminary section on the geometrical treatment of aberrations; this serves as an introduction to the classification of image errors used in the main discussion.

§ 2. GEOMETRICAL TREATMENT OF ABERRATIONS

The geometrical theory of aberrations is usually developed by methods due to Hamilton based on Fermat's principle of stationary path. This principle asserts that if light is propagated from a point P′ to a point P in a medium of which the refractive index at a typical point is denoted by μ, it will travel along a path for which the optical length

$$\int_{P'}^{P} \mu \, ds \qquad \ldots\ldots(2.1)$$

is stationary with respect to a small variation of that path. The integral (2.1) when taken along a natural ray may be regarded as a function of six variables, namely the rectangular Cartesian coordinates of P′ and P with respect to some fixed reference system. In consequence of Fermat's principle, this function, sometimes called the characteristic function of the medium and denoted by $V(x', y', z'; x, y, z)$ satisfies the relations

$$\left.\begin{aligned} \frac{\partial V}{\partial x'} &= -\mu' l', & \frac{\partial V}{\partial y'} &= -\mu' m', & \frac{\partial V}{\partial z'} &= -\mu' n', \\ \frac{\partial V}{\partial x} &= \mu l, & \frac{\partial V}{\partial y} &= \mu m, & \frac{\partial V}{\partial z} &= \mu n. \end{aligned}\right\} \qquad \ldots\ldots(2.2)$$

Here (l', m', n') and (l, m, n) denote the direction cosines of the rays at P′ and at P respectively, while μ' and μ are the refractive indices at these points.

Consider an optical instrument with a point source at P_0' (x_0', y_0', z_0') emitting monochromatic light. From (2.2) it follows that the equations

$$V(x_0', y_0', z_0'; x, y, z) = \text{constant} \qquad \ldots\ldots(2.3)$$

represent the orthogonal trajectories of the rays, i.e. the wave fronts of geometrical optics. Except in the ideal case when (2.3) represents concentric spherical surfaces in the image space of the system, the rays will intersect the image plane at different points. Geometrical optics identifies the light intensity at a typical element of the image plane with the density of these ray intersections. Apart from the details of structure, this gives an approximation to the actual light distribution when the deformations of the wave fronts are of a considerable amount.

To investigate the appearance of the image it is convenient to introduce an aberration function Φ which measures the deviation of the wave fronts from the spherical form. We consider a system with cylindrical symmetry and take a reference sphere centred on the Gaussian image P_0 of P_0', passing through the centre of O of the exit pupil. Further we denote by A a typical point on this sphere and by B the intersection of P_0 A with the wave which passes through O (Figure 1).

The aberration function, defined by the equation

$$\Phi = AB, \qquad \ldots\ldots(2.4)$$

may be expressed as function of the three rotational invariants σ^2, r^2 and $\sigma r \cos(\chi - \theta)$, where (r, θ), and (σ, χ) denote polar coordinates in the exit pupil and of the image point P_0 respectively. It may as a rule be expanded in the form *

$$\Phi(\sigma, r, \phi) = \sum_{i,j,k,} a_{ijk} \sigma^{2i+k} r^j \cos^k \phi, \qquad \ldots\ldots(2.5)$$

where $\phi = \chi - \theta$, i, j, k are non-negative integers, $j \geqslant k$, $j - k$ is even $(\neq 0)$, $2i + j + k \geqslant 4$, and the a's are constants.

Let P be the actual point of intersection with the image plane of a typical ray proceeding from the exit pupil. From Φ it is possible to determine with the help of Fermat's principle the ray aberration displacements P_0P. Those terms in (2.5) for which $2i + j + k = N$ are called the wave aberrations of the Nth order

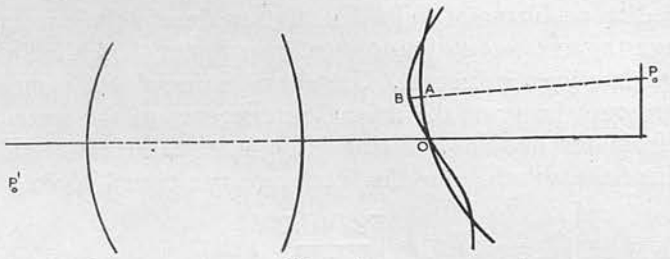


Figure 1.

The line BAP_0 does not necessarily lie in the plane $P_0' OP_0$. O is the centre of the exit pupil.

(N is always even) and give rise to ray aberrations of order † $N - 1$. The five aberrations of the lowest order ($N = 4$) are as a rule dominant in the paraxial region. The corresponding ray aberrations are known respectively as third order (sometimes primary or Seidel) spherical aberrations, coma, astigmatism, field curvature and distortion and have been discussed in the literature in much detail. Their physical meaning is too well known to need repeating here. For detailed discussions of geometrical aberrations we refer the reader to papers by Schwarzschild (1905), by Steward (1926) and by Nijboer (1943).

§ 3. THE DIFFRACTION THEORY OF ABERRATIONS

3.1. *A Historical Review*

We mentioned that the geometrical theory of aberrations identifies the intensity in the image plane with the ray density. This approximation, which in many cases gives an adequate picture of the light distribution in the image, gradually loses its validity as the aberrations become smaller. In the limiting case of perfectly spherical waves, for example, issuing from a circular opening,

* A different expansion recently proposed by Nijboer (1942, 1943) will be discussed later. The present notation is substantially due to him.

† A certain confusion exists in the literature about the terminology. Some earlier authors call first order ray aberrations those which we have denoted as ray aberrations of the third order. The classification used here has the advantage that it can be generalized to cover errors of focusing which are represented by first order terms. Such errors are often associated with the chromatic aberration.

geometrical optics predicts in the focal plane an infinite intensity at focus and zero intensity elsewhere; in reality the image consists of a bright central patch surrounded by rings—the familiar Airy pattern. It is therefore clear that in certain cases more refined investigations are needed. A direct application of electromagnetic theory presents considerable mathematical difficulties; instead one often uses approximations based on Kirchhoff's formula

$$U(P)=\frac{1}{4\pi}\iint_{\mathrm{D}}\left\{U\frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{r}\right)-\frac{e^{-ikr}}{r}\frac{\partial U}{\partial n}\right\}dS. \qquad \ldots\ldots(3.1)$$

This expresses the wave disturbance at any point P inside a closed region D as a surface integral involving the disturbance U and its gradient $\partial U/\partial n$ over the boundary. Here r represents the distance of a typical surface element from P, $\partial/\partial n$ denotes differentiation along the inward normal to D, $k=2\pi/\lambda$ and λ is the wavelength. In the investigations which we shall discuss, the scalar disturbance U is identified with a component of the electric or the magnetic vector or of a Hertz vector. From it the intensity I is derived by means of the approximate relation

$$I=\text{constant}\,|U|^2. \qquad \ldots\ldots(3.2)$$

When applied to problems of light distribution due to the passage of aberrant waves through the exit pupil of an optical instrument, Kirchhoff's formula with certain approximations reduces to Huyghens' principle of geometrical optics, taking into account however the mutual interference of the secondary wavelets proceeding from the opening. Using Huyghens' principle in this form, Airy (1835) obtained the solution for the ideal case mentioned above. It is usually written as

$$I(v)=\left[\frac{2J_1(v)}{v}\right]^2, \qquad \ldots\ldots(3.3)$$

where in suitable units $I(v)$ is the intensity in the pattern at distance v from its centre and J_1 is a Bessel function of the first kind.*

In two important papers Lommel (1885, 1886) extended Airy's classical result by deriving expressions for the intensity distribution due to spherical waves in planes other than the geometrical focal plane. For diffraction at a circular opening his solution takes the form †

$$I(u,v)=\left(\frac{2}{u}\right)^2[U_1^2(u,v)+U_2^2(u,v)], \qquad \ldots\ldots(3.4)$$

where the U's are two of the functions

$$U_n(u,v)=\sum_{s=0}^{\infty}(-1)^s\left(\frac{u}{v}\right)^{n+2s}J_{n+2s}(v), \qquad \ldots\ldots(3.5)$$

u is a parameter which specifies the position of the receiving plane and v has the same meaning as before. When $|u/v|>1$ the convergence of (3.5) is slow and for this case Lommel expressed the solution in terms of the functions

$$V_n(u,v)=\sum_{s=0}^{\infty}(-1)^s\left(\frac{v}{u}\right)^{n+2s}J_{n+2s}(v), \qquad \ldots\ldots(3.6)$$

* The derivation of (3.3) involved the assumption that the angular semi-aperture of the opening is small and that the disturbance over the converging wave is of uniform amplitude. In practice these assumptions may be far from valid but it has been shown by Hopkins (1943, 1944) that with a suitable interpretation of the argument v, Airy's formula is accurate to within a few per cent for systems with angular semi-aperture up to about 30°.

† The case of a circular opening was treated in an almost identical manner by Struve (1886). Formulae more convenient for calculating the intensity far from focus were given by Schwarzschild 1897).

which are related to the U_n functions by means of the relations

$$\left.\begin{aligned} U_{2n+1}(u,v)+V_{-2n+1}(u,v) &= (-1)^n \sin\frac{1}{2}\left(u+\frac{v^2}{u}\right) \\ -U_{2n+2}(u,v)+V_{-2n}(u,v) &= (-1)^n \cos\frac{1}{2}\left(u+\frac{v^2}{u}\right). \end{aligned}\right\} \quad \ldots\ldots(3.7)$$

The functions U_n and V_n which Lommel introduced and which now bear his name play an important part in a number of related problems. Lommel evaluated numerically the solution for a large number of particular cases. Full use of his results was however not made until forty years later when Berek (1926) obtained with their help a graphical representation of the three-dimensional light distribution near focus.* Recently, Zernike and Nijboer (1949) gave a more accurate and detailed diagram (see Figure 2) based on a different expansion of the diffraction integral.

The first investigations concerning diffraction images in the presence of monochromatic aberrations appear to be due to Rayleigh and Strehl. Rayleigh

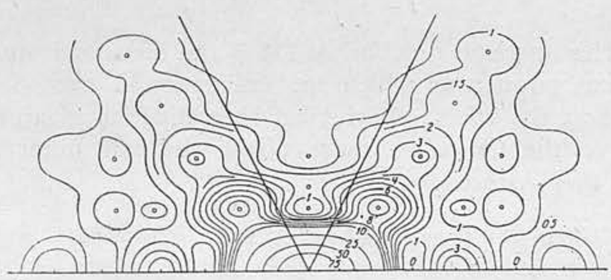


Figure 2. Isophotes (lines of equal intensity) near focus in a meridional plane in absence of aberrations. The straight lines indicate the boundary of the geometrical shadow. The numbers give intensity as percentage of intensity at focus. Axial minima and maxima are indicated by short strokes, others by small circles. *After* Zernike and Nijboer (1949).

(1879) studied images formed by cylindrical waves affected by a certain unsymmetrical aberration which are diffracted at a rectangular aperture, and also examined the effects of third order spherical aberration. In the latter case, however, he confined his investigations to determining the intensity at the centre of the pattern only. He formulated an important tolerance criterion, known in an extended form as Rayleigh's limit. This criterion asserts that the quality of an instrument is not sensibly affected by the presence of certain commonly occurring types of aberration if they are such that the waves in the image space do not deviate by more than a quarter of a wavelength from suitably chosen spherical surfaces. This criterion has since become widely used in formulating conditions concerning the maximum amount of aberrations which may be tolerated in the practical design of optical instruments.

Strehl, in his book *Théorie des Fernrohres* (1894) and in numerous papers mostly published between 1893 and 1930 in the *Zeitschrift für Instrumentkünde* and in *Zentralzeitung für Optik und Mechanik*, studied the effect of third order aberrations but he confined his researches mainly to investigating the variation of intensity along the principal ray. To him we owe the important

* Berek's figure is rather inaccurate. Later versions of it show the ouline of the geometrical shadow incorrectly as, for example, in the following books: *Handbuch der Physik*, 1929, **21**, 885 and Picht, J., 1931, *Optische Abbildung* (Braunschweig: Vieweg), p. 71.

concept now known as Strehl definition (Definitionshelligkeit), viz. the ratio of the maximum intensity in a particular receiving plane of an instrument to the intensity at the centre of the Airy disc in a perfect system of the same aperture and focal length. This quantity supplies in many cases a good measure of the quality of the system.

Conrady (1919), Buxton (1921, 1923) and Martin (1922) investigated the effects of various aberrations with the help of numerical integrations. They computed a number of particular cases and obtained valuable information concerning Rayleigh's criterion. Some work of these authors was also concerned with balancing spherical aberrations of different order against each other so as to obtain high intensity at the centre of the pattern; this question was later studied in a more general manner by Richter (1925).

In important papers Steward (1925) and Picht (1925, 1926) derived series expansions for the intensity distribution in typical diffraction images. In Steward's treatment the diffraction integral, derived by an immediate application of Huyghens' principle, gives the intensity I at a point P of the pattern in the form

$$I = \left| \iint e^{ik(V+d)}\, dS \right|^2, \qquad \ldots\ldots(3.8)$$

where V is a characteristic function and d is the distance from P to a typical point of the exit pupil over which the integration is carried out. Steward considered at first the effect of spherical aberration in an arbitrary receiving plane parallel to the Gaussian image plane, the exit pupil being assumed circular. (3.8) then reduces to

$$I(u,v) = \left| \int_0^1 \exp(\tfrac{1}{2}iut + i \sum_{s\geqslant 2} A_s t^s) \,.\, J_0(v\sqrt{t})\, dt \right|^2. \qquad \ldots\ldots(3.9)$$

The problem treated earlier by Lommel corresponds to the case $A_s = 0$ for all s. By an argument similar to Lommel's, Steward showed that (3.9) can be expressed in terms of so-called generalized Lommel functions, these again being series involving Bessel functions.

When other aberrations are present (3.8) takes a more complicated form, since simplifying symmetry conditions no longer exist. Steward restricted the rest of his analysis to third order aberrations. His paper contains two diagrams showing isophotes (lines of equal intensity) in diffraction images, one affected by a certain amount of third order coma, the other by third order astigmatism. Unfortunately these diagrams are rather incomplete and contain some errors [see Nijboer (1947, p. 619) and Kingslake (1948, p. 152)]. Steward also studied the effects of different forms of aperture on the resolution. In particular he examined the influence of an annular aperture when third order spherical aberration is present under out-of-focus conditions.

Picht (1925, 1926) took as the starting point of his researches a well-known result due to Debye that the effect of spherical waves emerging from an optical instrument may be represented by a superposition of plane waves with different phases, amplitudes and directions of propagation. By generalizing Debye's formulae, Picht showed that the intensity distribution associated with the passage of aberrant waves through an exit pupil may be written in the form

$$I(x,y,z) = \left| \iint \psi(m,n) \exp\{-ik[(x-\xi)l + (y-\eta)m + (z-\zeta)n]\} \frac{dm\, dn}{l} \right|^2$$

where $\quad \xi = \xi(m,n), \quad \eta = \eta(m,n), \quad \zeta = \zeta(m,n) \qquad \ldots\ldots(3.10)$

are parametric equations of a typical wave, (l, m, n) are direction cosines of its normals and ψ is an amplitude factor. In the presence of third order aberrations, Picht obtained the development of (3.10) into series of Bessel functions. From these series he computed, for a small amount of spherical aberration, the three-dimensional intensity distribution in the neighbourhood of the paraxial focus. He also studied a number of images affected by third order astigmatism in systems with a rectangular exit pupil.

Born (1932, 1938) studied the influence of very small aberrations of the third order. He derived an expression for the intensity, applicable whether one or more ~~observations~~ are present, in terms of the functions aberrations/.

$$K_0(v)=\left[\frac{J_1(v)}{v}\right]^2, \qquad K_1(v)=\frac{2J_1(v)J_2(v)}{v}, \qquad K_2(v)=\frac{2J_1(v)J_2(v)}{v^2}.$$

Born's formula applies however to distributions in the Gaussian image plane only.

No further substantial advances were made up to about 1940. The following years witnessed a renewed interest in this subject and led to a number of important developments which we shall now discuss.

3.2. *Advances since* 1940

3.21 *Images in the presence of small aberrations.*

Some of the researches so far described led to results which in many cases permitted the calculation of the light distribution in the diffraction images. These solutions, however, were not entirely satisfactory as they were either too restricted or involved very heavy computations.

In an important thesis* Nijboer (1942) took up this problem afresh and obtained a simpler and more satisfactory solution for cases where the wave deformation is small, only a fraction of a wavelength. Nijboer considered two series expansions of the aberration function Φ, viz.

$$\Phi(\sigma, r, \phi) = \sum_{l,m,n} b'_{lnm}\sigma^{2l+m} r^n \cos m\phi \qquad \ldots\ldots(3.11)$$

and

$$\Phi(\sigma, r, \phi) = \sum_{l,m,n} b_{lnm}\sigma^{2l+m} R_n{}^m(r) \cos m\phi, \qquad \ldots\ldots(3.12)$$

where l, m, n are non-negative integers, $n \geqslant m$, $n-m$ is even, $R_n{}^m(r)$ are certain polynomials which we discuss later, the other symbols having same meaning as in §2. A closer examination shows that the first expansion is particularly suitable for a geometrical treatment of aberrations, the second for a diffraction treatment. We recall that the traditional expansion with the same variables would contain terms $\cos^m\phi$ in place of $\cos m\phi$. The occurrence of the Fourier terms is essential in Nijboer's theory.

Some advantages of the new expansions can be illustrated by considering first the geometrical aberrations. It can be shown that in the presence of a single aberration

$$b'_{lnm}\sigma^{2l+m} r \cos m\phi^n \qquad \ldots\ldots(3.13)$$

the rays from the point (r, ϕ) of the exit pupil intersect the plane through the Gaussian image point perpendicular to the principal ray in a point (x, y) given by

$$x+iy=\tfrac{1}{2}b'_{lnm}\sigma^{2l+m} r^{n-1}\frac{R}{a}\{(n+m)e^{-i(m-1)\varphi}+(n-m)e^{i(m+1)\varphi}\}, \qquad \ldots\ldots(3.14)$$

* An account of this work will also be found in Nijboer (1943, 1947). As mentioned by Nijboer the method was to some extent suggested by earlier unpublished researches of Zernike.

where a and R denote the radius of the exit pupil and of the reference sphere respectively. From (3.14) it follows that the main features of the aberration figure (given by r=constant), such as symmetry, depend on m while details depend both on m and n. It will be shown further that m and n play a similar role in the appearance of the diffraction pattern for an aberration specified by the term

$$b_{lnm}\sigma^{2l+m}R_n{}^m(r)\cos m\phi. \qquad \ldots\ldots(3.15)$$

This suggests a new classification of image errors in which the value of m determines the general type. By analogy with Seidel aberrations, the terms with $m=0$, 1 and 2 may be called spherical aberration, coma and astigmatism respectively. Curvature and distortion now appear as degenerate cases of spherical aberration and of coma. The terms with $m \geqslant 3$ do not occur in Seidel theory. This classification considerably simplies the discussion of the higher order effects.

In Nijboer's diffraction treatment the expansion of the aberration function is of the form (3.12) where

$$R_n{}^m(r)\cos m\phi \qquad \ldots\ldots(3.16)$$

are so-called circle polynomials introduced by Zernike (1934) in his diffraction theory of phase contrast. These polynomials are orthogonal within the unit circle and the $R_n{}^m(r)$ are given by

$$R_n{}^m(r) = (-1)^{(n-m)/2}\binom{(n+m)/2}{m} r^m F\left[\tfrac{1}{2}(n+m+2), -\tfrac{1}{2}(n+m), -\tfrac{1}{2}(n-m), m+1, r^2\right]$$

$$= \sum_{s=0}^{(n-m)/2} (-1)^s \frac{(n-s)!}{s!\,\{\frac{1}{2}(n+m)-s\}!\,\{\frac{1}{2}(n-m)-s\}!}\, r^{n-2s}, \qquad \ldots\ldots(3.17)$$

where F is a hypergeometric function. The expansion in terms of circle polynomials is particularly advantageous for the balancing of very small aberrations against each other so as to obtain maximum Strehl definition. We observe that in (3.15) a number of terms of the customary form $\sigma^{2l+m}r^i\cos^j\phi$ are combined; with the help of the orthogonality relations for the circle polynomials it may be shown that they have been combined in such a way as to give maximum intensity in the centre of coordinates for a given value of the coefficient of

$$\sigma^{2l+m}r^n\cos^m\phi.$$

This may be illustrated by recalling that Richter (1925) showed (in a different notation) that in the presence of a small amount of spherical aberration of the form

$$\Phi = A_6r^6 + A_4r^4 + A_2r^2 + A_0, \qquad \ldots\ldots(3.18)$$

the Strehl definition is a maximum for a given value of A_6 if

$$A_4/A_6 = -3/2, \quad A_2/A_6 = 3/5. \qquad \ldots\ldots(3.19)$$

In Nijboer's diffraction treatment the fifth order spherical aberration is given by the term

$$b_{060}R_6{}^0(r) = b_{060}(20r^6 - 30r^4 + 12r^2 - 1), \qquad \ldots\ldots(3.20)$$

and Richter's condition is automatically satisfied. However, we must insist that this balancing of aberrations as achieved in Nijboer's theory holds for very small aberrations only; in some cases, as Figure 8 indicates, it no longer holds when the maximum wave deformation is only about $0{\cdot}6\lambda$.

With the usual approximations the Huyghens–Kirchhoff integral for the disturbance at a point (q, ψ) in a receiving plane perpendicular to the principal ray and specified by a parameter p, can be written in the form

$$U(p, q, \psi) = \frac{1}{\pi}\int_0^1\int_0^{2\pi} \exp\{ipr^2 + iqr\cos(\phi-\psi) - ik\Phi(\sigma, r, \phi)\} r\, dr\, d\phi. \quad \ldots\ldots(3.21)$$

In the presence of a single aberration Nijboer writes

$$\Phi(\sigma, r, \phi) = \frac{1}{k}\beta_{lnm}R_n^{\ m}(r)\cos m\phi,$$

where
$$\beta_{lnm} = k\sigma^{2l+m}b_{lnm} \quad \ldots\ldots(3.22)$$

and expands the term $\exp ik\Phi$ in powers of β. Neglecting fifth and higher powers of β in (3.21) and integrating with respect to ϕ Nijboer obtains

$$\begin{aligned}
U(p, q, \psi) = {} & 2\int_0^1 e^{ipr^2}J_0(qr)r\,dr - 2i\beta_{lnm}i^m\cos m\psi\int_0^1 e^{ipr^2}R_n^{\ m}(r)J_m(qr)r\,dr \\
& + \frac{(i\beta_{lnm})^2}{2!}\left\{\int_0^1 e^{ipr^2}\{R_n^{\ m}(r)\}^2 J_0(qr)r\,dr + i^{2m}\cos 2m\psi \right. \\
& \left. \times\int_0^1 e^{ipr^2}\{R_n^{\ m}(r)\}^2 J_{2m}(qr)r\,dr\right\} - \frac{(i\beta_{lnm})^3}{2.3!}\left\{3i^m\cos m\psi \right. \\
& \left. \times\int_0^1 e^{ipr^2}\{R_n^{\ m}(r)\}^3 J_m(qr)r\,dr + i^{3m}\cos 3m\psi\int_0^1 e^{ipr^2}\{R_n^{\ m}(r)\}^3 J_{3m}(qr)r\,dr\right\} \\
& + \frac{(i\beta_{lnm})^4}{2^2.4!}\left\{3\int_0^1 e^{ipr^2}\{R_n^{\ m}(r)\}^4 J_0(qr)r\,dr + 4i^{2m}\cos 2m\psi \right. \\
& \times\int_0^1 e^{ipr^2}\{R_n^{\ m}(r)\}^4 J_{2m}(qr)r\,dr + i^{4m}\cos 4m\psi \\
& \left. \times\int_0^1 e^{ipr^2}\{R_n^{\ m}(r)\}^4 J_{4m}(qr)r\,dr\right\}. \quad \ldots\ldots(3.23)
\end{aligned}$$

From this expression a number of symmetry properties of the diffraction pattern may immediately be deduced. It is seen for example that the p-axis is an m-fold axis of symmetry. Moreover if m is odd (e.g. as in the case of coma) the intensity distribution is symmetrical about the plane $p=0$. If m is even (but $m\neq 0$) the intensity at any point is equal to that in the point resulting from a reflection in the plane $p=0$ and the additional rotation about the p-axis through an angle π/m. For some special cases this rule has previously been stated by Strehl, by Picht and by Steward.

Nijboer outlined a method based on various properties of the circle polynomials which makes possible the development of the integrals in (3.23) into series of Bessel functions. The series terminate in the important case $p=0$. From these expansions the isophotes near focus may be determined for a small arbitrary single aberration. Some of the diagrams obtained in this way by Nijboer (1942, 1947) and Zernike and Nijboer (1949) are reproduced in Figures *3, 4, 5, 7. Additional cases of slightly larger aberrations were studied by Nienhuis and Nijboer (1949). This later investigation was confined to third order effects and contains new formulae and diagrams (Figures 6 and 8) for the

* Unless otherwise stated, all figures refer to systems with a circular exit pupil and Φ_{max} denotes the maximum deviation of the wave from the reference sphere, shown in Figure 1. The Figures are not all reproduced on the same scale.

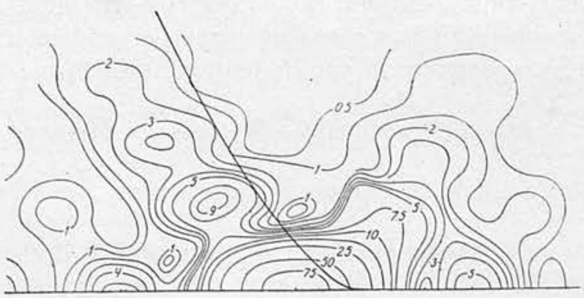
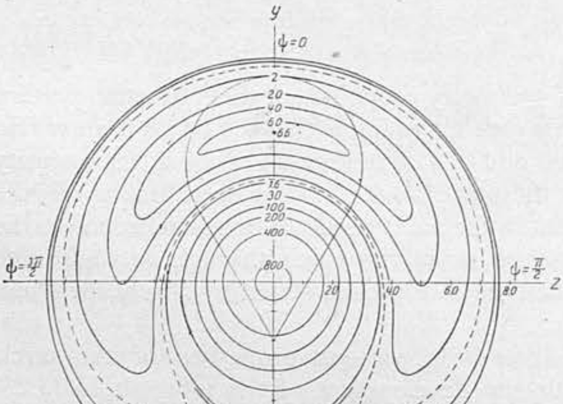


Figure 3. Isophotes as in Figure 2, in presence of third order spherical aberration of amount $\beta_{l40}=\frac{1}{2}$ ($\Phi_{\max}=0{\cdot}48\lambda$). The thick line indicates the geometrical caustic. Strehl definition 0·95. *After* Zernike and Nijboer (1949).

Figure 4. Isophotes as in Figure 2, in presence of spherical aberration of amount $\beta_{l60}=\frac{1}{2}$ ($1{\cdot}6\lambda$ of fifth order spherical aberration balanced against $2{\cdot}4\lambda$ of third order spherical aberration). The thick line indicates the geometrical caustic. Strehl definition 0·965. *After* Zernike and Nijboer (1949).



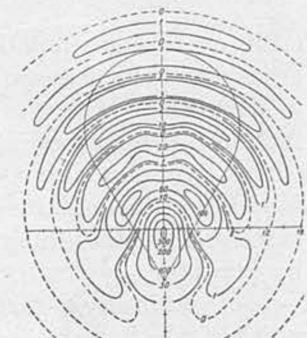


Figure 5. Isophotes in the Gaussian image plane in presence of third order coma of amount $\beta_{l31}=1$ ($\Phi_{\max}=0{\cdot}48\lambda$). Strehl definition 0·879. The dotted curve represents the line of zero intensity. The boundary of the geometrical confusion figure is also indicated. The numbers represent intensities corresponding to a value of 1,000 in the centre of the ideal Airy pattern. *After* Nijboer (1947).

Figure 6. Isophotes in the Gaussian image plane in presence of third order coma of amount $\beta_{l31}=3$ ($\Phi_{\max}=1{\cdot}4\lambda$). Strehl definition 0·306. The dotted curves represent the line of zero intensity. The boundary of the geometrical confusion figure is also indicated. *After* Nienhuis and Nijboer (1949).

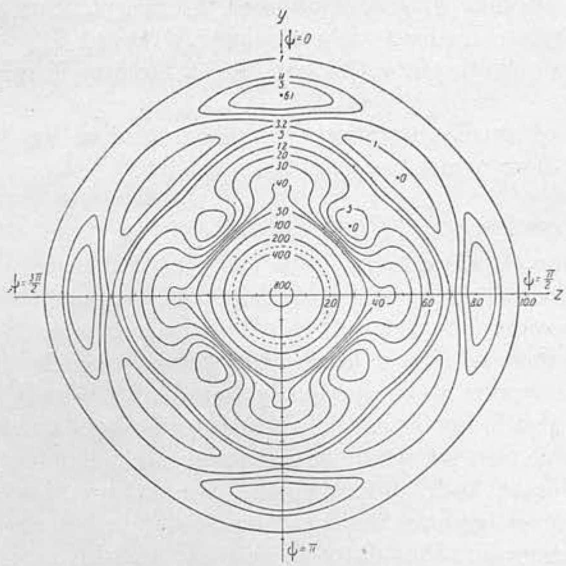


Figure 7. Isophotes in the central plane in presence of third order astigmatism of amount $\beta_{l31}=1(\Phi_{\max}=0{\cdot}16\lambda)$. Strehl definition 0·840. The dotted circle indicates the boundary of the geometrical confusion figure. *After* Nijboer (1947).

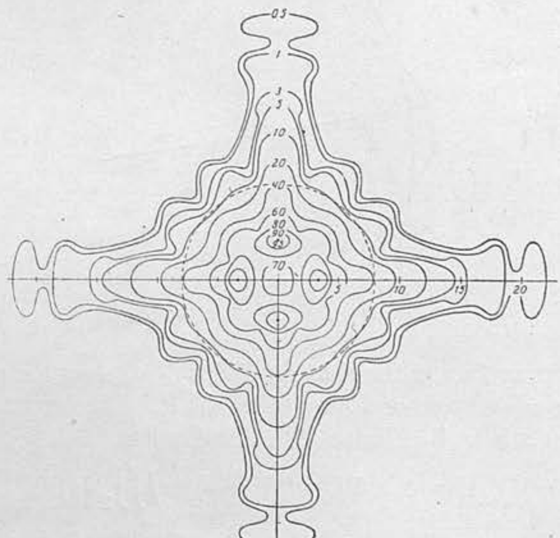


Figure 8. Isophotes in the central plane in presence of third order astigmatism of amount $\beta_{l22}=4(\Phi_{\max}=0{\cdot}64\lambda)$. Strehl definition 0·066. The dotted circle indicates the boundary of the geometrical confusion figure. *After* Nienhuis and Nijboer (1949).

light distribution in certain receiving planes. It was supplemented by photographs (see Plates *) of actual images showing a good agreement with the theoretical predictions. In this work the authors rediscover the formula

$$I=\left[\frac{2}{\beta}\sum_{s=0}^{\infty}J_{2s+1}(\beta)\right]^2,$$

originally given in a slightly different notation by Picht (1931, p. 202) for the Strehl definition in the central plane $p=0$ when third order astigmatism is present.

* ~~For all Plates see end of volume.~~

Some of the isophote diagrams involved the computation of intensity at as many as 600 points. This heavy labour is amply rewarded by the results which give the first real insight into the complex structure of important types of diffraction image.

A few cases of small aberrations were also studied by Maréchal and by Kingslake by methods which we shall discuss later.

3.22. *Tolerance conditions.*

Wang Ta-Hang (1941) discussed the location of the best axial focus in the presence of spherical aberration and deduced tolerance conditions for the maximum permissible amounts of third, fifth and seventh order spherical aberration for systems where the lower order terms are under control. Most of Wang Ta-Hang's results were previously found by Conrady, Buxton, Martin and Richter and also follow as special cases from Nijboer's researches.

Optimum correction of small aberrations and tolerance conditions were studied in an elegant and comprehensive manner by Maréchal (1944 a, b, c, 1947 a). These investigations are based on a relation between the intensity at points near the maximum of the diffraction pattern and the mean square deviation of the wave front from a certain reference sphere.

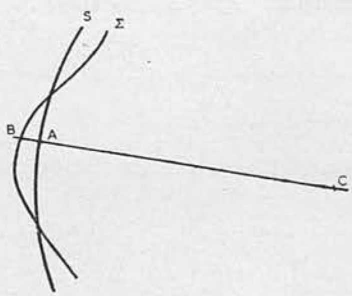


Figure 18.

Let Σ be a wave surface in the exit pupil of an instrument and let S be a sphere in its neighbourhood, centred on C and of radius R. Further let A be a point on S and let AC intersect Σ in B (Figure 18). Finally let

$$\mathrm{AB}=\Delta, \quad \mathrm{BC}=l, \quad d\omega=d\sigma\Big/\iint_{\Sigma}d\sigma, \qquad \ldots\ldots(3.24)$$

where $d\sigma$ is an element of Σ. (In general $\Delta\neq\Phi$ since C is not necessarily the Gauss image point and S does not necessarily pass through the centre of the exit pupil.)

For a given wave front and for fixed C, the mean square deviation

$$E=\iint_{\Sigma}\Delta^2 d\omega, \qquad \ldots\ldots(3.25)$$

of the wave front from the sphere is a function of the radius R. This function has a minimum

$$\left.\begin{aligned} E=E_0&=\iint_{\Sigma}l^2d\omega-\Big(\iint_{\Sigma}l\,d\omega\Big)^2, \\ R=R_0&=\iint_{\Sigma}l\,d\omega. \end{aligned}\right\} \qquad \ldots\ldots(3.26)$$

when

Maréchal showed, that when $|\Delta| < \lambda/4$ (a condition twice less severe than that imposed by Rayleigh's criterion), the intensity I (taken as unity at the centre of the Airy pattern) at C satisfies the relation *

$$I \sim \left[1 - \frac{2\pi^2}{\lambda^2} E_0\right]^2. \qquad \ldots\ldots (3.27)$$

Earlier authors found that in presence of certain aberrations of amount $\frac{1}{4}\lambda$, the central intensity takes on values between about 0·73 and 0·87, the exact amount depending on the type of aberration. With the help of (3.27) Maréchal was able to formulate a new tolerance criterion which unlike that of Rayleigh corresponds to a fixed lower limit for the central intensity.

A system is generally regarded as well corrected if the loss in intensity at the centre of the pattern does not exceed 20%, i.e. when $I \geqslant 0{\cdot}8$. For such a case (3.27) implies that $E_0 < \lambda^2/180$. Maréchal therefore concluded that for a well corrected system the root-mean-square deviation of the wave from the 'mean focal ~~square~~ sphere/' (given by $R = R_0$) is less than $\lambda/\sqrt{180}$. In accordance with this criterion, Maréchal then deduced the following tolerance conditions for defocusing and for aberrations of the third order:

Defocusing: $\lambda/2\alpha^2$.

Spherical aberration: $4\lambda/\alpha^2$.

Coma: $\left|\dfrac{\delta g}{g}\right| \leqslant 0{\cdot}63\,\dfrac{\lambda}{\sigma\alpha}$.

Astigmatism: $|t - s| \leqslant \lambda/\alpha^2\sqrt{2}$.

Astigmatism and curvature: $(t+s)^2 + 2(t-s)^2 \leqslant \lambda^2/\alpha^4$.

Combined aberrations: $\left(\dfrac{L}{L_{\max}}\right)^2 + \left(\dfrac{\delta g}{\delta g_{\max}}\right)^2 + \left(\dfrac{t-s}{(t-s)_{\max}}\right)^2 + \left(\dfrac{t+s}{(t+s)_{\max}}\right)^2 \leqslant 1.$

Here α denotes the maximum angular semi-aperture in the image space, L denotes the longitudinal spherical aberration, g the sine condition ratio, and t and s are the abscissae of the tangential and the sagittal focal lines. Further, $L_{\max}$ denotes the maximum amount of the longitudinal spherical aberration which may be tolerated when no other aberration is present, $\delta g_{\max}$, $(t-s)_{\max}$ and $(t+s)_{\max}$ having analogous meanings. For fifth order spherical aberration and fifth order coma Maréchal found that the tolerances may be increased by 50% provided these aberrations are suitably 'balanced'. A tolerance condition for third order coma substantially in agreement with Maréchal's was recently deduced by Kingslake (1948).

Toraldo di Francia (1946) derived an expression for the lower limit of intensity at the centre of the reference sphere in terms of the maximum wave aberration. In particular, this expression shows that when the aberration is less than $\lambda/4$ the intensity at the centre does not fall below 0·5.

Françon (1944, 1947 a, b, 1948 a) studied both theoretically and experimentally the efficiency of visual optical instruments suffering from spherical aberration. (Aberrations may be tolerated when the efficiency, defined as the ratio of the limiting resolvable separation of the unaided eye to that of the instrument–eye combination, is nearly 1.) He found that for a small pupil (0·8–0·9 mm. in diameter), the eye may be regarded as a perfect instrument and in this case a

* Väisälä (1922) derived a weaker condition $I \sim 1 - 4\pi^2 E_0/\lambda^2$.

high degree of correction is called for. For pupils of larger size poorer correction may be tolerated. Françon showed that for small pupils and objects of very low contrast ($\sim 0{\cdot}3$), aberrations of the order of $\lambda/16$ may be significant. Rayleigh's and Maréchal's criteria apply to a perfect eye, and contrast 1. Françon's investigations give, for the case of spherical aberration, an extension of Rayleigh's criterion for pupils up to 4 mm. and for a range of contrasts.

3.23. *Asymptotic behaviour of the diffraction integral.*

The effects of aberrations which are large compared with the wavelength may be investigated by examining the asymptotic behaviour of the diffraction integral for large values of $k = 2\pi/\lambda$. Van Kampen (1949) studied this problem by using a two-dimensional analogue of the principle of stationary phase (cf. van der Corput (1948), p. 206).

The Huyghens–Kirchhoff diffraction integral which describes the complex displacement at a point P in the image space of an instrument may be written in the form

$$U(\mathrm{P}) = \iint_{\mathrm{D}} kg(x, y) e^{ikf(x, y)}\, dx\, dy, \qquad \ldots\ldots(3.28)$$

where D is a domain bounded by a finite number of analytic curves C. Van Kampen found that, on account of the rapidity of fluctuation of the exponential term in (3.28), the value of $U(\mathrm{P})$ depends substantially on the behaviour of the integrand near a limited number of 'critical points' which may be of three kinds: (1) internal points of D at which f is stationary, i.e. where $\partial f/\partial x = \partial f/\partial y = 0$; (2) boundary points at which f is stationary, i.e. where $\partial f/\partial s = 0$, ds being element of C; (3) corner points or boundary points where two analytic curves join.

In the neighbourhood of a critical point of the first kind, the integrand of (3.28) can be expanded in the form

$$kge^{ikf} = k \exp ika_{00} \exp ik(a_{20}x^2 + a_{11}xy + a_{02}y^2 + \ldots)\,.\,(b_{00} + b_{10}x + \ldots). \qquad ..(3.29)$$

To find the corresponding contribution to the integral van Kampen ~~neglects in the exponent third and higher powers of x and y and~~ integrates term by term over the ranges $-\infty < x < \infty$, $-\infty < y < \infty$. The resulting expansion is a power series in k^{-1} beginning with a constant term. For the contribution of a critical point of the second kind, van Kampen obtains in an analogous manner a series with $k^{-1/2}$, $k^{-3/2}$, $k^{-5/2}$ and for the contribution of corner points, a power series in k^{-1}, in both these cases the constant term being absent. The total effect is then obtained by adding the contribution from each of the critical points.

The principal term of the final expansion is independent of the wavelength and comprises the constant terms of the expansions at critical points of the first kind. This part of the solution can be interpreted as a contribution of 'geometrical optics' taking, however, into account the mutual interference of the wave patches arriving at P from the immediate neighbourhood of critical points of the first kind.

The rest of van Kampen's paper deals with applications to third order aberrations in systems with a circular aperture. He then found for example that in the case of astigmatism points of the second kind give rise to an asteroid pattern; this pattern was also discussed by Nienhuis (1948).

Although van Kampen's investigations give some valuable results, his analysis cannot be regarded as satisfactory. For example, van Kampen integrates the expansions over infinite ranges, a procedure which introduces non-negligible errors and in some cases even leads to divergent integrals. To overcome the divergence difficulty he uses the well known but mathematically unsatisfactory artifice of multiplying the integrand by a term $e^{-\sigma\xi}$ and after integration sets $\sigma=0$. It appears that only the leading terms of van Kampen's expansion have any real claim to validity, and these themselves cannot be regarded as having been rigorously derived.

3.24. *Images formed by waves on non-uniform amplitude.**

When studying images which are formed by an instrument of the usual design it is as a rule permissible to assume that the waves are of uniform amplitude. Several authors have in recent years considered the possibility of improving the quality of the image by a suitable variation of the amplitude and sometimes also of the phase of the disturbance over the exit pupil. This may be realized in practice, for example by evaporating thin films of metallic or dielectric substances on to the lenses or by means of specially constructed filters.

With the usual approximations, it follows from Huyghens' principle that the complex displacement $U(x,y)$ in the receiving plane of a system imaging a point source may be written in the form of a double Fourier integral

$$U(x,y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(p,q)e^{2\pi i(px+qy)}\,dp\,dq, \qquad \ldots\ldots(3.30)$$

where $P(p,q)$ is the so-called pupil function (sometimes called transmission function), which describes, in terms of the optical direction cosines p and q in the image space, the complex displacement over a spherical reference surface centred at $x=y=0$. The integration is only formally carried out over an infinite domain since P is taken as zero when $p^2+q^2>\rho_m^2$, ρ_m being the maximum numerical aperture in the image space. In the case of rotational symmetry (3.30) may be written as

$$U(r)=2\pi\int_0^{\infty} P(\rho)J_0(2\pi r\rho)\rho\,d\rho, \qquad \ldots\ldots(3.31)$$

where ρ is the zonal numerical aperture in the image space and r is the distance from the optical axis in the receiving plane, measured in wavelengths. (The symbols r and P now denote different quantities from those in previous sections.)

Luneberg (1944) considered the case of rotational symmetry and studied the effects of variation of amplitude alone. He found that amongst all diffraction patterns of equal energy the highest central maximum is given by the normal pattern $P(\rho)=\text{constant}$. In particular, any pattern which gives improved resolution must therefore have a lower central maximum. By a variational argument Luneberg deduced an expression for the pupil function in terms of the radius r_0 of the first dark ring. It appears that for $r_0<0{\cdot}31(2\pi/\rho_0)$ (i.e. about a half of the radius r_A of the first dark ring in the Airy pattern) the resulting pattern has a rather low central maximum.

Luneberg also investigated how to chose $P(\rho)$ so that the fraction of the total illumination

$$L(r_0)=\frac{\int_0^{r_0}|U(r)|^2\,r\,dr}{\int_0^{\infty}|U(r)|^2\,r\,dr} \qquad \ldots\ldots(3.32)$$

* The investigations discussed in this section do not fall under the heading of diffraction theory of aberrations in the customary sense but are included since they are closely related to the problem of optimizing resolution by modifying the waves.

which reaches the circle of radius $r_0 < r_A$ is as large as possible. He showed that the corresponding pupil function is given as a solution of a certain integral equation. He also discussed the resolution of objects of periodic structure and found that in this case it is impossible to increase the resolving power by varying the amplitude only, but that the contrast of the image may be improved considerably by this means.

Couder (1944) discussed amplitude filters which consist of a plano-concave lens filled with an absorbing liquid. The corresponding pupil function is of the form $P(\rho) = 10^{-k\rho^2}$. It is perhaps worth noting that Straubel (1902, 1931) who appears to have been the first to study effects of non-uniformity of amplitude on resolution also considered pupil functions of this form. Couder found that with such a filter it is possible to redistribute the illumination in the diffraction pattern so as to facilitate the observation of double stars whose components differ widely in brightness. Amplitude filters were also discussed by Lansraux (1946, 1949).

Boughon, Dossier and Jacquinot (Boughon *et al.* 1946, Jacquinot *et al.* 1949) considered applications of amplitude variation to spectroscopy. For diffraction at a slit (3.31) and (3.32) are replaced by the equations

$$U(x) = \int_{-\infty}^{+\infty} P(x') e^{2\pi i x x'}\, dx', \qquad \ldots\ldots (3.33)$$

and

$$L(a) = \frac{\int_{-a}^{+a} [U(x)]^2\, dx}{\int_{-\infty}^{+\infty} [U(x)]^2\, dx}. \qquad \ldots\ldots (3.34)$$

The above authors used the analysis of trigonometrical polynomials to determine pupil functions which for a given value of the parameter a make the fraction of the total illumination $L(a)$ as large as possible. They also considered distributions for which $U(x)$ behaves asymptotically like x^{-m}, m being a positive integer. They showed that by the choice of a suitable trigonometric polynomial as pupil function the effect of diffraction fringes can be reduced, facilitating the detection of satellites to bright spectral lines.

Expansion of pupil functions into series of Hermitian polynomials were considered by Duffieux (1946 b). Expansions into series of Legendre polynomials were suggested by Slevogt (1949).

Lansraux (1947 a, b) discussed the evaluation of the diffraction integral (3.31) for cases where $P(\rho)$ can be expanded in a series of lambda functions

$$\Lambda_s(\rho) = [1 - (\rho/\rho_m)^2]^{s-1}. \qquad \ldots\ldots (3.35)$$

He showed that when

$$P(\rho) = \sum_{s=1}^{\infty} a_s \Lambda_s(\rho), \qquad \ldots\ldots (3.36)$$

where the a's are constants

$$U(\rho) = \pi \rho_m{}^2 \sum_{s=1}^{\infty} a_s 2^s (s-1)!\, \frac{J_s(2\pi\rho_m r)}{(2\pi\rho_m r)^s}. \qquad \ldots\ldots (3.37)$$

Lansraux asserts that for a small aberration (3.37) is rapidly convergent. That this is not necessarily so can be illustrated by considering the simplest aberration, namely defocusing, when (3.37) reduces to that solution of Lommel which involves the U_n series. These series converge rapidly only at points in the geometrical shadow (case $|u/v| < 1$ in § 3.1 above). In the direct beam of light even when very near focus the series (3.37) is therefore not suitable for computations.

Hopkins (1949) extended Lommel's analysis to waves of non-uniform amplitude, the non-uniformity considered being of the form $\alpha+\beta(\rho/\rho_m)^2$ where α and β are constants. He obtained a solution in terms of Lommel's original U_n and V_n functions and certain closely related functions and gave a number of intensity distribution curves for out-of-focus patterns. He showed that for most ordinary lens systems it is quite justifiable in evaluating the diffraction integral to assume the disturbance over the converging waves to be of uniform amplitude.

Osterberg and Wilkins (1949) studied theoretically the problem of coating the exit pupil with light absorbing and refracting materials so as to reduce the diameter of the central bright disc in the diffraction pattern. They considered pupil functions of the form (3.36) and derived Lansraux' expression very shortly by the application of Sonine's integral formula. In some cases choice can be

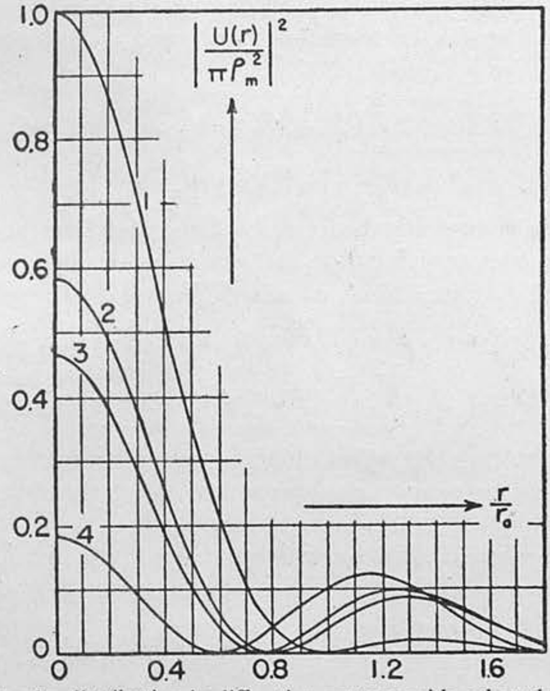


Figure 19. The intensity distribution in diffraction patterns with selected radii r of the first dark ring, the central maximum being the highest possible. Curves 2, 3, 4 correspond to $r=0{\cdot}8r_0$, $0{\cdot}75r_0$ and $0{\cdot}6r_0$ respectively, where r_0 denotes the radius of the first dark ring in the Airy pattern (curve 1). *After* Wilkins (1950).

made of the constants a_s to give pupil functions corresponding to diffraction patterns with prescribed properties. As Osterberg and Wilkins point out, however, it is not possible to go far in this direction because an arbitrary function $U(\rho)$ cannot, in general, be expressed in series of the type (3.37). The resolution of two particles in a bright field by Sonine type microscope objectives (microscope objectives for which $P(\rho)$ is given by (3.36)) was discussed in a paper by Osterberg and Wissler (1949).

Wilkins (1950) showed how to coat the exit pupil so as to obtain for any prescribed diameter of the central disc the highest possible central intensity. In Figure 19 we reproduce the intensity curves which he obtained for a few typical cases.

3.25. *Other researches.*

In his book *L'Intégrale de Fourier et ses Application a l'Optiques* and in a series of papers, Duffieux (1945 a, b, 1946 b, 1947, 1948) has, partly in collaboration with Lansraux (1945), attempted to apply the methods of Fourier transforms to the study of Fraunhofer diffraction.*

We recall that with a suitable interpretation of the variables p, q, x and y the complex disturbances $P(p, q)$ in the exit pupil and $U(x, y)$ in the 'pattern at infinity' due to a point source are related by the pair of double Fourier integrals

$$\left.\begin{aligned} U(x,y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(p,q)e^{-2\pi i(px+qy)}\,dp\,dq, \\ P(p,q) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} V(x,y)e^{2\pi i(px+qy)}\,dx\,dy. \end{aligned}\right\} \qquad \ldots\ldots(3.38)$$

If the system images an *extended incoherent object* in which the intensity distribution is given by $I(x, y)$, then the corresponding intensity distribution in the image is given by

$$\Pi(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I(x',y')E(x-x',y-y')\,dx'\,dy', \qquad \ldots\ldots(3.39)$$

where

$$E(x,y) = |U(x,y)|^2. \qquad \ldots\ldots(3.40)$$

By the application of Parseval's theorem, Duffieux and Lansraux (1945) showed that the double Fourier transform or the space frequency spectrum $T[\Pi(x,y)]$ is connected with that of the object, namely $T[I(x,y)]$, by the relations

$$T[\Pi(x,y)] = T[I(x,y)]\,T[E(x,y)] = D(p,q)\,T[I(x,y)] \qquad \ldots\ldots(3.41)$$

where

$$D(p,q) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E(x',y')e^{2\pi i(px+qy)}\,dx\,dy. \qquad \ldots\ldots(3.42)$$

From (3.41) it follows that $D(p, q)$ may be regarded as a transmission factor which characterizes the imaging properties of the system when the object is incoherently illuminated.

For an *extended coherent object* for which the amplitude distribution is denoted by $I'(x, y)$ the intensity distribution $\Pi'(x, y)$ in the image is given by

$$\Pi'(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I'(x,y)U(x-x',y-y')\,dx'\,dy' \qquad \ldots\ldots(3.43)$$

leading to a relation

$$T[\Pi'(x,y)] = T[I'(x,y)]\,T[U(x,y)] = P(p,q)\,T[I'(x,y)]. \qquad \ldots\ldots(3.44)$$

(3.44) shows that $P(p, q)$ may be regarded as a transmission factor when the object is coherently illuminated.

From D and P Duffieux and Lansraux derived other transmission factors which have a bearing on the theory of test objects for optical instruments. To one such factor $d(p)$, defined by

$$d(p) = \int_{-\infty}^{+\infty} D(p,q)\,dq, \qquad \ldots\ldots(3.45)$$

is related to resolving power in the image of non-coherent objects. By non-rigorous mathematical arguments Duffieux and Lansraux derived a number

* Additional references to papers dealing mainly with cases where the distribution in the image is a function of one variable only will be found in Duffieux (1946 a).

The existence of a Fourier integral relation between the disturbance over a spherical surface filling the exit pupil and the disturbance over its 'focal sphere' was first pointed out by Michelson (1905).

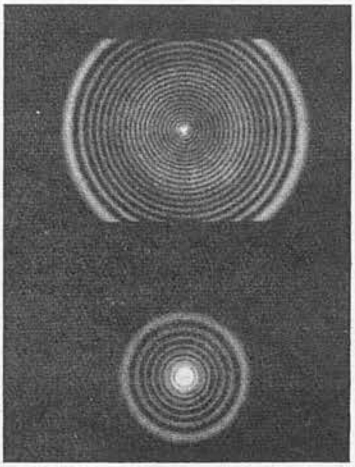


Figure 13. Images in presence of third order spherical aberration of amount $\Phi_{\max}=16\lambda$ at marginal focus and at circle of least confusion. *After* Nienhuis (1948).

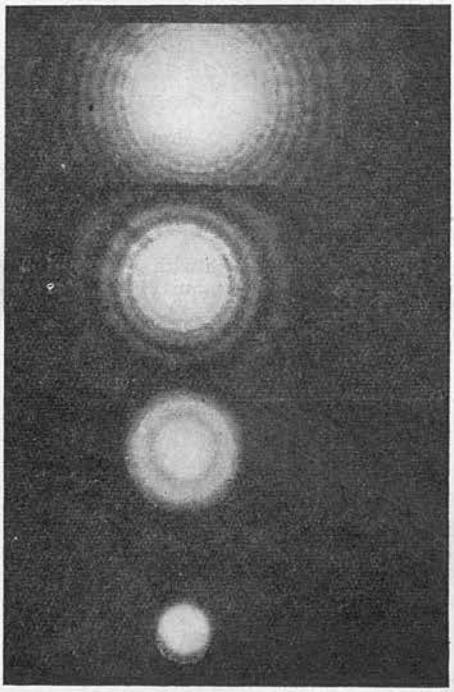


Figure 14. Images in plane of paraxial focus, in presence of third order spherical aberration of amounts $\Phi_{\max}=17\cdot5\lambda$, $8\cdot4\lambda$, $3\cdot7\lambda$, $1\cdot4\lambda$. *After* Nienhuis (1948).

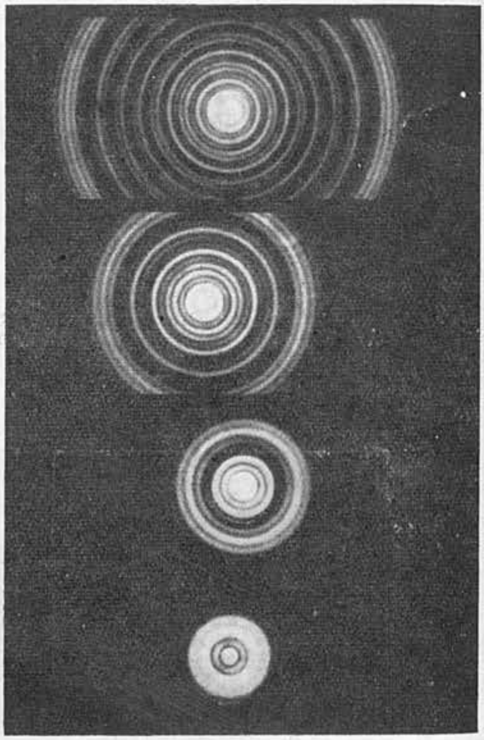


Figure 15. Images in plane of circle of least confusion in presence of third order spherical aberration of amounts $\Phi_{\max}=17\cdot5\lambda$, $8\cdot4\lambda$, $3\cdot7\lambda$, $1\cdot4\lambda$. Scale $3\times$ as in Figure 14. *After* Nienhuis (1948).

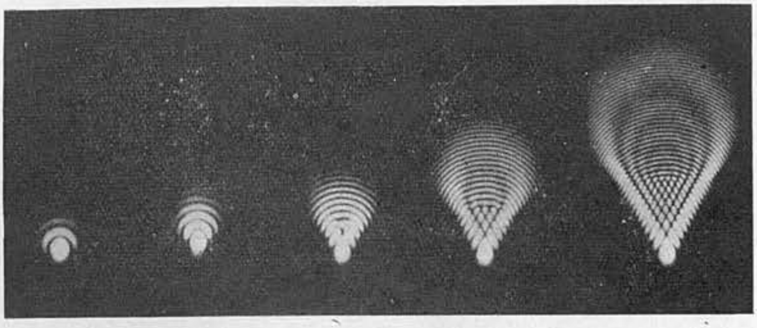


Figure 16. Images in Gaussian plane in presence of third order coma of amounts $\Phi_{max}=0{\cdot}3\lambda$, 1λ, $2{\cdot}4\lambda$, 5λ, 10λ,. *After* Nienhuis, *reprinted from* Zernike (1948).

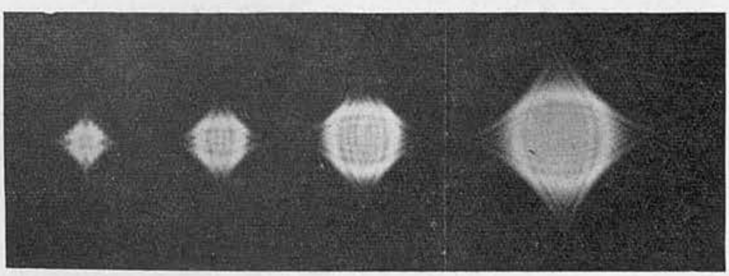


Figure 17. Images in central plane in presence of third order astigmatism of amounts $\Phi_{max}=1{\cdot}4\lambda$ $2{\cdot}7\lambda$, $3{\cdot}5\lambda$, $6{\cdot}5\lambda$. *After* Nienhuis, *reprinted from* Zernike (1948).

of relations connecting the various transmission factors and the intensity distribution in the image. One of these relations obtained by a formal application of Plancherel's theorem connects P with the moments of the energy distribution of the image. Duffieux (1948) stated that these moments when taken about the centre of gravity of the distribution are the same for the geometrical and the physical image. As has been pointed out by E. H. Linfoot *, Duffieux' analysis is rendered invalid by failure to take into account the divergence of certain

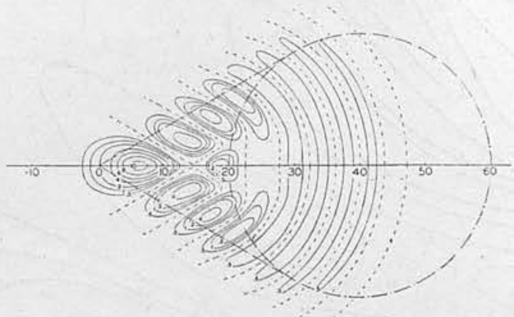


Figure 20. Isophotes in Gaussian plane in presence of third order coma, of amount $\Phi_{max}=3{\cdot}2\lambda$ *After* Kingslake (1948).

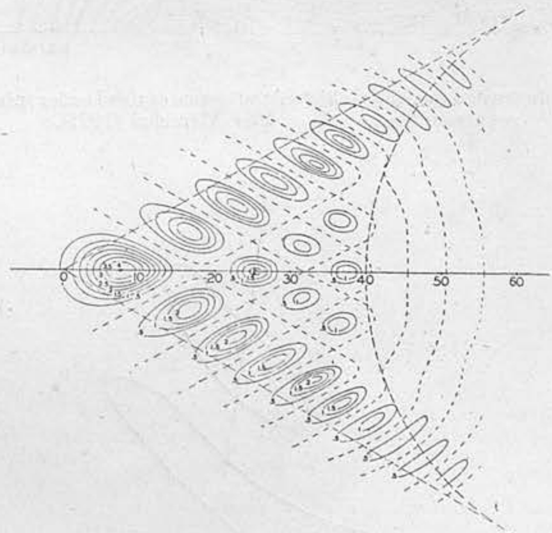


Figure 21. Isophotes in Gaussian plane in presence of third order coma, of amount $\Phi_{max}=6{\cdot}4\lambda$ *After* Kingslake (1948).

integrals; that the result also is incorrect can easily be seen by considering the moments of an error-free image. Several other conclusions of Duffieux are based on faulty mathematical arguments and appear to be incorrect.

Kingslake (1948) investigated the diffraction structure of third order comatic images by numerical integrations. He demonstrated that the image is of the size and shape determined by ordinary geometrical optics and that the main effect of diffraction is to break up the image into an elaborate fine structure of dots and lines of light. We give here two of his valuable diagrams (Figures 20 and 21) showing the isophotes for coma of fairly large amounts.

* *Math. Rev.*, 1946, **7**, 269; *Ibid.*, 1949, **10**, 220.

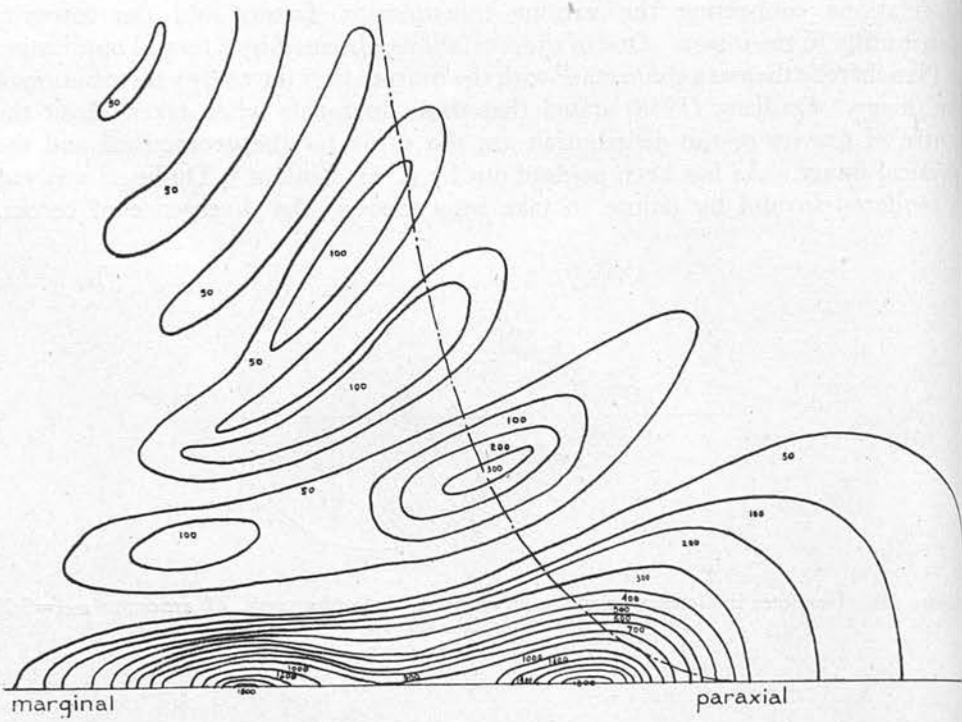


Figure 22. Isophotes in a meridional plane in presence of third order spherical aberration, of amount $\Phi_{max}=4\lambda$. *After* Maréchal (1951).

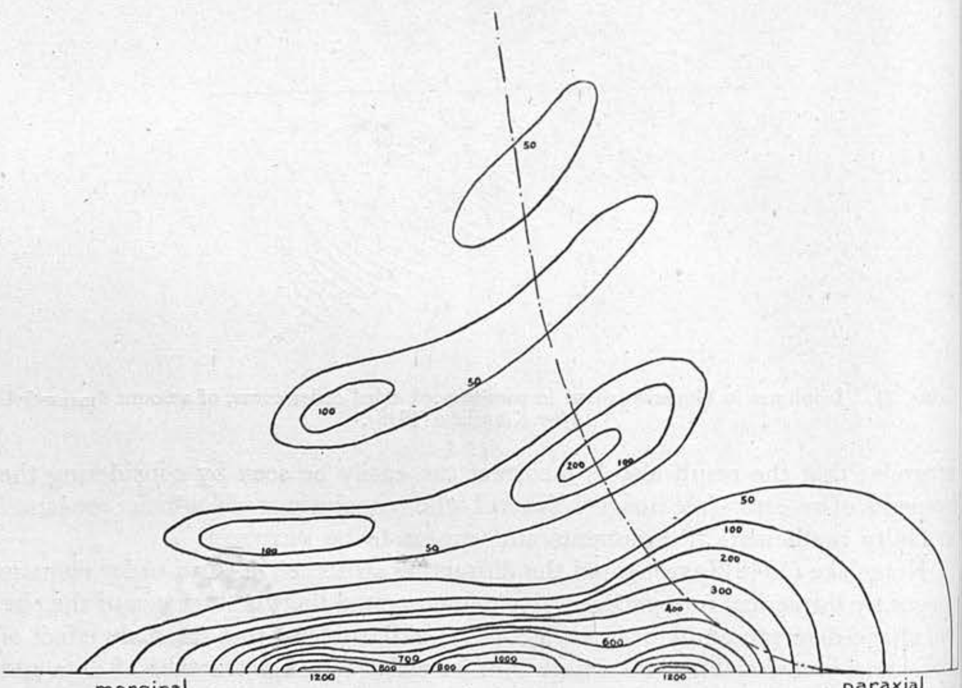


Figure 23. Isophotes in a meridional plane in presence of third order spherical aberration, of amount $\Phi_{max}=6\lambda$. *After* Maréchal (1951).

Maréchal (1947b, 1948) designed and constructed a machine for rapid evaluation of double integrals of the type occurring in the investigations of aberration diffraction effects. To determine the intensity at each point of the image, operations taking about 8–10 minutes are needed. The precision of the apparatus is quite adequate for practical purposes, the estimated inaccuracies ΔI being as follows ($I=1$ at the centre of the Airy pattern): $\Delta I<0{\cdot}01$ when $I\sim 1$, $\Delta I<0{\cdot}003$ when $I\sim 0{\cdot}1$, $\Delta I<0{\cdot}001$ when $I\sim 0{\cdot}01$. With the help of this

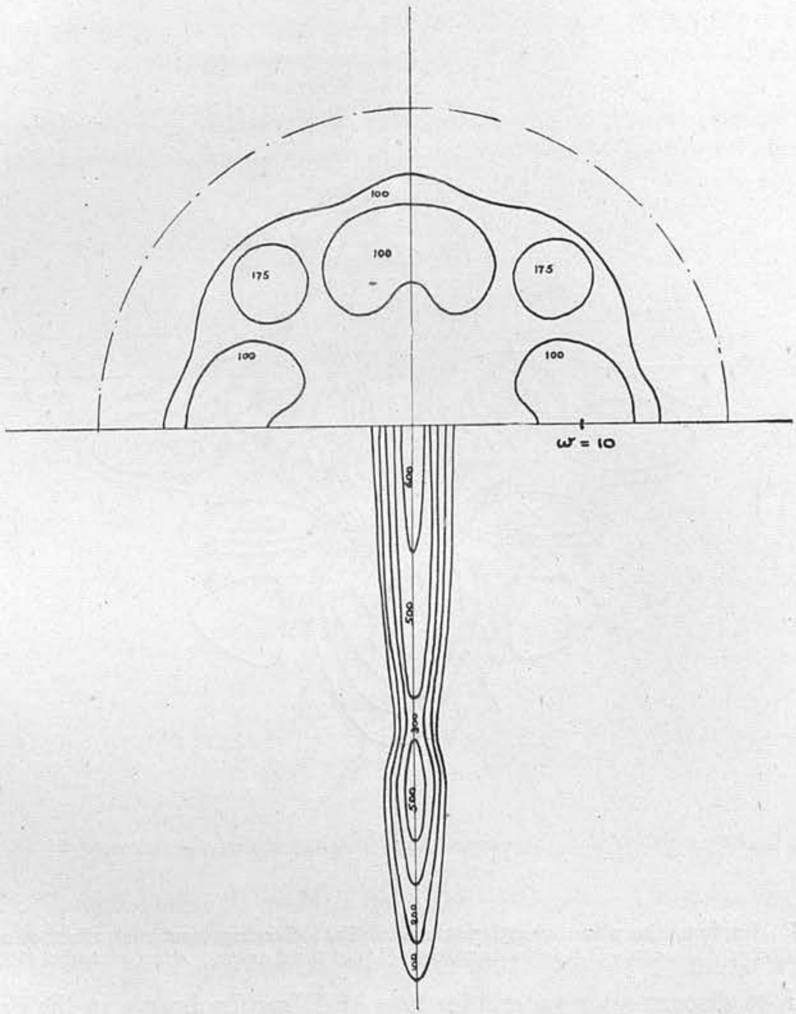


Figure 24. Isophotes in presence of third order astigmatism, of amount $\Phi_{max}=1{\cdot}5\lambda$ in central plane (upper half), in plane containing a focal line (lower half). *After* Maréchal (1951).

integrator Maréchal (1948, 1951) obtained a number of important isophote diagrams some of which are shown in Figures 22–25.* Most of these diagrams show the effects of aberrations in the range 1λ to 10λ for which available series expansions are not well adapted. The calculations performed with the aid of this integrator were also compared by Maréchal with photometric measurements of the intensity distribution in actual images; good agreement was found.

* I am greatly indebted to Dr Maréchal for giving me access to these figures before the publication of his own paper.

It is well known that in the approximations of Kirchhoff's scalar theory, the disturbance produced by the passage of spherical waves through an aperture opening may be regarded as due to the combined effect of 'geometrical waves' and 'diffraction waves' emerging from the aperture edge.* Nienhuis (1948) assumed that this result holds (and also that the 'diffraction waves' possess analogous properties) when the incident waves are not strictly spherical† and

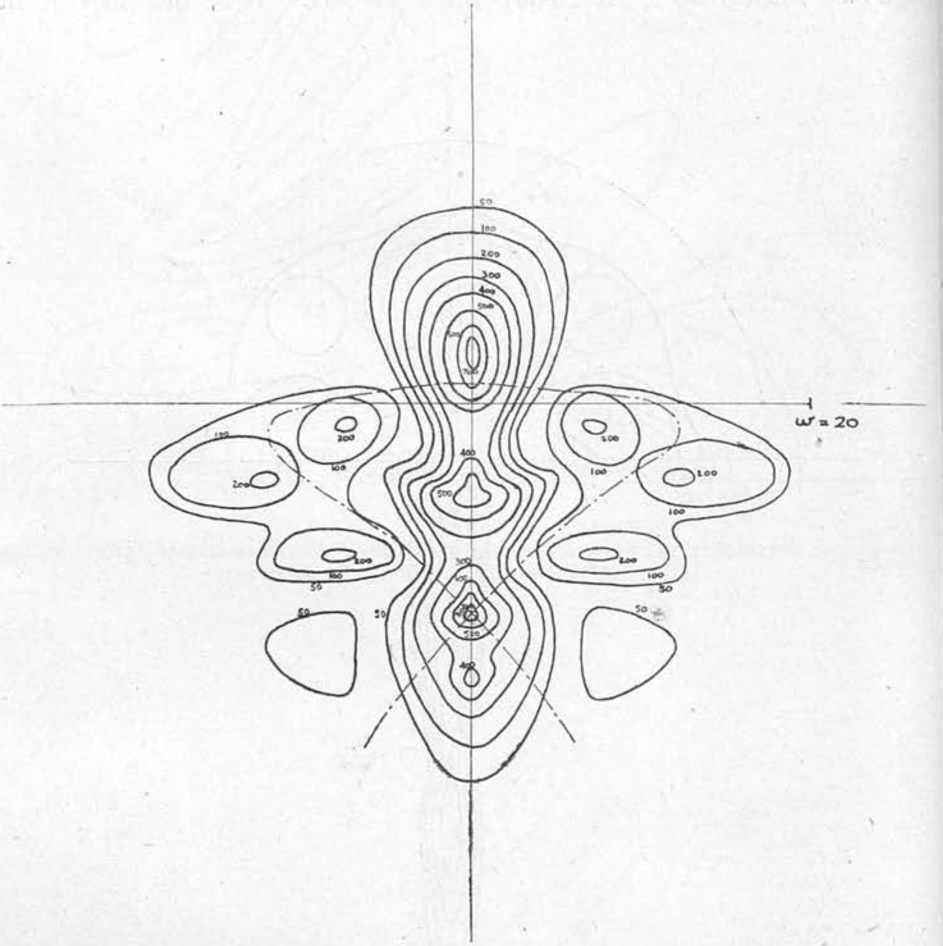


Figure 25. Isophotes in plane of astigmatic focal line, showing combined effect of spherical aberration (2λ), coma (1λ) and astigmatism (1λ) of third order. *After* Maréchal (1951).

applied it to discuss some general features of diffraction images in the presence of third order aberrations.

In the case of astigmatism, Nienhuis showed by a geometrical argument that the 'diffraction waves' from a circular edge give rise to an asteroid pattern. This pattern is partly obstructed by the effect of the 'geometrical waves'. Its

* Ramachandran (1945) showed that the integral expressing the edge effect can be derived in a very simple manner if the variation of the amplitude of the secondary wavelets with direction is neglected. This idea was also used by Kathawate (1945) to discuss the effect of diffraction of spherical waves at boundaries and obstacles of various forms.

† Many assertions in Nienhuis' thesis and in other literature give the impression that the transformation of the diffraction integral into an integral expressing the edge effect and a term expressing the geometrical effect has been carried out in the general case, but an examination of the relevant papers reveals that the published proofs relate to spherical and plane waves only.

shape is in the first approximation independent of the focal setting and is therefore more apparent in a receiving plane through one of the focal lines (see Figure 12).* In the presence of coma, rays from opposite points of each aperture zone meet the Gaussian plane at the same point. Nienhuis found that the fringe structure of the comatic pattern can be ascribed to the interference of these rays, in agreement with his observation that the fringes disappear when half of the light is blocked out by using a semicircular aperture (Figure 10). For spherical aberration Nienhuis found that in the receiving plane through the paraxial focus it is mainly the edge effects which determine the structure of the pattern whilst in other planes the interference of the 'geometrical waves' plays a considerable part.

In connection with this work it is of interest to note that Durand (1949) studied in a comprehensive manner the fringe structure of third order aberration

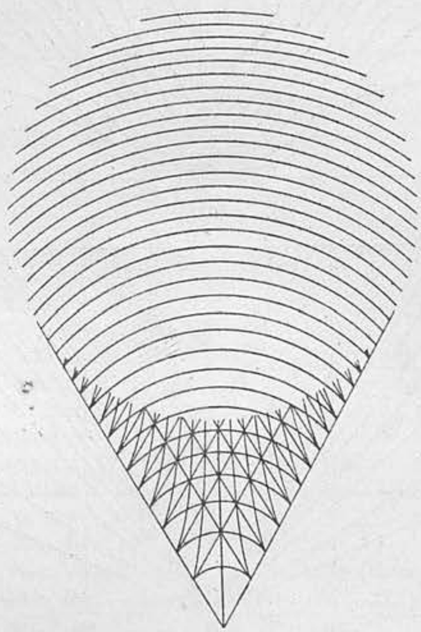


Figure 26. **Fringe structure of comatic pattern of amount $\Phi_{max}=8{\cdot}3\lambda$.** ***After*** **Durand (1949).**

figures. Though Durand's analysis does not take account of diffraction it has many points of contact with the work of Nienhuis. Durand's results may be expected to give good approximations to the appearance of the image whenever edge effects do not play an important part. This is particularly well illustrated by the case of coma (Figure 26).

Françon (1948 a, b) discussed the effects on the image due to a certain type of non-homogeneity in the glass of an objective. He found that in the presence of filaments or veins whose refractive index differs slightly from the bulk of the glass, the effects are negligible if the path differences they introduce are less than $\lambda/8$ for $p=50$, $\lambda/6$ for $p=100$ and $\lambda/4$ for $p=200$, p being the ratio of the

* Some of Nienhuis' photographs of diffraction images are also given by Zernike (1948). The corresponding values of the aberration constants differ in several cases and some appear to be incorrect. In particular it appears that the aberration constants for the first two comatic patterns of Figure 16 should be larger. Values of Φ_{max} based on Nienhuis' data are given in Figures 16 and 17.

diameter of the objective to the width of the filament. When one observes two neighbouring objects of widely differing brightness the tolerances are much more severe.

Epstein (1949) took up once more the problem treated by Lommel (1885) (see § 3.1) and found methods for the numerical evaluation of the associated diffraction integral, which possess some advantages over those of Conrady and Buxton.

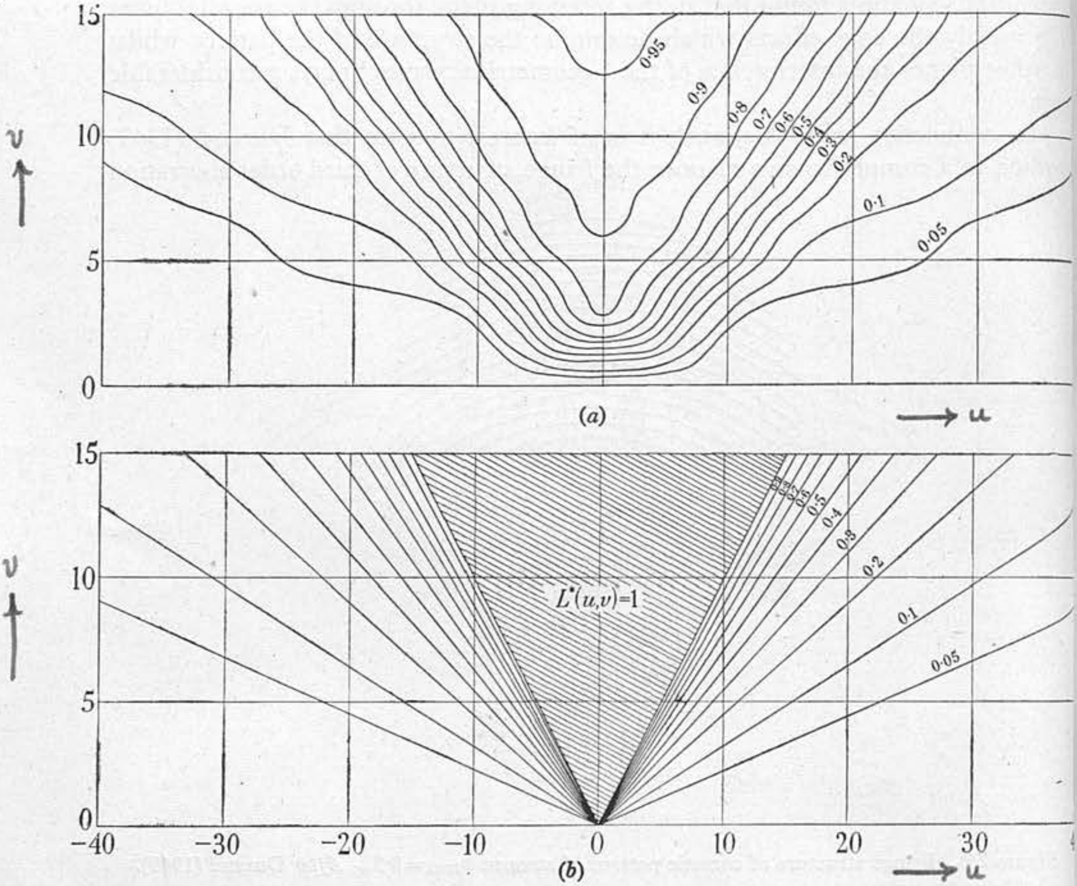


Figure 27.

$$u=\frac{2\pi}{\lambda}\left(\frac{a}{f}\right)^2 \Delta f, \quad v=\frac{2\pi}{\lambda}\frac{a}{f}r,$$

a = radius of exit pupil, f = focal length, Δf = amount of defocusing.

Contour lines for the fraction of the total illumination inside circles of radius r, centred on axis, in receiving planes near focus of spherical waves issuing from a circular aperture (*a*) according to diffraction theory, (*b*) according to geometrical optics. *After* Wolf (1950).

The distribution of the total illumination in out-of-focus images formed by spherical waves issuing from a circular aperture was investigated by Wolf (1951). He derived expressions for the fraction of the total illumination present inside circles of given radii in the receiving plane and applied them to study the distribution in a few selected planes near the focus. Wolf also gave a diagram (see Figure 27) showing the corresponding three-dimensional distribution and compared it with the prediction of geometrical optics. His results, together with those of Lommel, Zernike and Nijboer, provide a mathematical basis for a quantitative discussion of the imaging properties in optical systems where, as for example in a well corrected refracting telescope objective, the chromatic variation of focus is the only appreciable aberration.

Mme. Regnier (1950) studied the influence of spherical aberration on the limiting visibility of Foucault test objects. She used a relation due to Duffieux and Lansraux between the pupil function $P(p, q)$, the transmission factor $d(p)$ and the limiting visibility, and calculated $d(p)$ for spherical aberration of amounts $\lambda/4$, $\lambda/2$, $3\lambda/4$, λ and for different forms of exit pupil. The results yield conclusions about the dependence of the limiting visibility of test objects on the aberration of the objective.

ACKNOWLEDGMENTS

The author gratefully acknowledges the permission granted by various authors and publishing bodies for the reproduction of diagrams and photographs.

REFERENCES

AIRY, G. B., 1835, *Trans. Camb. Phil. Soc.*, **5**, 283.

BEREK, M., 1926, *Z. Phys.*, **40**, 421.

BORN, M., 1932, *Naturwissenschaften*, **20**, 921; 1933, *Optik* (Berlin: Springer), p. 202.

BOUGHON, P., DOSSIER, B., JACQUINOT, P., 1946, *C. R. Acad. Sci., Paris*, **223**, 661.

BUXTON, A., 1921, *Mon. Not. R. Astr. Soc.*, **81**, 547; 1923, *Ibid.*, **83**, 475.

CONRADY, A. E., 1919, *Mon. Not. R. Astr. Soc.*, **79**, 575.

VAN DER CORPUT, J. G., 1948, *Proc. K. Akad. wet. Amst.*, **51**, 6.

COUDER, A., 1944, *C. R. Acad. Sci., Paris*, **219**, 677.

DUFFIEUX, P. M., 1945 a, *C. R. Acad. Sci., Paris*, **220**, 846; 1945 b, *Ibid.*, **220**, 911; 1946 a, *L'Intégrale de Fourier et ses Applications a l'Optique* (Besançon: Faculté des Sciences); 1946 b, *C. R. Acad. Sci., Paris*, **222**, 1482; 1947, *Ann. Phys., Paris*, **2**, 95; 1948, *C. R. Acad. Sci., Paris*, **226**, 1257.

DUFFIEUX, P. M., and LANSRAUX, G., 1945, *Rev. Opt.* (*théor. instrum.*), **24**, 65, 151, 215.

DURAND, E., 1949, Contribution in *La Théorie des Images Optiques* (Paris : Éditions de la *Revue d'Optique*), p. 127.

EPSTEIN, L., 1949, *J. Opt. Soc. Amer.*, **39**, 226.

FRANÇON, M., 1944, *Cah. Phys.* No. 26, 15; 1947 a, *Rev. Opt.* (*théor. instrum.*), **26**, 354; 1947 b, *Ibid.*, **26**, 369; 1948 a, *Ibid.*, **27**, 157; 1948 b, *Ibid.*, **27**, 595, 761; 1949, *C. R. Acad. Sci., Paris*, **229**, 293.

HOPKINS, H. H., 1943, *Proc. Phys. Soc.*, **55**, 116; 1944, *Ibid.*, **56**, 48; 1949, *Proc. Phys. Soc.* B, **62**, 22.

JACQUINOT, P., BOUGHON, P., DOSSIER, B., 1949, Contribution in *La Théorie des Images Optiques* (Paris : Éditions de la *Revue d'Optique*), p. 183.

VAN KAMPEN, N. G., 1949, *Physica*, **14**, 575.

KATHAWATE, Y. V., 1945, *Proc. Indian. Acad. Sci.* A, **21**, 177.

KINGSLAKE, R., 1948, *Proc. Phys. Soc.*, **61**, 147.

KÖNIG, A., 1929, *Handb. Exp. Phys.*, **20**, Part II, 141.

KORRINGA, J., NIJBOER, B. R. A., and MARÉCHAL, A, Document S.O. 48-8, Commission Internationale d'Optique (Paris: Institute d'Optique).

LANSRAUX, G., 1946, *C. R. Acad. Sci., Paris*, **222**, 1434; 1947 a, *Rev. Opt.* (*théor. instrum.*), **26**, 24; 1947 b, *Ibid.*, **26**, 278; 1949, Contribution in *La Théorie des Images Optique* (Paris : Éditions de la *Revue d'Optique*), p. 194.

VON LAUE, M. V. 1928, *Handb. Exp. Phys.*, **18**, 211.

LOMMEL, E., 1885, *Abh. Bayer. Akad.*, **15**, Abth. 2, 229; 1886, *Ibid.*, **15**, Abth. 3, 529. Shortened account given in GRAY, A., MATHEWS, G. B., and MACROBERT, T. M., 1931, *A Treatise on Bessel Functions*, 2nd Edit. (Macmillan: London).

LUNEBERG, R. K., 1944, *Mathematical Theory of Optics* (Providence, R.I.: Brown University).

MARÉCHAL, A., 1944 a, *C. R. Acad. Sci., Paris*, **218**, 345; 1944 b, *Ibid.*, **218**, 395; 1944 c, *Cah. Phys.*, No. 26, 1; 1947 a, *Rev. Opt. (théor. instrum.)*, **26**, 257; 1947 b, *J. Opt. Soc. Amer.*, **37**, 403; 1948, *Rev. Opt. (théor instrum.)*, **27**, 73; 1951, *Ibid.*, in the press.
MARTIN, L. C., 1922, *Mon. Not. R. Astr. Soc.*, **82**, 310; 1946, *Photogr. J.*, **86 B**, 47.
MICHELSON, A. A., 1905, *Phil. Mag.*, **9**, 506.
NIENHUIS, K., 1948, *Thesis*, University of Groningen.
NIENHUIS, K., and NIJBOER, B. R. A., 1949, *Physica*, **14**, 590.
NIJBOER, B. R. A., 1942, *Thesis*, University of Groningen; 1943, *Physica*, **10**, 679; 1947, *Ibid.*, **13**, 605.
OSTERBERG, H., and WILKINS, J. E., 1949, *J. Opt. Soc. Amer.*, **39**, 553.
OSTERBERG, H., and WISSLER, F. C., 1949, *J. Opt. Soc. Amer.* **39**, 558.
PICHT, J., 1925, *Ann. Phys., Lpz.*, **77**, 685; 1926, *Ann. Phys., Lpz.*, **80**, 491. (Also 1931, *Optische Abbildung* (Braunschweig: Vieweg).)
RAMACHANDRAN, G. N., 1945, *Proc. Indian Acad. Sci.* A, **21**, 165.
RAYLEIGH, 1879, *Phil. Mag.*, **8**, 403. (Also *Scientific Papers*, **1**, 428.)
RÉGNIER SIMONE M. J., 1950, *Rev. Opt. (théor. instrum.)*, **29**, 315.
RICHTER, R., 1925, *Z. Instrumkde*, **45**, 1.
SCHWARZSCHILD, K., 1897, *S.B. München Akad. Wiss. (Math.-Phys. Kl.)*; **27**, 271; 1905, *Astro. Mitt. Köngl. Sternwarte, Göttingen*, **9**. (Also *Abh. K. Gess. Wiss., Göttingen, Math.-Phys. Kl.*, **4**.)
SLEVOGT, H., 1949, *Optik*, **4**, 349.
STREHL, K., 1894, *Theorie des Fernrohrs* (Leipzig: Barth).
STEWARD, G. C., 1925, *Phil. Trans. Roy. Soc.* A, **225**, 131; 1926, *Trans. Camb. Phil. Soc.*, **23**, 235. (Also 1928, *The Symmetrical Optical System*, Cambridge Tracts, No. 25.)
STRAUBEL, R., 1902, *Phys. Z.*, **4**, 16, 74; 1931, *8th International Congress on Photography*, Dresden (Leipzig: Barth), p. 353.
STRUVE, H., 1886, *Mém. Akad. Sci., St. Pétersbourg* (7), **34**, No. 5, 1.
TORALDO DI FRANCIA, G., 1946, *Atti Fond. " Giorgio Ronchi "*, **1**, 136.
VÄISÄLÄ, Y., 1922, *Annales Univ. Fennicae Aboensis*, A, **1**, 108.
WANG TA-HANG, 1941, *Proc. Phys. Soc.*, **53**, 157.
WILKINS, J. E., 1950, *J. Opt. Soc. Amer.*, **40**, 222.
WOLF, E., 1951, *Proc. Roy. Soc.* A, **204**, 533.
ZERNIKE, F., 1934, *Physica*, **1**, 689; 1948, *Proc. Phys. Soc.*, **61**, 158.
ZERNIKE, F., and NIJBOER, B. R. A., 1949, Contribution in *La Théorie des Images Optiques* (Paris: Éditions de la *Revue d'Optique*), p. 227.

1·5

Reprinted without change of pagination from the
Proceedings of the Royal Society, A, *volume* 212, p. 426, 1952

On the foundation of the scalar diffraction theory of optical imaging

BY O. THEIMER†
Department of Physics, Technische Hochschule, Graz, Austria,

G. D. WASSERMANN
Department of Mathematics, King's College, Newcastle-upon-Tyne

AND E. WOLF
Department of Mathematical Physics, University of Edinburgh

(*Communicated by M. Born, F.R.S.—Received* 30 *November* 1951)

In this paper the relationship between the rigorous vector theory and the approximate scalar theory is investigated and the usual procedure of calculating the intensity in optical images from a single scalar wave-function is justified. Unlike the earlier discussion of Picht (1925) the present analysis takes fully into account the polarization and amplitude properties which an actual light source and the optical instrument impose upon the radiation field.

In the calculations of diffraction images the effect of the non-monochromatic nature of natural light is usually disregarded. We give an estimate for the maximum wave-length spread of the source which will give rise to a diffraction pattern that may be calculated from the single monochromatic wave-function of the scalar theory.

We find that one of the main reasons for the validity of the scalar theory for wide class of optical instruments is the fact that each monochromatic component of the spectrum gives rise to **E** and **H** fields with the property that at any particular instant of time they are nearly constant over each of the geometrical wave-fronts which are not too close to the source or the image region.

1. INTRODUCTION

According to electromagnetic theory, the light intensity is defined as the time average of the energy which crosses per second a unit area which contains the electric vector **E** and the magnetic vector **H**. Since this energy flow is given by the Poynting vector

$$\mathbf{S} = \frac{c}{4\pi}(\mathbf{E} \wedge \mathbf{H}), \tag{1·1}$$

and since the **E** and **H** fields are generally determined from *three* independent scalar wave-functions (e.g. the components of a Hertzian vector), it follows that the intensity must also generally be calculated from three scalar wave-functions. It is well known, however, that for the purposes of instrumental optics the intensity may as a rule be taken as approximately equal to the square of the modulus of a *single* scalar wave-function which is usually identified with a component of one of the field vectors **E** or **H** or of a Hertzian vector. In spite of the fact that practically all diffraction calculations which relate to optical imaging are based on such an approximate treatment, no complete justification of this procedure has been given.

† On leave of absence at the Department of Mathematical Physics, University of Edinburgh.

The only previous investigation which deals with the relation between the rigorous theory based on three scalar functions (referred to as 'vector theory') and the approximate treatment which employs one scalar function (referred to as 'scalar theory') appears to be that due to Picht (1925). He identified the single wave-function of the scalar theory with a component of the Hertzian vector. A scrutiny of his analysis shows, however, that his proof is unsatisfactory and does not bring out the essential physical aspects of the problem. In the first place, Picht considered only the idealized, strictly monochromatic case; he did not take into account the polarization and the amplitude properties which an actual source and the optical instrument impose upon the radiation field in the image space. It will be evident from the analysis of the present paper that these properties are in fact essential for a complete justification of the validity of the scalar theory. Further, Picht confined his attention only to points in the image space through which a single ray passes; thus he excluded exactly those points which are of the greatest practical importance, namely, those in the region of focus. Apart from these considerations it must also be pointed out that the scalar function used by Picht is not formally equivalent to the scalar function employed by recent writers (cf. review by Wolf (1951)).

In this paper we shall investigate the relation between the two theories and justify the use of a single scalar wave-function for the calculation of intensity in a wide class of instruments. We find that one of the main reasons for the approximate validity of the scalar theory in such cases is the fact that each monochromatic component of the spectrum of the source gives rise to **E** and **H** fields with the property that at any particular instant of time they are nearly constant over each of the geometrical wave-fronts which are not too close to the source or the image region.

Owing to the generality of the problem, our arguments are of necessity of a qualitative rather than of a quantitative nature; in any particular case the errors involved can, however, be estimated. In the last section we give an estimate for the maximum wave-length spread of the source that will give rise to a diffraction pattern which may be calculated from a single monochromatic wave-function of the scalar theory.

2. Preliminary considerations

We consider a symmetrical optical system with a point source at Q_0 (figure 1) emitting natural, nearly monochromatic light of angular frequency ω_0. At Q_0 we choose a set of rectangular Cartesian axes (x_1, x_2, x_3) with the x_3 direction along the principal ray. The source may be regarded as a dipole of moment $\mathbf{P}(t)$ which varies both in magnitude and direction with time t. The components of $\mathbf{P}(t)$ in the three directions will be written in the form of Fourier integrals,

$$P_j(t) = \frac{1}{\surd(2\pi)}\int_{-\infty}^{+\infty} p_j(\omega)\exp \mathrm{i}\omega t\,\mathrm{d}\omega \quad (j = 1, 2, 3). \tag{2·1}$$

Since $\mathbf{P}(t)$ is real it follows that the complex quantities $p_j(t)$ satisfy the relations

$$p_j^*(\omega) = p_j(-\omega), \tag{2·2}$$

where the asterisk denotes the complex conjugate. Consequently (2·1) can be written as

$$P_j(t) = \mathscr{R}\sqrt{\frac{2}{\pi}}\int_0^\infty p_j(\omega)\exp i\omega t\,d\omega \quad (j = 1, 2, 3), \tag{2·3}$$

$\mathscr{R}$ denoting the real part.

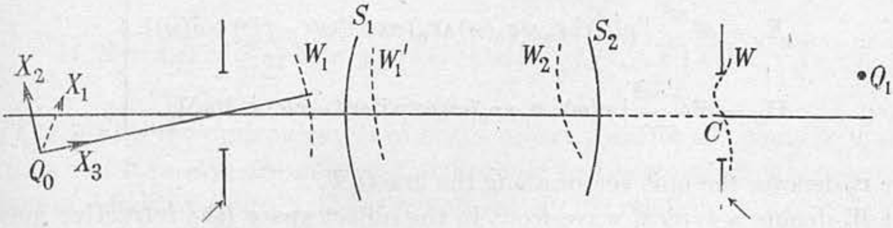


FIGURE 1

We denote by $|p_j(\omega)|$ and $\delta_j(\omega)$ the amplitude and the phase of $p_j(\omega)$,

$$p_j(\omega) = |p_j(\omega)|\exp i\,\delta_j(\omega). \tag{2·4}$$

Since the source is assumed to emit nearly monochromatic light the modulus $|p_j(\omega)|$ will, for each j, differ appreciably from zero only within a narrow interval $(\omega_0 - \frac{1}{2}\Delta\omega, \omega_0 + \frac{1}{2}\Delta\omega)$. The assumption that the light is natural implies that $\delta_j(\omega)$ are rapidly and irregularly varying functions over the ω range (cf. Planck 1900).

For a fixed j, the typical monochromatic component

$$\mathscr{R}\{p_j(\omega)\exp i\omega t\} = |p_j(\omega)|\cos[\omega t + \delta_j(\omega)] \tag{2·5}$$

may be regarded as representing a Hertzian oscillator with its axis along the x_j direction, giving rise to linearly polarized monochromatic light of frequency ω. The superposition of these components in (2·3) leads to the vector $\mathbf{P}(t)$ whose dependence on t is of complicated character; in consequence, no matter how small $\Delta\omega$ may be, the resultant field will no longer be polarized.

We shall assume throughout that the inclinations to the axis of the rays which pass through the instrument are not large, say not more than 10° or 15°. This condition is satisfied by a wide class of optical instruments. (In most telescopes, for example, the maximum inclination does not exceed 3°.)

3. The field produced by a single monochromatic oscillator

3·1. Since the electromagnetic field due to the actual light source may be regarded as a superposition of strictly monochromatic fields, it will be convenient to examine first the contributions to the $\mathbf{E}$ and $\mathbf{H}$ vectors arising from a typical Hertzian oscillator at Q_0 with sharp frequency ω. As the field of such an oscillator is weak in the neighbourhood of its axis, and as we assume that the angles which the diameters of the entry pupil subtend at Q_0 are small, it follows that only the components $P_1(t)$ and $P_2(t)$ of $\mathbf{P}(t)$ will substantially contribute to the field. We shall therefore take as our typical oscillator one which has its axis in the x_1x_2 plane.†

† This choice is of importance. The main conclusion of the present section holds only for fields produced by oscillators which have their axes almost at right angle to the principal ray.

Let $$\mathscr{R}\{p(\omega)\,\boldsymbol{\rho}_0(\omega)\exp \mathrm{i}\omega t\} \tag{3·1}$$

be the moment of this typical dipole, $\boldsymbol{\rho}_0(\omega)$ being a unit vector in the direction of its axis. Such a dipole will produce at a point T *in vacuo*, whose distance from Q_0 is large compared to the wave-length $\lambda = 2\pi c/\omega$, a field given by (cf. Joos 1947, p. 325)

$$\left.\begin{aligned} \mathbf{E}_\omega &= \mathscr{R}\frac{\omega^2}{c^2 r}\,|\,p(\omega)\,|\,\mathbf{r}_0\wedge[\boldsymbol{\rho}_0(\omega)\wedge\mathbf{r}_0]\exp \mathrm{i}[\omega(t-r/c)+\delta(\omega)],\\ \mathbf{H}_\omega &= \mathscr{R}\frac{-\omega^2}{c^2 r}\,|\,p(\omega)\,|\,\mathbf{r}_0\wedge\boldsymbol{\rho}_0(\omega)\exp \mathrm{i}[\omega(t-r/c)+\delta(\omega)], \end{aligned}\right\} \tag{3·2}$$

where $\mathbf{r}_0$ denotes the unit vector along the line Q_0T.

Let W_1 denote a typical wave-front in the object space (the refractive index of which is assumed to be unity) at distance from Q_0 which is large compared to the wave-length. Since we assume that the angles which the rays make with the axis of the system are small, it follows immediately from (3·2) that at any particular instant of time the vectors $\mathbf{E}_\omega$ and $\mathbf{H}_\omega$ do not vary appreciably in magnitude and direction over W_1. (For any given case the actual variation can be calculated without difficulty from (3·2).)

The effect of the first surface† S_1 on the field is twofold.‡ First, the amplitudes of the field vectors are diminished on account of reflexion losses; secondly, the planes of polarization are rotated. Fresnel's formulae show that both these effects depend mainly on the magnitudes of the angle of incidence. If this angle is small (say 10° or so), reflexion losses are also small ($\simeq 5\,\%$) and the rotations of the planes of polarization do not exceed a few degrees (cf. Born 1933, p. 34). Moreover, these effects are practically uniform over S_1. Since the instantaneous values of $\mathbf{E}_\omega$ and $\mathbf{H}_\omega$ do not vary appreciably over the wave-front W_1 they will consequently not vary appreciably over the wave-front W_1' which immediately follows the surface S_1 (see figure 1). The same applies to the behaviour of the two fields over any other wave-front in the space between S_1 and the second surface S_2. For, in the first place, as Friedlander (1947) showed, in a homogeneous medium the polarization along each ray remains constant. Secondly, since the wave-fronts are nearly spherical (centred on the Gaussian image of Q_0 by the first surface), the amplitudes will be diminished almost in the ratio of their paraxial radii of curvature.

Repeating these arguments we finally arrive at a wave-front W which passes through the centre C of the exit pupil (figure 1) and find again that the instantaneous values of $\mathbf{E}_\omega$ and $\mathbf{H}_\omega$ do not vary appreciably over it. With the help of this result we shall now be able to give a mathematical representation for the field vectors in the region of the image.

3·2. We choose a system of rectangular Cartesian axes (x, y, z) with origin at the Gaussian image point Q_1 of Q_0 formed by light of wave-length $\lambda_0 = 2\pi c/\omega_0$ and with

† We assume here that S_1 is a refracting surface. If S_1 is a mirror, no essential modifications of our argument are necessary as is seen by inspection of Fresnel's formulae.

‡ In ordinary optical systems absorption losses can be neglected. In recent years attempts have been made to improve the quality of the image by the introduction of artificial absorbing layers (cf. Osterberg & Wilkins 1949) or of special absorbing filters (cf. Couder 1944). Such special systems would require a more refined investigation and will not be considered here.

the z direction along the principal ray CQ_1. The field at all points except those in the immediate neighbourhood of the geometrical shadow can be approximately expressed in the form† (cf. Friedlander 1947)

$$\left.\begin{aligned} \mathbf{E}_\omega(x,y,z,t) &= \mathscr{R}\frac{\omega^2}{c^2}\mathbf{e}_\omega(x,y,z)\exp\mathrm{i}\left[\omega\left(t-\frac{1}{c}f_\omega(x,y,z)\right)+\delta(\omega)\right],\\ \mathbf{H}_\omega(x,y,z,t) &= \mathscr{R}\frac{\omega^2}{c^2}\mathbf{h}_\omega(x,y,z)\exp\mathrm{i}\left[\omega\left(t-\frac{1}{c}f_\omega(x,y,z)\right)+\delta(\omega)\right].\end{aligned}\right\} \tag{3·3}$$

Here $f_\omega(x,y,z)$ is the optical length from the object point to the point (x,y,z) and $\mathbf{e}_\omega(x,y,z)$ and $\mathbf{h}_\omega(x,y,z)$ are mutually orthogonal real vectors. In a homogeneous medium of refractive index n, these vectors satisfy the relation

$$|\mathbf{h}_\omega| = n\,|\mathbf{e}_\omega|. \tag{3·4}$$

We take a reference sphere Σ centred on Q_1 which passes through the centre C of the exit pupil (figure 2). We denote by R the radius CQ_1 of this sphere. In practice the distance, measured along a ray, between Σ and the wave front W will at no point of W exceed more than a few dozen wave-lengths. Consequently on Σ just as on W the vectors $\mathbf{e}_\omega$ and $\mathbf{h}_\omega$ will be practically constant in magnitude and direction.

Let $Q(X,Y,Z)$ be a point in the region of the image where the intensity is to be determined. If the angles which the diameters of the exit pupil subtend at Q are small, we may apply Kirchhoff's formula with the usual approximations and find on integrating (3·3) over that part Σ' of Σ which fills approximately the exit pupil,‡

$$\left.\begin{aligned} \mathbf{E}_\omega(X,Y,Z,t) &= \mathscr{R}\frac{\omega^3}{2\pi\mathrm{i}c^3}\exp\mathrm{i}[\omega t+\delta(\omega)]\\ &\qquad\times\iint_{\Sigma'}\frac{1}{s}\mathbf{e}_\omega(x',y',z')\exp\left\{-\frac{\mathrm{i}\omega}{c}[f_\omega(x',y',z')+s]\right\}\mathrm{d}\sigma',\\ \mathbf{H}_\omega(X,Y,Z,t) &= \mathscr{R}\frac{\omega^3}{2\pi\mathrm{i}c^3}\exp\mathrm{i}[\omega t+\delta(\omega)]\\ &\qquad\times\iint_{\Sigma'}\frac{1}{s}\mathbf{h}_\omega(x',y',z')\exp\left\{-\frac{\mathrm{i}\omega}{c}[f_\omega(x',y',z')+s]\right\}\mathrm{d}\sigma',\end{aligned}\right\} \tag{3·5}$$

where $s = s(x',y',z',X,Y,Z)$ denotes the distances from a typical point of the reference sphere to Q.

Since the vectors $\mathbf{e}_\omega(x',y',z')$ and $\mathbf{h}_\omega(x',y',z')$ do not vary appreciably over the surface of integration we may replace them, without introducing appreciable errors, by the values $\mathbf{e}_\omega(0,0,-R)$, and $\mathbf{h}_\omega(0,0,-R)$ which they take at the centre C of the

† Our equations differ from Friedlander's by the presence of the factor $(\omega^2/c^2)\exp\mathrm{i}\,\delta(\omega)$ and the omission of higher-order terms. This factor suggested by the comparison with (3·2) is unimportant for Friedlander's analysis but has to be included in our present formulae in view of the subsequent integration over the ω band.

‡ In our choice of the surface of integration we have followed recent writers (e.g. Nijboer 1947). We could, alternatively, have performed the integration directly over the wave-front W or over the exit pupil.

exit pupil. Since these vectors are mutually orthogonal and satisfy (3·4), we may set, if in addition we take $n = 1$,

$$\left.\begin{aligned}\mathbf{e}_\omega(0,0,-R) &= a(\omega)\,\boldsymbol{\alpha}(\omega),\\ \mathbf{h}_\omega(0,0,-R) &= a(\omega)\,\boldsymbol{\beta}(\omega),\end{aligned}\right\} \tag{3·6}$$

where $\boldsymbol{\alpha}(\omega)$ and $\boldsymbol{\beta}(\omega)$ are unit mutually orthogonal vectors in the plane perpendicular to the z direction. (3·5) then becomes

$$\left.\begin{aligned}\mathbf{E}_\omega(X,Y,Z,t) &= \mathscr{R}\frac{\omega^2}{c^2}u_\omega(X,Y,Z)\,a(\omega)\,\boldsymbol{\alpha}(\omega)\exp\mathrm{i}\{\omega t+\delta(\omega)\},\\ \mathbf{H}_\omega(X,Y,Z,t) &= \mathscr{R}\frac{\omega^2}{c^2}u_\omega(X,Y,Z)\,a(\omega)\,\boldsymbol{\beta}(\omega)\exp\mathrm{i}\{\omega t+\delta(\omega)\},\end{aligned}\right\} \tag{3·7}$$

u_ω denoting the scalar wave-function

$$u_\omega(X,Y,Z) = \frac{\omega}{2\pi\mathrm{i}c}\iint_{\Sigma'}\frac{1}{s}\exp\left\{-\frac{\mathrm{i}\omega}{c}[f_\omega(x',y',z')+s]\right\}\mathrm{d}\sigma'. \tag{3·8}$$

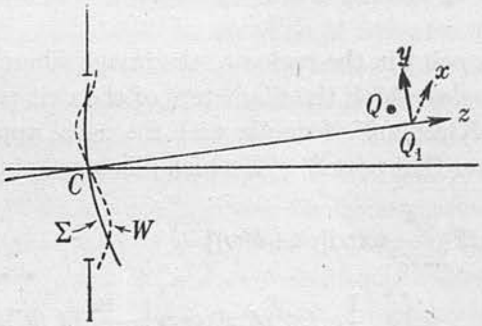


FIGURE 2

From (3·7) one can immediately show by calculating the Poynting vector $\mathbf{S}_\omega = \frac{c}{4\pi}[\mathbf{E}_\omega \wedge \mathbf{H}_\omega]$ and taking the time average, that the intensity at the point $Q(X,Y,Z)$ due to the single dipole (3·1) at Q_0 is proportional to the square of the modulus of the scalar wave-function $u_\omega(X,Y,Z)$. For a full justification of the validity of the scalar theory we must, however, carry out the time averaging not for the monochromatic field but for the total field produced by the actual light source. This will be done in the next section with the help of the results which we have just derived.

4. The total field in the image region

We saw that the contributions of each frequency component to the total field may be regarded as arising essentially from two dipoles at Q_0 with their axes along the x_1 and x_2 directions respectively. Hence it follows from (2·3) and (3·7) that the total field in the image region, due to a point source of natural light at P_0, may be expressed approximately in the form

$$\left.\begin{aligned}\mathbf{E}(X,Y,Z,t) &= \mathscr{R}\sqrt{\frac{2}{\pi}}\int_0^\infty \frac{\omega^2}{c^2}u_\omega(X,Y,Z)\\ &\qquad\times[a_1(\omega)\,\boldsymbol{\alpha}_1(\omega)\exp\mathrm{i}\,\delta_1(\omega)+a_2(\omega)\,\boldsymbol{\alpha}_2(\omega)\exp\mathrm{i}\,\delta_2(\omega)]\exp\mathrm{i}\omega t\,\mathrm{d}\omega,\\ \mathbf{H}(X,Y,Z,t) &= \mathscr{R}\sqrt{\frac{2}{\pi}}\int_0^\infty \frac{\omega^2}{c^2}u_\omega(X,Y,Z)\\ &\qquad\times[a_1(\omega)\,\boldsymbol{\beta}_1(\omega)\exp\mathrm{i}\,\delta_1(\omega)+a_2(\omega)\,\boldsymbol{\beta}_2(\omega)\exp\mathrm{i}\,\delta_2(\omega)]\exp\mathrm{i}\omega t\,\mathrm{d}\omega.\end{aligned}\right\}\tag{4·1}$$

Here suffixes 1 and 2 refer to the contributions from oscillators which have their axes along the x_1 and x_2 directions.

In order to determine the intensity at a typical point in the image region it will be convenient to write down separate expressions for each of the Cartesian components of **E** and **H**. Let $\theta_1(\omega)$ and $\theta_2(\omega)$ denote the angles which the unit vectors $\boldsymbol{\alpha}_1(\omega)$ and $\boldsymbol{\alpha}_2(\omega)$ make respectively with the x direction in the image space. Since $\boldsymbol{\alpha}_1(\omega)$ and $\boldsymbol{\beta}_1(\omega)$ (and $\boldsymbol{\alpha}_2(\omega)$ and $\boldsymbol{\beta}_2(\omega)$) are real, mutually orthogonal vectors which lie in a plane perpendicular to the z direction, it follows from (4·1) that the components of **E** and **H** are approximately given by†

$$\left.\begin{aligned}E_x(X,Y,Z,t) &= H_y(X,Y,Z,t) = \mathscr{R}\sqrt{\frac{2}{\pi}}\int_0^\infty u_\omega(X,Y,Z)f(\omega)\exp\mathrm{i}\omega t\,\mathrm{d}\omega,\\ E_y(X,Y,Z,t) &= -H_x(X,Y,Z,t) = \mathscr{R}\sqrt{\frac{2}{\pi}}\int_0^\infty u_\omega(X,Y,Z)\,g(\omega)\exp\mathrm{i}\omega t\,\mathrm{d}\omega,\\ E_z(X,Y,Z,t) &= H_z(X,Y,Z,t) = 0,\end{aligned}\right\}\tag{4·2}$$

where

$$\left.\begin{aligned}f(\omega) &= \frac{\omega^2}{c^2}[a_1(\omega)\cos\theta_1(\omega)\exp\mathrm{i}\,\delta_1(\omega)+a_2(\omega)\cos\theta_2(\omega)\exp\mathrm{i}\,\delta_2(\omega)],\\ g(\omega) &= \frac{\omega^2}{c^2}[a_1(\omega)\sin\theta_1(\omega)\exp\mathrm{i}\,\delta_1(\omega)+a_2(\omega)\sin\theta_2(\omega)\exp\mathrm{i}\,\delta_2(\omega)].\end{aligned}\right\}\tag{4·3}$$

For our subsequent analysis it will be convenient to express the right-hand sides in (4·2) as complete Fourier integrals over the range from $-\infty$ to $+\infty$. This can be done immediately by defining the functions u, f and g for negative ω by means of the relations

$$\left.\begin{aligned}f(-\omega) &= f^*(\omega),\\ g(-\omega) &= g^*(\omega),\\ u_{-\omega}(X,Y,Z) &= u^*_\omega(X,Y,Z).\end{aligned}\right\}\tag{4·4}$$

† It would be incorrect to conclude from (4·2) that the direction of the energy flow in the image region is everywhere parallel to z. For the relative errors in (4·2) may substantially affect the calculations of direction in regions where the intensity is small, e.g. in the neighbourhood of the dark rings in the Airy pattern.

(4·2) can then be rewritten in the equivalent form

$$\left.\begin{aligned} E_x(X,Y,Z,t) = H_y(X,Y,Z,t) &= \frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{+\infty} u_\omega(X,Y,Z) f(\omega) \exp \mathrm{i}\omega t \,\mathrm{d}\omega, \\ E_y(X,Y,Z,t) = -H_x(X,Y,Z,t) &= \frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{+\infty} u_\omega(X,Y,Z)\, g(\omega) \exp \mathrm{i}\omega t \,\mathrm{d}\omega, \\ E_z(X,Y,Z,t) = H_z(X,Y,Z,t) &= 0. \end{aligned}\right\} \tag{4·5}$$

From (4·5) (or (4·2)) it follows that the modulus of the Poynting vector

$$\mathbf{S} = \frac{c}{4\pi}[\mathbf{E} \wedge \mathbf{H}]$$

can be approximately expressed in the form

$$|\,\mathbf{S}\,| = \frac{c}{4\pi}[E_x^2 + E_y^2] = \frac{c}{4\pi}[H_x^2 + H_y^2]. \tag{4·6}$$

To obtain the required expression for the intensity, we must calculate the average of $|\,\mathbf{S}\,|$ over a time interval which is large compared to the period $\tau_0 = 2\pi/\omega_0$. This calculation can conveniently be carried out by means of the elegant method of Born & Jordan (1925)† (also Born 1951, p. 407).

We consider the radiation field only within a time interval $0 \leqslant t \leqslant T$ where $T \gg \tau_0$, and assume that $\mathbf{E} = \mathbf{H} = 0$ for $t > T$. This assumption involves no loss of generality, since for natural light the wave trains may be regarded as statistically independent over time intervals which are large compared to the period τ_0. It then follows from (4·5) by the Fourier inversion formula that

$$u_\omega(X,Y,Z) f(\omega) = \frac{1}{\sqrt{(2\pi)}}\int_0^T E_x(X,Y,Z,t) \exp(-\mathrm{i}\omega t)\,\mathrm{d}t, \tag{4·7}$$

with similar expressions involving the other field components.

On the assumption that the electromagnetic field is stationary, the intensity $I(X,Y,Z)$ can now be calculated easily by averaging over the time interval $0 \leqslant t \leqslant T$. Equation (4·6) gives

$$\begin{aligned} I(X,Y,Z) &= \frac{1}{T}\int_0^T |\,\mathbf{S}\,|\,\mathrm{d}t \\ &= \frac{c}{4\pi}\frac{1}{T}\int_0^T [E_x^2 + E_y^2]\,\mathrm{d}t. \end{aligned} \tag{4·8}$$

In order to evaluate this integral we write with the help of (4·5),

$$\int_0^T E_x^2\,\mathrm{d}t = \frac{1}{\sqrt{(2\pi)}}\int_0^T E_x\,\mathrm{d}t \int_{-\infty}^{+\infty} u_\omega f(\omega) \exp \mathrm{i}\omega t\,\mathrm{d}\omega. \tag{4·9}$$

† Alternatively, this could be done by the somewhat lengthy method used by Planck (1900) in his classical paper on black-body radiation.

On changing the order of integration in the last expression we obtain

$$\int_0^T E_x^2\,\mathrm{d}t = \int_{-\infty}^{+\infty} u_\omega f(\omega)\,\mathrm{d}\omega \frac{1}{\sqrt{(2\pi)}}\int_0^T E_x \exp \mathrm{i}\omega t\,\mathrm{d}t,$$

$$= \int_{-\infty}^{+\infty} u_\omega f(\omega)\, u_\omega^* f_\omega^*\,\mathrm{d}\omega \quad \text{(by (4·7) and (4·4))}$$

$$= 2\int_0^\infty |\, u_\omega \,|^2 \,|f(\omega)\,|^2\,\mathrm{d}\omega.$$

Similarly,

$$\int_0^T E_y^2\,\mathrm{d}t = 2\int_0^\infty |\, u_\omega\,|^2\,|\, g(\omega)\,|^2\,\mathrm{d}\omega. \tag{4·10}$$

On substituting from the last two equations into (4·8) we find

$$I(X,Y,Z) = \frac{c}{2\pi T}\int_0^\infty |\, u_\omega(X,Y,Z)\,|^2 [\,|f(\omega)\,|^2 + |\, g(\omega)\,|^2]\,\mathrm{d}\omega. \tag{4·11}$$

This expression contains the scalar wave-function $u_\omega(x,y,z)$ under the integral sign. Since in the scalar theory this function characterizes the diffraction pattern, we must consider more closely the manner in which it depends on ω. This is done in the appendix. It is shown there (A 16) that $|\, u_\omega\,|$ satisfies the relation

$$\frac{|\, u_\omega\,|}{\omega} = \frac{|u_{\omega_0}|}{\omega_0} \tag{4·12}$$

whenever the frequency spread of the source (i.e. the frequency range over which $|\, p_j(\omega)\,|$ $(j = 1, 2, 3)$ differ substantially from zero), fulfils the condition (A 14):

$$\frac{|\,\Delta\omega\,|}{\omega_0} = \frac{|\,\Delta\lambda\,|}{\lambda_0} \lesssim O\left\{1{\cdot}6\times 10^{-2}\left[\,0{\cdot}61B + \frac{|Z|}{\lambda_0} + \text{max.}\left|\frac{V}{\lambda_0} - \left(\frac{\partial V}{\partial \lambda}\right)_{\lambda_0}\right|\,\right]^{-1}\right\}. \tag{4·13}$$

Here V denotes the aberration of the wave in the exit pupil and max. denotes maximum on Σ'. B (which is of order 1 in most practical cases) gives in suitable units the distance of the point of observation $Q(X, Y, Z)$ from the principal ray.

Whenever (4·13) is satisfied, the expression (4·11) for the intensity may be written with the help of (4·12) in the form

$$I(X,Y,Z) = |\, u_{\omega_0}(X,Y,Z)\,|^2 \frac{c}{2\pi\omega_0^2 T}\int_0^\infty \omega^2[\,|f(\omega)\,|^2 + |\, g(\omega)\,|^2]\,\mathrm{d}\omega. \tag{4·14}$$

Clearly the term

$$\frac{c}{2\pi\omega_0^2 T}\int_0^\infty \omega^2[\,|f(\omega)\,|^2 + |\, g(\omega)\,|^2]\,\mathrm{d}\omega, \tag{4·15}$$

which is independent of (X, Y, Z), must also be independent of T (which is implicitly contained in f and g on account of (4·7)) if a stationary phenomenon is observed. Hence (4·15) must be a constant (C, say), and the intensity may therefore be finally written in the form

$$I(X, Y, Z) = C\,|\, u_{\omega_0}(X, Y, Z)\,|^2, \tag{4·16}$$

where, as before, u_ω denotes the scalar wave-function

$$u_\omega(X,Y,Z) = \frac{\omega}{2\pi i c}\iint_{\Sigma'}\frac{1}{s}\exp\left\{-\frac{i\omega}{c}[f_\omega(x',y',z')+s]\right\}d\sigma'. \tag{4·17}$$

Equation (4·16) establishes the approximate validity of the scalar diffraction theory and justifies, under the restrictions mentioned, the use of the single scalar function (4·17) in the calculation of intensity.

Appendix. Behaviour of the scalar wave-function u_ω over the frequency band

From (4·17) we see that the behaviour of u_ω depends essentially on the exponent $\{-i(\omega/c)(f+s)\}$ which in general varies over the domain of integration. The variation is, however, small compared with the actual value of the exponent. This suggests that we should separate the exponent into two parts, one of which remains constant during the integration. To do this we rewrite f_ω in the form

$$f_\omega(x',y',z') = d_0 + V_\omega(x',y',z'), \tag{A 1}$$

where d_0 denotes the optical distance for light of wave-length λ_0 from the source Q_0 to the centre C of the exit pupil. Since in an error-free system the optical distance from Q_0 to each point of the reference sphere is the same, it follows that $V_\omega(x',y',z')$ may be interpreted as the total aberration (i.e. the monochromatic wave aberration plus colour error) at the typical point (x',y',z') of the reference sphere. With the help of (A 1) u_ω may be written in the form

$$u_\omega(X,Y,Z) = \frac{\omega}{2\pi i c}\iint_{\Sigma'}\frac{1}{s}\exp\left[-\frac{i\omega}{c}(d_0+V_\omega+s)\right]d\sigma'. \tag{A 2}$$

We next set

$$\left.\begin{aligned} X &= \rho\cos\phi, \quad x' = \rho'\cos\phi',\\ Y &= \rho\sin\phi, \quad y' = \rho'\sin\phi',\end{aligned}\right\} \tag{A 3}$$

so that ρ and ρ' denote respectively the geometrical distances of the point of observation Q and of a typical point Q' on Σ' from the principal ray. We then find by simple calculation (cf. Nijboer 1947, p. 611) that

$$s = R+\psi, \tag{A 4}$$

where R is the radius of the reference sphere and

$$\psi \simeq \left[1-\frac{1}{2}\left(\frac{\rho'}{R}\right)^2\right]Z + \frac{\rho^2+Z^2}{2R} - \frac{\rho\rho'}{R}\cos(\phi-\phi'). \tag{A 5}$$

Hence (A 2) may be written in the form

$$u_\omega(X,Y,Z) = \frac{\omega}{2\pi i c}v_\omega(X,Y,Z)\exp\left[-i\frac{\omega}{c}(d_0+R)\right], \tag{A 6}$$

where

$$v_\omega(X,Y,Z) = \iint_{\Sigma'}\frac{1}{s}\exp\left[-i\frac{\omega}{c}(\psi+V_\omega)\right]d\sigma'. \tag{A 7}$$

Clearly v_ω will remain substantially independent of ω over the frequency range $(\omega_0 - \frac{1}{2}\Delta\omega, \omega_0 + \frac{1}{2}\Delta\omega)$ provided that the term $\frac{\omega}{c}(\psi + V_\omega)$ in the exponent does not vary by more than a few units per cent as ω takes on different values in the range, whilst X, Y, Z and x', y', z' are kept fixed.

We have, on expanding the exponent into a power series round ω_0 and on neglecting terms above the first power in $(\omega - \omega_0)$,

$$\frac{\omega}{c}(\psi + V_\omega) \simeq \frac{\omega_0}{c}(\psi + V_{\omega_0}) + \frac{\omega - \omega_0}{c}\left[\psi + V_{\omega_0} + \omega_0\left(\frac{\partial V_\omega}{\partial \omega}\right)_{\omega_0}\right]. \tag{A 8}$$

The above condition can therefore be written as

$$\left|\frac{\Delta\omega}{2c}\left\{\psi + V_{\omega_0} + \omega_0\left(\frac{\partial V_\omega}{\partial \omega}\right)_{\omega_0}\right\}\right| \leqslant \frac{A}{100}, \tag{A 9}$$

where A is a constant of the order unity and expresses as a percentage the maximum permissible variation of the exponent. We next substitute for ψ from (A 5) and obtain

$$\left|\frac{\Delta\omega}{2c}\right|\left|\left[1 - \frac{1}{2}\left(\frac{\rho'}{R}\right)^2\right]Z + \frac{\rho^2 + Z^2}{2R} - \frac{\rho\rho'}{R}\cos(\phi - \phi') + V_{\omega_0} + \omega_0\left(\frac{\partial V}{\partial \omega_0}\right)_{\omega_0}\right| \leqslant \frac{A}{100}. \tag{A 10}$$

We can replace (A 10) by the weaker condition

$$\frac{1}{2}\frac{|\Delta\omega|}{c}\left\{\left[1 + \frac{1}{2}\left(\frac{a}{R}\right)^2\right]|Z| + \frac{\rho^2 + Z^2}{2R} + \frac{a\rho}{R} + \text{max.}\left|V_{\omega_0} + \omega_0\left(\frac{\partial V}{\partial \omega}\right)_{\omega_0}\right|\right\} \leqslant \frac{A}{100}, \tag{A 11}$$

where a denotes the radius of the exit pupil and max. denotes the maximum value on the reference sphere Σ'.

In (A 11) we can evidently neglect the term $\frac{1}{2}(a/R)^2$ which will in practice be small compared to unity and also the term $(\rho^2 + Z^2)/2R$ which will be small compared to $a\rho/R$. ρ will as a rule be of the order of the resolution of the system, i.e.

$$\rho = B\frac{0{\cdot}61R\lambda_0}{a}, \tag{A 12}$$

where B is a quantity of order unity.† (A 11) can therefore be written as

$$\frac{1}{2}\frac{|\Delta\omega|}{c}\left\{|Z| + 0{\cdot}61B\lambda_0 + \text{max.}\left|V_{\omega_0} + \omega_0\left(\frac{\partial V}{\partial \omega}\right)_{\omega_0}\right|\right\} \leqslant \frac{A}{100}, \tag{A 13}$$

or finally, from (A 13), remembering that $\omega = 2\pi c/\lambda$ and that A is of order unity (we actually take $A = 5$),

$$\frac{|\Delta\omega|}{\omega_0} = \frac{|\Delta\lambda|}{\lambda_0} \lesssim O\left\{1{\cdot}6 \times 10^{-2}\left[0{\cdot}61B + \frac{|Z|}{\lambda_0} + \text{max.}\left|\frac{V}{\lambda_0} - \left(\frac{\partial V}{\partial \lambda}\right)_{\lambda_0}\right|\right]^{-1}\right\}. \tag{A 14}$$

Here $\Delta\lambda$ denotes the variation in the wave-length associated with the variation $\Delta\omega$ in the frequency.

† B may be interpreted as giving the distance of the point $Q(X, Y, Z)$ of observation from the principal ray, when the unit of length is taken to be the radius of the first dark ring in the Airy pattern for wave-length λ_0, in an error free system of the same focal ratio and focal length.

Equation (A 14) provides a rough measure for the wave-length spread over which $v_\omega(X, Y, Z)$ may be regarded as practically independent of ω. In any practical case the quantity max. $\left|\frac{V}{\lambda_0} - \left(\frac{\partial V}{\partial \lambda}\right)_{\lambda_0}\right|$ can be estimated from a ray trace or from power series expansions for the aberration function V; Z and B will be given by the position of the point of observation Q. As an example, let us consider the diffraction pattern within the circular region round Q_1, of radius equal to four times the radius of the first Airy ring ($B = 4$), in the plane $Z = 0$, of a reflecting system ($\partial V/\partial \lambda = 0$) for which the maximum aberration does not exceed $20\lambda_0$ (max. $|V/\lambda_0| = 20$). (A 14) shows that in this case v_ω will not vary by more than a few units per cent provided

$$\left|\frac{\Delta\lambda}{\lambda_0}\right| \lesssim O(7 \times 10^{-4}).$$

Using (A 6) we now find that whenever (A 14) is satisfied, u_ω may be written in the form

$$u_\omega(X, Y, Z) = \frac{\omega}{2\pi i c}[v_{\omega_0}(X, Y, Z)] \exp\left\{-\frac{i\omega}{c}(d_0 + R)\right\}. \tag{A15}$$

Hence in such a case $|u_\omega/\omega|$ is practically independant of ω over the interval $(\omega_0 - \frac{1}{2}\Delta\omega, \omega_0 + \frac{1}{2}\Delta\omega)$, and therefore in this interval u_ω satisfies the relation

$$\frac{|u_\omega(X, Y, Z)|}{\omega} = \frac{|u_{\omega_0}(X, Y, Z)|}{\omega_0}. \tag{A 16}$$

We wish to thank Professor Max Born, F.R.S., for his interest in this work and for valuable suggestions. We are also indebted to Dr L. M. Yang for some useful comments and criticisms.

References

Born, M. 1933 *Optik*. Berlin: Springer.
Born, M. 1951 *Atomic physics*, 5th ed. Glasgow: Blackie.
Born, M. & Jordan, P. 1925 *Z. Phys.* **33**, 479.
Couder, A. 1944 *C.R. Acad. Sci., Paris*, **219**, 677.
Friedlander, F. G. 1947 *Proc. Camb. Phil. Soc.* **43**, 284.
Joos, G. 1947 *Theoretical physics*. Glasgow: Blackie.
Nijboer, B. R. A. 1947 *Physica*, **13**, 605.
Osterberg, H. & Wilkins, J. E. 1949 *J. Opt. Soc. Amer.* **39**, 553.
Picht, J. 1925 *Ann. Phys., Lpz.*, **77**, 685; Also *Optische Albildung*, 1931, p. 135. Braunschweig: Vieweg.
Planck, M. 1900 *Ann. Phys., Lpz.*, **1**, 69.
Wolf, E. 1951 *Rep. Progr. Phys.* **14**, 95.

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, CAMBRIDGE
(BROOKE CRUTCHLEY, UNIVERSITY PRINTER)

1.6

Reprinted from JOURNAL OF THE OPTICAL SOCIETY OF AMERICA, Vol. 42, No. 8, 547–552, August, 1952
Printed in U. S. A.

On a New Aberration Function of Optical Instruments

E. WOLF
Department of Mathematical Physics, University of Edinburgh, Edinburgh, Great Britain
(Received February 1, 1952)

A new characteristic function is introduced which enables an immediate determination of the departures of imaging pencils of rays from homocentricity as well as of the deformations of the associated wave fronts from spherical form. This function is therefore well suited for investigations of aberration effects on the basis of geometrical optics or of diffraction theory.

Within the approximations of fourth-order theory, the new function is closely related to the Seidel eikonal of Schwarzschild, but unlike the latter it is defined in simple physical terms and is applicable whether the imaging of a point object or of an extended object is under investigation. The series development of this function is carried out in detail up to fourth order.

I. INTRODUCTION

THE term "aberration" is used in optics to describe either the lack of homocentricity in a pencil of rays or the deviation of a wave surface from a spherical form. Either of these notions can form a starting point for investigations into the imaging qualities of optical instruments. Although there exists a close connection between the different treatments, those based on the concept of a wave aberration have become increasingly more favored in recent years. One of the main reasons for this development is undoubtedly the fact that the concept of a wave aberration can be usefully employed in investigations within the domain of geometrical optics as well as that of diffraction theory.

An elegant geometrical optical analysis dealing with ray aberrations only was carried out by Schwarzschild.[1,2] Although Schwarzschild's main results do not go much beyond those obtained by Seidel in his classical papers, Schwarzschild's analysis is of great importance since it enables a rigorous and systematic separation into orders of the various terms that contribute to the ray aberrations. In his investigations Schwarzschild used Hamilton's method of characteristic functions. He introduced a new function, the so-called Seidel eikonal, of which the derivatives with respect to suitably chosen variables gave the deviations from the Gauss image point of the ray intersections with the image plane. Though very powerful, Schwarzschild's treatment possesses the disadvantage that it is based on a function for which no simple interpretation can be given in physical terms. Similar criticism may also be raised against later treatments by other authors, in which the variables themselves are defined by means of complicated and often rather clumsy expressions.

Investigations of the image defects based on the notion of the wave aberration were carried out chiefly by Nijboer[3,4] and Hopkins.[5] These authors use certain aberration functions that measure directly the deviation of the waves from spherical form. It is evident that these functions must also be closely related to Hamilton's characteristics; this relationship has so far, however, not been studied in the literature.

In the present paper we introduce a new generalized aberration function, which we call the aberration characteristic, and discuss its main properties. We show that from it the departures of the imaging pencils of rays from homocentricity (the ray aberrations) as well as the deformations of the associated wave fronts from spherical form (the wave aberrations) can immediately be determined. In Sec. III we investigate the form of this function up to fourth order for any centered system. We find that of its six fourth-order coefficients, five are simple multiples of the Seidel sums (representing primary spherical aberration, coma, astigmatism, curvature, and field distortion) and that the remaining fourth-order coefficient is closely related to the coefficient that represents the spherical aberration of the center of the pupil.

The new aberration characteristic possesses several advantages over the functions used by earlier authors. For example, unlike the Seidel eikonal, it is defined in simple physical terms. Further, unlike the aberration functions of Nijboer and Hopkins, the aberration characteristic is immediately applicable in investigations concerned with the imaging of point—as well as of extended—objects. It can be applied to centered systems consisting of spheric or aspheric surfaces and can be used in researches carried out on the basis of geometrical optics or of diffraction theory.

II. THE ABERRATION CHARACTERISTIC

We consider a rotationally symmetrical optical system. We denote by P_0 a typical point of the object plane (assumed for the moment to be at finite distance) and by P_0', P_1', and P_1 the points of intersection of a ray from P_0 with the entry pupil, the exit pupil, and the

[1] K. Schwarzschild, Astron. Mitt. Königl. Sternwarte, Göttingen, **9**, (1905).

[2] A detailed account of Schwarzschild's investigations will also be found in M. Born, *Optik* (Verlag. Julius Springer, Berlin, 1933). An extension of Schwarzschild's analysis to nonsymmetrical systems was made by G. D. Rabinovich, Akad. Nauk. SSSR, Zhurnal Eksper. Teor. Fiz. **16**, 161 (1946).

[3] B. R. A. Nijboer, Physica **10**, 679 (1943).

[4] B. R. A. Nijboer, Physica, **13**, 605 (1947).

[5] H. H. Hopkins, *Wave Theory of Aberrations* (Oxford, Clarendon Press, 1950).

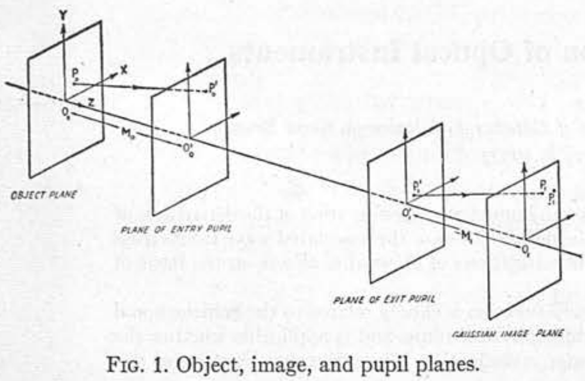


FIG. 1. Object, image, and pupil planes.

Gaussian image plane, respectively (Fig. 1). At the axial point of each of these planes we take mutually parallel Cartesian axes, the Z directions being taken along the axis of the system.

Let P_1^* be the Gaussian image of P_0; then $\langle P_1^*P_1\rangle$ will be called the *aberration of the ray* which passes through P_1'. Except in the special case where P_1 is a stigmatic image of P_0, the wave fronts in the image space associated with the bundle of rays from P_0 will deviate from spherical form. To describe these wave deviations we introduce a spherical reference surface centered at the Gaussian image point P_1^* and passing through the center O_1' of the exit pupil. We call this surface the *Gaussian reference sphere.*

Let Q and S, respectively, be the points of intersection of the ray with the wave front through O_1' and with the Gaussian reference sphere (see Fig. 2, where for simplicity the relevant points and lines are shown as coplanar); then $\langle QS\rangle$ will be called the *aberration of the wave element at* Q.

We define the new aberration characteristic $\Psi(X_0, Y_0, X, Y)$ as the optical length of the actual ray, from P_0 to S:

$$\Psi(X_0, Y_0, X, Y)=[P_0S], \tag{1}$$

where (X_0, Y_0) and (X, Y) denote the coordinates of P_0 and of S, respectively, and the square bracket denotes optical length.

We first show that from the knowledge of this function we can immediately determine the aberration $\Phi(X_0, Y_0, X, Y)=\langle QS\rangle$ (measured as positive in Fig. 2) of the wave element at Q.‡ We have

$$\Phi(X_0, Y_0, X, Y)=\langle QS\rangle=(1/n_1)\{[P_0S]-[P_0Q]\}. \tag{2}$$

But, since Q lies on the wave front through O_1',

$$[P_0Q]=[P_0O_1'] \tag{3}$$

so that

$$\Phi(X_0, Y_0, X, Y)=(1/n_1)\{[P_0S]-[P_0O_1']\}, \tag{4}$$

or, in terms of Ψ,

$$\Phi(X_0, Y_0, X, Y)=(1/n_1)\{\Psi(X_0, Y_0, X, Y)\\ -\Psi(X_0, Y_0, O, O)\}. \tag{5}$$

This simple relation expresses the *wave aberration in terms of the aberration characteristic.*

Next we show that from the knowledge of Ψ we can also easily determine the ray aberrations. In order to see clearly the degree of accuracy of our expressions it will be convenient to introduce a quantity§ μ defined by

$$\mu=|a_1/M_1|, \tag{6}$$

where a_1 denotes the radius of the exit pupil and M_1 (taken as negative in accordance with Schwarzschild's notation) is the distance between the exit pupil and the Gaussian image plane. μ will be regarded as a small quantity of the first order.

If $\bar{X}, \bar{Y}, \bar{Z}$ are the coordinates, referred to axes at O_1', of the typical point Q on the wave front through O_1' and if T denotes the point of intersection of the line QP_1^* with the Gaussian reference sphere, then (see Fig. 3)

$$\langle QP\rangle_1^{*2}=(R_1+\langle QT\rangle)^2,$$

i.e.,

$$(\bar{X}-X_1^*)^2+(\bar{Y}-Y_1^*)^2+(\bar{Z}+M_1)^2=\{R_1+\langle QT\rangle\}^2, \tag{7}$$

where

$$R_1=(M_1^2+X_1^{*2}+Y_1^{*2})^{\frac{1}{2}}=-M_1-\frac{1}{2M_1}(X_1^{*2}+Y_1^{*2})\\ +\frac{1}{8M_1^3}(X_1^{*2}+Y_1^{*2})^2+O(M_1\mu^6) \tag{8}$$

is the radius of the Gaussian reference sphere.

Now $\langle QT\rangle=O(M_1\mu^4)$ so that Eq. (7) becomes, on neglecting the term in $\langle QT\rangle^2$,

$$\bar{X}^2+\bar{Y}^2+\bar{Z}^2-2(\bar{X}X_1^*+\bar{Y}Y_1^*-\bar{Z}M_1)\\ -2R_1\langle QT\rangle+O(M_1^2\mu^8)=0. \tag{9}$$

In Eq. (9) we may replace $\langle QT\rangle$ by $\Phi(X_0, Y_0, X, Y)$ without changing the error term. We then obtain the

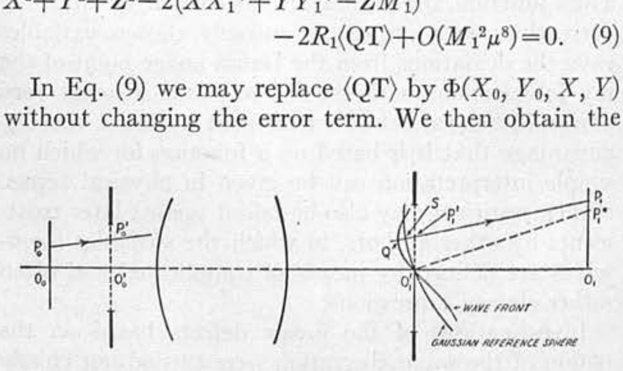


FIG. 2. Definition of aberrations.

‡ $n_1\Phi$, where n_1 denotes the refractive index of the image space, is essentially the W function of Hopkins (see reference 5), but is expressed in terms of coordinates taken in different planes. Up to fourth order Φ may also be identified with the aberration function V of Nijboer. (See reference 3.)

§ The use of such a quantity to make visible the order of accuracy is due to J. L. Synge, J. Opt. Soc. Am. **33**, 129 (1943). The great usefulness of such an artifice has been first fully recognized by E. H. Linfoot, who applied it in a number of papers concerning the aberration theory of the Schmidt camera [e.g. E. H. Linfoot, Monthly Notices Roy. Astron. Soc. **109**, 279 (1949)].

following approximate equation for the wave front:

$$\bar{X}^2+\bar{Y}^2+\bar{Z}^2-2(\bar{X}X_1^*+\bar{Y}Y_1^*-M_1\bar{Z})$$
$$-2R_1\Phi(X_0, Y_0, X, Y)+O(M_1^2\mu^8)=0. \quad (10)$$

Now $X=\bar{X}+O(M_1\mu^7)$, $Y=\bar{Y}+O(M_1\mu^7)$ so that the direction cosines of the ray QP_1 are, with inaccuracies $O(M_1\mu^7)$, in the ratios

$$\left(\bar{X}-X_1^*-R_1\frac{\partial\Phi}{\partial X}\right):\left(\bar{Y}-Y_1^*-R_1\frac{\partial\Phi}{\partial Y}\right):$$
$$(\bar{Z}+M_1). \quad (11)$$

The coordinates (X_1, Y_1) of P_1 are therefore given, with errors $O(M_1\mu^7)$ by the equations

$$\frac{X_1-\bar{X}}{\bar{X}-X_1^*-R_1(\partial\Phi/\partial X)}=\frac{Y_1-\bar{Y}}{\bar{Y}-Y_1^*-R_1(\partial\Phi/\partial Y)}=-1, \quad (12)$$

whence

$$\left.\begin{aligned} X_1-X_1^*&=R_1\frac{\partial\Phi(X_0, Y_0, X, Y)}{\partial X}+O(M_1\mu^7),\\ Y_1-Y_1^*&=R_1\frac{\partial\Phi(X_0, Y_0, X, Y)}{\partial Y}+O(M_1\mu^7).\end{aligned}\right\} \quad (13)$$

Finally, using Eq. (5) and writing

$$\Delta X_1=X_1-X_1^*, \quad \Delta Y_1=Y_1-Y_1^*, \quad (14)$$

it follows that *the ray aberration components are given by*

$$\Delta X_1+i\Delta Y_1=\frac{R_1}{n_1}\left(\frac{\partial}{\partial X}+i\frac{\partial}{\partial Y}\right)\psi(X_0, Y_0, X, Y)$$
$$+(1+i)O(M_1\mu^7), \quad (15)$$

where to the same degree of accuracy we may replace R_1 by $-M_1-(1/2M_1)(X_1^{*2}+Y_1^{*2})$.

On account of symmetry, the four variables enter Φ and Ψ only in the combinations of $X_0^2+Y_0^2$, X_0X+Y_0Y, and X^2+Y^2. If then we introduce polar coordinates

$$\begin{aligned} X_0&=r\cos\chi, \quad X=\rho\cos\theta,\\ Y_0&=r\sin\chi, \quad Y=\rho\sin\theta,\end{aligned} \quad (16)$$

and set $\varphi=\theta-\chi$, the aberration characteristic becomes a function of the three rotational invariants ρ^2, r^2 and $\rho r\cos\varphi$ only, and the two fundamental equations (5) and (15) for the wave aberrations and the ray aberrations can be written as||

$$\Phi(r, \rho, \varphi)=(1/n_1)\{\Psi(r, \rho, \varphi)-\Psi(r, 0, 0)\} \quad (17)$$

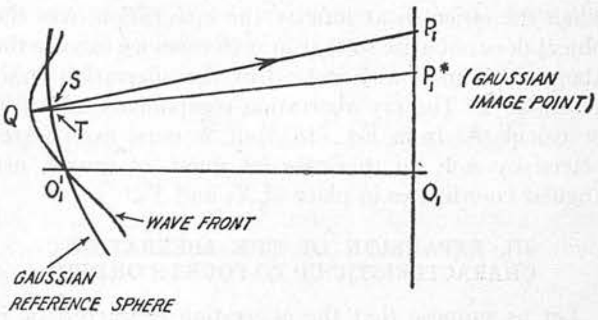


FIG. 3. Computation of aberrations.

and

$$\Delta X_1+i\Delta Y_1=\frac{R_1}{n_1}e^{i(\chi+\varphi)}$$
$$\times\left[\frac{\partial}{\partial\rho}+\frac{i}{\rho}\frac{\partial}{\partial\varphi}\right]\Psi(r, \rho, \varphi)+e^{i\pi/4}O(M_1\mu^7). \quad (18)$$

We observe that because of the error term the simple relations of the form (13) secure, in general, an order-to-order correlation between the wave aberrations and the ray aberration up to sixth order only. This fact appears to have escaped the attention of earlier authors.

The aberration characteristic is particularly suited for investigations of imaging properties on the basis of diffraction theory. Since the phase difference between a typical point S on the reference sphere and the object point P_0 is $(2\pi/\lambda)\Psi(X_0, Y_0, X, Y)$, where λ denotes the wavelength, it follows, by the usual application of Huygens-Fresnel principle and with the usual approximations, that the disturbance u_P at a point P in the neighborhood of the Gaussian focus P_1^* due to a point object P_0 is given by

$$u_P=\text{constant}\int\int_\Sigma g(\rho,r,\varphi)\frac{\exp ik[\psi(r,\rho,\varphi)+n_1R_1']}{R_1R_1'}d\sigma. \quad (19)$$

Here $g(r, \rho, \varphi)$ is an amplitude factor, which in most cases can be replaced by a constant $R_1'=\langle SP\rangle$ and $k=2\pi/\lambda$. The integration is carried out over that part Σ of the Gaussian reference sphere which approximately fills the exit pupil. In practice, however, it will as a rule be permissible to set $d\sigma=\rho d\rho d\varphi$ and integrate over $0\leq\rho\leq a$, $0\leq\varphi<2\pi$.

If we wish to investigate the imaging of an extended object, the total effect at P can be obtained from Eq. (19) by taking into account its coherence properties[6] and integrating over it.

The aberration characteristic Ψ is only defined when the object is at a finite distance. In the special case

|| It is sometimes desirable to determine the deviation of the wave surface from a surface other than the Gaussian sphere. This can be done by adding to the aberration characteristic terms which represent the distance between the Gaussian reference sphere and the actual comparison sphere. Such modifications are of importance in investigations concerned with the balancing of aberrations of different types and orders against each other to obtain optimum performance.

[6] H. H. Hopkins, Proc. Roy. Soc. (London) A208, 263 (1951).

when the object is at infinity the integration over the object does not arise so that in such cases we can use the aberration function Φ instead of the aberration characteristic Ψ. The ray aberration components may still be calculated from Eq. (18), but Ψ must now be replaced by $n_1\Phi$. In this case we must, of course, use angular coordinates in place of X_0 and Y_0.

III. EXPANSION OF THE ABERRATION CHARACTERISTIC UP TO FOURTH ORDER

Let us suppose that the aberration characteristic is expanded in a power series in terms of the three rotational invariants. Then

$$\Psi=\Psi^{(0)}+\Psi^{(2)}+\Psi^{(4)}\cdots+\Psi^{(2K)}+\cdots, \tag{20}$$

where $\Psi^{(2K)}$ denotes terms of $O(M_1\mu^{2K})$. We shall investigate the terms up to and including the fourth order.

From the definition of Ψ and from Fig. 2,

$$\Psi=[P_0P_1]-[SP_1]. \tag{21}$$

If $W(p_0, q_0, p_1, q_1)$ denotes the angle characteristic¶ of the instrument, i.e., the optical length of the ray between the feet of the perpendiculars dropped from the axial points O_0 and O_1 (see Fig. 1) on to the portions of the ray in the two external spaces, then

$$[P_0P_1]=W-n_0(p_0X_0+q_0Y_0)+n_1(p_1X_1+q_1Y_1); \tag{22}$$

here p_0, q_0 denote the cosines of the angles which the ray in the object space makes with the X and Y directions, p_1 and q_1, having analogous meanings. Also from the figure,

$$\begin{aligned}[SP_1]&=n_1F_1+n_1[p_1(X_1-X_1^*)\\&\quad+q_1(Y_1-Y_1^*)]+O(M_1\mu^6)\\&=-n_1M_1-\frac{n_1}{2M_1}(X_1^{*2}+Y_1^{*2})\\&\quad+\frac{n_1}{8M_1^3}(X_1^{*2}+Y_1^{*2})^2+n_1[p_1(X_1-X_1^*)\\&\quad+q_1(Y_1-Y_1^*)]+O(M_1\mu^6).\end{aligned} \tag{23}$$

On substituting from Eqs. (22) and (23) into Eq. (21), we find

$$\begin{aligned}\psi(X_0, Y_0, X, Y)&=W(p_0, q_0, p_1, q_1)\\&\quad-n_0(p_0X_0+q_0Y_0)+n_1(p_1X_1^*+q_1Y_1^*)\\&\quad+n_1M_1+\frac{n_1}{2M_1}(X_1^{*2}+Y_1^{*2})\\&\quad-\frac{n_1}{8M_1^3}(X_1^{*2}+Y_1^{*2})^2+O(M_1\mu^6).\end{aligned} \tag{24}$$

¶ We use here substantially Schwarzschild's notation.

We observe that since $X=X_1'+O(M_1\mu^3)$, $Y=Y_1'+O(M_1\mu^3)$, and $\partial\Psi/\partial X=O(\mu^3)$, $\partial\Psi/\partial Y=O(\mu^3)$ *we can replace X by X_1' and Y by Y_1' in Eq.(24) without changing the error term.*

Following Schwarzschild, we shall introduce the so-called *Seidel variables* which, within the accuracy of Gaussian dioptric, have very simple invariant properties. These variables are defined by:

$$\left.\begin{aligned}&x_0=\alpha(X_0/l_0), && y_0=\alpha(Y_0/l_0),\\&x_1=\alpha(X_1/l_1), && y_1=\alpha(Y_1/l_1),\\&\xi_0=(1/\lambda_0)(X_0+M_0p_0), && \eta_0=(1/\lambda_0)(Y_0+M_0p_0),\\&\xi_1=(1/\lambda_1)(X_1+M_1p_1), && \eta_1=(1/\lambda_1)(Y_1+M_1p_1),\end{aligned}\right\} \tag{25}$$

where l_1/l_0 denotes the magnification ratio between the image plane and the object plane, λ_1/λ_0 denotes the magnification ratio between the planes of the exit and the entry pupils, and

$$\alpha=\frac{n_0\lambda_0l_0}{M_0}=\frac{n_1\lambda_1l_1}{M_1}.$$

The Seidel variables may be interpreted as follows: if we assume that the refractive index of the object and image spaces are unity, i.e., $n_0=n_1=1$, then**

$$x_0=\lambda_0(X_0/M_0),\quad y_0=\lambda_0(Y_0/M_0),$$
$$x_1=\lambda_1(X_1/M_1),\quad y_1=\lambda_1(Y_1/M_1);$$

hence *x_0/λ_0 and y_0/λ_0 measure the angular distances of the object from the axis when observed from the center of the entry pupil. The quantities $-x_1/\lambda_1-y_1/\lambda_1$ have an analogous meaning. The number pairs $(\lambda_0\xi_0, \lambda_0\eta_0)$ and $(\lambda_1\xi_1, \lambda_1\eta_1)$ are, respectively, with errors $O(M_1\mu^3)$, the coordinates of the points P_0' and P_1' in which the typical ray intersects the planes of the pupils. With the same error terms $(\lambda_1\xi_1, \lambda_1\eta_1)$ may also be identified with the coordinates (X, Y) of the point S (see Fig. 2) in which the ray intersects the Gaussian reference sphere.*

In terms of the Seidel variables, the contributions to Eq. (24) become

$$\left.\begin{aligned}&n_0(p_0X_0+q_0Y_0)\\&\qquad=(\xi_0x_0+\eta_0y_0)-(M_0/n_0\lambda_0^2)(x_0^2+y_0^2),\\&n_1(p_1X_1^*+q_1Y_1^*)\\&\qquad=(\xi_1x_0+\eta_1y_0)-(M_1/n_1\lambda_1^2)(x_0x_1+y_0y_1),\\&(n_1/2M_1)(X_1^{*2}+Y_1^{*2})=(M_1/2n_1\lambda_1^2)(x_0^2+y_0^2),\\&(n_1/8M_1^3)(X_1^{*2}+Y_1^{*2})^2=(M_1/8n_1^3\lambda_1^4)(x_0^2+y_0^2)^2.\end{aligned}\right\} \tag{26}$$

** In the special case when the object is at infinity, the factor M_1/λ_1 has a very simple interpretation. We have, if we also take the arbitrary factor λ_0 equal to unity,

$$X_1=\frac{l_1}{l_0}X_0=\frac{M_1}{\lambda_1}\frac{X_0}{M_0},\quad Y_1=\frac{l_1}{l_0}Y_0=\frac{M_1}{\lambda_1}\frac{Y_0}{M_0}.$$

When the object is at infinity the factor of proportionality between X_1 and X_0/M_0 (and similarly between Y_1 and Y_0/M_0) equals to the focal length f of the instrument, taken with a negative sign; hence

$$\frac{M_1}{\lambda_1}=-f.$$

The negative sign arises from the fact that the image is inverted when the focal length is positive.

Hence Eq. (24) takes the form

$$\Psi=W+x_0(\xi_1-\xi_0)+y_0(\eta_1-\eta_0)+n_1M_1 +[(M_0/n_0\lambda_0^2)+(M_1/2n_1\lambda_1^2)](x_0^2+y_0^2) -(M_1/n_1\lambda_1^2)(x_0x_1+y_0y_1) -(M_1/8n_1^3\lambda_1^4)(x_0^2+y_0^2)^2+O(M_1\mu^6). \quad (27)$$

We can simplify Eq. (27) by observing that

$$\begin{aligned} x_0x_1&=\tfrac{1}{2}(x_1^2+2x_0x_1-x_1^2)\\ &=\tfrac{1}{2}x_1^2+\tfrac{1}{2}[x_0+(x_1-x_0)][x_0-(x_1-x_0)]\\ &=\tfrac{1}{2}x_1^2+\tfrac{1}{2}x_0^2+O(M_1\mu^6),\end{aligned}$$

and similarly

$$y_0y_1=\tfrac{1}{2}y_1^2+\tfrac{1}{2}y_0^2+O(M_1\mu^6).$$

On substituting these expressions in Eq. (27) we obtain

$$\Psi=W+x_0(\xi_1-\xi_0)+y_0(\eta_1-\eta_0)+n_1M_1 +(M_0/n_0\lambda_0^2)(x_0^2+y_0^2)-(M_1/2n_1\lambda_1^2)(x_1^2+y_1^2) -(M_1/8n_1^3\lambda_1^4)(x_0^2+y_0^2)^2+O(M_1\mu^6). \quad (28)$$

We now recall that Schwarzschild defined the so-called Seidel eikonal S by means of the expression††

$$S=W+(M_0/2n_0\lambda_0^2)(x_0^2+y_0^2) -(M_1/2n_1\lambda_1^2)(x_1^2+y_1^2) +x_0(\xi_1-\xi_0)+y_0(\eta_1-\eta_0), \quad (29)$$

of which the total differential is

$$dS=(\xi_1-\xi_0)dx_0+(\eta_1-\eta_0)dy_0 +(x_0-x_1)d\xi_1+(y_0-y_1)d\eta_1. \quad (30)$$

Remembering that in the accuracy of Gaussian dioptric $x_0=x_1$, $y_0=y_1$, $\xi_0=\xi_1$, $\eta_0=\eta_1$, it immediately follows from Eq. (30) that‡‡

$$S=S^{(0)}+S^{(4)}+\cdots. \quad (31)$$

Schwarzschild showed further that

$$S^{(4)}=-(A/4)r_0^2-(B/4)\rho_1^2-Cr_0^2\rho_1^2\cos^2\varphi -(D/2)r_0^2\rho_1^2+Er_0^3\rho_1\cos\varphi+Fr_0\rho_1^3\cos\varphi, \quad (32)$$

where

$$r_0^2=x_0^2+y_0^2,\quad \rho_1^2=\xi_1^2+\eta_1^2$$

and

$$r_0\rho_1\cos\varphi=x_0\xi_1+y_0\eta_1 \quad (33)$$

are the three Seidel rotational invariants and B, C, D, E, and F are the five well-known Seidel coefficients of primary spherical aberration, coma, astigmatism, curvature, and field distortion, respectively.

A comparison of Eqs. (28) and (29) shows that

$$\psi=S+n_1M_1+\frac{M_0}{2n_0\lambda_0^2}r_0^2-\frac{M_1}{8n_1^3\lambda_1^4}r_0^4+O(M_1\mu^6). \quad (34)$$

We now separate into orders. Remembering that $S^{(2)}=0$, we immediately obtain, if we also observe that $\Psi(0,0,0,0)=[O_1O_1']$,

$$\left.\begin{aligned}\Psi^{(0)}&=[O_0O_1'],\\ \Psi^{(2)}&=(M_0/2n_0\lambda_0^2)r_0^2,\\ \Psi^{(4)}&=S^{(4)}-(M_1/8n_1^3\lambda_1^4)r_0^4.\end{aligned}\right\} \quad (35)$$

Since $\Psi^{(4)}$ differs from $S^{(4)}$ only by the term $(M_1/8n_1^3\lambda_1^4)r_0^4$, it follows that when $\Psi^{(4)}$ is expressed in a form analogous to Eq. (32), it only differs in the coefficient of r_0^4:

$$\psi^{(4)}=-\frac{1}{4}\left(A+\frac{M_1}{2n_1^3\lambda_1^4}\right)r_0^4-\frac{B}{4}\rho_1^4-Cr_0^2\rho_1^2\cos^2\varphi -\frac{D}{2}r_0^2\rho_1^2+Er_0^3\rho_1\cos\varphi+Fr_0\rho_1^3\cos\varphi. \quad (36)$$

Hence, in this representation five of the coefficients of $\Psi^{(4)}$ are identical with the Seidel sums. Moreover, since

$$\psi^{(4)}(r_0,O,\varphi)=\frac{1}{4}\left(A+\frac{M_1}{2n_1^3\lambda_1^4}\right)r_0^4,$$

the remaining coefficient is related to the spherical aberration of the axial point of the pupil.

Finally, in order to express the aberration characteristic in terms of the ordinary rotational invariants of Sec. II rather than in terms of the Seidel rotational invariants r_0^2, ρ_1^2, and $r_0\rho_1\cos\varphi$ we observe that

$$r_0^2=\left(\frac{n_0\lambda_0}{M_0}\right)^2r^2,\quad \rho_1^2=\frac{1}{\lambda_1^2}[\rho^2+O(M_1^2\mu^4)]$$

and

$$r_0\rho_1\cos\varphi=\frac{n_0\lambda_0}{M_0\lambda_1}[r\rho\cos\varphi+O(X_0M_1\mu^3)].$$

It is seen on substituting these expressions into Eqs. (35) and (36) that the error terms will give contributions of order higher than fourth. Hence, *the required expression for the aberration characteristic up to fourth order is obtained by replacing in Eqs. (35) and (36) the Seidel rotational invariants by* $(n_0\lambda_0/M_0)^2r^2$, $(1/\lambda_1^2)\rho^2$, *and* $(n_0\lambda_0/M_0\lambda_1)r\rho\cos\varphi$, *respectively*. We then finally obtain, if in addition we *set* the arbitrary factor λ_0 *equal to unity*,

$$\left.\begin{aligned}\Psi^{(0)}&=[O_0O_1'],\\ \Psi^{(2)}&=\tfrac{1}{2}(n_0/M_0)r^2,\\ \Psi^{(4)}&=-\tfrac{1}{4}A'r^4-\tfrac{1}{4}B'\rho^4-C'r^2\rho^2\cos^2\varphi-\tfrac{1}{2}D'r^2\rho^2 +E'r^3\rho\cos\varphi+F'r\rho^3\cos\varphi,\end{aligned}\right\} \quad (37)$$

†† Schwarzschild was led to the introduction of the S function by observing that in terms of the Seidel variables

$$dW=x_0\left(d\xi_0-\frac{M_0}{n_0}\frac{dx_0}{\lambda_0^2}\right)+y_0\left(d\eta_0-\frac{M_0}{n_0}\frac{dy_0}{\lambda_0^2}\right) -x_1\left(d\xi_1-\frac{M_1}{n_1}\frac{dx_1}{\lambda_1^2}\right)-y_1\left(d\eta_1-\frac{M_1}{n_1}\frac{dy_1}{\lambda_1^2}\right)$$

and that the terms on the right-hand side are part of the simple total differential (30).

‡‡ Schwarzschild (see reference 1) draws the incorrect conclusion that not only $S^{(2)}$ but also $S^{(0)}$ is zero. This error does, however, not vitiate his subsequent analysis. It is seen from Eqs. (34), and (35), that $S^{(0)}=[O_0O_1]$.

where§§

$$\begin{aligned} A'&=(n_0/M_0)^4(A+M_1/2n_1^3\lambda_1^4),\\ B'&=(1/\lambda_1^4)B,\\ C'&=(n_0/M_0)^2(1/\lambda_1^2)C,\\ D'&=(n_0/M_0)^2(1/\lambda_1^2)D,\\ E'&=(n_0/M_0)^3(1/\lambda_1)E,\\ F'&=(n_0/M_0)(1/\lambda_1^3)F. \end{aligned} \tag{38}$$

We also immediately find from Eqs. (17), (20), and (37) that

$$\Phi=\Phi^{(4)}+\Phi^{(6)}+\cdots, \tag{39}$$

where

$$n_1\Phi^{(4)}=-\tfrac{1}{4}B'\rho^4-C'r^2\rho^2\cos^2\varphi -\tfrac{1}{2}D'r^2\rho^2+E'r^3\rho\cos\varphi+F'r\rho^3\cos\varphi. \tag{40}$$

This equation shows that the five aberration coefficients B', C', D', E', and F' can be interpreted in a natural way when they are defined with relation to the deformations of the wave fronts from spherical form.

ACKNOWLEDGMENT

In conclusion it is a pleasure to thank Professor Max Born, F.R.S., for many stimulating discussions and for his interest in this work.

§§ When the object is at infinity we use, as explained at the end of Sec. II, the Φ function only. In that case Φ is still given by Eq. (40) but $(r/M_0)^2$ must be used in place of the variable r^2.

PART II

A GENERAL THEORY OF INTERFERENCE AND DIFFRACTION OF LIGHT

2.1

Reprinted without change of pagination from the
Proceedings of the Royal Society, A, *volume* 225, pp. 96–111, 1954

A macroscopic theory of interference and diffraction of light from finite sources
I. Fields with a narrow spectral range

By E. Wolf

Department of Mathematical Physics, University of Edinburgh

(*Communicated by M. Born, F.R.S.—Received* 27 *January* 1954)

A macroscopic theory of interference and diffraction of light in stationary fields produced by finite sources which emit light within a finite spectral range is formulated. It is shown that a generalized Huygens principle may be obtained for such fields, which involves only observable quantities. The generalized Huygens principle expresses the *intensity* at a typical point of the field in terms of an integral taken twice independently over an arbitrary surface, the integral involving the intensity distribution over the surface and the values of a certain correlation factor, which is found to be the 'degree of coherence' previously introduced by Zernike. Next it is shown that under fairly general conditions, this correlation factor is essentially the normalized integral over the source of the Fourier (frequency) transform of the spectral intensity function of the source, and that it may be determined from simple interference experiments. Further, it is shown that in regions where geometrical optics is a valid approximation, the coherence factor itself then obeys a simple geometrical law of propagation. Several results on partially coherent fields, established previously by Van Cittert, Zernike, Hopkins and Rogers, follow as special cases from these theorems.

The results have a bearing on many optical problems and can also be applied in investigations concerned with other types of radiation.

1. Introduction

In the usual treatments of interference and diffraction of light, the source is assumed to be of vanishingly small dimensions (a point source), emitting strictly monochromate radiation. Such treatments correspond essentially to an idealized wave field created by a (classical) oscillator. Huygens's principle, in the extended formulation of Fresnel, may be regarded as an approximate propagation law for such fields.

In recent years remarkable advances have been made in practical optics in connexion with interference and diffraction of light and electrons; in particular, the phase-contrast method (Zernike 1934 *a*, *b*), the method of the coherent background (Zernike 1948) and the method of reconstructed wave-fronts (Gabor 1949, 1951) must be mentioned. These discoveries, as well as numerous problems in both theoretical and practical optics, make it highly desirable to extend the theory of interference and diffraction to fields produced by an actual source, i.e. a source of finite extension and one which emits light within a finite frequency range.

First steps towards formulating such a theory were made by Berek (1926 *a*, *b*, *c*, *d*), Van Cittert (1934, 1939), Zernike (1938) and Hopkins (1951, 1953). The experimental counterpart of these investigations dates back to Michelson† (1890, 1891 *a*, *b*, 1892, 1920).‡ In the theoretical papers just referred to correlation factors for light disturbances at two arbitrary points or at two instants of time were introduced and applied to particular problems, notably by Hopkins (1953). However, only a moderate progress was achieved in formulating the general laws relating to such fields and in fact the subject presents to-day a somewhat confused picture. This is mainly because each of the four authors introduced a formally different correlation factor, which in turn led to many disconnected results.

In the present investigation a systematic study is made of interference and diffraction in stationary§ optical fields produced by finite sources which emit light within a finite frequency range. Part I is mainly restricted to the case when the effective frequency range is sufficiently narrow. In § 2 it is shown that a Huygens principle may be formulated for such fields, which involves only observable quantities. In this generalized form, the Huygens principle expresses the *intensity* at an arbitrary point of the field in terms of an integral taken twice independently over an arbitrary surface, the integrand involving (1) the values of the intensity at all points of this surface and (2) a correlation factor, which turns out to be the complex form of the 'degree of coherence' introduced by Zernike. In this formulation the Huygens principle is subject to similar restrictions on its range of validity as encountered in connexion with its usual form, but a rigorous formulation is possible and will be given in part II of this investigation.

In § 3 the significance of the coherence factor is discussed, and it is shown that it may be determined from simple interference experiments. In § 4 it is shown that under fairly general conditions the coherence factor is essentially the normalized integral over the source of the Fourier (frequency) transform of the intensity function $j(\boldsymbol{\xi}, \nu)$ of the source. This relation takes a particularly simple form when the frequency range of the radiation is sufficiently narrow and when some further simplifying conditions are satisfied. Under these restrictions several of the earlier

† See also Michelson & Pease (1921) and Pease (1931).

‡ A fuller historical survey is given in my article in *Vistas in astronomy* (Wolf 1954). In addition to the literature quoted there, reference to a discussion of a more abstract kind may be added: Wiener (1930), chapter III, § 9.

§ By a stationary field we mean here a field of which all observable properties are constant in time. This definition includes as special case the usual case of high frequency sinusoidal time dependence; or the field constituted by the steady flux of (polychromatic) radiation through an optical system. But it excludes fields for which the time average over a macroscopic time-interval of the flux of radiation depends on time.

results of Van Cittert (1934), Zernike (1938) and Hopkins (1951) are then shown to follow as special cases. It is also found that in regions where the approximations of geometrical optics hold, the coherence factor itself then obeys a simple geometrical law of propagation. In § 6 it is shown that when the source is small, the generalized Huygens principle may be expressed in a simple form in which the properties of the source and the transmission properties of the medium are completely separated.

The results have an immediate bearing on many optical problems and can also be applied in investigations concerned with other types of radiation.

2. A generalized Huygens's principle

We shall be concerned with stationary optical fields and begin by considering the propagation of a beam of natural, nearly monochromatic light from a finite source Σ. For reasons of convergence we assume that the radiation field exists only between the instants $t = -T$ and $t = +T$. It is easy to pass to the limit $T \to \infty$ subsequently.

Let $V(\mathbf{x}, t)$ denote the disturbance at a point specified by the position vector $\mathbf{x}$, at time t. We shall represent V in the form of a Fourier integral:

$$V(\mathbf{x}, t) = \int_{-\infty}^{+\infty} v(\mathbf{x}, \nu)\, \mathrm{e}^{-2\pi i\nu t}\, \mathrm{d}\nu. \tag{2·1}$$

Then
$$v(\mathbf{x}, \nu) = \int_{-T}^{T} V(\mathbf{x}, t)\, \mathrm{e}^{2\pi i\nu t}\, \mathrm{d}t. \tag{2·2}$$

Since the light is assumed to be almost monochromatic, $|\, v(\mathbf{x}, \nu) \,|$ will differ appreciably from zero only in a narrow frequency range $\nu_0 - \Delta\nu \leqslant \nu \leqslant \nu_0 + \Delta\nu$.

Let us take a surface $\mathscr{A}$ cutting across the beam and consider the intensity at a point $P(\mathbf{x})$ on that side of $\cancel{\Sigma}$ towards which the light is advancing (figure 1). $\mathscr{A}$/

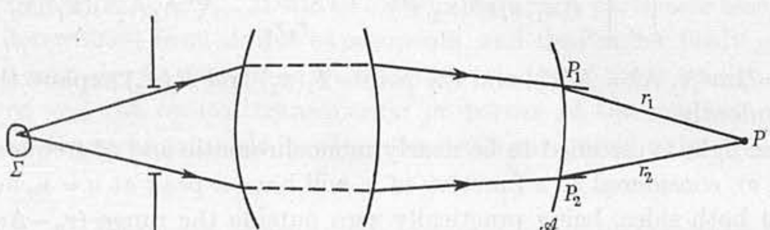


Figure 1

Each Fourier component of (2·1) represents a perfectly monochromatic wave, and therefore (under the usual restrictions on its range of validity) obeys Huygens's principle in the usual form

$$v(\mathbf{x}, \nu) = \int_{\mathscr{A}} v(\mathbf{x}_1, \nu)\, \frac{\mathrm{e}^{ikr_1}}{r_1}\, \Lambda_1\, \mathrm{d}\mathbf{x}_1. \tag{2·3}$$

Here $\mathbf{x}_1$ is the position vector of a typical point P_1 on the surface $\mathscr{A}$, r_1 is the distance from P_1 to P, $k = \dfrac{2\pi\nu}{c} = \dfrac{2\pi}{\lambda}$, c being the vacuum velocity of light and λ the wavelength. Further, Λ_1 is the usual inclination factor of Huygens's principle:

$$\Lambda_1 = \frac{\mathrm{i}}{2\lambda}[\cos\phi_1' - \cos\phi_1]; \tag{2·4}$$

the meaning of the angles ϕ_1 and ϕ_1' is shown in figure 2.

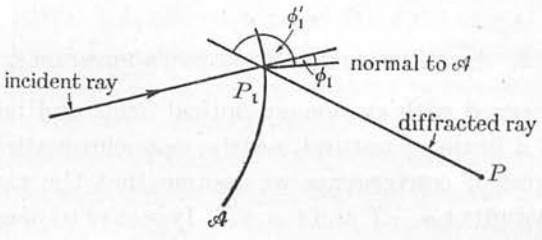


Figure 2

From (2·1) and (2·3) it follows that

$$V(\mathbf{x}, t) = \int_{-\infty}^{+\infty} \mathrm{d}\nu\, \mathrm{e}^{-2\pi\mathrm{i}\nu t} \int_{\mathscr{A}} v(\mathbf{x}_1, \nu) \frac{\mathrm{e}^{\mathrm{i}kr_1}}{r_1} \Lambda_1\, \mathrm{d}\mathbf{x}_1. \tag{2·5}$$

The intensity at P is then given by

$$I(\mathbf{x}) = \langle V(\mathbf{x}, t)\, V^*(\mathbf{x}, t)\rangle, \tag{2·6}$$

where asterisks denote the complex conjugate and the brackets denote average over the interval $-T \leqslant t \leqslant T$. Substituting from (2·5) into (2·6) we obtain

$$I(\mathbf{x}) = \Big\langle \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathrm{d}\nu\, \mathrm{d}\nu' \exp\{-2\pi\mathrm{i}(\nu - \nu')\, t\} \times \int_{\mathscr{A}}\int_{\mathscr{A}} v(\mathbf{x}_1, \nu)\, v^*(\mathbf{x}_2, \nu') \frac{\exp\{\mathrm{i}[kr_1 - k'r_2]\}}{r_1 r_2} \Lambda_1 \Lambda_2'^*\, \mathrm{d}\mathbf{x}_1\, \mathrm{d}\mathbf{x}_2 \Big\rangle, \tag{2·7}$$

where $k' = 2\pi\nu'/c$, $\Lambda' = \Lambda(k')$, and the points $P_1(\mathbf{x}_1)$ and $P_2(\mathbf{x}_2)$ explore the surface $\mathscr{A}$ independently.

Since the light is assumed to be nearly monochromatic and of frequency ν_0 the curve $v(\mathbf{x}, \nu)$, considered as a function of ν, will have a peak at $\nu = \nu_0$ and fall off rapidly on both sides, being practically zero outside the range $(\nu_0 - \Delta\nu, \nu_0 + \Delta\nu)$. Under these conditions we may replace both k and k' in the term

$$\exp\{\mathrm{i}[kr_1 - k'r_2]\}\, \Lambda_1 \Lambda_2'^*$$

by $k^{(0)} = 2\pi\nu_0/c$. (2·7) then becomes

$$I(\mathbf{x}) = \int_{\mathscr{A}}\int_{\mathscr{A}} \Gamma(\mathbf{x}_1, \mathbf{x}_2) \frac{\exp\{\mathrm{i}k^{(0)}(r_1 - r_2)\}}{r_1 r_2} \Lambda_1^{(0)} \Lambda_2^{(0)*}\, \mathrm{d}\mathbf{x}_1\, \mathrm{d}\mathbf{x}_2, \tag{2·8}$$

where

$$\Gamma(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2T}\int_{-T}^{T} \mathrm{d}t \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} v(\mathbf{x}_1, \nu)\, v^*(\mathbf{x}_2, \nu') \exp\{-2\pi\mathrm{i}(\nu - \nu')\, t\}\, \mathrm{d}\nu\, \mathrm{d}\nu'. \tag{2·9}$$

Since
$$\operatorname*{Lim}_{T\to\infty}\int_{-T}^{+T}\exp\{-2\pi i(\nu-\nu')t\}\,dt=\delta(\nu-\nu'),$$

where δ is the Dirac Delta function, it follows that, for sufficiently large T, (2·9) reduces to
$$\Gamma(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{2T}\int_{-\infty}^{+\infty}v(\mathbf{x}_1,\nu)\,v^*(\mathbf{x}_2,\nu)\,d\nu, \tag{2·10}$$

or, using the convolution (Faltung) theorem,
$$\Gamma(\mathbf{x}_1,\mathbf{x}_2)=\langle V(\mathbf{x}_1,t)\,V^*(\mathbf{x}_2,t)\rangle. \tag{2·11}$$

It will be useful to normalize Γ by setting
$$\gamma(\mathbf{x}_1,\mathbf{x}_2)=\frac{\Gamma(\mathbf{x}_1,\mathbf{x}_2)}{\sqrt{\{\Gamma(\mathbf{x}_1,\mathbf{x}_1)\,\Gamma(\mathbf{x}_2,\mathbf{x}_2)\}}}. \tag{2·12}$$

Now $\Gamma(\mathbf{x}_s,\mathbf{x}_s)$ $(s=1,2)$ is nothing but the intensity I_s at the point $\mathbf{x}_s$, so that (2·12) may be written as
$$\gamma(\mathbf{x}_1,\mathbf{x}_2)=\frac{\Gamma(\mathbf{x}_1,\mathbf{x}_2)}{\sqrt{(I_1I_2)}}. \tag{2·13}$$

Substituting from (2·13) into (2·8), we finally obtain
$$I(\mathbf{x})=\int_{\mathscr{A}}\int_{\mathscr{A}}\sqrt{(I_1I_2)}\,\gamma_{12}\frac{\exp\{ik^{(0)}(r_1-r_2)\}}{r_1r_2}\Lambda_1^{(0)}\Lambda_2^{(0)*}\,d\mathbf{x}_1\,d\mathbf{x}_2, \tag{2·14}$$

where γ_{12} has been written for $\gamma(\mathbf{x}_1,\mathbf{x}_2)$.

Equation (2·14) may be regarded as a *generalized Huygens principle*. In the usual formulation (2·3), Huygens's principle applies only to strictly coherent radiation and expresses the (non-observable) *disturbance* at a point in the wave field as sum of contributions from each element $d\mathbf{x}_1$ of the primary wave (or, more generally, of an arbitrary surface). In the present formulation, the restriction of strict coherence is dropped and the *intensity* is calculated by summing over all products $d\mathbf{x}_1 d\mathbf{x}_2$ of the surface, each contribution being weighted by the appropriate value of the correlation factor γ_{12}. It will be shown that in any particular case, this factor may be determined from simple experiments, and that under fairly general conditions it may also be calculated from the knowledge of the intensity function of the source and the optical transmission properties of the medium. Hence our generalized Huygens principle involves *observable quantities*† only.

3. Determination of the γ factor from experiment

The relations (2·13) and (2·11) are formally equivalent to relations (4) and (5) of Zernike's (1938) paper and show that $\Gamma(\mathbf{x}_1,\mathbf{x}_2)$ is the *mutual intensity* and $\gamma(\mathbf{x}_1,\mathbf{x}_2)$

† An earlier formulation of Huygens's principle in terms of observable quantities, due to Gabor (1952; Private communication), must also be mentioned: 'Set up a coherent radiation field and apply to it a small perturbation by introducing objects in the path of radiation which do not destroy the coherence. If the *absolute amplitudes* of the perturbed field are known in one cross-section, they are thereby determined in all cross-sections.' This formulation is, however, not sufficiently general, being restricted to strictly coherent radiation.

the complex degree of coherence,† two important concepts introduced by Zernike. The degree of coherence has been previously defined in a different way and under more restrictive conditions by Van Cittert (1934).

In general γ is complex. It is easily seen that its absolute value is less than unity. For one has, using the well-known modulus inequality for integrals,

$$\left|\int_{-T}^{T} V(\mathbf{x}_1, t)\, V^*(\mathbf{x}_2, t)\,\mathrm{d}t\right| \leqslant \left\{\int_{-T}^{T} |\, V(\mathbf{x}_1, t)\, V^*(\mathbf{x}_2, t)\,|\,\mathrm{d}t\right\}. \tag{3·1}$$

Moreover, by Schwarz's inequality,

$$\left\{\int_{-T}^{T} |\, V(\mathbf{x}_1, t)\, V^*(\mathbf{x}_2, t)\,|\,\mathrm{d}t\right\}^2 \leqslant \int_{-T}^{T} |\, V(\mathbf{x}_1, t)\,|^2\,\mathrm{d}t \int_{-T}^{T} |\, V(\mathbf{x}_2, t)\,|^2\,\mathrm{d}t. \tag{3·2}$$

From (3·1) and (3·2),

$$\left|\int_{-T}^{T} V(\mathbf{x}_1, t)\, V^*(\mathbf{x}_2, t)\,\mathrm{d}t\right| \leqslant \left\{\int_{-T}^{T} |\, V(\mathbf{x}_1, t)\,|^2\,\mathrm{d}t\right\}^{\frac{1}{2}} \left\{\int_{-T}^{T} |\, V(\mathbf{x}_2, t)\,|^2\,\mathrm{d}t\right\}^{\frac{1}{2}}, \tag{3·3}$$

or, in terms of Γ, $\quad |\,\Gamma(\mathbf{x}_1, \mathbf{x}_2)\,| \leqslant \sqrt{\{\Gamma(\mathbf{x}_1, \mathbf{x}_1)\,\Gamma(\mathbf{x}_2, \mathbf{x}_2)\}}$,

whence

$$|\,\gamma(\mathbf{x}_1, \mathbf{x}_2)\,| \leqslant 1. \tag{3·4}$$

In order to see the significance of γ and also to confirm our earlier statement that it is an observable quantity, we shall apply our generalized Huygens principle to a simple interference experiment.

Figure 3 shows the arrangement. Light from a finite source Σ falls either directly or via an optical system on to a screen $\mathscr{A}$ which has small openings at P_1 and P_2. The resulting interference fringes are observed on a second screen $\mathscr{A}'$.

If $\mathrm{d}\mathscr{A}_1$ and $\mathrm{d}\mathscr{A}_2$ denote the areas of the openings at P_1 and P_2, our generalized Huygens principle (2·14) reduces to the following expression when integration is taken over the non-illuminated side of the screen $\mathscr{A}$:

$$\begin{aligned} I \sim I_1\gamma_{11}\left(\frac{1}{r_1}\right)^2 \Lambda_1\Lambda_1^*(\mathrm{d}\mathscr{A}_1)^2 + \sqrt{(I_1 I_2)}\,\gamma_{12}\,\frac{\exp\{ik(r_1-r_2)\}}{r_1 r_2}\,\Lambda_1\Lambda_2^*\,\mathrm{d}\mathscr{A}_1\,\mathrm{d}\mathscr{A}_2 \\ + I_2\gamma_{22}\left(\frac{1}{r_2}\right)^2 \Lambda_2\Lambda_2^*(\mathrm{d}\mathscr{A}_2)^2 + \sqrt{(I_2 I_1)}\,\gamma_{21}\,\frac{\exp\{ik(r_2-r_1)\}}{r_2 r_1}\,\Lambda_2\Lambda_1^*\,\mathrm{d}\mathscr{A}_2\,\mathrm{d}\mathscr{A}_1. \end{aligned} \tag{3·5}$$

The upper index zero has now been omitted on k, Λ_1 and Λ_2^*. Since $\gamma_{11} \equiv 1$ it follows that the first term gives precisely the value $I^{(1)}$ of the intensity which would be obtained at P if the opening at P_1 alone was open ($\mathrm{d}\mathscr{A}_2 = 0$), the third term having a similar interpretation:

$$\left.\begin{aligned} I^{(1)}(\mathbf{x}) &= \frac{I_1}{r_1^2}\,|\,\Lambda_1\,|^2\,(\mathrm{d}\mathscr{A}_1)^2, \\ I^{(2)}(\mathbf{x}) &= \frac{I_2}{r_2^2}\,|\,\Lambda_2\,|^2\,(\mathrm{d}\mathscr{A}_2)^2. \end{aligned}\right\} \tag{3·6}$$

† In the first part of this paper, Zernike defined the degree of coherence of two light vibrations as the visibility of the interference fringes that may be obtained from them under the best circumstances, i.e. when both intensities are made equal and only small path-differences introduced. The analytical definition given by equation (4) of his paper [equivalent to (2·12) above] is however more general and applies whether or not the two intensities are equal; nor is it restricted to small path differences.

The second and the fourth term in (3·5) are complex conjugates of each other. Hence using the identity

$$z+z^* = 2\mathscr{R}(z) \tag{3·7}$$

($\mathscr{R}$ denoting the real part), the sum of the two terms may be written as

$$2\sqrt{(I_1 I_2)}\frac{1}{r_1}\frac{1}{r_2}|\Lambda_1||\Lambda_2||\gamma_{12}|\cos[\arg\gamma_{12}+k(r_1-r_2)]\,\mathrm{d}\mathscr{A}_1\mathrm{d}\mathscr{A}_2$$
$$= 2\sqrt{(I^{(1)}I^{(2)})}\,|\gamma_{12}|\cos[\arg\gamma_{12}+k(r_1-r_2)], \tag{3·8}$$

and (3·5) then reduces to

$$I = I^{(1)}+I^{(2)}+2\sqrt{(I^{(1)}I^{(2)})}\,|\gamma_{12}|\cos[\arg\gamma_{12}+k(r_1-r_2)]. \tag{3·9}$$

We see that in the limiting cases $|\gamma_{12}| = 1$ and $|\gamma_{12}| = 0$ (3·9) reduces to the usual laws for the combination of *perfectly coherent* and *completely incoherent* disturbances. Hence (3·9) may be regarded as a *generalized interference law* in which the factor γ_{12} is a measure of the degree of correlation between the disturbances at P_1 and P_2. This law was derived previously in a different manner by Zernike

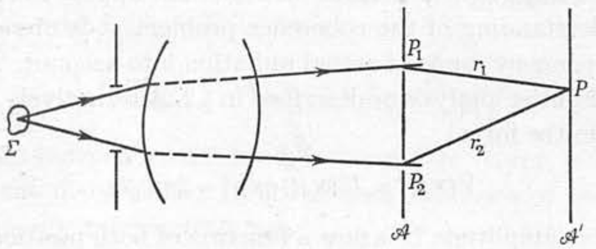


FIGURE 3. An interference experiment with a finite source.

(1938). It has also been derived under a more restrictive definition of the coherence factor by Hopkins (1951).

The generalized interference law (3·9) enables the calculation of $I(\mathbf{x})$ when $I^{(1)}$, $I^{(2)}$ and γ_{12} is known. Conversely when $I^{(1)}$, $I^{(2)}$ and I are known γ_{12} may be determined from measurements of intensities. One has only to measure $I^{(1)}$, $I^{(2)}$ and I for different values of r_1 and r_2 and solve the resulting equations obtained from (3·9) for $|\gamma_{12}|$ and $\arg\gamma_{12}$. Hence the coherence factor may be considered to be an *observable quantity*.

As pointed out by Zernike and Hopkins, the coherence factor is closely related to the *visibility*† and the *position* of the fringes. By (3·9) one has

$$\left.\begin{aligned} I_{\text{max.}} &= I^{(1)}+I^{(2)}+2\sqrt{(I^{(1)}I^{(2)})}\,|\gamma_{12}|,\\ I_{\text{min.}} &= I^{(1)}+I^{(2)}-2\sqrt{(I^{(1)}I^{(2)})}\,|\gamma_{12}|.\end{aligned}\right\} \tag{3·10}$$

† The visibility $\mathscr{V}(\mathbf{x})$ of interference fringes, a concept due to Michelson, is defined by

$$\mathscr{V} = \frac{I_{\text{max.}}-I_{\text{min.}}}{I_{\text{max.}}+I_{\text{min.}}},$$

where $I_{\text{max.}}$ is the maximum of the intensity at the centre of the brightest fringe near $P(\mathbf{x})$ and $I_{\text{min.}}$ is the intensity at the centre of the adjacent dark fringe.

Hence the visibility $\mathscr{V}$ is given by

$$\mathscr{V} = \frac{2}{\sqrt{\frac{I^{(1)}}{I^{(2)}}} + \sqrt{\frac{I^{(2)}}{I^{(1)}}}} \, |\gamma_{12}|. \tag{3·11}$$

Moreover, it is seen from (3·9) that $\arg \gamma_{12}$ appears formally as the 'phase difference' between the disturbances at the two points; it is equal (in suitable units) to the amount of lateral displacement of the fringe system from the position which it would occupy if $\arg \gamma_{12}$ was equal to zero. In practice one can often set $I^{(1)} = I^{(2)}$. (3·11) shows that the visibility is then simply equal to the absolute value of the coherence factor.

In the paper already referred to, Hopkins (1951) claimed that Zernike's results relating to partially coherent fields may be derived without any resort to statistical analysis. This claim is not justified; Hopkins's analysis, unlike Zernike's, is applicable only to the limiting case of vanishingly narrow frequency range. For a similar reason, a claim by Van Cittert (1939) as to the equivalence of his correlation factor with the degree of coherence of Zernike also does not appear to be justified.

For a full understanding of the coherence problem, it is obviously essential to take the finite frequency range of actual radiation into account. This may be done with the help of Fourier analysis as described in § 2. Alternatively, one may express the disturbance in the form

$$V(\mathbf{x}, t) = U(\mathbf{x}, t) \exp\{-2\pi i \nu_0 t\}, \tag{3·12}$$

where the complex amplitude U is now a function of both position and time. Since the light is assumed to be nearly monochromatic and of frequency ν_0, U will, however, at each point vary slowly (and irregularly) in comparison with the frequency ν_0, and will remain practically constant over an interval of time depending on the coherence length of the light.†

The path differences encountered in instrumental optics are, as a rule, small compared to the coherence length. Hopkins's theory will in these cases lead substantially to the same results as the present analysis. When large path differences are involved Hopkins's theory can, however, no longer be expected to be applicable.

The representation (3·12) was used by Zernike (1938) and also previously by Berek (1926 (*a*)–(*d*)). For a quantitative treatment the Fourier integral approach seems, however, to be more appropriate and has also the attractive feature that it brings the optical coherence problems within the scope of methods well established in connexion with other statistical problems encountered in physics.

Experimental investigations, using small path differences, were recently carried out by Baker (1953) and Arnulf, Dupuy & Flamant (1953). Good agreement with the existing theories was obtained.

† The coherence length may be defined as the maximum path difference between the interfering beams for which interference fringes may be obtained. The finite value of the coherence length arises mainly from (1) natural broadening of spectral lines due to the finite lifetime of atomic states, (2) broadening due to atomic collisions and (3) Doppler broadening due to the thermal motion of the atoms. For an excellent account of the coherence length see Born (1933).

Finally a paper by Duffieux (1953) may also be mentioned. It contains some criticisms of the earlier theories and discusses the connexion between the coherence factor and the transmission functions introduced by him and Lansraux in earlier investigations.

4. Expressions for the mutual intensity function

We shall now derive an explicit expression for the mutual intensity function Γ_{12} in terms of the intensity function of the source and the transmission function of the medium. No restriction on the frequency range will now be imposed.

We assume that the source Σ is a radiating element of a plane surface and divide it into elements $\mathrm{d}\Sigma_1, \mathrm{d}\Sigma_2, \ldots$, whose linear dimensions are small compared to the wave-length. If $V_m(\mathbf{x}, t)$ denotes the disturbance at the point $P(\mathbf{x})$ and time t due to the mth element of the source, then

$$V(\mathbf{x}, t) = \sum_m V_m(\mathbf{x}, t),$$

and (2·11) becomes

$$\begin{aligned}\Gamma_{12} &= \langle \sum_m V_m(\mathbf{x}_1, t) \sum_n V_n^*(\mathbf{x}_2, t)\rangle \\ &= \langle \sum_m V_m(\mathbf{x}_1, t)\, V_m^*(\mathbf{x}_2, t)\rangle + \langle \sum_{m \neq n} V_m(\mathbf{x}_1, t)\, V_n^*(\mathbf{x}_2, t)\rangle, \end{aligned} \tag{4·1}$$

where m and n run independently through all possible values. Now to a very good approximation, the disturbances from different elements of the source may be treated as statistically independent, i.e.

$$\langle V_m(\mathbf{x}_1, t)\, V_n^*(\mathbf{x}_2, t)\rangle = 0 \quad \text{when} \quad m \neq n,$$

so that the last term in (4·1) vanishes. If we also introduce the Fourier inverse $v_m(\mathbf{x}, \nu)$ of $V_m(\mathbf{x}, t)$

$$v_m(\mathbf{x}, \nu) = \int_{-T}^{T} V_m(\mathbf{x}, t)\, \mathrm{e}^{2\pi i \nu t}\, \mathrm{d}t,$$

and use the convolution theorem, we then obtain the following expression for Γ_{12}:

$$\Gamma_{12} = \langle \sum_m V_m(\mathbf{x}_1, t)\, V_m^*(\mathbf{x}_2, t)\rangle \tag{4·2a}$$

$$= \frac{1}{2T} \sum_m \int_{-\infty}^{+\infty} v_m(\mathbf{x}_1, \nu)\, v_m^*(\mathbf{x}_2, \nu)\, \mathrm{d}\nu. \tag{4·2b}$$

Let $$K(\mathbf{x}', \mathbf{x}'', \nu) = a(\mathbf{x}', \mathbf{x}'') \exp\{2\pi i (\nu/c) \mathscr{S}(\mathbf{x}', \mathbf{x}'')\} \tag{4·3}$$

(a and $\mathscr{S}$ real) be the *transmission function* of the medium, defined as the disturbance at $P''(\mathbf{x}'')$ due to a monochromatic point source of frequency ν and of unit strength† at $P'(\mathbf{x}')$. (Strictly a and $\mathscr{S}$ also depend on the frequency, but this dependence may here be neglected.) In particular, if the points are situated in a region where diffraction effects are not dominant, then the phase function $\mathscr{S}$ is simply the

† We define a source of unit strength as one which *in vacuo* would give rise to a disturbance of unit amplitude at a unit distance from it.

Hamilton point characteristic function of the medium, i.e. the optical length of the natural ray joining $\mathbf{x}'$ to $\mathbf{x}''$, and the amplitude a may be obtained from the usual conservation law of geometrical optics. *In vacuo*, for example,

$$K = \frac{\exp\{ik\,|\,\mathbf{x}'-\mathbf{x}''\,|\}}{|\,\mathbf{x}'-\mathbf{x}''\,|}. \tag{4·4}$$

If $\boldsymbol{\xi}_m$ denotes the position vector of the element $d\Sigma_m$ of the source, then the disturbance at a point $P(\mathbf{x})$ due to a typical Fourier component $v_m(\mathbf{x},\nu)$ of $V_m(\mathbf{x},t)$ may be written as

$$v_m(\mathbf{x},\nu) = v_m(\boldsymbol{\xi}_m,\nu)\,K(\boldsymbol{\xi}_m,\mathbf{x},\nu). \tag{4·5}$$

Hence (4·2b) becomes

$$\begin{aligned}\Gamma_{12} &= \frac{1}{2T}\sum_m \int_{-\infty}^{+\infty} |\,v_m(\boldsymbol{\xi}_m,\nu)\,|^2\,K(\boldsymbol{\xi}_m,\mathbf{x}_1,\nu)\,K^*(\boldsymbol{\xi}_m,\mathbf{x}_2,\nu)\,d\nu \\ &= \frac{1}{2T}\sum_m a(\boldsymbol{\xi}_m,\mathbf{x}_1)\,a(\boldsymbol{\xi}_m,\mathbf{x}_2)\int_{-\infty}^{+\infty} |\,v_m(\boldsymbol{\xi}_m,\nu)\,|^2 \exp\{2\pi i\nu\tau_{12}(\boldsymbol{\xi}_m)\}\,d\nu, \end{aligned} \tag{4·6}$$

where

$$\tau_{12}(\boldsymbol{\xi}_m) = \frac{1}{c}[\mathscr{S}(\boldsymbol{\xi}_m,\mathbf{x}_1)-\mathscr{S}(\boldsymbol{\xi}_m,\mathbf{x}_2)] \tag{4·7}$$

represents the difference in the time needed for light to travel to $\mathbf{x}_1$ and $\mathbf{x}_2$ from the element $d\Sigma_m$ of the source. Applying the convolution theorem to (4·6) it follows that Γ may also be written in the form

$$\Gamma_{12} = \sum_m a(\boldsymbol{\xi}_m,\mathbf{x}_1)\,a(\boldsymbol{\xi}_m,\mathbf{x}_2)\,\langle V_m(\boldsymbol{\xi}_m,t-\tau_{12})\,V_m^*(\boldsymbol{\xi}_m,t)\rangle. \tag{4·8}$$

Instead of $|\,v_m(\boldsymbol{\xi}_m,\nu)\,|$ which is defined for only a discontinuous set of values $\boldsymbol{\xi}_m$, we may introduce a function $j(\boldsymbol{\xi},\nu)$ which represents the *intensity per unit area* of the source, *per unit frequency range*;† it is defined for the whole continuous set of $\boldsymbol{\xi}$ values. Absorbing the factor $1/2T$ in our definition of j, (4·6) then becomes

$$\Gamma_{12} = \int_\Sigma d\boldsymbol{\xi}\,a(\boldsymbol{\xi},\mathbf{x}_1)\,a(\boldsymbol{\xi},\mathbf{x}_2)\int_{-\infty}^{+\infty} j(\boldsymbol{\xi},\nu)\exp\{2\pi i\nu\tau_{12}(\boldsymbol{\xi})\}\,d\nu. \tag{4·9}$$

The amplitude factor $a(\boldsymbol{\xi},\mathbf{x}_s)$ $(s=1,2)$ of the transmission function will as a rule be a slowly varying function of $\boldsymbol{\xi}$. Also in most applications the linear dimensions of the source will be small compared to the distance from the source to $\mathbf{x}_s$. Hence in (4·8) and (4·9) we may as a rule replace $a(\boldsymbol{\xi},\mathbf{x}_s)$ by $a(0,\mathbf{x}_s)$ without introducing an appreciable error. Finally, normalizing Γ as before, we obtain

$$\gamma_{12} = \frac{\sum_m \langle V_m(\boldsymbol{\xi}_m,t-\tau_{12})\,V_m^*(\boldsymbol{\xi}_m,t)\rangle}{\sum_m \langle V_m(\boldsymbol{\xi}_m,t)\,V_m^*(\boldsymbol{\xi}_m,t)\rangle} \tag{4·10a}$$

$$= \frac{\int_\Sigma d\boldsymbol{\xi}\int_{-\infty}^{+\infty} j(\boldsymbol{\xi},\nu)\exp\{2\pi i\nu\tau_{12}(\boldsymbol{\xi})\}\,d\nu}{\int_\Sigma d\boldsymbol{\xi}\int_{-\infty}^{+\infty} j(\boldsymbol{\xi},\nu)\,d\nu}, \tag{4·10b}$$

i.e. *the coherence factor is essentially the normalized integral over the source of the Fourier (frequency) transform of the intensity function of the source.*

† We neglect here the variation of the intensity with direction. The effect of this variation can also be taken into account by introducing a more general intensity function $j(\boldsymbol{\xi},\mathbf{p},\nu)$, $\mathbf{p}$ being a directional variable.

5. Some approximate expressions for the coherence factor

We shall now consider the form which the expressions for γ take when the effective frequency range is sufficiently narrow. We shall show that they lead, in special cases, to several results obtained previously by Van Cittert (1934), Zernike (1938), Hopkins (1951) and Rogers (1953).

Let us assume that the effective frequency range is so narrow that ν may be replaced by ν_0 in the integral in (4·9). This will be permissible, if

$$|\Delta\nu\tau_{12}(\boldsymbol{\xi})| \ll 1,$$

or, since $\Delta\nu = -c\Delta\lambda/\lambda^2$ and $\tau_{12}(\boldsymbol{\xi}) = \frac{1}{c}[\mathscr{S}(\boldsymbol{\xi}, \mathbf{x}_1) - \mathscr{S}(\boldsymbol{\xi}, \mathbf{x}_2)]$, if

$$\frac{|\Delta\lambda|}{\lambda} \ll \frac{\lambda}{|\mathscr{S}(\boldsymbol{\xi}, \mathbf{x}_1) - \mathscr{S}(\boldsymbol{\xi}, \mathbf{x}_2)|}. \tag{5·1}$$

We also set
$$\int_{-\infty}^{\infty} j(\boldsymbol{\xi}, \nu)\, d\nu = J(\boldsymbol{\xi}); \tag{5·2}$$

(4·9) then gives on normalizing

$$\gamma_{12} = \frac{1}{\sqrt{(I_1 I_2)}} \int_\Sigma J(\boldsymbol{\xi})\, a(\boldsymbol{\xi}, \mathbf{x}_1)\, a(\boldsymbol{\xi}, \mathbf{x}_2) \exp\{2\pi i \nu_0 \tau_{12}(\boldsymbol{\xi})\}\, d\boldsymbol{\xi}, \tag{5·3}$$

where
$$I_s = \Gamma_{ss} = \int_\Sigma J(\boldsymbol{\xi})\, a^2(\boldsymbol{\xi}, \mathbf{x}_s)\, d\boldsymbol{\xi} \quad (s = 1, 2). \tag{5·3a}$$

In particular, *in vacuo* (air) one has (cf. (4·4))

$$a(\boldsymbol{\xi}, \mathbf{x}_s) = \frac{1}{R_s}, \quad \mathscr{S}(\boldsymbol{\xi}, \mathbf{x}_s) = R_s,$$

with
$$R_s = |\mathbf{x}_s - \boldsymbol{\xi}|.$$

(5·3) then reduces to

$$\gamma_{12} = \frac{1}{\sqrt{(I_1 I_2)}} \int_\Sigma \frac{J(\boldsymbol{\xi})}{R_1 R_2} \exp\{ik^{(0)}(R_1 - R_2)\}\, d\boldsymbol{\xi}, \tag{5·4}$$

where
$$I_s = \int_\Sigma \frac{J(\boldsymbol{\xi})}{R_s^2}\, d\boldsymbol{\xi} \quad (s = 1, 2).$$

Integrals of the form (5·4) are well known in optics. They represent the complex amplitude at the point P_2 in a diffraction pattern around P_1, when diffraction takes place at an aperture which is identical in form with Σ, the amplitude in the diffracting aperture being proportional to $J(\boldsymbol{\xi})$. This expression for the coherence factor was first obtained by Zernike (1938) and later by Hopkins (1951).

If, as before, we neglect the variation of the amplitude factor with $\boldsymbol{\xi}$, we obtain from (5·3)

$$\gamma_{12} = \frac{\int_\Sigma J(\boldsymbol{\xi}) \exp\{2\pi i \nu_0 \tau_{12}(\boldsymbol{\xi})\}\, d\boldsymbol{\xi}}{\int_\Sigma J(\boldsymbol{\xi})\, d\boldsymbol{\xi}}. \tag{5·5}$$

This relation may be further simplified if either the linear dimensions of the source, or the distance between P_1 and P_2 are small compared to the distance from the source to P_1 and P_2. In the first case one may expand $\tau_{12}(\boldsymbol{\xi})$ at a suitable point O ($\boldsymbol{\xi}=0$) of the source and neglect higher order terms:

$$\tau_{12}(\boldsymbol{\xi}) \sim \frac{1}{c}[\mathscr{S}(0,\mathbf{x}_1)-\mathscr{S}(0,\mathbf{x}_2)]+\frac{1}{c}\boldsymbol{\xi}\,.\,\frac{\partial}{\partial\boldsymbol{\xi}}[\mathscr{S}(\boldsymbol{\xi},\mathbf{x}_1)-\mathscr{S}(\boldsymbol{\xi},\mathbf{x}_2)]_{\boldsymbol{\xi}=0}. \qquad (5\cdot6)$$

Now by the fundamental property of the $\mathscr{S}$ function,

$$\frac{\partial}{\partial\boldsymbol{\xi}}[\mathscr{S}(\boldsymbol{\xi},\mathbf{x}_s)]_{\xi=0} = -\mathbf{p}_s \quad (s=1,2), \qquad (5\cdot7)$$

where $\mathbf{p}_s$ denotes the ray vector† at O of the ray $\overrightarrow{OP_s}$. Hence

$$\tau_{12}(\boldsymbol{\xi}) \sim \frac{1}{c}[\mathscr{S}(0,\mathbf{x}_1)-\mathscr{S}(0,\mathbf{x}_2)]-\frac{1}{c}\boldsymbol{\xi}\,.\,(\mathbf{p}_1-\mathbf{p}_2), \qquad (5\cdot8)$$

and (5·5) reduces to

$$\gamma_{12}=\frac{\int_\Sigma J(\boldsymbol{\xi})\exp\{-\mathrm{i}k^{(0)}\boldsymbol{\xi}\,.\,(\mathbf{p}_1-\mathbf{p}_2)\}\,d\boldsymbol{\xi}}{\int_\Sigma J(\boldsymbol{\xi})\,d\boldsymbol{\xi}}\exp\{\mathrm{i}k^{(0)}[\mathscr{S}(0,\mathbf{x}_1)-\mathscr{S}(0,\mathbf{x}_2)]\}. \qquad (5\cdot9)$$

Hence *when the effective frequency range is sufficiently narrow and the source small enough, the coherence factor is equal to the product of the term*

$$\exp\{\mathrm{i}k^{(0)}[\mathscr{S}(0,\mathbf{x}_1)-\mathscr{S}(0,\mathbf{x}_2)]\}$$

and the complex amplitude in an associated Fraunhofer diffraction pattern.

Similar analysis may be used when the distance between P_1 and P_2 is small compared to the distance from the source to these points. In place of (5·6) we now have

$$\begin{aligned}\tau_{12}(\boldsymbol{\xi}) &\sim -\frac{1}{c}(\mathbf{x}_2-\mathbf{x}_1)\left[\frac{\partial\mathscr{S}(\boldsymbol{\xi},\mathbf{x})}{\partial\mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}_1}\\ &= -\frac{1}{c}(\mathbf{x}_2-\mathbf{x}_1)\,.\,\mathbf{p}_1'(\boldsymbol{\xi}),\end{aligned} \qquad (5\cdot10)$$

where $\mathbf{p}_1'(\boldsymbol{\xi})$ is the ray vector at P_1 of the ray $\boldsymbol{\xi}\to\mathbf{x}_1$. (5·5) now reduces to

$$\gamma_{12}=\frac{\int_\Sigma J(\boldsymbol{\xi})\exp\{-\mathrm{i}k^{(0)}(\mathbf{x}_2-\mathbf{x}_1)\,.\,\mathbf{p}_1'(\boldsymbol{\xi})\}\,d\boldsymbol{\xi}}{\int_\Sigma J(\boldsymbol{\xi})\,d\boldsymbol{\xi}}. \qquad (5\cdot11)$$

In particular, assume that P_1 and P_2 are in a plane parallel to the source and illuminated directly by it and that the medium between the source and the plane is homogeneous and of refractive index $n=1$. If we choose as origin of the position vector the point P_1, then $\mathbf{p}_1'(\boldsymbol{\xi}) = -\boldsymbol{\xi}/|\boldsymbol{\xi}|$. Moreover, it will often be permissible to

† A ray vector $\mathbf{p}$ at a point P is defined by the relation $\mathbf{p}=n\mathbf{s}$, $\mathbf{s}$ being the unit vector along the ray at P and n the value of the refractive index at that point.

replace $|\boldsymbol{\xi}|$ by the distance d between the plane of the source and the plane containing P_1 and P_2. (5·11) then reduces to

$$\gamma_{12} = \frac{\int_\Sigma J(\boldsymbol{\xi}) \exp\{ik^{(0)}\mathbf{x}_2 . \boldsymbol{\xi}/d\}\, d\boldsymbol{\xi}}{\int_\Sigma J(\boldsymbol{\xi})\, d\boldsymbol{\xi}}; \tag{5·12}$$

hence *the coherence factor is again expressed in the form of a complex amplitude in an associated Fraunhofer diffraction pattern*. This result was first obtained, for a source of circular or rectangular form, under a somewhat different definition of the γ factor by Van Cittert (1934). It is of importance in the theory of the stellar interferometer (cf. Hopkins 1951).

From (5·9) one can easily derive a simple '*propagation law*' for the coherence factor, valid (subject to the restrictions mentioned) in regions where diffraction effects are not dominant.† Consider two pairs of points $P_1(\mathbf{x}_1)$, $P_2(\mathbf{x}_2)$ and $P_1'(\mathbf{x}_1')$, $P_2'(\mathbf{x}_2')$ in the field, such that P_1' lies on the ray from O to P_1 and P_2' lies on the ray from O to P_2 (figure 4). Then it immediately follows from (5·9) that

$$\gamma(\mathbf{x}_1', \mathbf{x}_2') = \gamma(\mathbf{x}_1, \mathbf{x}_2) \exp\{ik^{(0)}[\mathscr{S}(\mathbf{x}_1, \mathbf{x}_1') - \mathscr{S}(\mathbf{x}_2, \mathbf{x}_2')]\}. \tag{5·13}$$

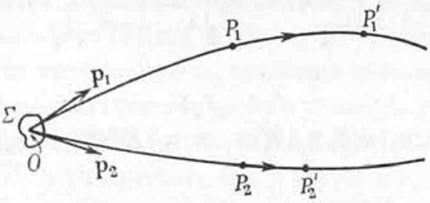


FIGURE 4. Propagation of 'coherence' in regions where geometrical optics is a valid approximation (small source and narrow frequency range assumed).

In particular, if the optical path $[P_1P_1']$ equals the optical path $[P_2P_2']$, e.g. when P_1 and P_2 are one wave-front and P_1' and P_2' on another wave-front, then $\mathscr{S}(\mathbf{x}_1, \mathbf{x}_1') = \mathscr{S}(\mathbf{x}_2, \mathbf{x}_2')$ and (5·13) reduces to

$$\gamma(\mathbf{x}_1', \mathbf{x}_2') = \gamma(\mathbf{x}_1, \mathbf{x}_2). \tag{5·14}$$

We may therefore say that *under the restrictions mentioned, the coherence is propagated in accordance with the laws of geometrical optics.*

It was shown by Hopkins (1951) that the absolute value of the coherence factor for two points in the entry pupil of an optical system and of the corresponding conjugate points in the exit pupil is the same. The result is seen to be an immediate consequence of the theorem just established.

It was also pointed out by Rogers (1953) that the degree of coherence is conserved in a beam of light from image plane to image plane in Gaussian systems. This result too is a special case of the law expressed by (5·13).

† A more general propagation law valid everywhere will be discussed in part II of this investigation.

6. The generalized Huygens principle in the special case of a small source

In this section we shall reformulate the generalized Huygens principle by making use of the relation which expresses the coherence factor in terms of the intensity function of the source, the source being assumed to be sufficiently small. Before doing this, we shall, however, generalize (2·14) by dropping the restriction that the medium between the surface of integration and the point of observation is homogeneous.

If the medium between $\mathscr{A}$ and P is heterogeneous or contains refracting or reflecting surfaces, then clearly the factor $\exp\{ik^{(0)}r_s)/r_s\}$ (with $s = 1, 2$) in (2·14) has to be replaced by the appropriate transmission function $K(\mathbf{x}_s, \mathbf{x}, \nu_0)$, giving

$$I(\mathbf{x}) = \int_{\mathscr{A}}\int_{\mathscr{A}} \sqrt{(I_1 I_2)}\,\gamma_{12} K(\mathbf{x}_1, \mathbf{x}, \nu_0)\,K^*(\mathbf{x}_2, \mathbf{x}, \nu_0)\,\Lambda_1^{(0)}\Lambda_2^{(0)*}\,\mathrm{d}\mathbf{x}_1\,\mathrm{d}\mathbf{x}_2. \tag{6·1}$$

Assume that the conditions under which (5·9) was derived are satisfied. Then, from (5·3a),

$$I_s = \Gamma_{ss} = a^2(0, \mathbf{x}_s) \int_\Sigma J(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = a^2(0, \mathbf{x}_s)\,\mathscr{J}, \tag{6·2}$$

where
$$\mathscr{J} = \int_\Sigma J(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \int_\Sigma \mathrm{d}\boldsymbol{\xi} \int_{-\infty}^{+\infty} j(\boldsymbol{\xi}, \nu)\,\mathrm{d}\nu. \tag{6·3}$$

(6·1) then becomes

$$I(\mathbf{x}) = \mathscr{J} \int_{\mathscr{A}}\int_{\mathscr{A}} \gamma_{12} a(0, \mathbf{x}_1)\,a(0, \mathbf{x}_2)\,K(\mathbf{x}_1, \mathbf{x}, \nu_0)\,K^*(\mathbf{x}_2, \mathbf{x}, \nu_0)\,\Lambda_1^{(0)}\,\Lambda_2^{(0)*}\,\mathrm{d}\mathbf{x}_1\,\mathrm{d}\mathbf{x}_2. \tag{6·4}$$

Substituting from (5·9), setting

$$\frac{\displaystyle\int_\Sigma J(\boldsymbol{\xi}) \exp\{-ik^{(0)}\boldsymbol{\xi}\,.\,(\mathbf{p}_1 - \mathbf{p}_2)\}\,\mathrm{d}\boldsymbol{\xi}}{\displaystyle\int_\Sigma J(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}} = \sigma(\mathbf{p}_1 - \mathbf{p}_2), \tag{6·5}$$

and using (4·3) we obtain

$$I(\mathbf{x}) = \mathscr{J} \int_{\mathscr{A}}\int_{\mathscr{A}} \sigma(\mathbf{p}_1 - \mathbf{p}_2)\,K(0, \mathbf{x}_1, \nu_0)\,K^*(0, \mathbf{x}_2, \nu_0) \times K(\mathbf{x}_1, \mathbf{x}, \nu_0)\,K^*(\mathbf{x}_2, \mathbf{x}, \nu_0)\,\Lambda_1^{(0)}\,\Lambda_2^{(0)*}\,\mathrm{d}\mathbf{x}_1\,\mathrm{d}\mathbf{x}_2. \tag{6·6}$$

In this formula, *the effect of the source and the transmission properties of the medium are completely separated.* The source is characterized by the factor $\sigma(\mathbf{p}_1 - \mathbf{p}_2)$ which is the normalized Fourier transform of $J(\boldsymbol{\xi})$, and the medium is characterized by the transmission function K which in practice can be calculated by methods well known in optical designing (e.g. from a ray trace). The ray vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ occurring in the source factor σ and the position vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ of points on the surface of integration are connected by the canonical relations (5·7). It is clear that the integration over the surface $\mathscr{A}$ may be replaced by integration over the solid angle which the entry pupil of the instrument subtends at the source. Also instead of integration *twice* over *one* surface (e.g. the exit pupil) one may reformulate (6·6) so as to involve *one* integration over each of *two* different surfaces (e.g. the entry and the exit pupils). Similar formulae were found by Hopkins (1951, 1953).

Within the accuracy of the present analysis (6·6) shows that *in a given medium, sources which have identical σ factors will give rise to the same intensity distribution.* This is probably the main reason why, in Gabor's method of reconstructed wavefronts (Gabor 1949, 1951), one may use a different source for the reconstruction from that employed in taking the hologram. In practice, the intensity function will be often practically independent of the position of the radiating element, and it then follows that provided the geometrical shapes of the sources are the same the σ factors will be identical.

7. Concluding remarks

Our generalized Huygens principle makes it possible to obtain solutions to a variety of problems encountered in light optics, and with suitable modifications may also be applied in investigations concerned with other kinds of radiation (electron beams, X-rays, micro-waves). Some possible applications of our results to astronomical investigations and their relation to Michelson's pioneering researches on the application of interference methods to astronomy have already been briefly discussed elsewhere (Fürth & Finlay-Freundlich 1954; Wolf 1954).

It is well known that in applications to problems of image formation in optical systems with low numerical aperture, Huygens's principle gives results in excellent agreement with experiments, but it fails at higher apertures. Now (5·9) and (5·11) show that in systems with high aperture the γ factor may vary appreciably over the domain of integration. Consequently it may be expected that our generalized Huygens principle, which takes into account this variation, will have a wider range of validity than Huygens's principle in its usual form. Since in our formulation the Huygens principle involves observable quantities only, it should be possible to determine its range of validity from experiment.

It can be shown that in the limiting case as $|\gamma_{12}| \rightarrow 1$, (6·1) reduces to usual expressions for the intensity due to an ideal monochromatic point source. Thus (6·1) contains practically the whole elementary diffraction theory of image formation as a special case, and together with (4·9) leads, as we have seen, to many of the previously derived results concerning partially coherent fields. However, the limitations of our analysis must also be stressed, by summarizing the main assumptions under which these formulae were derived: The generalized Huygens principle has been established on the assumption that the usual conditions for the validity of Huygens's principle are satisfied and that the effective frequency range is sufficiently narrow; in the derivation of the expression (4·9) for the Γ factor it has been assumed that the effect of the medium is described with a sufficient accuracy by a transmission function of the form (4·3). In (6·6) it was assumed, in addition to both these conditions, that the source is sufficiently small. Some of these restrictions will be removed in part II of this investigation, where it will be shown that the coherence factors of Van Cittert, Zernike and Hopkins are special cases of a more general correlation function which rigorously obeys the wave equation and with the help of which precise propagation laws may be formulated.

In conclusion, I wish to thank Professor Max Born, F.R.S., for stimulating and helpful discussions, and Professor E. Finlay-Freundlich for valuable information concerning possible applications to astronomy. I am also indebted to Dr A. B. Bhatia and Mr G. Weeden for useful suggestions. Finally, I wish to thank Mr and Mrs R. M. Sillitto for having drawn my attention to some important aspects of Michelson's work.

Part of this work was carried out during the tenure of an Imperial Chemical Industries Research Fellowship and was also supported by a Research Grant from the Carnegie Trust for the Universities of Scotland, both of which are gratefully acknowledged.

References

Arnulf, A., Dupuy, O. & Flamant, F. 1953 *Rev. Opt. (théor. instrum.)*, **32**, 529.
Baker, L. R. 1953 *Proc. Phys. Soc.* B, **66**, 975.
Berek, M. 1926*a* *Z. Phys.* **36**, 675.
Berek, M. 1926*b* *Z. Phys.* **36**, 824.
Berek, M. 1926*c* *Z. Phys.* **37**, 287.
Berek, M. 1926*d* *Z. Phys.* **40**, 420.
Born, M. 1933 *Optik*, §42, 132. Berlin: Springer.
Duffieux, P. 1953 *Rev. Opt. (théor. instrum.)*, **32**, 129.
Fürth, R. & Finlay-Freundlich, E. 1954 Contribution in *Vistas in astronomy*. London: Pergamon Press.
Gabor, D. 1949 *Proc. Roy. Soc.* A, **197**, 454.
Gabor, D. 1951 *Proc. Phys. Soc.* B, **64**, 449.
Hopkins, H. H. 1951 *Proc. Roy. Soc.* A, **208**, 263.
Hopkins, H. H. 1953 *Proc. Roy. Soc.* A, **217**, 408.
Michelson, A. A. 1890 *Phil. Mag.* **30**, 1.
Michelson, A. A. 1891*a* *Phil. Mag.* **31**, 256.
Michelson, A. A. 1891*b* *Phil. Mag.* **31**, 338.
Michelson, A. A. 1892 *Phil. Mag.* **34**, 280.
Michelson, A. A. 1920 *Astrophys. J.* **51**, 257.
Michelson, A. A. & Pease, F. G. 1921 *Astrophys. J.* **53**, 249.
Pease, F. G. 1931 Contribution to *Ergebnisse der Exacten Naturwissenschaften*. Berlin: Springer, **10**, 84.
Rogers, G. L. 1953 *Nature, Lond.*, **172**, 118.
Van Cittert, P. H. 1934 *Physica*, **1**, 201.
Van Cittert, P. H. 1939 *Physica*, **6**, 1129.
Wiener, N. 1930 *Acta Math.* **55**, 182.
Wolf, E. 1954 Contribution in *Vistas in astronomy*. London: Pergamon Press.
Zernike, F. 1934*a* *Physica*, **1**, 689.
Zernike, F. 1934*b* *Mon. Not. R. Astr. Soc.* **94**, 377.
Zernike, F. 1938 *Physica*, **5**, 785.
Zernike, F. 1948 *Proc. Phys. Soc.* **61**, 158.

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, CAMBRIDGE
(BROOKE CRUTCHLEY, UNIVERSITY PRINTER)

Reprinted without change of pagination from the
Proceedings of the Royal Society, A, *volume* 230, pp. 246–265, 1955

A macroscopic theory of interference and diffraction of light from finite sources

II. Fields with a spectral range of arbitrary width

By E. Wolf
The Physical Laboratories, University of Manchester

(*Communicated by M. Born, F.R.S.—Received* 29 *November* 1954)

The results of part I of this investigation are generalized to stationary fields with a spectral range of arbitrary width. For this purpose it is found necessary to introduce in place of the mutual intensity function of Zernike a more general correlation function

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \langle \hat{V}(\mathbf{x}_1, t+\tau)\, \hat{V}^*(\mathbf{x}_2, t)\rangle,$$

which expresses the correlation between disturbances at any two given points $P_1(\mathbf{x}_1)$, $P_2(\mathbf{x}_2)$ in the field, the disturbance at P_1 being considered at a time τ later than at P_2. It is shown that $\hat{\Gamma}$ is an observable quantity. Expressions for $\hat{\Gamma}$ in terms of functions which specify the source and the transmission properties of the medium are derived.

Further, it is shown that *in vacuo* the correlation function obeys rigorously the two wave equations

$$\nabla_s^2 \hat{\Gamma} = \frac{1}{c^2}\frac{\partial^2 \hat{\Gamma}}{\partial \tau^2} \quad (s = 1, 2),$$

where ∇_s^2 is the Laplacian operator with respect to the co-ordinates (x_s, y_s, z_s) of $P_s(\mathbf{x}_s)$. Using this result, a formula is obtained which expresses rigorously the correlation between disturbances at P_1 and P_2 in terms of the values of the correlation and of its derivatives at all pairs of points on an arbitrary closed surface which surrounds P_1 and P_2. A special case of this formula ($P_2 = P_1$, $\tau = 0$) represents a rigorous formulation of the generalized Huygens principle, involving observable quantities only.

1. Introduction

In part I of this investigation (Wolf (1954*a*), to be referred to as I), interference and diffraction of light in stationary fields produced by finite sources which emit light within a small but finite spectral range were studied. Whilst the results obtained are of interest in connexion with a variety of problems, they are inadequate in the treatment of problems where the path differences between the interfering beams are sufficiently large (cf. appendix 2 below). The present paper removes this restriction, and extends the results to stationary fields of any spectral range.

In the investigation of I, the mutual intensity function

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2) = \langle \hat{V}(\mathbf{x}_1, t)\, \hat{V}^*(\mathbf{x}_2, t)\rangle \tag{1·1}$$

of Zernike, played an essential part. In (1·1), $\hat{V}(\mathbf{x}, t)$ represents the complex disturbance at a point specified by the position vector $\mathbf{x}$, at time t, asterisks denoting the complex conjugate and sharp brackets the time average. In order to extend the analysis to fields with a spectral range of arbitrary width, it is found necessary to introduce in place of (1·1) the more general correlation function

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \langle \hat{V}(\mathbf{x}_1, t+\tau)\, \hat{V}^*(\mathbf{x}_2, t)\rangle. \tag{1·2}$$

With the help of this function which is shown to represent an observable quantity, the generalized Huygens principle derived in I and the generalized interference law (I, (3·9)) of Zernike & Hopkins are extended to the wider class of fields under consideration. Expressions for $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$ in terms of quantities which specify the source and the medium are also derived.

In § 7, it is shown that *in vacuo* the correlation function (1·2) obeys rigorously the wave equations

$$\left.\begin{aligned} \nabla_1^2 \hat{\Gamma} &= \frac{1}{c^2}\frac{\partial^2 \hat{\Gamma}}{\partial \tau^2}, \\ \nabla_2^2 \hat{\Gamma} &= \frac{1}{c^2}\frac{\partial^2 \hat{\Gamma}}{\partial \tau^2}, \end{aligned}\right\} \tag{1·3}$$

where ∇_1^2 and ∇_2^2 are the Laplacian operators with respect to the co-ordinates of the points specified by the position vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively.

In § 8, a formula is derived which expresses rigorously the correlation between the disturbances at two given points $P_1(\mathbf{x}_1)$ and $P_2(\mathbf{x}_2)$ in the field in terms of the correlation, and its derivatives, between the disturbances at all pairs of points on an arbitrary closed surface surrounding P_1 and P_2. A special case of this formula ($\mathbf{x}_1 = \mathbf{x}_2$, $\tau = 0$) represents a rigorous formulation of our generalized Huygens principle.

As in I, polarization effects are not considered in the present paper. They will be taken into account in part III of this investigation, where it will be shown that a natural generalization of our results leads to a unified treatment of partial coherence and partial polarization and to a formulation of a wide branch of optics in terms of observable quantities only.†

2. Preliminary: a complex representation of real, polychromatic fields

When dealing with a real, monochromatic (or nearly monochromatic) field, it is usual to employ a complex representation, the field variable being identified with the real part of an appropriate complex function.

† A brief preliminary account of some of these results will be found in Wolf (1954*b*).

In the present paper we shall be concerned with polychromatic fields, i.e. with fields which cover a finite spectral range. It will be convenient to use also in this case a complex representation, this being a natural extension of the representation used in connexion with monochromatic fields.

Let $F(t)$ be a real function, defined for all values of t ($-\infty < t < \infty$) which possess a Fourier integral representation:

$$F(t) = \int_0^\infty (a_\nu \cos 2\pi\nu t + b_\nu \sin 2\pi\nu t)\,\mathrm{d}\nu. \tag{2·1}$$

With F we associate the (generally complex) function $\hat{F}$, defined as

$$\hat{F}(t) = \int_0^\infty (a_\nu + ib_\nu)\,\mathrm{e}^{-2\pi i\nu t}\,d\nu. \tag{2·2}$$

It is seen that

$$F(t) = \mathscr{R}\hat{F}(t), \tag{2·3}$$

where $\mathscr{R}$ denotes the real part.

$\hat{F}(t)$ will be referred to as the *half-range complex function* associated with the real function $F(t)$; it is characterized by the property that it may be represented by a Fourier integral which contains no terms of negative frequencies.† F defines $\hat{F}$ uniquely and vice versa.

Throughout this paper a real function and the associated half-range complex function will be denoted by the same symbol, the latter being distinguished by a circumflex. The use of half-range complex functions in place of real functions considerably shortens some of our calculations, and enables the resulting formulae to be expressed in a form which closely resembles those obtained in connexion with almost monochromatic fields in I.

3. A space-time correlation function of stationary fields

In order to extend the analysis of I to stationary fields the spectral range of which is arbitrarily wide, it is necessary, as will be seen below, to introduce, in place of the mutual intensity function (1·1) of Zernike, the more general correlation function‡

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \langle \hat{V}(\mathbf{x}_1, t+\tau)\,\hat{V}^*(\mathbf{x}_2, t)\rangle, \tag{3·1}$$

† *Added in proof* 21 *March* 1955: Since this was written I find that the same complex representation of real fields was introduced previously by Gabor (1946, p. 432) in his interesting investigations in communication theory. Gabor also points out that the real and imaginary parts of such 'half-range complex functions' are Hilbert transforms of each other.

‡ The auto-correlation function, which in our notation would be written as $\hat{\Gamma}(\mathbf{x}, \mathbf{x}, \tau)$, was previously employed in optics by a number of authors, e.g. Wiener (1930), Van Cittert (1939) and Parke (1948). Wiener (1930, p. 119) points out that this function played a fundamental part in Schuster's theory of white light.

Added in proof 21 *March* 1955: In a very interesting paper which was published since this was written, Blanc-Lapierre & Dumontet (1955) applied the general theory of random functions to the optical coherence problems. In their treatment which leads to several new results some of which are closely related to ours, the cross-correlation function (3·2) plays also a basic role.

which expresses the correlation between disturbances at the points $P_1(\mathbf{x}_1)$ and $P_2(\mathbf{x}_2)$, the disturbance at $\mathbf{x}_1$ being considered at a time τ later than at $\mathbf{x}_2$. The sharp brackets denote the time average.†

By a straightforward calculation carried out fully in appendix 1, it may be shown that if $\hat{V}$ is a half-range complex function (in the sense defined in the previous section) the correlation function defined by (3·1) is likewise a half-range complex function, the knowledge of $\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau)$ being equivalent to the knowledge of the real function

$$\Gamma(\mathbf{x}_1,\mathbf{x}_2,\tau) = \mathscr{R}\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau) = 2\langle V(\mathbf{x}_1,t+\tau)\,V(\mathbf{x}_2,t)\rangle. \tag{3·2}$$

We shall need an expression for $\hat{\Gamma}$ in terms of the Fourier components of $\hat{V}$. Let

$$\left.\begin{aligned}\hat{V}(\mathbf{x},t) &= \int_0^\infty v(\mathbf{x},\nu)\,\mathrm{e}^{-2\pi i\nu t}\,d\nu;\\ \text{then}\qquad v(\mathbf{x},\nu) &= \int_{-T}^{T}\hat{V}(\mathbf{x},t)\,\mathrm{e}^{2\pi i\nu t}\,\mathrm{d}t.\end{aligned}\right\} \tag{3·3}$$

It follows from (3·1), on using the convolution theorem, that

$$\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_0^\infty v(\mathbf{x}_1,\nu)\,v^*(\mathbf{x}_2,\nu)\,\mathrm{e}^{-2\pi i\nu\tau}\,\mathrm{d}\nu. \tag{3·4}$$

We note a useful relation between $\hat{\Gamma}(\mathbf{x}_2,\mathbf{x}_1,-\tau)$ and $\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau)$. From (3·4),

$$\hat{\Gamma}(\mathbf{x}_2,\mathbf{x}_1,-\tau) = \hat{\Gamma}^*(\mathbf{x}_1,\mathbf{x}_2,\tau). \tag{3·5}$$

We shall normalize $\hat{\Gamma}$ by setting

$$\hat{\gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau) = \frac{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau)}{\sqrt{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_1,0)}\sqrt{\hat{\Gamma}(\mathbf{x}_2,\mathbf{x}_2,0)}} = \frac{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau)}{\sqrt{I(\mathbf{x}_1)}\sqrt{I(\mathbf{x}_2)}}, \tag{3·6}$$

where $$I(\mathbf{x}_s) = \hat{\Gamma}(\mathbf{x}_s,\mathbf{x}_s,0) = \langle\hat{V}(\mathbf{x}_s,t)\,\hat{V}^*(\mathbf{x}_s,t)\rangle \quad (s=1,2) \tag{3·7}$$

is the intensity at the point $P_s(\mathbf{x}_s)$. On applying the modulus inequality for integrals and the Schwarz inequality to (3·4), one has, by a similar argument as in I, § 3,

$$|\hat{\gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau)| \leqslant 1. \tag{3·8}$$

4. An approximate propagation law for $\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau)$ and an extension of the generalized Huygens principle to fields with spectral range of arbitrary width

As in I we consider the propagation of a beam of light from a finite source Σ. The field will again be assumed to be stationary, but no restriction on the width of the spectral range will now be imposed.

† In I, the time average was taken over the finite range $-T\leqslant t\leqslant T$, but it is more convenient mathematically (although it is not more significant physically) to allow the range to become infinite by proceeding to the limit $T\to\infty$, as customary. We shall now understand the time average in this sense:

$$\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\hat{V}(\mathbf{x}_1,t+\tau)\,\hat{V}^*(\mathbf{x}_2,t)\,\mathrm{d}t.$$

The integration over the frequency range was taken in I from $-\infty$ to $+\infty$, but as explained in the preceding section we may set $v(\mathbf{x},\nu)=0$ for $\nu<0$ and hence integrate over the positive range $0\leqslant\nu<\infty$ only. The quantities which were denoted by $V(\mathbf{x},t)$, $\Gamma(\mathbf{x}_1,\mathbf{x}_2)$ and $\gamma(\mathbf{x}_1,\mathbf{x}_2)$ in I would in the present notation be therefore written as $\hat{V}(\mathbf{x},t)$, $\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2)$ and $\hat{\gamma}(\mathbf{x}_1,\mathbf{x}_2)$.

Let $\mathscr{A}$ be a surface cutting across the beam (see figure 1) and let $P_1(\mathbf{x}_1)$ and $P_2(\mathbf{x}_2)$ be two points on that side of $\mathscr{A}$ towards which the light is advancing. We shall derive an expression for the correlation $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$ between the disturbances at P_1 and P_2, in terms of the values which $\hat{\Gamma}$ takes at all pairs of points on the surface $\mathscr{A}$.

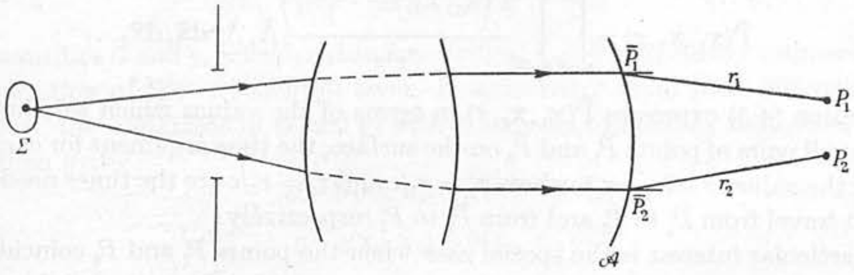


FIGURE 1. Derivation of a propagation law for $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$.

Each Fourier term $v(\mathbf{x}, \nu)\exp\{-\mathrm{i}\,2\pi\nu t\}$ of $\hat{V}$ represents a perfectly monochromatic wave. Hence the value of v at P_1 may be expressed (under the usual restrictions) in terms of the values which v takes at all points on the surface $\mathscr{A}$, by means of the ordinary Huygens principle:

$$v(\mathbf{x}_1, \nu) = \int_{\mathscr{A}} v(\bar{\mathbf{x}}_1, \nu)\frac{\mathrm{e}^{\mathrm{i}kr_1}}{r_1}\Lambda_1 \mathrm{d}\bar{\mathbf{x}}_1. \tag{4·1}$$

Here $\bar{\mathbf{x}}_1$ is the position vector at a typical point $\bar{P}_1$ on the surface $\mathscr{A}$, r_1 is the distance from $\bar{P}_1$ to P_1 (see figure 1), $k = 2\pi\nu/c = 2\pi/\lambda$, c being the vacuum velocity of light and λ the wave-length. Λ denotes as before the usual inclination factor (see I, (2·4)).

In a similar way, the values $v(\mathbf{x}_2, \nu)$ at P_2 may be expressed in the form

$$v(\mathbf{x}_2, \nu) = \int_{\mathscr{A}} v(\bar{\mathbf{x}}_2, \nu)\frac{\mathrm{e}^{\mathrm{i}kr_2}}{r_2}\Lambda_2 \mathrm{d}\bar{\mathbf{x}}_2. \tag{4·2}$$

From (4·1) and (4·2),

$$v(\mathbf{x}_1, \nu)\, v^*(\mathbf{x}_2, \nu) = \int_{\mathscr{A}}\int_{\mathscr{A}} v(\bar{\mathbf{x}}_1, \nu)\, v^*(\bar{\mathbf{x}}_2, \nu)\frac{\mathrm{e}^{\mathrm{i}k(r_1-r_2)}}{r_1 r_2}\Lambda_1\Lambda_2^* \mathrm{d}\bar{\mathbf{x}}_1 \mathrm{d}\bar{\mathbf{x}}_2, \tag{4·3}$$

the points $\bar{P}_1(\bar{\mathbf{x}}_1)$ and $\bar{P}_2(\bar{\mathbf{x}}_2)$ exploring the surface $\mathscr{A}$ independently. Next we multiply (4·3) by $\frac{1}{2T}\exp\{-2\pi\mathrm{i}\nu\tau\}$, integrate over the frequency range and proceed to the limit $T \to \infty$. We then obtain the following expression for $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$:

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \int_{\mathscr{A}}\int_{\mathscr{A}}\left\{\lim_{T\to\infty}\frac{1}{2T}\int_0^\infty v(\bar{\mathbf{x}}_1, \nu)\, v^*(\bar{\mathbf{x}}_2, \nu)\exp\left\{-2\pi\mathrm{i}\nu\left[\tau - \frac{r_1 - r_2}{c}\right]\right\}\Lambda_1\Lambda_2^*\mathrm{d}\nu\right\}\frac{1}{r_1}\frac{1}{r_2}\mathrm{d}\bar{\mathbf{x}}_1 \mathrm{d}\bar{\mathbf{x}}_2. \tag{4·4}$$

Now in the integral over the frequency range, the factors Λ_1 and Λ_2^* are well-behaved functions, depending on the frequency only through a multiplicate factor ν.

We shall take Λ_1 and Λ_2^* outside the ν integration in mean values† (denoted by $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2^*$). The remaining part of the frequency integral gives, in the limit $T \to \infty$, precisely $\hat{\Gamma}\left(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \tau - \frac{r_1 - r_2}{c}\right)$. Hence (4·4) reduces to the relatively simple law

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \int_{\mathscr{A}} \int_{\mathscr{A}} \frac{\hat{\Gamma}\left(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \tau - \frac{r_1 - r_2}{c}\right)}{r_1 r_2} \tilde{\Lambda}_1 \tilde{\Lambda}_2^* \, d\bar{\mathbf{x}}_1 \, d\bar{\mathbf{x}}_2. \tag{4·5}$$

Equation (4·5) expresses $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$ in terms of the values which this function takes at all pairs of points $\bar{P}_1$ and $\bar{P}_2$ on the surface, the time argument for each pair having the values $\tau - (\tau_1 - \tau_2)$, where $\tau_1 = r_1/c$ and $\tau_2 = r_2/c$ are the times needed for light to travel from $\bar{P}_1$ to P_1 and from $\bar{P}_2$ to P_2 respectively.

Of particular interest is the special case when the points P_1 and P_2 coincide and when, in addition, $\tau = 0$. Denoting the common point by $P(\mathbf{x})$, the left-hand side of (4·5) reduces to the intensity $I(\mathbf{x})$ at P, and, if (3·6) is also used, one obtains

$$I(\mathbf{x}) = \int_{\mathscr{A}} \int_{\mathscr{A}} \frac{\sqrt{I(\bar{\mathbf{x}}_1)}\sqrt{I(\bar{\mathbf{x}}_2)}}{r_1 r_2} \hat{\gamma}\left(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \frac{r_2 - r_1}{c}\right) \tilde{\Lambda}_1 \tilde{\Lambda}_2^* \, d\bar{\mathbf{x}}_1 \, d\bar{\mathbf{x}}_2, \tag{4·6}$$

$I(\bar{\mathbf{x}}_1)$ and $I(\bar{\mathbf{x}}_2)$ being the intensities at two typical points $\bar{P}_1$ and $\bar{P}_2$ of the surface $\mathscr{A}$.

In (4·6) $\hat{\gamma}$ may be replaced by $\gamma = \mathscr{R}\hat{\gamma}$, since the imaginary part of $\hat{\gamma}$ contributes nothing to the integral, as I and $\tilde{\Lambda}_1 \tilde{\Lambda}_2^*$ are real. That the integral is actually real may be shown formally by verifying that it remains unchanged when its complex conjugate is taken. This result follows immediately on using (3·5) and interchanging the independent variables $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$. In place of (4·6) we may therefore write

$$I(\mathbf{x}) = \int_{\mathscr{A}} \int_{\mathscr{A}} \frac{\sqrt{I(\bar{\mathbf{x}}_1)}\sqrt{I(\bar{\mathbf{x}}_2)}}{r_1 r_2} \gamma\left(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \frac{r_2 - r_1}{c}\right) \tilde{\Lambda}_1 \tilde{\Lambda}_2^* \, d\bar{\mathbf{x}}_1 \, d\bar{\mathbf{x}}_2. \tag{4·6a}$$

Equation (4·6a) (or (4·6)) may be regarded as a *generalized Huygens principle for stationary fields of an arbitrary spectral range*. It expresses the *intensity* at the point $P(\mathbf{x})$ in terms of the intensity distribution over an arbitrary surface $\mathscr{A}$, the contribution from each pair of elements of the surface being weighed by the appropriate value of the correlation factor $\gamma(\mathbf{x}_1, \mathbf{x}_2, \tau)$. This factor, which is a generalization of the degree of coherence of Zernike, may, like the latter, be determined from experiments (cf. § 5 below). It may also be calculated from the knowledge of an (observable) correlation function of the source and of the transmission properties of the medium (cf. § 6 below). Hence our extended formulation of the generalized Huygens principle involves again observable quantities only.

† This step of the analysis, though formally correct, is somewhat unsatisfactory, since the mean values will depend not only on the geometrical situation, but also on the form of v, as function of frequency. It should, however, be borne in mind that the inclination factor Λ of the ordinary Huygens principle represents only a rough approximation, and is, in most practical cases, simply replaced by a constant. The difficulty disappears in the rigorous formulation given in § 8 below.

In order to see more clearly the connexion between the formula just derived, and some of the formulae given previously, consider the special case when the light is practically monochromatic, of frequency ν_0. If we express $\hat{\Gamma}$ and $\hat{\gamma}$ in the form

$$\left.\begin{aligned}\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) &= G(\mathbf{x}_1, \mathbf{x}_2, \tau)\,\mathrm{e}^{-2\pi \mathrm{i}\nu_0\tau},\\ \hat{\gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) &= g(\mathbf{x}_1, \mathbf{x}_2, \tau)\,\mathrm{e}^{-2\pi \mathrm{i}\nu_0\tau},\end{aligned}\right\} \tag{4·7}$$

the quantities G and g, considered as functions of τ, will vary slowly compared with the variation of the exponential term. If sufficiently small path differences are involved, the variations of G and g, with τ, may be completely neglected, i.e. we may then write

$$\left.\begin{aligned}G &\sim G(\mathbf{x}_1, \mathbf{x}_2) \sim \hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, 0),\\ g &\sim g(\mathbf{x}_1, \mathbf{x}_2) \sim \hat{\gamma}(\mathbf{x}_1, \mathbf{x}_2, 0).\end{aligned}\right\} \tag{4·8}$$

G is essentially Zernike's mutual intensity (denoted by $\Gamma(\mathbf{x}_1, \mathbf{x}_2)$ in I) and g his complex degree of coherence (denoted by $\gamma(\mathbf{x}_1,\mathbf{x}_2)$ in I). With this substitution, (4·5) reduces to Zernike's propagation law (Zernike 1938, equation (9)).†

$$G(\mathbf{x}_1, \mathbf{x}_2) \sim \int_{\mathscr{A}}\int_{\mathscr{A}} \frac{G(\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2)}{r_1 r_2}\,\mathrm{e}^{\mathrm{i}k^{(0)}(r_1 - r_2)}\,\Lambda_1^{(0)}\Lambda_2^{(0)*}\,\mathrm{d}\overline{\mathbf{x}}_1\,\mathrm{d}\overline{\mathbf{x}}_2, \tag{4·9}$$

where $k^{(0)} = 2\pi\nu_0/c$, $\Lambda^{(0)} = \Lambda(\nu_0)$; and with the same approximation (4·6) reduces to the more restricted formulation of the generalized Huygens principle given in I.

5. The generalized interference law and determination of the correlation function from experiments

5·1. *The case* $\mathbf{x}_1 \neq \mathbf{x}_2$

In order to determine the correlation functions from experiment we may use a procedure similar to that described in I in connexion with the determination of the less general correlation functions of Zernike.

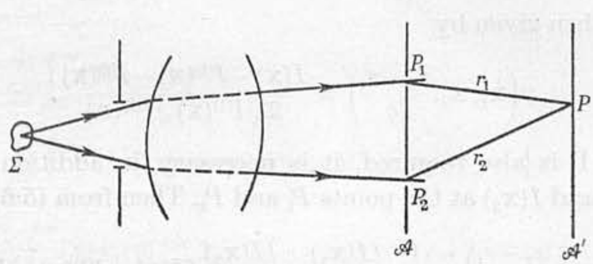


Figure 2. Experimental determination of $\Gamma(\mathbf{x}_1, \mathbf{x}_2, \tau)$.

We place a screen $\mathscr{A}$ across the field so as to pass through the points $P_1(\mathbf{x}_1)$ and $P_2(\mathbf{x}_2)$, small openings $\mathrm{d}\mathscr{A}_1$ and $\mathrm{d}\mathscr{A}_2$ being made at these points. The resulting intensity distribution is observed on a second screen $\mathscr{A}'$ (figure 2).

† Zernike actually neglected the variation of the inclination factor over the surface of integration, setting as usual, $\Lambda = \mathrm{i}/\lambda$.

Taking the integrals (4·6a) over the non-illuminated side of the screen $\mathscr{A}$, we obtain for the intensity $I(\mathbf{x})$ at the point $P(\mathbf{x})$ the expression

$$\begin{aligned}I(\mathbf{x}) \sim \frac{I(\mathbf{x}_1)}{r_1^2}\,|\,\tilde{\Lambda}_1\,|^2\,(\mathrm{d}\mathscr{A}_1)^2 + \frac{I(\mathbf{x}_2)}{r_2^2}\,|\,\tilde{\Lambda}_2\,|^2\,(\mathrm{d}\mathscr{A}_2)^2 \\ + \frac{\sqrt{I(\mathbf{x}_1)}\sqrt{I(\mathbf{x}_2)}}{r_1 r_2}\gamma\left(\mathbf{x}_1, \mathbf{x}_2, \frac{r_2-r_1}{c}\right)\tilde{\Lambda}_1\tilde{\Lambda}_2^*\,\mathrm{d}\mathscr{A}_1\,\mathrm{d}\mathscr{A}_2 \\ + \frac{\sqrt{I(\mathbf{x}_2)}\sqrt{I(\mathbf{x}_1)}}{r_2 r_1}\gamma\left(\mathbf{x}_2, \mathbf{x}_1\frac{r_1-r_2}{c}\right)\tilde{\Lambda}_2\tilde{\Lambda}_1^*\,\mathrm{d}\mathscr{A}_2\,\mathrm{d}\mathscr{A}_1. \quad (5{\cdot}1)\end{aligned}$$

Now the first term on the right is precisely the intensity $I^{(1)}(\mathbf{x})$ which would be obtained at P if the openings at P_1 alone was open ($\mathrm{d}\mathscr{A}_2 = 0$):

$$I^{(1)}(\mathbf{x}) = \frac{I(\mathbf{x}_1)}{r_1^2}\,|\,\tilde{\Lambda}_1\,|^2\,(\mathrm{d}\mathscr{A}_1)^2, \quad (5{\cdot}2)$$

the second term
$$I^{(2)}(\mathbf{x}) = \frac{I(\mathbf{x}_2)}{r_2^2}\,|\,\tilde{\Lambda}_2\,|^2\,(\mathrm{d}\mathscr{A}_2)^2 \quad (5{\cdot}3)$$

having a similar interpretation. Also, on account of (3·5) and remembering that γ is the real part of $\hat{\gamma}$,

$$\gamma\left(\mathbf{x}_2, \mathbf{x}_1, \frac{r_1-r_2}{c}\right) = \gamma\left(\mathbf{x}_1, \mathbf{x}_2, \frac{r_2-r_1}{c}\right). \quad (5{\cdot}4)$$

Further, since $\tilde{\Lambda}$ is purely imaginary, $\tilde{\Lambda}_2\tilde{\Lambda}_1^* = \tilde{\Lambda}_1\tilde{\Lambda}_2^* = |\,\tilde{\Lambda}_1\,|\,|\,\tilde{\Lambda}_2\,|$. Hence (5·1) reduces to

$$I(\mathbf{x}) = I^{(1)}(\mathbf{x}) + I^{(2)}(\mathbf{x}) + 2\sqrt{I^{(1)}(\mathbf{x})}\sqrt{I^{(2)}(\mathbf{x})}\gamma\left(\mathbf{x}_1, \mathbf{x}_2, \frac{r_2-r_1}{c}\right). \quad (5{\cdot}5)$$

(5·5) represents *a general interference law for stationary fields.*

It is seen that in order to determine $\gamma(\mathbf{x}_1, \mathbf{x}_2, \tau)$ it is only necessary to take the distances r_1 and r_2 such that $\frac{r_2-r_1}{c} = \tau$, and to measure the intensities $I(\mathbf{x})$, $I^{(1)}(\mathbf{x})$ and $I^{(2)}(\mathbf{x})$. γ is then given by

$$\gamma\left(\mathbf{x}_1, \mathbf{x}_2, \frac{r_2-r_1}{c}\right) = \frac{I(\mathbf{x}) - I^{(1)}(\mathbf{x}) - I^{(2)}(\mathbf{x})}{2\sqrt{I^{(1)}(\mathbf{x})}\sqrt{I^{(2)}(\mathbf{x})}}. \quad (5{\cdot}6)$$

If the value of Γ is also required, it is necessary, in addition, to measure the intensities $I(\mathbf{x}_1)$ and $I(\mathbf{x}_2)$ at the points P_1 and P_2. Then from (5·6) and (3·6),

$$\Gamma\left(\mathbf{x}_1, \mathbf{x}_2, \frac{r_2-r_1}{c}\right) = \frac{1}{2}\sqrt{\frac{I(\mathbf{x}_1)}{I^{(1)}(\mathbf{x})}}\sqrt{\frac{I(\mathbf{x}_2)}{I^{(2)}(\mathbf{x})}}\,[I(\mathbf{x}) - I^{(1)}(\mathbf{x}) - I^{(2)}(\mathbf{x})]. \quad (5{\cdot}7)$$

In the special case when the effective frequency range is sufficiently narrow (path differences small enough), (5·5) reduces, on substitution from (4·7), to the formula I, (3·9) of Zernike and Hopkins:

$$I(\mathbf{x}) = I^{(1)}(\mathbf{x}) + I^{(2)}(\mathbf{x}) + 2\sqrt{I^{(1)}(\mathbf{x})}\sqrt{I^{(2)}(\mathbf{x})}\,|\,g(\mathbf{x}_1, \mathbf{x}_2)\,|\cos[\arg g(\mathbf{x}_1, \mathbf{x}_2) + k^{(0)}(r_1 - r_2)]. \quad (5{\cdot}8)$$

5·2. *The case* $\mathbf{x}_1 = \mathbf{x}_2$

It has been assumed so far that the two points P_1 and P_2 are distinct. We must now consider the special case when they coincide.

When $P_1 = P_2$, (3·1) and (3·4) reduce to

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_1, \tau) = \langle \hat{V}(\mathbf{x}_1, t+\tau)\, \hat{V}^*(\mathbf{x}_1, t)\rangle = \lim_{T\to\infty} \frac{1}{2T}\int_0^\infty |\, v(\mathbf{x}_1, \nu)\,|^2\, \mathrm{e}^{-2\pi \mathrm{i}\nu\tau}\, \mathrm{d}\nu. \tag{5·9}$$

Hence $\hat{\Gamma}$ may now be determined by measuring the spectral intensity function $I(\mathbf{x}_1, \nu) = \lim_{T\to\infty} \frac{1}{2T} |\, v(\mathbf{x}_1, \nu)\,|^2$ at P_1 and evaluating the Fourier integral. (If, in addition to $\mathbf{x}_1 = \mathbf{x}_2$, one also has $\tau = 0$, $\hat{\Gamma}$ reduces, of course, to the intensity at P_1.)

A more direct determination of $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$ appears to be possible, in principle, by the use of some interferometer based on the principle of division of amplitude (cf. Williams 1941).

Suppose that the beam of light is divided at the point $P_1(\mathbf{x}_1)$ into two beams (for example, in the Michelson interferometer), which then proceed via different paths and are reunited again at a point $P(\mathbf{x})$. Let the transmission functions (cf. I, § 4) of the two paths for strictly monochromatic radiation of frequency ν be

$$K(\mathbf{x}_1, \mathbf{x}, \nu) = |\, K_1\,|\, \mathrm{e}^{2\pi \mathrm{i}\nu\phi_1}, \quad K(\mathbf{x}_2, \mathbf{x}, \nu) = |\, K_2\,|\, \mathrm{e}^{2\pi \mathrm{i}\nu\phi_2}. \tag{5·10}$$

Then the Fourier component $v(\mathbf{x}, \nu)$ at P is related to that at P_1 by

$$v(\mathbf{x}, \nu) = (K_1\Lambda_1 + K_2\Lambda_2)\, v(\mathbf{x}_1, \nu)\, \delta\mathscr{A}. \tag{5·11}$$

where $\delta\mathscr{A}$ is the element of area around P_1 which reflects and transmit the incident light. Hence the intensity at P is given by

$$\begin{aligned} I(\mathbf{x}) &= \lim_{T\to\infty} \frac{1}{2T}\int_0^\infty |\, v(\mathbf{x}, \nu)\,|^2\, \mathrm{d}\nu \\ &= I^{(1)}(\mathbf{x}) + I^{(2)}(\mathbf{x}) + \mathscr{J}(\mathbf{x}), \end{aligned} \tag{5·12}$$

where

$$\left.\begin{aligned} I^{(1)}(\mathbf{x}) &= \lim_{T\to\infty} \frac{1}{2T}\int_0^\infty |\, v(\mathbf{x}_1, \nu)\,|^2\, |\, K_1\,|^2\, |\, \Lambda_1\,|^2 (\delta\mathscr{A})^2\, \mathrm{d}\nu, \\ I^{(2)}(\mathbf{x}) &= \lim_{T\to\infty} \frac{1}{2T}\int_0^\infty |\, v(\mathbf{x}_1, \nu)\,|^2\, |\, K_2\,|^2\, |\, \Lambda_2\,|^2 (\delta\mathscr{A})^2\, \mathrm{d}\nu, \\ \mathscr{J}(\mathbf{x}) &= \lim_{T\to\infty} \frac{1}{T}\int_0^\infty |\, v(\mathbf{x}_1, \nu)\,|^2\, |\, K_1\,|\,|\, K_2\,|\,|\, \Lambda_1\,|\,|\, \Lambda_2\,|\, (\delta\mathscr{A})^2 \cos 2\pi\nu(\phi_2 - \phi_1)\, \mathrm{d}\nu. \end{aligned}\right\} \tag{5·13}$$

In general, $|\, K_1\,|$, $|\, K_2\,|$, ϕ_1 and ϕ_2 will depend on the frequency, since owing to the refraction at the individual elements of the interferometer, light of different frequencies will proceed along slightly different paths. In many cases, however, this effect will be negligible; $|\, K_1\,|$ and $|\, K_2\,|$ may then be taken outside the integrals in (5·13) and $\mathscr{J}$ becomes, apart from a multiplicative factor, the Fourier cosine transform of the spectral intensity function, this being equal to Γ according to

(5·9). The expression (5·12) for the intensity may then be expressed in a form strictly analogous to (5·5):

$$I(\mathbf{x}) = I^{(1)}(\mathbf{x}) + I^{(2)}(\mathbf{x}) + 2\sqrt{I^{(1)}(\mathbf{x})}\sqrt{I^{(2)}(\mathbf{x})}\,\gamma\left(\mathbf{x}_1, \mathbf{x}_1, \frac{\phi_2-\phi_1}{c}\right), \tag{5·14}$$

with

$$\left.\begin{aligned} I^{(1)}(\mathbf{x}) &= |K_1|^2\,|\tilde{\Lambda}_1|^2\,(\delta\mathscr{A})^2\,I(\mathbf{x}_1),\\ I^{(2)}(\mathbf{x}) &= |K_2|^2\,|\tilde{\Lambda}_2|^2\,(\delta\mathscr{A})^2\,I(\mathbf{x}_1),\\ I(\mathbf{x}_1) &= \lim_{T\to\infty}\frac{1}{2T}\int_0^\infty |v(\mathbf{x}_1,\nu)|^2\,\mathrm{d}\nu = \langle \hat{V}(\mathbf{x}_1,t)\,\hat{V}^*(\mathbf{x}_1,t)\rangle. \end{aligned}\right\} \tag{5·15}$$

The term $I^{(1)}(\mathbf{x})$ represents the intensity which is obtained at the point $P(\mathbf{x})$ if the second beam is excluded ($K_2 = 0$), the term $I^{(2)}(\mathbf{x})$ having a similar interpretation. Hence to measure $\gamma(\mathbf{x}_1, \mathbf{x}_1, \tau)$ it is only necessary (provided $|K_1|$, $|K_2|$, ϕ_1 and ϕ_2 may be treated as independent of ν) to set the interferometer so that $(\phi_2-\phi_1)/c = \tau$. γ is then given by a formula analogous to (5·6):

$$\gamma\left(\mathbf{x}_1, \mathbf{x}_1, \frac{\phi_2-\phi_1}{c}\right) = \frac{I(\mathbf{x}) - I^{(1)}(\mathbf{x}) - I^{(2)}(\mathbf{x})}{2\sqrt{I^{(1)}(\mathbf{x})}\sqrt{I^{(2)}(\mathbf{x})}}. \tag{5·16}$$

And Γ is given by

$$\Gamma\left(\mathbf{x}_1, \mathbf{x}_1, \frac{\phi_2-\phi_1}{c}\right) = I(\mathbf{x}_1)\,\gamma\left(\mathbf{x}_1, \mathbf{x}_1, \frac{\phi_2-\phi_1}{c}\right). \tag{5·17}$$

The investigations of Zernike (1938, 1948; see also I, § 3) brought out the close connexion which exists between the visibility factor of Michelson and the *correlation* (characterized by $\hat{\gamma}(\mathbf{x}_1, \mathbf{x}_2, 0)$) of disturbances *at two points in the field, at the same instant of time*. These investigations interpreted, also from a new point of view, Michelson's method for the determination of the intensity of radiation across a radiating source from the measurements of the visibility of fringes, a method which in recent years has become of fundamental importance in radio astronomy (cf. Ryle 1950; Smith 1952). With the help of the preceding analysis, it will now be shown that there is a complementary relation between the visibility function and the *correlation* (characterized by $\hat{\gamma}(\mathbf{x}_1, \mathbf{x}_1, \tau)$) of disturbances *at two instants of time, at the same point in the field*. This result leads to a new interpretation of Michelson's well-known method (Michelson 1891, 1892) for the determination of the energy distribution in spectral lines from measurements of the visibility.†

Let ν_0 be the mean frequency of the spectral distribution, assumed now to be confined within a narrow frequency range $\nu_0 - \Delta\nu \leqslant \nu \leqslant \nu_0 + \Delta\nu$ and set

$$\hat{\gamma}(\mathbf{x}_1, \mathbf{x}_1, \tau) = h(\mathbf{x}_1, \tau)\,\mathrm{e}^{-2\pi i \nu_0 \tau}. \tag{5·18}$$

Further, assume, as is usually the case, that $I^{(1)}(\mathbf{x}) \sim I^{(2)}(\mathbf{x})$. Equation (5·14) then becomes

$$I(\mathbf{x}) = 2I^{(1)}(\mathbf{x})\,\{1 + |h(\mathbf{x}_1,\tau)|\cos[\arg h(\mathbf{x}_1,\tau) - 2\pi\nu_0\tau]\}. \tag{5·19}$$

Since $\Delta\nu/\nu_0 \ll 1$, h, considered as a function of τ, will change very slowly in comparison with $\cos 2\pi\nu_0\tau$ and $\sin 2\pi\nu_0\tau$, so that the minima and maxima of the intensity are effectively given by τ values which satisfy the relation

$$\sin[\arg h(\mathbf{x}_1,\tau) - 2\pi\nu_0\tau] = 0. \tag{5·20}$$

† In this connexion see also the paper by Van Cittert (1939).

The corresponding maxima and minima have the values

$$\left.\begin{aligned} I_{\max.} &= 2I^{(1)}(\mathbf{x})\,[1+|\,h(\mathbf{x}_1,\tau)\,|],\\ I_{\min.} &= 2I^{(1)}(\mathbf{x})\,[1-|\,h(\mathbf{x}_1,\tau)\,|].\end{aligned}\right\} \tag{5·21}$$

The visibility $\mathscr{V}$ of the fringes is therefore

$$\mathscr{V} = \frac{I_{\max.}-I_{\min.}}{I_{\max.}+I_{\min.}} = |\,h(\mathbf{x}_1,\tau)\,| = |\,\hat{\gamma}(\mathbf{x}_1,\mathbf{x}_1,\tau)\,|, \tag{5·22}$$

i.e. the visibility is equal to the absolute value of the complex correlation factor $\hat{\gamma}(\mathbf{x}_1,\mathbf{x}_1,\tau)$. Hence according to (5·9), using the Fourier inversion formula, it is possible to obtain information about the intensity distribution in the spectrum from measurements of the visibility, provided that suitable assumptions about the associated phase $[\arg\hat{\gamma}(\mathbf{x}_1,\mathbf{x}_1,\tau)]$ are made.†

It is now seen that the two methods of Michelson correspond essentially to the two limiting cases $\tau\to 0$ and $\mathbf{x}_2\to\mathbf{x}_1$ of our theory.

6. Expressions for $\hat{\Gamma}$ in terms of quantities which specify the source and the medium

We shall now derive explicit expressions for the $\hat{\Gamma}$ factor in terms of quantities which specify the source and the transmission properties of the medium.

As in I we assume the source Σ to be a radiating plane area and divide it into elements $\delta\Sigma_1, \delta\Sigma_2, \ldots$, which are small in linear dimensions in comparison with the optical wave-lengths. Let

$$\hat{V}_m(\mathbf{x},t) = \int_0^\infty v_m(\mathbf{x},\nu)\,\mathrm{e}^{-2\pi\mathrm{i}\nu t}\,\mathrm{d}\nu \tag{6·1}$$

be the disturbance due to the mth element; the total disturbance $\hat{V}(\mathbf{x},t)$ is then given by

$$\hat{V}(\mathbf{x},t) = \sum_m \hat{V}_m(\mathbf{x},t). \tag{6·2}$$

Hence

$$\begin{aligned}\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau) &= \langle \hat{V}(\mathbf{x}_1,t+\tau)\,\hat{V}^*(\mathbf{x}_2,t)\rangle\\ &= \sum_m\sum_n \hat{\Gamma}_{mn}(\mathbf{x}_1,\mathbf{x}_2,\tau),\end{aligned} \tag{6·3}$$

where

$$\begin{aligned}\hat{\Gamma}_{mn}(\mathbf{x}_1,\mathbf{x}_2,\tau) &= \langle \hat{V}_m(\mathbf{x}_1,t+\tau)\,\hat{V}^*_n(\mathbf{x}_2,t)\rangle\\ &= \lim_{T\to\infty}\frac{1}{2T}\int_0^\infty v_m(\mathbf{x}_1,\nu)\,v^*_n(\mathbf{x}_2,\nu)\,\mathrm{e}^{-2\pi\mathrm{i}\nu\tau}\,\mathrm{d}\nu.\end{aligned} \tag{6·4}$$

In most cases of practical interest (e.g. for a gas discharge or incandescent solid) it will be permissible to assume that the radiation from the different elements of the source is mutually incoherent, i.e. that for all values of $\mathbf{x}_1$, $\mathbf{x}_2$ and τ

$$\hat{\Gamma}_{mn}(\mathbf{x}_1,\mathbf{x}_2,\tau) = 0 \quad \text{when} \quad m\neq n. \tag{6·5}$$

† As is evident from (5·19), the phase may in principle be obtained from the measurement of the position of the fringes. This has been pointed out already by Rayleigh (1892) in an open letter to Michelson, in which he discussed the question of a complete determination of the intensity distribution from Michelson's experiments.

There are, however, important cases when this assumption does not hold. For example, if the source is not a 'natural' source, but is a secondary source obtained by imagining a source of natural light by a lens of a finite aperture, then on account of diffraction there will exist a finite degree of correlation in the plane of the secondary source at points which are sufficiently close to each other. Accordingly, we shall first consider the general case when $\hat{\Gamma}_{mn} \neq 0$.

Let $a_m(\nu)$ be the strength of the radiation from the element $\delta\Sigma_m$, at frequency ν, and let $K(\mathbf{x}', \mathbf{x}'', \nu)$ be the transmission function of the medium. Then†

$$v_m(\mathbf{x}, \nu) = a_m(\nu)\, K(\boldsymbol{\xi}_m, \mathbf{x}, \nu), \tag{6·6}$$

where $\boldsymbol{\xi}_m$ denotes the position vector of the mth element of the source. (6·3) then becomes

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \sum_m \sum_n \int_0^\infty J_{mn}(\nu) L(\boldsymbol{\xi}_m, \boldsymbol{\xi}_n; \mathbf{x}_1, \mathbf{x}_2; \nu)\, \mathrm{e}^{-2\pi i\nu\tau}\, \mathrm{d}\nu, \tag{6·7}$$

with

$$\left.\begin{aligned} J_{mn}(\nu) &= \lim_{T\to\infty} \frac{1}{2T} [a_m(\nu)\, a_n^*(\nu)], \\ L(\boldsymbol{\xi}_m, \boldsymbol{\xi}_n; \mathbf{x}_1, \mathbf{x}_2; \nu) &= K(\boldsymbol{\xi}_m, \mathbf{x}_1, \nu)\, K^*(\boldsymbol{\xi}_n, \mathbf{x}_2, \nu). \end{aligned}\right\} \tag{6·8}$$

and

Let

$$\int_0^\infty J_{mn}(\nu)\, \mathrm{e}^{-2\pi i\nu u}\, \mathrm{d}\nu = \hat{\Gamma}_{mn}(u). \tag{6·9}$$

We also introduce the frequency transform of L:

$$\int_0^\infty L(\boldsymbol{\xi}_m, \boldsymbol{\xi}_n; \mathbf{x}_1, \mathbf{x}_2; \nu)\, \mathrm{e}^{-2\pi i\nu u}\, \mathrm{d}\nu = M(\boldsymbol{\xi}_m, \boldsymbol{\xi}_n; \mathbf{x}_1, \mathbf{x}_2; u). \tag{6·10}$$

The relation (6·7) then becomes, on using the convolution theorem,

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \sum_m \sum_n \int_{-\infty}^{+\infty} \hat{\Gamma}_{mn}(u)\, M(\boldsymbol{\xi}_m, \boldsymbol{\xi}_n; \mathbf{x}_1, \mathbf{x}_2; \tau - u)\, \mathrm{d}u. \tag{6·11}$$

(6·11) expresses $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$ in terms of $\hat{\Gamma}_{mn}(u)$ and M. The former specifies the source and may be determined from experiments described in § 5. The latter specifies the medium and may be obtained from calculations based on a ray trace.

For incoherent sources (i.e. sources for which (6·5) holds), the double summation in (6·11) reduces to a single summation:

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \sum_m \int_0^\infty J_{mn}(\nu)\, L(\boldsymbol{\xi}_m, \boldsymbol{\xi}_n; \mathbf{x}_1, \mathbf{x}_2; \nu)\, \mathrm{e}^{-2\pi i\nu\tau}\, \mathrm{d}\nu \tag{6·12}$$

$$= \sum_m \int_{-\infty}^{+\infty} \hat{\Gamma}_{mn}(u)\, M(\boldsymbol{\xi}_m, \boldsymbol{\xi}_n; \mathbf{x}_1, \mathbf{x}_2; \tau - u)\, \mathrm{d}u. \tag{6·13}$$

If the elements $\delta\Sigma_1, \delta\Sigma_2, \ldots$ are taken small enough, one may replace the summations by integrations over the source, provided that obvious modifications are made: One introduces in place of $\hat{\Gamma}_{mn}$ and J_{mn} which are defined only for the discrete set of $\boldsymbol{\xi}$ values, the functions $\hat{\Omega}$ and j defined for a continuous range of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, such that

$$\hat{\Gamma}_{mn}(u) = \hat{\Omega}(\boldsymbol{\xi}, \boldsymbol{\xi}', u)\, \mathrm{d}\Sigma\, \mathrm{d}\Sigma', \tag{6·14}$$

$$J_{mn}(\nu) = j(\boldsymbol{\xi}, \boldsymbol{\xi}', \nu)\, \mathrm{d}\Sigma\, \mathrm{d}\Sigma', \tag{6·15}$$

† We neglect here the variation of the source strength with direction.

it being assumed that $m \neq n$. $\hat{\Omega}$ and j are, on account of (6·9), Fourier transforms of each other. In place of (6·7) and (6·11) one then obtains

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \int_\Sigma \int_\Sigma \mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\boldsymbol{\xi}' \int_0^\infty j(\boldsymbol{\xi}, \boldsymbol{\xi}', \nu)\, L(\boldsymbol{\xi}, \boldsymbol{\xi}'; \mathbf{x}_1, \mathbf{x}_2; \nu)\, \mathrm{e}^{-2\pi\mathrm{i}\nu\tau}\,\mathrm{d}\nu \tag{6·16}$$

$$= \int_\Sigma \int_\Sigma \mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\boldsymbol{\xi}' \int_{-\infty}^{+\infty} \hat{\Omega}(\boldsymbol{\xi}, \boldsymbol{\xi}', u)\, M(\boldsymbol{\xi}, \boldsymbol{\xi}'; \mathbf{x}_1, \mathbf{x}_2; \tau - u)\,\mathrm{d}u, \tag{6·17}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ explore the surface Σ of the source independently.

In the case of an incoherent source, we set

$$\hat{\Gamma}_{mm}(u) = \hat{\Omega}(\boldsymbol{\xi}, u)\,\mathrm{d}\Sigma, \tag{6·18}$$

$$J_{mm}(u) = j(\boldsymbol{\xi}, \nu)\,\mathrm{d}\Sigma. \tag{6·19}$$

$j(\boldsymbol{\xi}, \nu)$ is nothing but the spectral intensity function of the source; it represents the intensity per unit area of the source per unit frequency range. In place of (6·12) and (6·13) one then obtains

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \int_\Sigma \mathrm{d}\boldsymbol{\xi} \int_0^\infty j(\boldsymbol{\xi}, \nu)\, L(\boldsymbol{\xi}, \boldsymbol{\xi}; \mathbf{x}_1, \mathbf{x}_2; \nu)\mathrm{e}^{-2\pi\mathrm{i}\nu\tau}\,\mathrm{d}\nu \tag{6·20}$$

$$= \int_\Sigma \mathrm{d}\boldsymbol{\xi} \int_{-\infty}^{+\infty} \hat{\Omega}(\boldsymbol{\xi}, u)\, M(\boldsymbol{\xi}, \boldsymbol{\xi}; \mathbf{x}_1, \mathbf{x}_2; \tau - u)\,\mathrm{d}u. \tag{6·21}$$

As an example, consider the case of an incoherent source in a homogeneous medium. The transmission function of a homogeneous medium is

$$K(\boldsymbol{\xi}, \mathbf{x}, \nu) = \frac{\exp\left\{2\pi\mathrm{i}\frac{\nu}{c}|\mathbf{x} - \boldsymbol{\xi}|\right\}}{|\mathbf{x} - \boldsymbol{\xi}|}, \tag{6·22}$$

so that

$$L(\boldsymbol{\xi}, \boldsymbol{\xi}; \mathbf{x}_1, \mathbf{x}_2, \nu) = \frac{\exp\left\{2\pi\mathrm{i}\frac{\nu}{c}(R_1 - R_2)\right\}}{R_1 R_2}, \tag{6·23}$$

with

$$R_1 = |\mathbf{x}_1 - \boldsymbol{\xi}|, \quad R_2 = |\mathbf{x}_2 - \boldsymbol{\xi}|. \tag{6·24}$$

(6·20) then gives the following expression for $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$:

$$\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \int_\Sigma \mathrm{d}\boldsymbol{\xi} \int_0^\infty \frac{j(\boldsymbol{\xi}, \nu)}{R_1 R_2} \exp\left\{-2\pi\mathrm{i}\nu\left[\tau - \frac{R_1 - R_2}{c}\right]\right\}\mathrm{d}\nu \tag{6·25}$$

$$= \int_\Sigma \frac{\hat{\Omega}\left(\boldsymbol{\xi}, \tau - \dfrac{R_1 - R_2}{c}\right)}{R_1 R_2}\,\mathrm{d}\boldsymbol{\xi}. \tag{6·26}$$

7. Differential equations for the correlation function $\hat{\Gamma}$

In its usual form, Huygens's principle describes within a certain degree of accuracy the propagation of the light disturbance $\hat{V}(\mathbf{x}, t)$. As is well known, this principle may be regarded as an approximate formulation of a rigorous theorem due to Kirchhoff (see, for example, Baker & Copson 1950), this theorem being a consequence of the fact that $\hat{V}$ obeys rigorously the wave equation.

In the present investigation we have found a generalization of the Huygens principle which applies to the intensity rather than to the complex disturbance, and, more generally, we have found a kind of Huygens principle for the propagation of the correlation function $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$. These results suggest that $\hat{\Gamma}$ itself obeys certain differential equations and that our propagation laws are essentially some approximate formulations of the associated 'Kirchhoff's theorems'. We shall now show that this indeed is the case.

In vacuo, the complex disturbance $\hat{V}(\mathbf{x}, t)$ satisfies the wave equation

$$\nabla^2 \hat{V} - \frac{1}{c^2}\frac{\partial^2 \hat{V}}{\partial t^2} = 0. \tag{7·1}$$

Consequently each Fourier component $v(\mathbf{x}, \nu)$ obeys the equation

$$\nabla^2 v + \left(\frac{2\pi\nu}{c}\right)^2 v = 0. \tag{7·2}$$

Let

$$\nabla_1^2 \equiv \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2} \tag{7·3}$$

be the Laplacian operator with respect to the co-ordinates x_1, y_1, z_1 of the point $P_1(\mathbf{x}_1)$. It then follows from (3·4) and (7·2) that

$$\begin{aligned}\nabla_1^2 \hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) &= \lim_{T\to\infty} \frac{1}{2T}\int_0^{+\infty} [\nabla_1^2 v(\mathbf{x}_1, \nu)]\, v^*(\mathbf{x}_2, \nu)\, \mathrm{e}^{-2\pi i\nu\tau}\, \mathrm{d}\nu \\ &= -\frac{2\pi^2}{c^2}\lim_{T\to\infty}\frac{1}{T}\int_0^{\infty} \nu^2 v(\mathbf{x}_1, \nu)\, v^*(\mathbf{x}_2, \nu)\, \mathrm{e}^{-2\pi i\nu\tau}\, \mathrm{d}\nu. \end{aligned} \tag{7·4}$$

Also from (3·4),

$$\frac{\partial^2}{\partial \tau^2}\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau) = -2\pi^2 \lim_{T\to\infty}\frac{1}{T}\int_0^{\infty} \nu^2 v(\mathbf{x}_1, \nu)\, v^*(\mathbf{x}_2, \nu)\, \mathrm{e}^{-2\pi i\nu\tau}\, \mathrm{d}\nu. \tag{7·5}$$

Comparison of (7·4) and (7·5) shows that

$$\nabla_1^2 \hat{\Gamma} - \frac{1}{c^2}\frac{\partial^2 \hat{\Gamma}}{\partial \tau^2} = 0. \tag{7·6}$$

Similarly, if ∇_2^2 denotes the Laplacian operator with respect to the co-ordinates x_2, y_2, z_2 of the point $P_2(\mathbf{x}_2)$, then

$$\nabla_2^2 \hat{\Gamma} - \frac{1}{c^2}\frac{\partial^2 \hat{\Gamma}}{\partial \tau^2} = 0. \tag{7·7}$$

Hence, in vacuo, *the correlation function* $\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \tau)$ *obeys rigorously the two wave equations* (7·6) *and* (7·7).

Each of the two wave equations describes the variation of the correlation when one of the points (P_2 or P_1) is fixed whilst the other point as well as the parameter τ varies. It will be recalled that τ denotes a time difference; in all experiments it will play the part of the difference in the optical path (divided by c). The 'actual' time makes no appearance in our formulae. This is a most desirable aspect of the theory, since true time variations are not observed in optical fields.

8. A rigorous formulation of the propagation law for $\hat{\Gamma}$ and of the generalized Huygens principle

We are now in a position to derive a rigorous propagation law for $\hat{\Gamma}$ and also to formulate rigorously the generalized Huygens principle.

Let $P(\mathbf{x})$ and $A(\mathbf{a})$ be any two points in the field and let $\mathscr{A}$ be any closed surface surrounding P; the point A may be either inside or outside this surface.

If ∇^2 denotes the Laplacian operator with respect to the co-ordinates of $P(\mathbf{x})$, then, according to (7·6),

$$\nabla^2\hat{\Gamma}(\mathbf{x},\mathbf{a},\tau)-\frac{1}{c^2}\frac{\partial^2\hat{\Gamma}(\mathbf{x},\mathbf{a},\tau)}{\partial\tau^2}=0. \tag{8·1}$$

Hence, using Kirchhoff's integral formula (cf. Baker & Copson 1950, p. 37), we may express $\hat{\Gamma}(\mathbf{x},\mathbf{a},\tau)$ in the following form:

$$\hat{\Gamma}(\mathbf{x},\mathbf{a},\tau)=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f[\hat{\Gamma}]^-+g\left[\frac{\partial}{\partial\tau}\hat{\Gamma}\right]^-+h\left[\frac{\partial}{\partial n}\hat{\Gamma}\right]^-\right\}\mathrm{d}\bar{\mathbf{x}}, \tag{8·2}$$

where

$$f=\frac{\partial}{\partial n}\left(\frac{1}{r}\right),\quad g=-\frac{1}{cr}\frac{\partial r}{\partial n},\quad h=-\frac{1}{r}, \tag{8·3}$$

r being the distance from a typical point $\bar{P}(\bar{\mathbf{x}})$ on the surface to $P(\mathbf{x})$ (see figure 3a), $\partial/\partial n$ denoting differentiation along the inward normal to $\mathscr{A}$; and the brackets $[\dots]^-$ denote retarded values, i.e. values obtained by replacing τ by $\tau-r/c$, e.g.

$$[\hat{\Gamma}]^-=\hat{\Gamma}(\bar{\mathbf{x}},\mathbf{a},\tau-r/c). \tag{8·4}$$

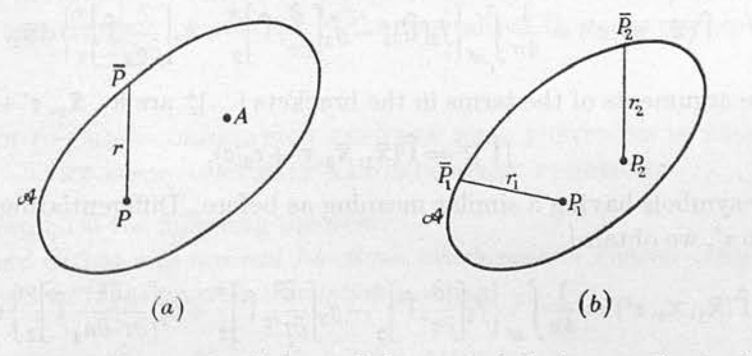


Figure 3. Illustrating Kirchhoff's integral theorem and the 'propagation law' for $\hat{\Gamma}$.

It may be shown by an argument similar to that used to derive Kirchhoff's formula that the integral (8·2) with advanced values ($\tau+r/c$ in place of $\tau-r/c$) also satisfies the wave equation, provided g is replaced by $-g$. If expressions for $\hat{V}$ in terms of retarded terms only are admitted, as it is reasonable to do on physical grounds, then Kirchhoff's integral for $\hat{\Gamma}(\mathbf{x},\mathbf{a},\tau)$ in terms of the retarded values of $\hat{\Gamma}$ clearly represents the physical solution. The situation is, however, different in the case of $\hat{\Gamma}(\mathbf{a},\mathbf{x},\tau)$. For if again only expressions in terms of retarded values are admitted for $\hat{V}$, Kirchhoff's integral for $\hat{\Gamma}(\mathbf{a},\mathbf{x},\tau)$ must involve terms like

$$\begin{aligned} f\langle\hat{V}(\mathbf{a},t+\tau)\,\hat{V}^*(\bar{\mathbf{x}},t-r/c)\rangle &= f\langle\hat{V}(\mathbf{a},t+\tau+r/c)\,\hat{V}^*(\bar{\mathbf{x}},t)\rangle \\ &= f\hat{\Gamma}(\mathbf{a},\bar{\mathbf{x}},\tau+r/c), \end{aligned} \tag{8·5}$$

i.e. it must involve 'advanced' values of $\hat{\Gamma}$. The full formula for $\hat{\Gamma}(\mathbf{a},\mathbf{x},\tau)$ therefore is

$$\Gamma(\mathbf{a},\mathbf{x},\tau)=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f[\hat{\Gamma}]^{+}-g\left[\frac{\partial}{\partial\tau}\hat{\Gamma}\right]^{+}+h\left[\frac{\partial}{\partial n}\hat{\Gamma}\right]^{+}\right\}\mathrm{d}\bar{\mathbf{x}}, \tag{8·6}$$

the brackets $[\ldots]^{+}$ denoting the advanced values, e.g.

$$[\hat{\Gamma}]^{+}=\hat{\Gamma}(\mathbf{a},\bar{\mathbf{x}},\tau+r/c). \tag{8·7}$$

We now apply (8·2) and (8·6) to derive an expression for $\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\tau)$ in terms of an integral which involves the values of $\hat{\Gamma}$ and its derivatives at all pairs of points $\bar{P}_1$, $\bar{P}_2$ on an arbitrary closed surface $\mathscr{A}$ surrounding P_1 and P_2 (see figure 3*b*).

We set $\mathbf{x}=\mathbf{x}_1$, $\mathbf{a}=\mathbf{x}_2$, $\bar{\mathbf{x}}=\bar{\mathbf{x}}_1$ in (8·2). This gives

$$\Gamma(\mathbf{x}_1,\mathbf{x}_2,\tau)=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f_1[\hat{\Gamma}]_1^{-}+g_1\left[\frac{\partial}{\partial\tau}\hat{\Gamma}\right]_1^{-}+h_1\left[\frac{\partial}{\partial n_1}\hat{\Gamma}\right]_1^{-}\right\}\mathrm{d}\bar{\mathbf{x}}_1, \tag{8·8}$$

where the arguments in the brackets $[\ldots]_1^{-}$ are $\bar{\mathbf{x}}_1$, $\mathbf{x}_2$, $\tau-r_1/c$, e.g.

$$[\hat{\Gamma}]_1^{-}=\hat{\Gamma}(\bar{\mathbf{x}}_1,\mathbf{x}_2,\tau-r_1/c); \tag{8·9}$$

r_1 being the distance from $\bar{P}_1$ to P_1, f_1, g_1 and h_1 the appropriate values of f, g and h $\partial/\partial n_1$ denoting differentiation at P_1 along the inward normal to $\mathscr{A}$.

Next we express each of the retarded terms in (8·8) in terms of Kirchhoff's integral over the surface. Setting $\mathbf{a}=\bar{\mathbf{x}}_1$, $\mathbf{x}=\mathbf{x}_2$, $\bar{\mathbf{x}}=\bar{\mathbf{x}}_2$ in (8·6) and writing τ' in place of τ where τ' is arbitrary for the present, we obtain

$$\hat{\Gamma}(\bar{\mathbf{x}}_1,\mathbf{x}_2,\tau')=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f_2[\hat{\Gamma}]_2^{+}-g_2\left[\frac{\partial}{\partial\tau'}\hat{\Gamma}\right]_2^{+}+h_2\left[\frac{\partial}{\partial n_2}\hat{\Gamma}\right]_2^{+}\right\}\mathrm{d}\bar{\mathbf{x}}_2, \tag{8·10}$$

where the arguments of the terms in the brackets $[\ldots]_2^{+}$ are $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, $\tau'+r_2/c$, e.g.

$$[\hat{\Gamma}]_2^{+}=\hat{\Gamma}(\bar{\mathbf{x}}_1,\bar{\mathbf{x}}_2,\tau'+r_2/c), \tag{8·11}$$

the other symbols having a similar meaning as before. Differentiating (8·10) with respect to τ', we obtain

$$\frac{\partial}{\partial\tau'}\hat{\Gamma}(\bar{\mathbf{x}}_1,\mathbf{x}_2,\tau')=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f_2\left[\frac{\partial}{\partial\tau'}\hat{\Gamma}\right]_2^{+}-g_2\left[\frac{\partial^2}{\partial\tau'^2}\hat{\Gamma}\right]_2^{+}+h_2\left[\frac{\partial^2}{\partial\tau'\,\partial n_2}\hat{\Gamma}\right]_2^{+}\right\}\mathrm{d}\bar{\mathbf{x}}_2. \tag{8·12}$$

Differentiation of (8·10) with respect to n_1 gives

$$\frac{\partial}{\partial n_1}\hat{\Gamma}(\bar{\mathbf{x}}_1,\mathbf{x}_2,\tau')=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f_2\left[\frac{\partial}{\partial n_1}\hat{\Gamma}\right]_2^{+}-g_2\left[\frac{\partial^2}{\partial n_1\,\partial\tau'}\hat{\Gamma}\right]_2^{+}+h_2\left[\frac{\partial^2}{\partial n_1\,\partial n_2}\hat{\Gamma}\right]_2^{+}\right\}\mathrm{d}\bar{\mathbf{x}}_2. \tag{8·13}$$

Setting $\tau'=\tau-r_1/c$ in (8·10) to (8·13) and substituting into (8·8), we finally obtain

$$\begin{aligned}\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2\,\tau)=\frac{1}{(4\pi)^2}\int_{\mathscr{A}}\int_{\mathscr{A}}\bigg\{&f_1f_2[\hat{\Gamma}]-f_1g_2\left[\frac{\partial}{\partial\tau}\hat{\Gamma}\right]+f_1h_2\left[\frac{\partial}{\partial n_2}\hat{\Gamma}\right]\\&+g_1f_2\left[\frac{\partial}{\partial\tau}\hat{\Gamma}\right]-g_1g_2\left[\frac{\partial^2}{\partial\tau^2}\hat{\Gamma}\right]+g_1h_2\left[\frac{\partial^2}{\partial\tau\,\partial n_2}\hat{\Gamma}\right]\\&+h_1f_2\left[\frac{\partial}{\partial n_1}\hat{\Gamma}\right]-h_1g_2\left[\frac{\partial^2}{\partial n_1\,\partial\tau}\hat{\Gamma}\right]+h_1h_2\left[\frac{\partial^2}{\partial n_1\,\partial n_2}\hat{\Gamma}\right]\bigg\}\,\mathrm{d}\bar{\mathbf{x}}_1\,\mathrm{d}\bar{\mathbf{x}}_2,\end{aligned} \tag{8·14}$$

where the arguments of the terms in the brackets are $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \tau-\frac{r_1-r_2}{c}$, e.g.

$$[\hat{\Gamma}] = \hat{\Gamma}\left(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \tau-\frac{r_1-r_2}{c}\right). \tag{8·15}$$

The formula (8·14) may be regarded as a rigorous formulation of the propagation law for $\hat{\Gamma}$. It expresses the correlation between disturbances at P_1 and P_2 in terms of the correlations and its derivatives at all pairs of points on an arbitrary closed surface surrounding P_1 and P_2.

In the special case when P_1 and P_2 coincide ($\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}$ (say)) and when in addition $\tau = 0$, (8·14) reduces to, when we also substitute from (3·6),

$$\begin{aligned} I(\mathbf{x}) = \frac{1}{(4\pi)^2}\int_{\mathscr{A}}\int_{\mathscr{A}} & \sqrt{\bar{I}_1}\sqrt{\bar{I}_2}\left\{f_1 f_2[\hat{\gamma}] + (f_2 g_1 - f_1 g_2)\left[\frac{\partial}{\partial\tau}\hat{\gamma}\right] - g_1 g_2\left[\frac{\partial^2}{\partial\tau^2}\hat{\gamma}\right]\right\} \\ & + \sqrt{\bar{I}_1}\left\{f_1 h_2 \frac{\partial}{\partial n_2}(\sqrt{\bar{I}_2}[\hat{\gamma}]) + g_1 h_2 \frac{\partial}{\partial n_2}\left(\sqrt{\bar{I}_2}\left[\frac{\partial}{\partial\tau}\hat{\gamma}\right]\right)\right\} \\ & + \sqrt{\bar{I}_2}\left\{f_2 h_1 \frac{\partial}{\partial n_1}(\sqrt{\bar{I}_1}[\hat{\gamma}]) - g_2 h_1 \frac{\partial}{\partial n_1}\left(\sqrt{\bar{I}_1}\left[\frac{\partial}{\partial\tau}\hat{\gamma}\right]\right)\right\} \\ & + h_1 h_2 \frac{\partial^2}{\partial n_1 \partial n_2}(\sqrt{\bar{I}_1}\sqrt{\bar{I}_2}[\hat{\gamma}])\,\mathrm{d}\bar{\mathbf{x}}_1\,\mathrm{d}\bar{\mathbf{x}}_2, \end{aligned} \tag{8·16}$$

where $\bar{I}_1 = I(\bar{\mathbf{x}}_1)$, $\bar{I}_2 = I(\bar{\mathbf{x}}_2)$ and $[\hat{\gamma}] = \hat{\gamma}\,(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, (r_2-r_1)/c)$. By a similar argument to that used in connexion with (4·6), $\hat{\gamma}$ may be replaced in (8·16) by γ. (8·16) may be regarded as a rigorous formulation of the generalized Huygens principle.

Appendix 1. Cross-correlation between real functions in terms of the associated half-range complex functions

We shall establish the following theorem:

Let $F(t)$ and $G(t)$ be any two real functions, which possess Fourier integral representations and let $P(\tau)$ be the cross correlation function

$$P(\tau) = 2\int_{-\infty}^{+\infty} F(t+\tau)\,G(t)\,\mathrm{d}t. \tag{A 1·1}$$

Then the half-range complex function $\hat{P}(\tau)$ associated with $P(\tau)$ is given by

$$\hat{P}(\tau) = \int_{-\infty}^{+\infty} \hat{F}(t+\tau)\,\hat{G}^*(t)\,\mathrm{d}t, \tag{A 1·2}$$

where $\hat{F}$ and $\hat{G}$ are the half-range complex functions associated with F and G respectively.

Proof. Let a_ν and b_ν be the Fourier coefficients of F and c_ν and d_ν the Fourier coefficients of G:

$$F(t) = \int_0^\infty (a_\nu \cos 2\pi\nu t + b_\nu \sin 2\pi\nu t)\,\mathrm{d}\nu, \tag{A 1·3}$$

$$G(t) = \int_0^\infty (c_\nu \cos 2\pi\nu t + d_\nu \sin 2\pi\nu t)\,\mathrm{d}\nu. \tag{A 1·4}$$

In accordance with the definition of § 2, the associated half-range complex functions are then given by

$$\hat{F}(t) = \int_0^\infty (a_\nu + \mathrm{i}b_\nu)\,\mathrm{e}^{-2\pi\mathrm{i}\nu t}\,\mathrm{d}\nu, \tag{A 1·5}$$

$$\hat{G}(t) = \int_0^\infty (c_\nu + \mathrm{i}d_\nu)\,\mathrm{e}^{-2\pi\mathrm{i}\nu t}\,\mathrm{d}\nu. \tag{A 1·6}$$

Consider now the Fourier representations of $P(\tau)$ and $\hat{P}(\tau)$. Substitution from (A 1·4) into (A 1·1) gives

$$P(\tau) = 2\int_{-\infty}^{+\infty} \mathrm{d}t\,F(t+\tau)\int_0^\infty (c_\nu \cos 2\pi\nu t + d_\nu \sin 2\pi\nu t)\,\mathrm{d}\nu. \tag{A 1·7}$$

Now
$$\begin{aligned}\int_{-\infty}^{\infty} F(t+\tau)\cos 2\pi\nu t\,\mathrm{d}t &= \int_{-\infty}^{+\infty} F(u)\cos 2\pi\nu(u-\tau)\,\mathrm{d}u \\ &= \cos 2\pi\nu\tau\int_{-\infty}^{+\infty} F(u)\cos 2\pi\nu u\,\mathrm{d}u + \sin 2\pi\nu\tau\int_{-\infty}^{+\infty} F(u)\sin 2\pi\nu u\,\mathrm{d}u \\ &= \tfrac{1}{2}(a_\nu \cos 2\pi\nu\tau + b_\nu \sin 2\pi\nu\tau),\end{aligned} \tag{A 1·8}$$

where the Fourier inversion formula was used. Similarly

$$\int_{-\infty}^{\infty} F(t+\tau)\sin 2\pi\nu t\,\mathrm{d}t = \tfrac{1}{2}(b_\nu \cos 2\pi\nu\tau - a_\nu \sin 2\pi\nu\tau). \tag{A 1·9}$$

Hence, interchanging the order of integration in (A 1·7) and using the last two relations, one obtains

$$P(\tau) = \int_0^\infty \{[a_\nu c_\nu + b_\nu d_\nu]\cos 2\pi\nu\tau + [b_\nu c_\nu - a_\nu d_\nu]\sin 2\pi\nu\tau\}\,\mathrm{d}\nu. \tag{A 1·10}$$

The associated half-range complex function $\hat{P}(\tau)$ is therefore given by

$$\begin{aligned}\hat{P}(\tau) &= \int_0^\infty \{[a_\nu c_\nu + b_\nu d_\nu] + \mathrm{i}[b_\nu c_\nu - a_\nu d_\nu]\}\,\mathrm{e}^{-2\pi\mathrm{i}\nu\tau}\,\mathrm{d}\nu \\ &= \int_0^\infty (a_\nu + \mathrm{i}b_\nu)(c_\nu - \mathrm{i}d_\nu)\,\mathrm{e}^{-2\pi\mathrm{i}\nu\tau}\,\mathrm{d}\nu.\end{aligned} \tag{A 1·11}$$

Next consider the integral on the right of (A 1·2). Substitution from (A 1·6) gives

$$\int_{-\infty}^{\infty} \hat{F}(t+\tau)\,\hat{G}^*(t)\,\mathrm{d}t = \int_{-\infty}^{+\infty} \mathrm{d}t\,\hat{F}(t+\tau)\int_0^\infty (c_\nu - \mathrm{i}d_\nu)\,\mathrm{e}^{2\pi\mathrm{i}\nu t}\,\mathrm{d}\nu. \tag{A 1·12}$$

But
$$\begin{aligned}\int_{-\infty}^{\infty} \hat{F}(t+\tau)\,\mathrm{e}^{2\pi\mathrm{i}\nu t}\,\mathrm{d}t &= \int_{-\infty}^{+\infty} \hat{F}(u)\,\mathrm{e}^{2\pi\mathrm{i}\nu(u-\tau)}\,\mathrm{d}u \\ &= (a_\nu + \mathrm{i}b_\nu)\,\mathrm{e}^{-2\pi\mathrm{i}\nu\tau},\end{aligned} \tag{A 1·13}$$

where the Fourier inversion formula was used. Interchanging the order of integration in (A 1·12) and using (A 1·13), one obtains

$$\int_{-\infty}^{+\infty} \hat{F}(t+\tau)\, \hat{G}^*(t)\, \mathrm{d}t = \int_0^{\infty} (a_\nu + \mathrm{i}b_\nu)\,(c_\nu - \mathrm{i}d_\nu)\, \mathrm{e}^{-2\pi \mathrm{i}\nu\tau}\, \mathrm{d}\nu. \qquad (\mathrm{A}\,1{\cdot}14)$$

On comparing (A 1·14) with (A 1·11), the theorem follows.

Appendix 2. The range of validity of the restricted theory of paper I

We shall now investigate the range of validity of the restricted formulation of the generalized Huygens principle and of the generalized interference law of Zernike and Hopkins, as given in the preceding paper of this series.

The essential approximation which was made in the derivation of these formulae in I, was the replacement of the exponential term

$$\exp\{\mathrm{i}(kr_1 - k'r_2)\} \quad \text{by} \quad \exp\{\mathrm{i}k^{(0)}(r_1 - r_2)\}$$

in the expression† I, (2·7):

$$I(\mathbf{x}) = \frac{1}{2T}\int_{-T}^{T} \mathrm{d}t \int_0^\infty\int_0^\infty \mathrm{d}\nu\, \mathrm{d}\nu' \exp\{-2\pi\mathrm{i}(\nu-\nu')\,t\}$$
$$\times \int_{\mathscr{A}}\int_{\mathscr{A}} v(\mathbf{x}_1,\nu)\, v^*(\mathbf{x}_2,\nu')\, \frac{\exp\{\mathrm{i}(kr_1 - k'r_2)\}}{r_1 r_2}\, \Lambda_1\Lambda_2^*\, \mathrm{d}\mathbf{x}_1\, \mathrm{d}\mathbf{x}_2. \qquad (\mathrm{A}\,2{\cdot}1)$$

To see what restriction this implies, we note that

$$kr_1 - k'r_2 = k^{(0)}(r_1 - r_2) + (k - k^{(0)})\,(r_1 - r_2) + (k - k')\,r_2. \qquad (\mathrm{A}\,2{\cdot}2)$$

(A 2·1) may therefore be written as

$$I(\mathbf{x}) = \int_{\mathscr{A}}\int_{\mathscr{A}} G(\mathbf{x}_1,\mathbf{x}_2,r_2)\, \frac{\exp\{\mathrm{i}[k^{(0)}(r_1 - r_2) + (k - k^{(0)})\,(r_1 - r_2)]\}}{r_1 r_2}\, \Lambda_1\Lambda_2^*\, \mathrm{d}\mathbf{x}_1\, \mathrm{d}\mathbf{x}_2, \qquad (\mathrm{A}\,2{\cdot}3)$$

where

$$G(\mathbf{x}_1,\mathbf{x}_2,r_2) = \frac{1}{2T}\int_{-T}^{T} \mathrm{d}t \int_0^\infty\int_0^\infty v(\mathbf{x}_1,\nu)\, v^*(\mathbf{x}_2,\nu') \exp\{-2\pi\mathrm{i}(\nu-\nu')\,(t - r_2/c)\}\, \mathrm{d}\nu\, \mathrm{d}\nu'. \qquad (\mathrm{A}\,2{\cdot}4)$$

We now change the variable of integration in (A 2·4) from t to $t' = t - r_2/c$. Then, since

$$\lim_{T'\to\infty} \int_{-T'}^{T'} \exp\{-2\pi\mathrm{i}(\nu-\nu')\,t'\}\, \mathrm{d}t' = \delta(\nu-\nu'), \qquad (\mathrm{A}\,2{\cdot}5)$$

where δ is the Dirac delta function, it follows, that in the limit $T \to \infty$, G becomes independent of r_2 and is equal to

$$\Gamma(\mathbf{x}_1,\mathbf{x}_2) = \lim_{T\to\infty} G(\mathbf{x}_1,\mathbf{x}_2,r_2) = \lim_{T\to\infty} \frac{1}{2T}\int_0^\infty v(\mathbf{x}_1,\nu) v^*(\mathbf{x}_2,\nu) \mathrm{d}\nu \qquad (\mathrm{A}\,2{\cdot}6)$$

$$= \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} V(\mathbf{x}_1,t)\, V^*(\mathbf{x}_2,t)\, \mathrm{d}t. \qquad (\mathrm{A}\,2{\cdot}7)$$

† In I, the lower limits of the integrals over the frequency range were actually taken as $-\infty$. We replace them here by 0, since as explained in § 2 above, $v(\mathbf{x},\nu) = 0$ for $\nu < 0$.

Hence in the limit $T \to \infty$, (A 2·3) differs from the approximate expression I, (2·8) by the presence of the term $(k-k^{(0)})(r_1-r_2)$ in the exponential term and by the factor $\Lambda_1^{(0)}, \Lambda_2^{(0)*}$ in place of $\Lambda_1 \Lambda_2^*$. Setting $k-k^{(0)} = \Delta k$, $r_1-r_2 = \Delta r$, and replacing the inclination factors by their mean values, it follows that I, (2·8) will be a valid approximation if

$$|\Delta k| \, |\Delta r| \ll 2\pi, \tag{A 2·8}$$

or, since $k = 2\pi/\lambda$, $\Delta k = -2\pi \Delta\lambda/\lambda^2$, and the condition becomes†

$$|\Delta r| \ll \frac{\lambda^2}{|\Delta\lambda|}. \tag{A 2·9}$$

Hence *the more restricted formulation of the generalized Huygens principle and of the generalized interference law given in paper I is applicable if the path differences between the interfering beams are small compared to* $\lambda^2/|\Delta\lambda|$.

The quantity $\lambda^2/|\Delta\lambda|$ has a simple physical interpretation: it represents (apart from a multiplicative factor of the order of unity), the coherence length of the radiation. A proof of this result will be found in Born (1933, p. 137) and Kahan (1952, p. 310).

This work was carried out during the tenure of an Imperial Chemical Industries Research Fellowship, the award of which, by the University of Manchester is gratefully acknowledged. It is also a pleasure to thank Dr F. D. Kahn for some helpful discussions.

References

Baker, B. B. & Copson, E. T. 1950 *The mathematical theory of Huygens's principle*, 2nd ed. Oxford: Clarendon Press.
Blanc-Lapierre, A. & Dumontet, P. 1955 *Rev. Opt. (théor. instrum.)*, **34**, 1.
Born, M. 1933 *Optik*. Berlin: Springer.
Gabor, D. 1946 *J. Inst. Elect. Engrs.* **93**, Part III, 429.
Kahan, T. 1952 *Suppl. Nuovo Cim.* **9**, 310.
Michelson, A. A. 1891 *Phil. Mag.* **31**, 338.
Michelson, A. A. 1892 *Phil. Mag.* **34**, 280.
Parke III, N. G. 1948 Ph.D. thesis, Massachusetts Institute of Technology. (See also *Tech. Rep. Electron., Mass. Inst. Tech.* no. 95 (1949).)
Rayleigh, Lord 1892 *Phil. Mag.* **34**, 407.
Ryle, M. 1950 *Rep. Progr. Phys.* **13**, 184.
Smith, F. G. 1952 *Mon. Not. R. Astr. Soc.* **112**, 497.
Van Cittert, P. H. 1939 *Physica*, **6**, 1129.
Williams, W. E. 1941 *Applications of interferometry*, 2nd ed. London: Methuen and Co.
Wiener, N. 1930 *Acta Math. Uppsala*, **55**, 117.
Wolf, E. 1954*a* *Proc. Roy. Soc.* A, **225**, 96.
Wolf, E. 1954*b* *Nuovo Cim.* **12**, 884.
Zernike, F. 1938 *Physica*, **5**, 785.
Zernike, F. 1948 *Proc. Phys. Soc.* **61**, 158.

† It is to be noted that this is also essentially the condition I, (5·1) under which the approximate expressions of Van Cittert, Zernike and Hopkins for the space-correlation factor were derived in the previous paper.

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, CAMBRIDGE
(BROOKE CRUTCHLEY, UNIVERSITY PRINTER)

2.3

ASTRONOMICAL CONTRIBUTIONS FROM THE
UNIVERSITY OF MANCHESTER

SERIES III, No. 23

OPTICS IN TERMS OF OBSERVABLE QUANTITIES

by

E. WOLF

Reprinted from

IL NUOVO CIMENTO

PRINTED IN ITALY BY TIPOGRAFIA COMPOSITORI - BOLOGNA

IL NUOVO CIMENTO Vol. XII, N. 6 1° Dicembre 1954

Optics in Terms of Observable Quantities.

E. Wolf

Department of Astronomy, The University, Manchester, England

(ricevuto il 26 Settembre 1954)

Summary. — Space-time correlation functions are defined which express the correlation between components of the electromagnetic field vectors in stationary fields. These functions form sets of 3×3 matrices, the individual elements of which obey the wave equation. Unlike the field vectors which are not measurable at the high frequencies encountered in Optics our correlation functions may be determined with the help of standard optical instruments. The results enable a unified treatment of theories of partial coherence and partial polarization to be obtained, and suggest a formulation of a wide branch of Optics in terms of observable quantities only.

In all Optical experiments the only quantities which are observable are the averages of certain quadratic functions of the field components. It is therefore tempting to try to formulate the laws of Optical fields directly in terms of such quantities rather than in terms of the unmeasurable field vectors as has been customary in the past.

It was as early as 1852 that Stokes (1) showed that a nearly monochromatic (plane) light wave may be characterized at each point by four parameters which now bear his name (*). If

$$E_x = a_1(\boldsymbol{x}, t) \cos \{2\pi\nu_0 t - \alpha_1(\boldsymbol{x}, t)\} , \qquad E_y = a_2(\boldsymbol{x}, t) \cos \{2\pi\nu_0 t - \alpha_2(\boldsymbol{x}, t)\}, \tag{1}$$

(1) G. G. Stokes: *Trans. Camb. Phil. Soc.*, **9**, 399 (1852). Also his *Mathematical and Physical Papers* (Cambridge, 1901), vol. III, p. 233.

(*) Very good accounts of the Stokes parameters may be found in Chandrasekhar (2) and Walker (3).

are the components of the electric vector of such a wave in two mutually orthogonal directions at right angles to the direction of propagation, the Stokes parameters are defined by

$$(2) \qquad \begin{cases} P = \langle a_1^2 + a_2^2 \rangle\,, & Q = \langle a_1^2 - a_2^2 \rangle\,, \\ U = \langle 2a_1 a_2 \cos(\alpha_1 - \alpha_2) \rangle\,, & V = \langle 2a_1 a_2 \sin(\alpha_1 - \alpha_2) \rangle\,, \end{cases}$$

the brackets $\langle .... \rangle$ denoting time average. Then the intensity $I(\psi, \varepsilon)$ associated with vibrations in the direction which makes an angle ψ with the x-direction, when retardation ε is introduced between the two components, is given by (see CHANDRASEKHAR (2), p. 29)

$$(3) \qquad I(\psi, \varepsilon) = \tfrac{1}{2}[P + Q\cos 2\psi + (U\cos\varepsilon - V\sin\varepsilon)\sin 2\psi]\,.$$

By measuring I for different values of ψ and ε, the four parameters may be determined.

Let us now introduce in place of E_x and E_y, the associated complex vectors

$$(4) \qquad \hat{E}_x = a_1(\boldsymbol{x}, t)\exp\{i[2\pi\nu_0 t - \alpha_1(\boldsymbol{x}, t)]\}, \quad \hat{E}_y = a_2(\boldsymbol{x}, t)\exp\{i[2\pi\nu_0 t - \alpha_2(\boldsymbol{x}, t)]\}.$$

Next we construct the four functions

$$(5) \qquad \mathcal{E}_{ij} = \langle \hat{E}_i(\boldsymbol{x}, t)\hat{E}_j^*(\boldsymbol{x}, t) \rangle,$$

where i and j can each take on the value x or y and asterisk denotes the complex conjugate. The knowledge of these four quantities is equivalent to the knowledge of the four Stokes parameters; in fact the Stokes parameters are simple linear combinations of the $\mathcal{E}_{ij}$'s. If (3) is expressed in terms of these quantities, one finds after a simple calculation that

$$(6) \qquad I(\psi, \varepsilon) = I_x(\psi) + I_y(\psi) + 2\sqrt{I_x(\psi)}\sqrt{I_y(\psi)}\,|\gamma_{xy}|\cos[\arg\gamma_{xy} + \varepsilon],$$

where

$$(7) \qquad I_x(\psi) = \mathcal{E}_{xx}\cos^2\psi\,, \qquad I_y(\psi) = \mathcal{E}_{yy}\sin^2\psi\,,$$

and

$$(8) \qquad \gamma_{xy} = \frac{\mathcal{E}_{xy}}{\sqrt{\mathcal{E}_{xx}}\sqrt{\mathcal{E}_{yy}}}\;;$$

(2) S. CHANDRASEKHAR: *Radiative Transfer* (Oxford, 1950).
(3) M. J. WALKER: *Amer. Journ. Phys.*, **22**, 170 (1954).

and the well-known inequality $P^2 \geqslant Q^2 + U^2 + V^2$ becomes simply $|\gamma_{xy}| \leqslant 1$. Equation (6) is formally identical with the *generalized interference law* derived in recent years in the theory of partially coherent scalar fields (ZERNIKE [4], HOPKINS [5], WOLF [6]). In the special case when $\gamma = 0$, (6) reduces to the usual law for the combination of completely incoherent fields; when $|\gamma| = 1$ it reduces to the ordinary interference law for fields which are perfectly coherent. Partially coherent and partially polarized fields are characterized by the intermediate values of $|\gamma|$.

Now the Stokes parameters (or the 2 by 2 matrix whose elements are defined by (5)) give the intensity and express the correlation between the components of $\boldsymbol{E}$ at the *same* point in space and at the *same* instant of time. Moreover they are defined only for a plane wave whose effective frequency range is sufficiently narrow. In order to characterize a general stationary field (*), we introduce, by analogy with the scalar case (see WOLF [7]), correlation functions between field components at different points in space and at different instants of time.

Let

$$E_i(\boldsymbol{x}, t) = \int_0^\infty a_{\nu i}(\boldsymbol{x}) \cos \{2\pi\nu t - \alpha_{\nu i}(\boldsymbol{x})\} \, \mathrm{d}\nu\,, \qquad (i = x,\ y \text{ or } z) \tag{9}$$

be the Fourier representation of a typical field component over the time interval $-T \leqslant t \leqslant T$, $\boldsymbol{E}$ being formally assumed to be zero outside this range, and define the functions

$$\hat{E}_i(\boldsymbol{x}, t) = \int_0^\infty a_{\nu i}(\boldsymbol{x}) \exp \{i[2\pi\nu t - \alpha_{\nu i}(\boldsymbol{x})\} \, \mathrm{d}\nu\,. \tag{10}$$

We now introduce a 3×3 correlation matrix $\mathcal{E}$ whose elements are

$$\mathcal{E}_{ij}(\boldsymbol{x}_1, \boldsymbol{x}_2, \tau) = \langle \hat{E}_i(\boldsymbol{x}_1, t+\tau) \hat{E}_j^*(\boldsymbol{x}_2, t) \rangle\,. \tag{11}$$

Similar considerations to those employed in connection with scalar fields of

---

([4]) F. ZERNIKE: *Physica*, **5**, 785 (1938).

([5]) H. H. HOPKINS: *Proc. Roy. Soc.*, A **208**, 263 (1951).

([6]) E. WOLF: *Proc. Roy. Soc.*, A **225**, 96 (1954).

(*) By a stationary field we mean here a field of which all *observable* properties are constant in time. This includes as a special case the usual case of high frequency periodic time-dependence; or the field constituted by the steady flux of polychromatic radiation through an optical system. But it excludes fields for which the time average over a macroscopic time interval of the flux of radiation depends on time.

([7]) E. WOLF: *Proc. Roy. Soc.*, in press.

arbitrary frequency range and with vector fields characterized by the Stokes parameters indicate, that the expressions for the electric energy density appropriate to various experimental conditions are simple functions of the elements of the $\mathcal{E}$-matrix; and moreover that all these elements may be determined from experiments by means of standard optical instruments.

Since the electric vector satisfies the wave equation, it can readily be shown that each element of the $\mathcal{E}$-matrix satisfies the two equations,

$$\text{(12)} \qquad \begin{cases} \nabla_1^2 \mathcal{E}_{ij} = \dfrac{1}{c^2}\dfrac{\partial^2 \mathcal{E}_{ij}}{\partial \tau^2}\,, \\[2ex] \nabla_2^2 \mathcal{E}_{ij} = \dfrac{1}{c^2}\dfrac{\partial^2 \mathcal{E}_{ij}}{\partial \tau^2}\,, \end{cases}$$

where ∇_1^2 and ∇_2^2 are the Laplacian operators with respect to the coordinates of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and c is the velocity of light in the vacuum. Thus not only the unmesurable field vectors, but also *the observable correlation functions* here introduced *obey rigorous propagation laws*. This result should prove particularly useful in connection with scattering problems.

In addition to the $\mathcal{E}$-matrix, one can introduce similar matrices involving components of the other field vectors ($\boldsymbol{H}$, $\boldsymbol{D}$ and $\boldsymbol{B}$) and also matrices involving mixed pairs like E_i and H_j. On account of Maxwell's equations, these matrices are related by a set of first order partial differential equations with respect to the variables $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and τ. In the analysis of all optical experiments $c\tau$ will play the part of an optical path difference. The actual time, like the frequency has been eliminated.

Unlike the $\mathcal{E}$-matrix, the $\mathcal{H}$ matrix is not likely to be of any interest in Optics, since no radiation detectors appear to be available at Optical wavelengths which would respond to the magnetic rather than the electric field. It may, however, prove useful in connection with applications to other types of partially coherent radiation, e.g. in Radio Astronomy. The matrix containing the mixed pairs should prove useful in experiments where the (averaged) flux of energy rather than the energy density is measured.

The matrices here introduced may be expected to play a role in Electromagnetic field theory which is somewhat analogous to that which the Density matrix of von NEUMAN [8] plays in Quantum Mechanics. An analogy between the Stokes parameters and the Density Matrix has been noted previously (PERRIN [9], FALKOFF and MACDONALD [10]; see also FANO [11]). It is, how-

[8] J. v. NEUMANN: *Gött. Nachr.*, 245 (1927).

[9] F. PERRIN, *Journ. Chem. Phys.*, **10**, 415 (1942).

[10] D. L. FALKOFF and J. E. MACDONALD: *Journ. Opt. Soc. Amer.*, **41**, 861 (1951).

[11] U. FANO: *Phys. Rev.*, **93**, 121 (1954).

ever, evident that only by considering more general correlation functions, such as those here introduced, does one obtain an adequate tool for the study of propagation problems in a general stationary electromagnetic field.

A fuller discussion of the subject matter of this note will be published at a later date.

This work was carried out during the tenure of an Imperial Chemical Industries Research Fellowship and was also supported by a grant from the Carnegie Trust for the Universities of Scotland, both of which are gratefully acknowledged.

RIASSUNTO (*)

Si definiscono funzioni di correlazione spazio-tempo che esprimono la correlazione fra componenti dei vettori del campo elettromagnetico in campi stazionari. Queste funzioni formano gruppi di 3×3 matrici i cui elementi individuali soddisfano all'equazione d'onda. A differenza dei vettori di campo che, alle alte frequenze che intervengono in Ottica, non sono misurabili, le nostre funzioni di correlazione possono essere determinate con l'ausilio degli ordinari apparecchi ottici. I risultati consentono un trattamento unificato delle teorie della coerenza parziale e della polarizzazione parziale e suggeriscono una formulazione di un ampio settore dell'Ottica in termini di sole grandezze osservabili.

(*) *Traduzione a cura della Redazione.*

PART III

STUDIES IN ELECTROMAGNETIC THEORY

REPRINTED FROM THE
PROCEEDINGS OF THE PHYSICAL SOCIETY, A, VOL. LXVI, p. 1129, 1953

A Scalar Representation of Electromagnetic Fields

BY H. S. GREEN† AND E. WOLF
Department of Mathematical Physics, University of Edinburgh

Communicated by Max Born; MS. received 16th June 1953, and in amended form 31st August 1953

Abstract. It is shown that in a region which is free of currents and charges, any electromagnetic field may be rigorously derived from a single, generally complex, scalar wave function $V(\mathbf{x}, t)$. In terms of this function the momentum density $\mathbf{g}(\mathbf{x}, t)$ and the energy density $w(\mathbf{x}, t)$ of the field may be defined in such a way that they are represented by expressions analogous to the formulae for the probability current and the probability density in quantum mechanics; in a homogeneous isotropic medium

$$\mathbf{g}(\mathbf{x}, t) = -\frac{1}{8\pi\mu_0 c}[\dot{V}^*\nabla V + \dot{V}\nabla V^*],$$

$$w(\mathbf{x}, t) = \frac{1}{8\pi}\left[\frac{\epsilon_0}{c^2}\dot{V}\dot{V}^* + \frac{1}{\mu_0}\nabla V.\nabla V^*\right].$$

The densities defined in this way differ from those given by the usual expressions, but the choice is justified since the differences disappear on integration over any arbitrary macroscopic domain. (The corresponding Lagrangian densities differ by a four divergence.) When V is of the form $V_0(\mathbf{x})e^{-i\omega t}$, $\mathbf{g}$ is found to form a solenoidal field which is orthogonal to the co-phasal surfaces arg V_0 = constant.

§ 1. INTRODUCTION

In a region of space which is free of charges and currents an electromagnetic field is fully specified by the magnetic vector potential $\mathbf{A}$. From it the electric and the magnetic field vectors may be derived by means of well-known relations.

In a wide class of problems, particularly in those encountered in optics, the actual behaviour of the field vectors is of little interest. What one primarily wishes to know is the average energy or the average flux, and one is led to wonder whether for such purposes the derivation from a vector potential is really the most suitable one. Except in the so-called 'rigorous' diffraction theory, one does in fact often employ in optical considerations of energy a single, generally complex, scalar wave function (usually called the disturbance, or the complex amplitude), whose squared modulus is taken as the measure of the light intensity. This simple procedure has been employed in optics since about the time of Fresnel (the disturbance then being considered to be the displacement of a particle in an 'elastic ether') and has been the subject of much criticism, in spite of the fact that under fairly general conditions it gives results which are found to be in excellent agreement with experiment. The validity of the scalar

† On leave of absence from the University of Adelaide, Adelaide, South Australia.

theory has been justified, at least as an approximation or as a time average, for a number of special cases only: (*a*) for certain two-dimensional problems and for monochromatic fields with rotational symmetry (cf. Braunbek 1951), (*b*) in applications to diffraction problems encountered in optical instruments of usual design and working with non-polarized and almost monochromatic light (Picht 1931, Luneberg 1947–8, Theimer, Wassermann and Wolf 1952).

Now it is well known that the energy density and the momentum density (and consequently the light intensity) are not uniquely defined by the electric and the magnetic field vectors. One may always add to the Lagrangian density a four divergence which gives no contribution to the field equations, though it may alter the local (unobservable) values of the densities of the energy and the momentum. Such alteration must of course lead to no observable changes in the *total* amount of energy and momentum contained in any extended (macroscopic) domain. The question is therefore open as to whether one could not define the energy and momentum densities in such a way that they would always be expressible in a simple manner in terms of a single complex scalar wave function, the field vectors remaining unchanged. In the present paper it is shown that this in fact is possible.

We find that in regions where no currents or charges are present the magnetic potential may be rigorously derived from a single, generally complex, scalar wave function $V(\mathbf{x}, t)$, which we call the *complex potential*. In terms of this function the momentum density $\mathbf{g}(\mathbf{x}, t)$ and the energy density $w(\mathbf{x}, t)$ in a homogeneous isotropic medium may be defined by means of formulae analogous to the expressions for the probability current and the probability density in quantum mechanics:

$$\left.\begin{aligned} \mathbf{g}(\mathbf{x}, t) &= -\frac{1}{8\pi\mu_0 c}[\dot{V}^*\nabla V + \dot{V}\nabla V^*], \\ \text{and} \qquad w(\mathbf{x}, t) &= \frac{1}{8\pi}\left[\frac{\epsilon_0}{c^2}\dot{V}\dot{V}^* + \frac{1}{\mu_0}\nabla V \,.\, \nabla V^*\right]; \end{aligned}\right\} \qquad \ldots\ldots(1.1)$$

these quantities satisfy a conservation law of the usual form:

$$\frac{1}{c}\frac{dw}{dt} + \nabla \,.\, \mathbf{g} = 0. \qquad \ldots\ldots(1.2)$$

The corresponding Lagrangian density L is

$$L(\mathbf{x}, t) = \frac{1}{8\pi}\left\{\frac{\epsilon_0}{c^2}\dot{V}\dot{V}^* - \frac{1}{\mu_0}\nabla V \,.\, \nabla V^*\right\}; \qquad \ldots\ldots(1.3)$$

here ϵ_0, μ_0 are the dielectric constant and the magnetic permeability and c is the vacuum velocity of light.

An interesting consequence of the present analysis is the result that in the case when the time enters V only through the factor $e^{-i\omega t}$, i.e. when V is of the form

$$V(\mathbf{x}, t) = v(\mathbf{x}) \exp\{i[k\mathscr{S}(\mathbf{x}) - \omega t]\} \quad (v \text{ and } \mathscr{S} \text{ real}) \qquad \ldots\ldots(1.4)$$

(ω = frequency, $k = \omega/c$), $\mathbf{g}$ and w become independent of time and (1.1) reduces to

$$\mathbf{g}(\mathbf{x}) = \frac{k^2}{4\pi\mu_0} v^2 \nabla \mathscr{S}. \qquad \ldots\ldots(1.5)$$

The conservation law now becomes

$$\nabla \,.\, \mathbf{g}(\mathbf{x}) = 0. \qquad \ldots\ldots(1.6)$$

Equation (1.5) shows that the energy flow is orthogonal to the surfaces of constant phase of the complex potential and (1.6) expresses the fact that the vector field $\mathbf{g}$ is solenoidal. One is thus led rigorously to the concept of 'electromagnetic rays', and one obtains for a class of electromagnetic fields a simple model which may be regarded as a natural generalization of geometrical optics.

§ 2. Definition of the Complex Potential

We consider an electromagnetic field in a homogeneous, isotropic medium. In regions free of currents and charges the field quantities may be derived from a single vector potential $\mathfrak{A}(\mathbf{x}, t)$ which satisfies the homogeneous wave equation and also the divergence condition

$$\nabla \cdot \mathfrak{A} = 0. \qquad \ldots\ldots(2.1)$$

In general, currents and charges will, of course, be present in some parts of the $\mathbf{x}$ space. It will be convenient to imagine these enclosed in boxes and to consider in place of $\mathfrak{A}$ the function $\mathbf{A}$ defined by

$$\begin{aligned} \mathbf{A}(\mathbf{x}, t) &= \mathfrak{A}(\mathbf{x}, t) \text{ outside and on the boxes} \\ &= 0 \qquad \text{inside the boxes.} \end{aligned}$$

$\mathbf{A}$ is formally defined over the whole $\mathbf{x}$ space and may be represented by a Fourier integral

$$\mathbf{A}(\mathbf{x}, t) = \int [\mathbf{a}(\mathbf{k}, t) \cos(\mathbf{k} \cdot \mathbf{x}) + \mathbf{b}(\mathbf{k}, t) \sin(\mathbf{k} \cdot \mathbf{x})]\, d\mathbf{k}, \qquad \ldots\ldots(2.2)$$

where, since $\mathbf{A}$ is real, the integration is carried over half of $\mathbf{k}$ space (e.g. $k_x \geqslant 0$). On account of (2.1) the (real) vectors $\mathbf{a}(\mathbf{k}, t)$ and $\mathbf{b}(\mathbf{k}, t)$ are orthogonal to $\mathbf{k}$,

$$\mathbf{a} \cdot \mathbf{k} = \mathbf{b} \cdot \mathbf{k} = 0. \qquad \ldots\ldots(2.3)$$

With each $\mathbf{k}$ we associate two unit vectors $\mathbf{l}_1(\mathbf{k})$ and $\mathbf{l}_2(\mathbf{k})$ such that $\mathbf{l}_1$, $\mathbf{l}_2$ and $\mathbf{k}$ form a right-handed orthogonal triad. This may be done for example by choosing a constant vector $\mathbf{n}$ and taking as $\mathbf{l}_1$ a unit vector perpendicular to the $(\mathbf{k}, \mathbf{n})$ plane and as $\mathbf{l}_2$ a unit vector perpendicular to $\mathbf{k}$ and $\mathbf{l}_1$:

$$\left.\begin{aligned} \mathbf{l}_1(\mathbf{k}) &= \frac{\mathbf{n} \wedge \mathbf{k}}{|\mathbf{n} \wedge \mathbf{k}|}, \\ \mathbf{l}_2(\mathbf{k}) &= \frac{\mathbf{k} \wedge \mathbf{l}_1}{|\mathbf{k} \wedge \mathbf{l}_1|} = \frac{k^2\mathbf{n} - (\mathbf{n} \cdot \mathbf{k})\mathbf{k}}{|k^2\mathbf{n} - (\mathbf{n} \cdot \mathbf{k})\mathbf{k}|}. \end{aligned}\right\} \qquad \ldots\ldots(2.4)$$

Equations (2.3) show that $\mathbf{a}$ and $\mathbf{b}$ lie in the plane of $\mathbf{l}_1$ and $\mathbf{l}_2$ and may therefore be resolved along these vectors:

$$\left.\begin{aligned} a_1 &= \mathbf{a} \cdot \mathbf{l}_1, \qquad & a_2 &= \mathbf{a} \cdot \mathbf{l}_2, \\ b_1 &= \mathbf{b} \cdot \mathbf{l}_1, \qquad & b_2 &= \mathbf{b} \cdot \mathbf{l}_2. \end{aligned}\right\} \qquad \ldots\ldots(2.5)$$

Next we form the complex combinations

$$\left.\begin{aligned} \alpha(\mathbf{k}, t) &= a_1(\mathbf{k}, t) + i a_2(\mathbf{k}, t), \\ \beta(\mathbf{k}, t) &= b_1(\mathbf{k}, t) + i b_2(\mathbf{k}, t), \end{aligned}\right\} \qquad \ldots\ldots(2.6)$$

and regard α and β as Fourier coefficients of a function V:

$$V(\mathbf{x}, t) = \int [\alpha(\mathbf{k}, t) \cos(\mathbf{k} \cdot \mathbf{x}) + \beta(\mathbf{k}, t) \sin(\mathbf{k} \cdot \mathbf{x})]\, d\mathbf{k}. \qquad \ldots\ldots(2.7)$$

We call $V(\mathbf{x}, t)$ the *complex potential* of the field.

We have now replaced the magnetic vector potential by a complex scalar function. Conversely it is easily seen that from the complex potential V, the magnetic vector potential $\mathbf{A}$ and consequently the field vectors may be uniquely derived. For in the first place, from the knowledge of V the quantities α and β may be obtained by applying the Fourier inversion formula. Then a_1, a_2, b_1 and b_2 may be determined from (2.6). From them the Fourier components $\mathbf{a}$ and $\mathbf{b}$ of the complex potential are obtained:

$$\mathbf{a} = a_1\mathbf{l}_1 + a_2\mathbf{l}_2, \qquad \mathbf{b} = b_1\mathbf{l}_1 + b_2\mathbf{l}_2. \qquad \ldots\ldots(2.8)$$

Finally, if one forms the combinations $\mathbf{a}\cos(\mathbf{k}\,.\,\mathbf{x}) + \mathbf{b}\sin(\mathbf{k}\,.\,\mathbf{x})$ and integrates over all $\mathbf{k}$, one obtains the vector potential $\mathbf{A}$. Hence the scalar $V(\mathbf{x}, t)$ completely specifies the field.*

It is easily seen that in a homogeneous isotropic medium of dielectric constant ϵ_0 and magnetic permeability μ_0 the complex potential satisfies the wave equation

$$\nabla^2 V - \frac{\epsilon_0\mu_0}{c^2}\ddot{V} = 0. \qquad \ldots\ldots(2.9)$$

For, since the vector potential $\mathbf{A}$ satisfies the wave equation

$$\nabla^2\mathbf{A} - \frac{\epsilon_0\mu_0}{c^2}\ddot{\mathbf{A}} = 0, \qquad \ldots\ldots(2.10)$$

one must have, for each component,

$$\mathbf{A}_\mathbf{k} = \mathbf{a}(\mathbf{k}, t)\cos(\mathbf{k}\,.\,\mathbf{x}) + \mathbf{b}(\mathbf{k}, t)\sin(\mathbf{k}\,.\,\mathbf{x}) \qquad \ldots\ldots(2.11)$$

of $\mathbf{A}$,

$$\ddot{\mathbf{A}}_\mathbf{k} + \frac{k^2c^2}{\epsilon_0\mu_0}\mathbf{A}_\mathbf{k} = 0. \qquad \ldots\ldots(2.12)$$

It then follows from (2.12), (2.5) and (2.6) that the corresponding terms V_K of V, i.e.

$$V_\mathbf{k} = \alpha(\mathbf{k}, t)\cos(\mathbf{k}\,.\,\mathbf{x}) + \beta(\mathbf{k}, t)\sin(\mathbf{k}\,.\,\mathbf{x}), \qquad \ldots\ldots(2.13)$$

must satisfy the scalar wave equation

$$\ddot{V}_\mathbf{k} + \frac{k^2c^2}{\epsilon_0\mu_0}V_\mathbf{k} = 0. \qquad \ldots\ldots(2.14)$$

Consequently the complex potential satisfies the homogeneous wave equation (2.9).

§ 3. The Momentum Density and the Energy Density

3.1. We now derive expressions for the momentum density and the energy density of the field in terms of the complex potential. We restrict our discussion to a homogeneous isotropic medium.

In terms of the vector potential, the electric and the magnetic field vectors are given by

$$\mathbf{E} = -\frac{1}{c}\dot{\mathbf{A}}, \qquad \mu_0\mathbf{H} = \nabla\wedge\mathbf{A}, \qquad \ldots\ldots(3.1)$$

and the total momentum $\mathbf{G}$ of the field is

$$\mathbf{G}(t) = \frac{1}{4\pi}\int(\mathbf{E}\wedge\mathbf{H})\,d\mathbf{x} = -\frac{1}{4\pi\mu_0 c}\int\dot{\mathbf{A}}\wedge(\nabla\wedge\mathbf{A})\,d\mathbf{x}, \qquad \ldots\ldots(3.2)$$

the integration being taken throughout the $\mathbf{x}$ space.

* *Note added in proof.* Since this paper was written our attention has been drawn to a paper by E. T. Whittaker (*Proc. Lond. Math. Soc.*, 1904, **1**, 367), where it is shown that *in vacuo* at points not occupied by electrons a field produced by any number of electrons moving in any arbitrary manner may be derived from two real scalar wave functions.

The connection between Whittaker's analysis and that of the present section is being investigated and it is hoped to publish the results in a later communication.

To express $\mathbf{G}$ in terms of V we first substitute into (3.2) from (2.2). We have

$$\left.\begin{aligned}\dot{\mathbf{A}} &= \int [\dot{\mathbf{a}}\cos(\mathbf{k}\,.\,\mathbf{x}) + \dot{\mathbf{b}}\sin(\mathbf{k}\,.\,\mathbf{x})]\,d\mathbf{k},\\ \nabla\wedge\mathbf{A} &= \int \mathbf{k}\wedge[-\mathbf{a}\sin(\mathbf{k}\,.\,\mathbf{x}) + \mathbf{b}\cos(\mathbf{k}\,.\,\mathbf{x})]\,d\mathbf{k}.\end{aligned}\right\} \qquad \ldots\ldots(3.3)$$

Hence

$$\mathbf{G} = -\frac{1}{4\pi\mu_0 c}\int d\mathbf{x}\iint [\dot{\mathbf{a}}(\mathbf{k}',t)\cos(\mathbf{k}'\,.\,\mathbf{x}) + \dot{\mathbf{b}}(\mathbf{k}',t)\sin(\mathbf{k}'\,.\,\mathbf{x})]$$
$$\wedge[\mathbf{k}\wedge(-\mathbf{a}(\mathbf{k},t)\sin(\mathbf{k}\,.\,\mathbf{x}) + \mathbf{b}(\mathbf{k},t)\cos(\mathbf{k}\,.\,\mathbf{x})]\,d\mathbf{k}\,d\mathbf{k}'. \qquad \ldots\ldots(3.4)$$

This becomes, with the help of the Fourier integral theorem,

$$\mathbf{G} = -\frac{\pi^2}{2\mu_0 c}\int \dot{\mathbf{a}}(\mathbf{k},t)\wedge[\mathbf{k}\wedge\mathbf{b}(\mathbf{k},t)] - \dot{\mathbf{b}}(\mathbf{k},t)\wedge[\mathbf{k}\wedge\mathbf{a}(\mathbf{k},t)]\,d\mathbf{k}, \qquad \ldots\ldots(3.5)$$

or, using (2.8),

$$\begin{aligned}\mathbf{G} &= \frac{\pi^2}{2\mu_0 c}\int [(a_1\dot{b}_1 + a_2\dot{b}_2) - (b_1\dot{a}_1 + b_2\dot{a}_2)]\mathbf{k}\,d\mathbf{k}\\ &= \frac{\pi^2}{2\mu_0 c}\int [(\alpha^*\dot{\beta} + \alpha\dot{\beta}^*) - (\dot{\alpha}^*\beta + \dot{\alpha}\beta^*)]\mathbf{k}\,d\mathbf{k}\\ &= -\frac{1}{8\pi\mu_0 c}\int (\dot{V}^*\nabla V + \dot{V}\nabla V^*)\,d\mathbf{x}. \qquad \ldots\ldots(3.6)\end{aligned}$$

Hence the density $\mathbf{g}$ of the momentum may be defined as

$$\mathbf{g}(\mathbf{x},t) = -\frac{1}{8\pi\mu_0 c}[\dot{V}^*\nabla V + \dot{V}\nabla V^*]. \qquad \ldots\ldots(3.7)$$

The electric and the magnetic energy densities may also easily be defined in terms of the complex potential. The total *electric energy* W_e is

$$\begin{aligned}W_e(t) &= \frac{\epsilon_0}{8\pi}\int E^2\,d\mathbf{x}\\ &= \frac{\epsilon_0}{8\pi c^2}\int d\mathbf{x}\iint \{\underline{\dot{a}}(\mathbf{k},t)\cos(\mathbf{k}\,.\,\mathbf{x}) + \underline{\dot{b}}(\mathbf{k},t)\sin(\mathbf{k}\,.\,\mathbf{x})\}\\ &\qquad .\{\underline{\dot{a}}(\mathbf{k}',t)\cos(\mathbf{k}'\,.\,\mathbf{x}) + \underline{\dot{b}}(\mathbf{k}',t)\sin(\mathbf{k}'\,.\,\mathbf{x})\}\,d\mathbf{k}\,d\mathbf{k}'\\ &= \frac{\pi^2\epsilon_0}{2c^2}\int (\underline{\dot{a}}^2 + \underline{\dot{b}}^2)\,d\mathbf{k}\\ &= \frac{\pi^2\epsilon_0}{2c^2}\int (\dot{a}_1{}^2 + \dot{a}_2{}^2 + \dot{b}_1{}^2 + \dot{b}_2{}^2)\,d\mathbf{k}\\ &= \frac{\pi^2\epsilon_0}{2c^2}\int (\dot{\alpha}\dot{\alpha}^* + \dot{\beta}\dot{\beta}^*)\,d\mathbf{k}\\ &= \frac{\epsilon_0}{8\pi c^2}\int (\dot{V}\dot{V}^*)\,d\mathbf{x}. \qquad \ldots\ldots(3.8)\end{aligned}$$

The total *magnetic energy* W_m is

$$W_m(t) = \frac{\mu_0}{8\pi}\int \mathbf{H}^2\,dx,$$

and one finds, by a similar calculation as above,

$$W_m(t) = \frac{1}{8\pi\mu_0}\int (\nabla V\,.\,\nabla V^*)\,d\mathbf{x}. \qquad \ldots\ldots(3.9)$$

Equations (3.8) and (3.9) show that one may define the electric and magnetic densities w_e and w_m by the expressions

$$w_e(\mathbf{x}, t) = \frac{\epsilon_0}{8\pi c^2} \dot{V}\dot{V}^*, \qquad w_m(\mathbf{x}, t) = \frac{1}{8\pi\mu_0} (\nabla V . \nabla V^*). \qquad \ldots\ldots(3.10)$$

It is easily verified, with the help of the wave equation (2.9), that the total energy density $w = w_e + w_m$ and the energy flux $\mathbf{S} = c\mathbf{g}$ satisfy the conservation law

$$\frac{dw}{dt} + \nabla . \mathbf{S} = 0. \qquad \ldots\ldots(3.11)$$

3.2. V is in general a complex function. Let v denote its amplitude and ϕ its phase:

$$V(\mathbf{x}, t) = v(\mathbf{x}, t) \exp\{i\phi(\mathbf{x}, t)\}. \qquad \ldots\ldots(3.12)$$

If one substitutes from (3.12) into the wave equation (2.9) and separates real and imaginary parts, one obtains the following two equations:

$$\nabla^2 v - v(\nabla\phi)^2 - \frac{\epsilon_0\mu_0}{c^2}(\ddot{v} - v\dot{\phi}^2) = 0, \qquad \ldots\ldots(3.13)$$

$$2(\nabla v) . (\nabla\phi) + v\nabla^2\phi - \frac{\epsilon_0\mu_0}{c^2}(2\dot{v}\dot{\phi} + v\ddot{\phi}) = 0. \qquad \ldots\ldots(3.14)$$

In terms of v and ϕ the momentum density $\mathbf{g}$ and the energy densities w_e and w_m are

$$\left.\begin{aligned} \mathbf{g}(\mathbf{x}, t) &= -\frac{1}{4\pi\mu_0 c}(\dot{v}\nabla v + v^2\dot{\phi}\nabla\phi), \\ w_e(\mathbf{x}, t) &= \frac{\epsilon_0}{8\pi c^2}(\dot{v}^2 + \dot{\phi}^2 v^2), \\ w_m(\mathbf{x}, t) &= \frac{1}{8\pi\mu_0}\{(\nabla v)^2 + v^2(\nabla\phi)^2\}. \end{aligned}\right\} \qquad \ldots\ldots(3.15)$$

Of particular interest is the case when the time enters only through a factor† $e^{-i\omega t}$. Then v is independent of time and ϕ is of the form

$$\phi(\mathbf{x}, t) = k\mathcal{S}(\mathbf{x}) - \omega t \qquad \ldots\ldots(3.16)$$

$(k = \omega/c)$. The eqns. (3.13) and (3.14) reduce to

$$(\nabla\mathcal{S})^2 - \frac{1}{k^2 v}\nabla^2 v = n^2, \qquad \ldots\ldots(3.17)$$

$$\nabla v . \nabla\mathcal{S} + \tfrac{1}{2} v\nabla^2\mathcal{S} = 0, \qquad \ldots\ldots(3.18)$$

where

$$n^2 = \epsilon_0\mu_0. \qquad \ldots\ldots(3.19)$$

The expressions (3.7) and (3.10) for the momentum densities and the energy densities become

$$\mathbf{g}(\mathbf{x}) = \frac{\epsilon_0 k^2}{4\pi n^2} v^2 \nabla\mathcal{S}, \qquad \ldots\ldots(3.20)$$

$$w_e(\mathbf{x}) = \frac{\epsilon_0 k^2}{8\pi} v^2, \qquad \ldots\ldots(3.21)$$

$$w_m(\mathbf{x}) = \frac{\epsilon_0 k^2}{8\pi n^2} v^2 \left[(\nabla\mathcal{S})^2 + \frac{1}{k^2}(\nabla \log v)^2 \right]. \qquad \ldots\ldots(3.22)$$

† *Note added in proof*: A general monochromatic wave must be represented by a V function of the form

$$V(\mathbf{x}, t) = V_1(\mathbf{x})e^{-i\omega t} + V_2(\mathbf{x})e^{i\omega t},$$

where V_1 and V_2 are complex functions of positions. It appears that the case here considered $(V_2 \equiv 0)$ represents a monochromatic wave of arbitrary shape but circularly polarized, at least in the sense of a space average over a macroscopic domain.

Equation (3.20) shows that the energy flow is orthogonal to the surfaces

$$\mathscr{S}(\mathbf{x}) = \text{constant}. \qquad \ldots\ldots(3.23)$$

Moreover, (3.18) may be written in the form

$$\nabla . (v^2 \nabla \mathscr{S}) = 0; \qquad \ldots\ldots(3.24)$$

this relation (which also follows from the conservation law (3.11)) shows that the vector field $\mathbf{g}$ is solenoidal.

The two results which we have just established show a close analogy with the model presented by geometrical optics. The surfaces $\mathscr{S}$ = constant may be regarded as generalized wave fronts and the energy may be considered as being propagated along the curves (rays) orthogonal to these surfaces.

To determine the wave fronts $\mathscr{S}$ = constant one must in general solve two equations (3.17) and (3.18). But we note that the second term in (3.17) contains as multiplicative factor the second power of the small quantity $1/k$. Except in special regions this term will be very small in comparison with $n^2 = \epsilon_0 \mu_0$ and may therefore be neglected. This implies that the geometrical wave fronts given by the solution of the eikonal equation

$$(\nabla \mathscr{S})^2 = n^2 \qquad \ldots\ldots(3.25)$$

are in general a very good approximation to the waves associated with the complex potential.† Better approximations may be obtained by the application of the W.K.B. method, or with the help of Huygens' principle (or Kirchhoff's integral formula).

One can also easily write down an expression for the variation of the amplitude along each 'electromagnetic ray' (orthogonal trajectory to the $\mathscr{S}$ surfaces). For if $\partial/\partial\tau$ denotes differentiation along a particular ray, $\partial/\partial\tau = \nabla\mathscr{S} . \nabla$, and (3.18) gives $\partial v/\partial\tau + \frac{1}{2}(\nabla^2 \mathscr{S})v = 0$; hence

$$v(\tau) = v(\tau_0) \exp\left[-\tfrac{1}{2} \int_{\tau_0}^{\tau} \nabla^2 \mathscr{S} \, d\tau \right]. \qquad \ldots\ldots(3.26)$$

§4. A Comparison with the Classical Theory

The energy and the momentum of the electromagnetic field may be derived by a variational principle from a Lagrangian density L' which is Lorentz invariant:

$$L' = \frac{1}{8\pi} \left\{ \frac{\epsilon_0}{c^2} \dot{\mathbf{A}}^2 - \frac{1}{\mu_0} (\nabla \wedge \mathbf{A})^2 \right\}. \qquad \ldots\ldots(4.1)$$

The variational principle leads to the vector wave equation (2.10) for the vector potential $\mathbf{A}$.

In the present theory we have, in place of (2.10), the scalar wave equation (2.9) for the complex potential V. This equation may also be derived from the variational principle by replacing L' by the Lagrangian density‡

$$L = \frac{1}{8\pi} \left\{ \frac{\epsilon_0}{c^2} \dot{V} \dot{V}^* - \frac{1}{\mu_0} \nabla V . \nabla V^* \right\}. \qquad \ldots\ldots(4.2)$$

† In this connection mention must be made of recent researches of Luneberg (cf. 1949, particularly pp. 55-56), which have shown that for small wavelengths geometrical optics, in its extended form, including also description of vectorial properties of the field, gives a very good first approximation to a steady state field.

‡ It is of interest to note the analogy with the Lagrangian density for a scalar meson with zero rest mass.

Since L' and L lead to the same field equations, their difference must be of the form

$$L - L' = \dot{P} + \operatorname{div} \mathbf{Q}. \qquad \ldots\ldots (4.3)$$

In consequence, the momentum density $\mathbf{g}$ and the energy density w of the present theory will differ from the usual expressions $\mathbf{g}'$ and w' associated with (4.1). In fact one has relations of the type

$$\mathbf{g} - \mathbf{g}' = \operatorname{grad} \hat{P}, \qquad w - w' = -\operatorname{div} \hat{\mathbf{Q}}, \qquad \ldots\ldots (4.4)$$

where $\hat{P} = P$ and $\hat{\mathbf{Q}} = \mathbf{Q}$ if P and $\mathbf{Q}$ do not involve the field variables. The effect of these differences will be negligible on integration over an arbitrary domain which is large compared with the wavelength.

That the usual definition of energy flow, based on the Poynting vector, is not the only possible one is of course well known.* Our present definition has several advantages. We saw, for example, that in a field represented by a complex potential of the form

$$V(\mathbf{x}, t) = V_0(\mathbf{x}) e^{-i\omega t} \qquad \ldots\ldots (4.5)$$

$\mathbf{g}$, w_e and w_m are independent of time and $\mathbf{g}$ is orthogonal to the surfaces of constant phase of the complex potential. But even in more complicated fields the densities as defined by the present theory seem to lead to simpler results. To illustrate this point we consider an example (essentially due to Braunbek 1951):

Suppose that the vector potential of a field *in vacuo* ($\epsilon_0 = \mu_0 = 1$) is

$$\mathbf{A}(\mathbf{x}, t) = C\{\mathbf{n}_2 \cos[k(\mathbf{n}_1 . \mathbf{x} - ct)] + \mathbf{n}_3 \cos[k(\mathbf{n}_2 . \mathbf{x} - ct)], \quad \ldots\ldots (4.6)$$

where C is a constant and $\mathbf{n}_1$, $\mathbf{n}_2$, $\mathbf{n}_3$ is a right-handed orthogonal triad of unit vectors. A straightforward calculation gives for the time average

$$\langle \mathbf{S}' \rangle = \left\langle \frac{c(\mathbf{E} \wedge \mathbf{H})}{4\pi} \right\rangle$$

of the Poynting vector:

$$\langle \mathbf{S}' \rangle = \frac{ck^2 C}{8\pi} [(\mathbf{n}_1 + \mathbf{n}_2) - \mathbf{n}_2 \cos[k(\mathbf{n}_1 - \mathbf{n}_2) . \mathbf{x}]. \qquad \ldots\ldots (4.7)$$

This expression does not satisfy the relation $\langle \mathbf{S}' \rangle . \operatorname{curl} \langle \mathbf{S}' \rangle = 0$ which would obtain if $\langle \mathbf{S}' \rangle$ possessed orthogonal trajectories (cf. Weatherburn 1930, p. 217). On the other hand one easily finds that the average energy flux $\mathbf{S} = c\mathbf{g}$ as defined by (3.7) possess orthogonal trajectories. For one has in this case (taking $\mathbf{n} = \mathbf{n}_1 + \mathbf{n}_2$)

$$V(\mathbf{x}, t) = C\left\{\frac{1}{\sqrt{2}}(1-i)\cos[k(\mathbf{n}_1 . \mathbf{x} - ct] + i\cos[k(\mathbf{n}_2 . \mathbf{x} - ct)]\right\}. \quad \ldots\ldots (4.8)$$

From (4.8) and (3.7) one obtains, by a straightforward calculation,

$$\langle \mathbf{S} \rangle = \langle c\mathbf{g} \rangle = \frac{ck^2 C}{8\pi} [\mathbf{n}_1 + \mathbf{n}_2] \left[1 - \frac{1}{\sqrt{2}} \cos(k(\mathbf{n}_1 - \mathbf{n}_2) . \mathbf{x})\right], \quad \ldots\ldots (4.9)$$

and it follows from (4.9) that $\langle \mathbf{S} \rangle$ satisfied the required condition for the existence of orthogonal trajectories: $\langle \mathbf{S} \rangle . \operatorname{curl} \langle \mathbf{S} \rangle = 0.$

It is seen that $\langle \mathbf{S}' \rangle$ and $\langle \mathbf{S} \rangle$ differ by a space-periodic function, to which clearly no physical meaning can be attached. For $\operatorname{div}\{\langle \mathbf{S}' \rangle - \langle \mathbf{S} \rangle\} = 0$, and consequently the (time-averaged) energy which crosses any closed surface in the field, will be the same whether $\mathbf{S}$ or $\mathbf{S}'$ is taken to define the energy flow.

* cf. Stratton 1941, p. 134; also Hines 1952.

Finally we give a summary of our main formula and the corresponding formulae of the classical theory:

	Usual definition	Present definition
Basic field quantity	$\mathbf{A}$ $\left(\mathbf{E}=-\frac{1}{c}\dot{\mathbf{A}},\ \mu_0\mathbf{H}=\nabla\wedge\mathbf{A}\right)$	V
Lagrangian density	$\frac{1}{8\pi}\left\{\frac{\epsilon_0}{c^2}\dot{\mathbf{A}}^2-\frac{1}{\mu_0}(\nabla\wedge\mathbf{A})^2\right\}$	$\frac{1}{8\pi}\left\{\frac{\epsilon_0}{c^2}\dot{V}\dot{V}^*-\frac{1}{\mu_0}\nabla V\,.\,\nabla V^*\right\}$
Field equation	$\nabla^2\mathbf{A}-\frac{\epsilon_0\mu_0}{c^2}\ddot{\mathbf{A}}=0$	$\nabla^2 V-\frac{\epsilon_0\mu_0}{c^2}\ddot{V}=0$
Momentum density	$\frac{1}{4\pi}(\mathbf{E}\wedge\mathbf{H})$	$-\frac{1}{8\pi\mu_0 c}(\dot{V}^*\nabla V+\dot{V}\nabla V^*)$
Electric energy density	$\frac{\epsilon_0}{8\pi}\,\mathbf{E}^2$	$\frac{\epsilon_0}{8\pi c^2}\dot{V}\dot{V}^*$
Magnetic energy density	$\frac{\mu_0}{8\pi}\,\mathbf{H}^2$	$\frac{1}{8\pi\mu_0}\nabla V\,.\,\nabla V^*$

Acknowledgment

In conclusion we wish to thank Professor Max Born for his interest in this work and for stimulating discussions.

References

Braunbek, W., 1951, *Z. Naturforsch.*, **6a,** 12.

Hines, C. O., 1952, *Canad. J. Phys.*, **30,** 123.

Luneberg, R. K., 1947–8, *Propagation of Electromagnetic Waves* (mimeographed notes of lectures delivered at New York University); 1949, *Asymptotic Development of Steady State Electromagnetic Fields, New York University, Mathematics Research Group,* Report No. EM-14.

Picht, J., 1931, *Optische Abbildung* (Braunschweig : Vieweg).

Stratton, J. A., 1941, *Electromagnetic Theory* (New York : McGraw-Hill).

Theimer, O., Wassermann, G. D., and Wolf, E., 1952, *Proc. Roy. Soc.* A, **212,** 426.

Weatherburn, C. E., 1930, *Differential Geometry of Three Dimensions,* Vol. II (Cambridge : University Press).

3.2

Reprinted from the *Proceedings of the Cambridge Philosophical Society*,
Volume 50, Part 4, pp. 614–622, 1954.
PRINTED IN GREAT BRITAIN

ON LINEARLY POLARIZED ELECTROMAGNETIC WAVES OF ARBITRARY FORM

BY A. NISBET AND E. WOLF

Communicated by N. KEMMER

Received 1 February 1954

ABSTRACT. Two simple laws connecting the amplitude and phase functions of a monochromatic electromagnetic wave of arbitrary form are derived, holding in the case when one of the field vectors is linearly polarized. The first is a generalized Fermat's principle which enables determination of the phase when the amplitude is known; the second expresses the propagation of the (vector) amplitude along the curves orthogonal to the co-phasal surfaces. Some other general properties of linearly polarized fields are also discussed, and illustrative examples are given.

1. *Introduction.* In recent years researches on the foundations of geometrical optics, and in particular the important investigations of Luneberg (6), have shown that it is possible to formulate simple transport laws relating not only to the propagation of energy but also the propagation of amplitude and direction of polarization of monochromatic electromagnetic fields in the limit of infinitely high frequency (negligible wave-length).

The present investigation is concerned with the question of extending these laws to fields of any frequency. It is shown that rigorous generalizations holding for any monochromatic electromagnetic field in which at least one of the field vectors is linearly polarized† may in fact be obtained. In particular, the following two laws are established for such fields:

(1) A relation between the vector amplitude and the phase which is of the form of a generalized eikonal equation. The associated Fermat's variational problem then enables the determination of the phase when the amplitude is known.

(2) A transport equation for the propagation of the vector amplitude along the orthogonal trajectories ('electromagnetic rays') of the co-phasal surfaces. This is found to be formally identical with the Luneberg-Friedlander transport equation.

Certain other general properties of linearly polarized fields are discussed and the results are illustrated with reference to the classical Sommerfeld half-plane problem, to the magnetic dipole field and to a wider class of fields of practical interest.

2. *Maxwell's equations in terms of amplitude and phase functions.* 2·1. We begin with a few preliminary remarks concerning polarization, amplitudes and phases of a monochromatic vector field.

A real vector field $\mathbf{G}(\mathbf{r},t)$ which is harmonic in the time t, with circular frequency ω, may be written

$$\begin{aligned}\mathbf{G}(\mathbf{r},t) &= \mathbf{p}(\mathbf{r})\cos\omega t+\mathbf{q}(\mathbf{r})\sin\omega t\\ &= \Re\mathbf{g}(\mathbf{r})\,e^{-i\omega t},\end{aligned} \tag{2·1}$$

† By a linearly polarized monochromatic wave we mean one in which the field vector at any particular point remains, with varying time, always in the same direction; but unlike the case of a homogeneous plane wave, this direction may be different at different points of the field. As will be shown below, many fields of practical interest are of this type.

where $\mathbf{g} = \mathbf{p} + i\mathbf{q}$, $\mathbf{p}$ and $\mathbf{q}$ being real vector functions of position $\mathbf{r}$, and $\Re$ denotes the real part. At any given point the vector $\mathbf{G}$ lies in a plane, the plane of $\mathbf{p}$ and $\mathbf{q}$; and its end-point describes with varying time an ellipse (the polarization ellipse) of which $\mathbf{p}$ and $\mathbf{q}$ are conjugate semi-diameters and $\mathbf{r}$ the centre. In special cases the ellipse may of course degenerate into a circle or a straight line, the field then being circularly or linearly polarized. It is important to realize that, contrary to the case of a plane wave, *the state of polarization may be different at different points of the field; and in particular that if* $\mathbf{G}$ *is linearly polarized throughout a given domain, it is not necessarily in the same direction at every point of the domain.* Such a variation is exhibited even by relatively simple fields, as, for example, that of an electric or magnetic dipole (see §4·3 below).

The field $\mathbf{G}$ may be represented in an infinite number of ways, namely,

$$\mathbf{G}(\mathbf{r}, t) = \Re\mathbf{g}'(\mathbf{r})\, e^{i\{k\mathscr{S}(\mathbf{r}) - \omega t\}}, \tag{2·2}$$

where

$$\mathbf{g}'\, e^{ik\mathscr{S}} = \mathbf{g}. \tag{2·3}$$

Here the amplitude function $\mathbf{g}' = \mathbf{p}' + i\mathbf{q}'$, $\mathbf{p}'$ and $\mathbf{q}'$ are any other pair of conjugate semi-diameters of the polarization ellipse, $\mathscr{S}$ is the associated phase function and k is the usual factor ω/c, c being the velocity of light *in vacuo*.

We shall mainly be concerned with linearly polarized waves. Then $\mathbf{p}$ and $\mathbf{q}$ are in the same direction, i.e.

$$\mathbf{q} = f\mathbf{p}, \tag{2·4}$$

where f is a real scalar function of position. In this case we may choose the amplitude function to be real and write†

$$\mathbf{G}(\mathbf{r}, t) = \mathbf{a}(\mathbf{r})\, e^{i\{k\mathscr{S}(\mathbf{r}) - \omega t\}}, \tag{2·5}$$

where

$$\mathbf{a} = \mathbf{p}\, \sqrt{(1 + f^2)} \tag{2·6}$$

and

$$\tan k\mathscr{S} = f. \tag{2·7}$$

We shall call (2·5) the *principal form of* $\mathbf{G}$, and $\mathbf{a}$ and $\mathscr{S}$ the *principal amplitude function* and *principal phase function* respectively.

We can also define the principal form of $\mathbf{G}$ for the general case of elliptic polarization as that corresponding to the choice of $\mathbf{p}'$ and $\mathbf{q}'$ along principal axes of the polarization ellipse. Definition (2·5) then corresponds to the limiting case when the minor axis is zero.

2·2. We now consider a monochromatic electromagnetic field in an isotropic non-conducting medium, of dielectric 'constant' $\epsilon(\mathbf{r})$ and magnetic permeability $\mu(\mathbf{r})$, free of currents and charges. We write the field vectors (in general elliptically polarized) in the form (2·2):

$$\left.\begin{aligned} \mathbf{E}(\mathbf{r}) &= \mathbf{e}(\mathbf{r})\, e^{i\{k\mathscr{S}_e(\mathbf{r}) - \omega t\}}, \\ \mathbf{H}(\mathbf{r}) &= \mathbf{h}(\mathbf{r})\, e^{i\{k\mathscr{S}_h(\mathbf{r}) - \omega t\}}. \end{aligned}\right\} \tag{2·8}$$

Substitution of these expressions into Maxwell's equations

$$\nabla \wedge \mathbf{E} - ik\mu\mathbf{H} = 0, \tag{2·9}$$

$$\nabla \wedge \mathbf{H} + ik\epsilon\mathbf{E} = 0, \tag{2·10}$$

$$\nabla \cdot (\mu\mathbf{H}) = 0, \tag{2·11}$$

$$\nabla \cdot (\epsilon\mathbf{E}) = 0, \tag{2·12}$$

† We omit the symbol $\Re$ in (2·5), and subsequently when there is no risk of confusion.

leads to
$$\mathbf{h}\, e^{ik\mathscr{S}_h} = \left\{\frac{1}{\mu}\nabla\mathscr{S}_e \wedge \mathbf{e} + \frac{1}{ik\mu}\nabla\wedge\mathbf{e}\right\} e^{ik\mathscr{S}_e}, \tag{2·13}$$
$$\mathbf{e}\, e^{ik\mathscr{S}_e} = \left\{-\frac{1}{\epsilon}\nabla\mathscr{S}_h \wedge \mathbf{h} - \frac{1}{ik\epsilon}\nabla\wedge\mathbf{h}\right\} e^{ik\mathscr{S}_h}, \tag{2·14}$$
$$\mathbf{h}\,.\nabla\mathscr{S}_h + \frac{1}{ik\mu}\nabla\,.(\mu\mathbf{h}) = 0, \tag{2·15}$$
$$\mathbf{e}\,.\nabla\mathscr{S}_e + \frac{1}{ik\epsilon}\nabla\,.(\epsilon\mathbf{e}) = 0. \tag{2·16}$$
Elimination of the magnetic vector gives the wave equation for the electric vector, which, in terms of $\mathbf{e}$ and $\mathscr{S}_e$, has the form
$$\mathscr{K}(\mathbf{e}, \nabla\mathscr{S}_e, n) + \frac{i}{k}\mathscr{L}(\mathbf{e}, \nabla\mathscr{S}_e, n, \mu) + \frac{1}{k^2}\mathscr{M}(\mathbf{e}, \epsilon, \mu) = 0, \tag{2·17}$$
where
$$n^2 = \epsilon\mu, \tag{2·18}$$
and
$$\left.\begin{aligned}\mathscr{K}(\mathbf{e}, \nabla\mathscr{S}_e, n) &= \{n^2 - (\nabla\mathscr{S}_e)^2\}\,\mathbf{e},\\ \mathscr{L}(\mathbf{e}, \nabla\mathscr{S}_e, n, \mu) &= \{\nabla^2\mathscr{S}_e - \nabla\mathscr{S}_e.\nabla\log\mu\}\,\mathbf{e} + 2(\mathbf{e}.\nabla\log n)\nabla\mathscr{S}_e + 2(\nabla\mathscr{S}_e.\nabla)\mathbf{e},\\ \mathscr{M}(\mathbf{e}, \epsilon, \mu) &= \nabla^2\mathbf{e} + (\nabla\log\mu)\wedge(\nabla\wedge\mathbf{e}) + \nabla(\mathbf{e}.\nabla\log\epsilon).\end{aligned}\right\} \tag{2·19}$$
The corresponding relation between $\mathbf{h}$ and $\mathscr{S}_h$ is
$$\mathscr{K}(\mathbf{h}, \nabla\mathscr{S}_h, n) + \frac{i}{k}\mathscr{L}(\mathbf{h}, \nabla\mathscr{S}_h, n, \epsilon) + \frac{1}{k^2}\mathscr{M}(\mathbf{h}, \mu, \epsilon) = 0. \tag{2·20}$$

We shall now investigate the implications of these relations for electromagnetic fields in which the electric vector is linearly polarized. Results for the complementary case of linearly polarized magnetic vector are of course strictly analogous, and may be obtained by interchange of $\mathbf{E}$ and $\mathbf{H}$, $\mathbf{e}$ and $\mathbf{h}$, $\mathscr{S}_e$ and $\mathscr{S}_h$, and ϵ and $-\mu$ in the appropriate formulae.

3. *Electromagnetic fields with linearly polarized* **E**-*vector*. When the electric vector is linearly polarized we use the principal form (2·5); then $\mathbf{e}(\mathbf{r}) = \mathbf{a}(\mathbf{r})$, where $\mathbf{a}$ is real. Separating real and imaginary parts in (2·17) then gives†
$$\mathscr{K}(\mathbf{a}, \nabla\mathscr{S}, n) + \frac{1}{k^2}\mathscr{M}(\mathbf{a}, \epsilon, \mu) = 0, \tag{3·1}$$
and
$$\mathscr{L}(\mathbf{a}, \nabla\mathscr{S}, n, \mu) = 0. \tag{3·2}$$
Also, the divergence condition (2·16) gives
$$\mathbf{a}\,.\nabla\mathscr{S} = 0 \tag{3·3}$$
and
$$\nabla\,.\mathbf{a} + \mathbf{a}\,.\nabla\log\epsilon = 0. \tag{3·4}$$
We now show that these equations have interesting interpretations.

3·1. *A generalization of Fermat's principle.* Written explicitly, (3·1) is
$$\{n^2 - (\nabla\mathscr{S})^2\}\,\mathbf{a} + \frac{1}{k^2}\{\nabla^2\mathbf{a} + (\nabla\log\mu)\wedge(\nabla\wedge\mathbf{a}) + \nabla(\mathbf{a}.\nabla\log\epsilon)\} = 0, \tag{3·5}$$

† In this section we write $\mathscr{S}$ in place of $\mathscr{S}_e$, as no confusion is likely to occur.

giving, after scalar multiplication by **a** and use of (3·4),

$$(\nabla\mathscr{S})^2 = n^2 - \frac{\mu}{k^2a^2}\mathbf{a}\,.\left\{\nabla\wedge\left(\frac{1}{\mu}\nabla\wedge\mathbf{a}\right)\right\}. \tag{3·6}$$

We note that as $k\to\infty$, (3·6) reduces correctly to the *eikonal equation* of geometrical optics:

$$(\nabla\mathscr{S})^2 = n^2. \tag{3·7}$$

In this limiting case, therefore, $\mathscr{S}$ is completely specified (apart from boundary conditions) by the refractive index function n alone. The surfaces $\mathscr{S}(\mathbf{r})$ = constant are then the wave fronts of geometrical optics. Their orthogonal trajectories are the geometrical optics rays, which are the solution of Fermat's variational problem

$$\delta\int n\,ds = 0. \tag{3·8}$$

In general, however, the second term on the right-hand side of (3·6) cannot be neglected; but if $\mathbf{a}(\mathbf{r})$ is assumed to be known, e.g. from experiment, the principal phase function may even then be determined—with the help of a (generalized) Fermat principle. For if we introduce a modified refractive index function

$$N(\mathbf{r}) = \sqrt{\left[n^2 - \frac{\mu}{k^2a^2}\mathbf{a}\,.\left\{\nabla\wedge\left(\frac{1}{\mu}\nabla\wedge a\right)\right\}\right]}, \tag{3·9}$$

(3·6) takes the form of a generalized eikonal equation

$$(\nabla\mathscr{S})^2 = N^2. \tag{3·10}$$

Hence if the principal amplitude $\mathbf{a}(\mathbf{r})$ is regarded as known, *the principal phase function $\mathscr{S}$ is nothing but Hamilton's characteristic function of the variational problem*

$$\delta\int N\,ds = 0. \tag{3·11}$$

We note the simple forms taken by N in the following special cases:

$$\left.\begin{aligned} &\text{(i) } \mu = 1: && N^2 = n^2 - \frac{1}{k^2a^2}\mathbf{a}\,.\{\nabla\wedge(\nabla\wedge\mathbf{a})\},\\ &\text{(ii) } \epsilon \text{ and } \mu \text{ constant:} && N^2 = n^2 + \frac{1}{k^2a^2}\mathbf{a}\,.\nabla^2\mathbf{a}. \end{aligned}\right\} \tag{3·12}$$

3·2. *Transport equation for the principal amplitude.* Next we consider the implication of (3·2), which is in full

$$(\nabla^2\mathscr{S} - \nabla\mathscr{S}\,.\,\nabla\log\mu)\,\mathbf{a} + 2(\mathbf{a}\,.\,\nabla\log n)\,\nabla\mathscr{S} + 2(\nabla\mathscr{S}\,.\,\nabla)\,\mathbf{a} = 0. \tag{3·13}$$

Introducing the operator

$$\partial/\partial\tau = \nabla\mathscr{S}\,.\,\nabla, \tag{3·14}$$

so that τ is a parameter specifying position along the orthogonal trajectories ('electromagnetic rays') to the surfaces $\mathscr{S}$ = constant, (3·13) becomes

$$\frac{\partial\mathbf{a}}{\partial\tau} + \tfrac{1}{2}\mathbf{a}\mu\nabla\,.\left(\frac{1}{\mu}\nabla\mathscr{S}\right) + (\mathbf{a}\,.\,\nabla\log n)\,\nabla\mathscr{S} = 0. \tag{3·15}$$

This gives the variation in the principal amplitude **a** along each ray. It is formally identical with the transport equation of Luneberg (6)† for the propagation of

† An account of some of Luneberg's analysis is also given in a paper by Kline (5). Luneberg's proof is based on the full Maxwell's equations, but it has been shown by Copson (3) that the transport equation is a consequence of the second order wave equation for each field vector.

discontinuities in an electromagnetic field, and with that of Friedlander (4)† valid within the accuracy of geometrical optics, for the propagation of the vector amplitude of a monochromatic field. However, in these papers the $\mathscr{S}$-functions satisfy rigorously the eikonal equation (3·7), whereas ours satisfies the more general equation (3·10).

The transport equation (3·15) may be regarded as complementary to the generalized eikonal equation (3·10). For (3·15) enables the determination of $\mathbf{a}$ when $\mathscr{S}$ is known, whereas (3·10) enables the determination of $\mathscr{S}$ when $\mathbf{a}$ is known. The content of (3·13) (or (3·15)) can be conveniently discussed in two stages, one relating to the variation of the squared amplitude along each 'ray', the other concerned with the variation of the direction of polarization.‡

We note first that (3·3) implies that *the electric vector lies in the tangent plane to the surface $\mathscr{S}$ = constant*. Multiplying (3·13) scalarly by $\mathbf{a}$ and using this result, it follows that

$$\nabla.\left(\frac{a^2}{\mu}\nabla\mathscr{S}\right)=0. \tag{3·16}$$

It will be seen later (§ 3·3) that (3·16) expresses the *law of conservation of the average energy flow (Poynting vector)*.

Equation (3·16) may also be written in the form

$$\frac{\partial}{\partial\tau}\log\frac{a^2}{\mu}=-\nabla^2\mathscr{S}. \tag{3·17}$$

Hence if a^2/μ is known at a particular point $\tau=\tau_0$ of a ray its value at any point on the ray is given by

$$\left(\frac{a^2}{\mu}\right)_\tau=\left(\frac{a^2}{\mu}\right)_{\tau_0}\exp\left\{-\int_{\tau_0}^{\tau}\nabla^2\mathscr{S}\,d\tau\right\}, \tag{3·18}$$

or, using (3·14), by the equivalent form

$$\left(\frac{a^2}{\mu}\right)_{\mathscr{S}}=\left(\frac{a^2}{\mu}\right)_{\mathscr{S}_0}\exp\left\{-\int_{\mathscr{S}_0}^{\mathscr{S}}\frac{\nabla^2\mathscr{S}}{(\nabla\mathscr{S})^2}d\mathscr{S}\right\}, \tag{3·19}$$

integration being along the ray in each case.

Next we derive from (3·15) a simple equation for the propagation along each ray of the unit vector $\mathbf{u}$ which specifies the direction of the electric vector. Substituting $\mathbf{a}=\mathbf{u}\sqrt{(\mathbf{a}^2)}$ into (3·15), and using (3·17), gives, after a simple calculation, the transport equation

$$\frac{\partial\mathbf{u}}{\partial\tau}+(\mathbf{u}.\nabla\log n)\nabla\mathscr{S}=0. \tag{3·20}$$

In particular, if the medium is homogeneous, $\nabla\log n=0$, and (3·20) gives $\partial\mathbf{u}/\partial\tau=0$, implying that in a homogeneous medium the direction of polarization remains constant along each ray.

3·3. *The average Poynting vector*. Equation (2·13) with $\mathbf{e}=\mathbf{a}$, where $\mathbf{a}$ is real, shows that in general the magnetic vector will be elliptically polarized even when the electric

† Though not explicitly stated by Friedlander, the later part of his paper is also restricted to linearly polarized fields.

‡ Our discussion of (3·15) is similar to that of Luneberg and Friedlander but is included here for the sake of completeness and also because our standpoint is different. For we are here concerned neither with a pulse solution (Luneberg) nor with geometrical optics (Friedlander), but with a (rigorous) solution of Maxwell's equations representing a monochromatic field with linearly polarized electric (or magnetic) field vector.

vector is everywhere linearly polarized. An example of this situation is the field of a magnetic Hertzian oscillator (see §4·3). It is, however, easy to see that the time average of $\mathbf{E}.\mathbf{H}$ over an interval of time which is large compared to the period $2\pi/\omega$ is zero, implying that $\mathbf{E}$ *and* $\mathbf{H}$ *are 'on the average' mutually orthogonal*. For, by (2·13),

$$\overline{\mathbf{E}.\mathbf{H}} = \tfrac{1}{2}\Re\mathbf{E}.\mathbf{H}^* = \tfrac{1}{2}\Re\left\{\frac{1}{\mu}\mathbf{a}.(\nabla\mathscr{S}\wedge\mathbf{a}) - \frac{1}{ik\mu}\mathbf{a}.(\nabla\wedge\mathbf{a})\right\} = 0. \tag{3·21}$$

Further, the time average of the Poynting vector $\mathbf{S}$ is given by

$$\overline{\mathbf{S}} = \frac{c}{8\pi}\Re\mathbf{E}\wedge\mathbf{H}^* = \frac{c}{8\pi}\Re\left\{\mathbf{a}\wedge\left(\frac{1}{\mu}\nabla\mathscr{S}\wedge\mathbf{a} - \frac{1}{ik\mu}\nabla\wedge\mathbf{a}\right)\right\} = \frac{c}{8\pi\mu}a^2\nabla\mathscr{S}, \tag{3·22}$$

showing that *in a field where* $\mathbf{E}$ *is linearly polarized the mean Poynting vector is orthogonal to the co-phasal surfaces* $\mathscr{S} = constant$ of the electric vector. Equations (3·22) and (3·16) then give the conservation law

$$\nabla.\overline{\mathbf{S}} = 0. \tag{3·23}$$

4. *Examples*. 4·1. Included in the class of problems to which the results of the previous section apply are what are essentially orthogonal two-parameter problems. Let $\theta_1, \theta_2, \theta_3$ be orthogonal curvilinear coordinates, defined by the metric

$$ds^2 = \gamma_1^2 d\theta_1^2 + \gamma_2^2 d\theta_2^2 + \gamma_3^2 d\theta_3^2, \tag{4·1}$$

where $$\gamma_j = \gamma_j(\theta_1, \theta_2) \quad (j = 1, 2, 3). \tag{4·2}$$

Then if $\mathbf{E}$, $\mathbf{H}$, ϵ and μ are also independent of θ_3, Maxwell's equations split up into two independent sets. The first set involves E_3, H_1 and H_2 only, and therefore represents a field with a linearly polarized $\mathbf{E}$-vector, while the other involves H_3, E_1 and E_2 only, and therefore represents a field with a linearly polarized $\mathbf{H}$-vector. The results of the previous section (with appropriate modification in the second case) are therefore immediately applicable to each of these two fields.

4·2. An interesting example in this category is the field obtained by the diffraction of an infinite plane wave by a semi-infinite perfectly reflecting screen, first solved rigorously by Sommerfeld in a well-known paper. Recent diagrams of Braunbek and Laukien (2) (Figs. 1*a*, *b*) based on Sommerfeld's solution, illustrate the general conclusion (3·22) that the mean Poynting vector is orthogonal to the co-phasal surfaces of the appropriate field vector†.

4·3. Another simple example is furnished by the field of the Hertzian oscillator. For a magnetic dipole of moment $\mathbf{m}\,e^{-i\omega t}$ *in vacuo* ($\epsilon = \mu = 1$), we have (see any standard textbook)

$$\mathbf{E} = \frac{k^2}{r}\left(1 + \frac{i}{kr}\right)\mathbf{m}\wedge\hat{\mathbf{r}}\,e^{i(kr-\omega t)}, \tag{4·3}$$

$$\mathbf{H} = \frac{k^2}{r}\left\{\left(1 + \frac{i}{kr} - \frac{1}{k^2r^2}\right)\mathbf{m} - \left(1 + \frac{3i}{kr} - \frac{3}{k^2r^2}\right)(\mathbf{m}.\hat{\mathbf{r}})\hat{\mathbf{r}}\right\}e^{i(kr-\omega t)}, \tag{4·4}$$

† For the case of 'plane fields' and fields with rotational symmetry the orthogonality of the mean Poynting vector to the co-phasal surfaces associated with a component of one of the field vectors has been pointed out previously by Braunbek (1).

where r is the distance from the dipole and $\hat{\mathbf{r}}$ is the unit radial vector. The electric field is here linearly polarized, though the magnetic field is in general elliptically polarized. The principal form of $\mathbf{E}$ is immediately seen to be

$$\mathbf{E} = \mathbf{a}\, e^{i(k\mathscr{S}_e - \omega t)}, \tag{4·5}$$

with

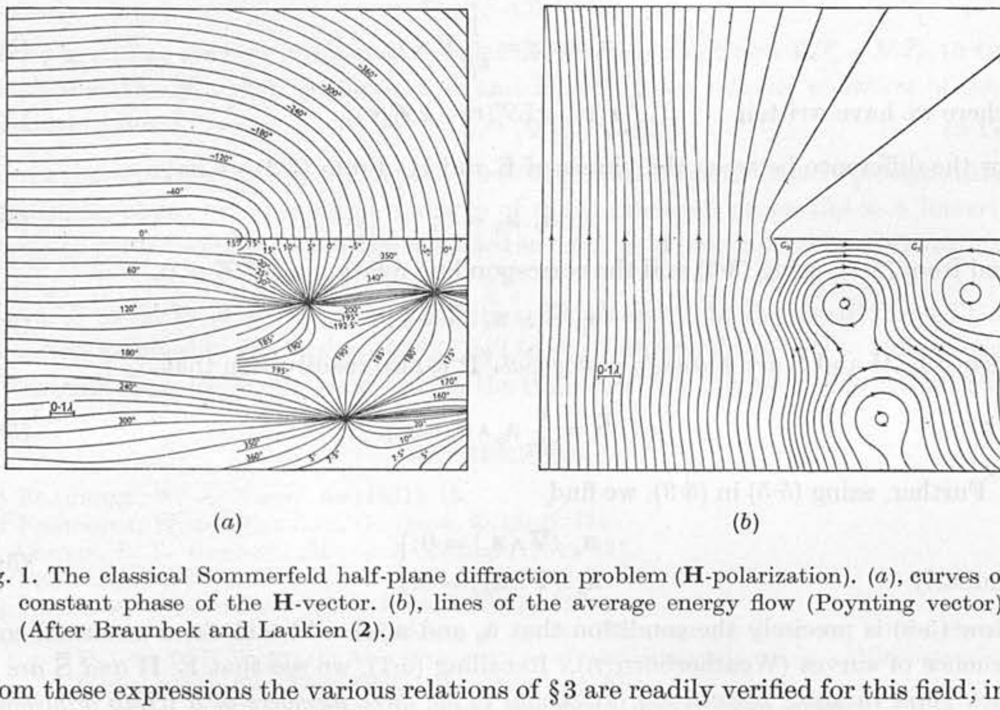
$$\mathbf{a} = \frac{k^2}{r}\left(1 + \frac{1}{k^2 r^2}\right)^{\frac{1}{2}} \mathbf{m} \wedge \hat{\mathbf{r}} \tag{4·6}$$

and

$$k\mathscr{S}_e = kr + \tan^{-1}(1/kr), \quad (0 < \tan^{-1}(1/kr) < \tfrac{1}{2}\pi). \tag{4·7}$$

In terms of the phase function $\mathscr{S}_e$ we also have

$$\mathbf{H} = \frac{k^2}{r}\left(1 + \frac{1}{k^2 r^2}\right)^{-\frac{1}{2}} \left\{\left(1 + \frac{i}{k^3 r^3}\right)\mathbf{m} - \left(1 + \frac{2i}{kr} + \frac{3i}{k^3 r^3}\right)(\mathbf{m}\,.\,\hat{\mathbf{r}})\,\hat{\mathbf{r}}\right\} e^{i(k\mathscr{S}_e - \omega t)}. \tag{4·8}$$



(a) (b)

Fig. 1. The classical Sommerfeld half-plane diffraction problem (**H**-polarization). (*a*), curves of constant phase of the **H**-vector. (*b*), lines of the average energy flow (Poynting vector). (After Braunbek and Laukien (**2**).)

From these expressions the various relations of § 3 are readily verified for this field; in particular, the time-averaged Poynting vector is given by

$$\overline{\mathbf{S}} = \frac{c}{8\pi}\,\Re\,\mathbf{E} \wedge \mathbf{H}^* = \frac{c}{8\pi}\frac{k^4}{r^2}(\mathbf{m} \wedge \hat{\mathbf{r}}) \wedge \{\mathbf{m} - (\mathbf{m}\,.\,\hat{\mathbf{r}})\,\hat{\mathbf{r}}\}$$

$$= \frac{c}{8\pi}\frac{k^4}{r^2}(\mathbf{m} \wedge \hat{\mathbf{r}})^2\,\hat{\mathbf{r}}. \tag{4·9}$$

Using (4·6) and (4·7), it then follows that everywhere (not only in the *distant* zone)

$$\overline{\mathbf{S}} = \frac{c}{8\pi} a^2 \nabla \mathscr{S}_e, \tag{4·10}$$

in agreement with (3·22).

5. *Electromagnetic fields with both* **E** *and* **H** *linearly polarized*. Finally, we consider the special case when the field vectors **E** and **H** are both linearly polarized. Then

$\mathbf{e} = \mathbf{a}_e$ and $\mathbf{h} = \mathbf{a}_h$ (say), where $\mathbf{a}_e$ and $\mathbf{a}_h$ are real, and all the results of §3 hold, not only for the **E** field but also for the **H** field (with appropriate changes). In addition, as we now show, this field possesses some further simplifying features, analogous to that of an ordinary plane wave.

The average Poynting vector may now be expressed in the two alternative forms:

$$\mathbf{S} = \frac{c}{8\pi\mu} a_e^2 \nabla \mathscr{S}_e = \frac{c}{8\pi\epsilon} a_h^2 \nabla \mathscr{S}_h, \tag{5·1}$$

implying that $\nabla\mathscr{S}_e$ and $\nabla\mathscr{S}_h$ are everywhere parallel.

Multiplying (2·13) by $e^{-ik\mathscr{S}_e}$ and equating real and imaginary parts gives

$$\mathbf{a}_h \cos\eta = \frac{1}{\mu} \nabla\mathscr{S}_e \wedge \mathbf{a}_e, \tag{5·2}$$

$$\mathbf{a}_h \sin\eta = \frac{1}{k\mu} \nabla \wedge \mathbf{a}_e, \tag{5·3}$$

where we have written
$$\eta(\mathbf{r}) = k\mathscr{S}_e(\mathbf{r}) - k\mathscr{S}_h(\mathbf{r}) \tag{5·4}$$

for the difference between the phases of **E** and **H**. From (5·2) we have

$$\mathbf{a}_e . \mathbf{a}_h = 0; \tag{5·5}$$

and from (5·1), using (3·3) and the corresponding relation $\mathbf{a}_h . \nabla\mathscr{S}_h = 0$,

$$\mathbf{a}_e . \overline{\mathbf{S}} = \mathbf{a}_h . \overline{\mathbf{S}} = 0. \tag{5·6}$$

Hence **E**, **H** and $\overline{\mathbf{S}}$ *are mutually orthogonal*. It is also readily seen that

$$\overline{\mathbf{S}} = \frac{c}{8\pi} \mathbf{a}_e \wedge \mathbf{a}_h \cos\eta. \tag{5·7}$$

Further, using (5·5) in (5·3), we find

$$\left.\begin{aligned} &\mathbf{a}_e . (\nabla \wedge \mathbf{a}_e) = 0; \\ \text{similarly}\quad &\mathbf{a}_h . (\nabla \wedge \mathbf{a}_h) = 0. \end{aligned}\right\} \tag{5·8}$$

Now (5·8) is precisely the condition that $\mathbf{a}_e$ and $\mathbf{a}_h$ should each form a normal congruence of curves (Weatherburn (7)). Recalling (5·1), we see that **E**, **H** *and* $\overline{\mathbf{S}}$ *are at each point in space respectively orthogonal to the three members of a triply orthogonal system of surfaces passing through that point.*

Eliminating $\mathbf{a}_h$ between (5·2) and the analogous relation

$$\mathbf{a}_e \cos\eta = -\frac{1}{\epsilon} \nabla\mathscr{S}_h \wedge \mathbf{a}_h \tag{5·9}$$

gives
$$\nabla\mathscr{S}_e . \nabla\mathscr{S}_h = \epsilon\mu \cos^2\eta. \tag{5·10}$$

With (5·1) this leads to
$$|\,\overline{\mathbf{S}}\,| = \frac{c}{8\pi} a_e a_h |\cos\eta|. \tag{5·11}$$

Further, the average energy density $\overline{W}$ is

$$\overline{W} = \frac{1}{8\pi} \overline{(\epsilon E^2 + \mu H^2)} = \frac{1}{16\pi} (\epsilon a_e^2 + \mu a_h^2), \tag{5·12}$$

so that the rate of flow, v, of the average energy density is given by

$$v = \frac{|\overline{\mathbf{S}}|}{\overline{W}} = \frac{c}{\sqrt{(\epsilon\mu)}}\left\{\frac{2\sqrt{(\epsilon\mu)}\,a_e a_h}{\epsilon a_e^2 + \mu a_h^2}\right\}|\cos\eta| \tag{5·13}$$

$$\leqslant \frac{c}{\sqrt{(\epsilon\mu)}}. \tag{5·14}$$

In the case when $\sin\eta = 0$, (5·4) and (5·1) give

$$\epsilon a_e^2 = \mu a_h^2, \tag{5·15}$$

so that the electric and magnetic energy densities are now equal; only in this case does the equality sign in (5·14) hold. Further, from (5·3) with $\sin\eta = 0$, we have, in place of (5·8), the stronger condition

$$\left.\begin{aligned} &\nabla\wedge\mathbf{a}_e = 0; \\ \text{similarly}\qquad &\nabla\wedge\mathbf{a}_h = 0. \end{aligned}\right\} \tag{5·16}$$

Hence $\mathbf{a}_e$ and $\mathbf{a}_h$ are now gradients of scalar fields. Also, since now $\nabla\mathscr{S}_e = \nabla\mathscr{S}_h$, (5·10) shows that the co-phasal surfaces of $\mathbf{E}$ and $\mathbf{H}$ satisfy the eikonal equation of geometrical optics:

$$(\nabla\mathscr{S}_e)^2 = \epsilon\mu, \quad (\nabla\mathscr{S}_h)^2 = \epsilon\mu. \tag{5·17}$$

We see therefore that an electromagnetic field with $\mathbf{E}$ and $\mathbf{H}$ both linearly polarized and differing in phase by an integral multiple of π has the same properties as a linearly polarized *plane* wave, and so can be regarded as its natural (curvilinear) generalization.

One of us (E. W.)† is indebted to Edinburgh University for the award of an I.C.I. Research Fellowship for Session 1953–4 and to the Carnegie Trust for the Universities of Scotland for a Research Grant during the time in which this work was carried out.

REFERENCES

(1) BRAUNBEK, W. *Z. Naturf.* 6*a* (1951), 12.
(2) BRAUNBEK, W. and LAUKIEN, G. *Optik*, 9 (1952), 174.
(3) COPSON, E. T. *Commun. pure appl. Math.* 4 (1951), 427.
(4) FRIEDLANDER, F. G. *Proc. Camb. phil. Soc.* 43 (1947), 284.
(5) KLINE, M. *Commun. pure appl. Math.* 4 (1951), 225.
(6) LUNEBERG, R. K. *Mathematical theory of optics.* (Mimeographed lecture notes, Brown University R.T., 1944). Also *Propagation of electromagnetic waves.* (Mimeographed lecture notes, New York University, 1947–8, p. 91.)
(7) WEATHERBURN, C. E. *Differential geometry*, vol. 2 (Cambridge, 1930), p. 217.

DEPARTMENT OF MATHEMATICAL PHYSICS
UNIVERSITY OF EDINBURGH

† Present address: Department of Astronomy, University of Manchester.

PART IV

RELATED INVESTIGATIONS

Reprinted from the *Proceedings of the Cambridge Philosophical Society*,
Volume 50, Part 1, pp. 40–48, 1954.
PRINTED IN GREAT BRITAIN

ON THE CIRCLE POLYNOMIALS OF ZERNIKE AND RELATED ORTHOGONAL SETS

BY A. B. BHATIA AND E. WOLF

Communicated by MAX BORN

Received 19 January 1953

ABSTRACT. The paper is concerned with the construction of polynomials in two variables, which form a complete orthogonal set for the interior of the unit circle and which are 'invariant in form' with respect to rotations of axes about the origin of coordinates. It is found that though there exist an infinity of such sets there is only one set which in addition has certain simple properties strictly analogous to that of Legendre polynomials. This set is found to be identical with the set of the circle polynomials of Zernike which play an important part in the theory of phase contrast and in the Nijboer-Zernike diffraction theory of optical aberrations.

The results make it possible to derive explicit expressions for the Zernike polynomials in a simple, systematic manner. The method employed may also be used to derive other orthogonal sets. One new set is investigated, and the generating functions for this set and for the Zernike polynomials are also given.

1. *Introduction*. In his paper on the phase-contrast test Zernike (7) introduced certain polynomials $U_n^m(x, y)$ which form a complete orthogonal set for the interior of the unit circle $x^2+y^2 \leqslant 1$ (x and y real). Apart from their usefulness in the theory of phase contrast, these polynomials play a fundamental role in the diffraction theory of optical aberrations as developed by Nijboer, partly in collaboration with Zernike (Nijboer (5) and (6), Zernike and Nijboer (9).)

Zernike obtained the circle polynomials as eigenfunctions of a certain second-order partial differential equation which is invariant with respect to rotations of axes about the origin of coordinates $x = y = 0$. Although his procedure leads to a set of polynomials with the required properties his derivation appears to be somewhat arbitrary. In a later paper Zernike and Brinkman (8) showed that the circle polynomials also occur in connexion with axially symmetrical solutions of the potential equation in four-dimensional space. The properties of these polynomials have been investigated in detail by Nijboer (5).

Now it is easy to see (cf. §2) that there exist an infinite number of complete sets of polynomials which are orthogonal for the interior of the unit circle. It is, therefore, natural to inquire what the properties are which distinguish the polynomials of Zernike from the other sets. This and some more general questions concerning polynomials that are orthogonal over the unit circle are discussed in the present paper. Of the great variety of possible sets only those are considered which transform into themselves by a representation of the two-dimensional rotation group. In the present case, however, it is possible to avoid the abstract formalism of group theory with the help of a kind of normalization; one determines sets which contain only such polynomials as are *invariant in form*† with respect to rotations of axes about the

† By 'invariance in form' we mean that the polynomials satisfy the invariance relation (2·3). This is in agreement with the definition of invariance given by Madelung ((4), p. 98 (5)). We use the phrase 'invariance in form' to avoid confusion with the more customary definition corresponding to the special case of (2·3) with $G(\phi) \equiv 1$.

origin. Such sets (provided they are complete) will clearly be useful in applications.

It is found that although there exist an infinity of complete orthogonal sets which contain only polynomials of the above form, there is only one set which, in addition, has certain properties that are analogous to those of the set of Legendre polynomials. This special set is found to be identical with that of the Zernike polynomials. With the help of this result and using some well-known properties of Jacobi polynomials explicit expressions for the Zernike polynomials are then derived in a simple and systematic manner.

The same method may be used to derive other orthogonal sets. One new set is discussed, and the generating functions for this set and for that of Zernike polynomials are given.

2. *Some orthogonal sets.* We wish to find polynomials (not necessarily real) in two real variables x and y, which form a complete orthogonal set for the interior of the unit circle $x^2+y^2 \leqslant 1$. The orthogonality condition will be written in the form

$$\iint\limits_{x^2+y^2\leqslant 1} V^*_{(\alpha)}(x,y)\, V_{(\beta)}(x,y)\, dx\, dy = A_{\alpha\beta}\delta_{\alpha\beta}, \tag{2·1}$$

where, for the present, $V_{(\alpha)}$ and $V_{(\beta)}$ denote typical polynomials of the set, an asterisk denotes the complex conjugate, $\delta_{\alpha\beta}$ is the Kronecker symbol and $A_{\alpha\beta}$ are normalization constants to be chosen later.

Clearly there are many such sets. This can be seen, for example, by applying the orthogonalization process of Schmidt to all monomials $x^i y^j$ $(i \geqslant 0, j \geqslant 0)$, and noting that the resulting set depends on the order† in which the monomials are orthogonalized. We must therefore specify more closely the set which we wish to derive.

2·1. We shall first inquire whether a complete orthogonal set exists which is such that it contains only polynomials that are invariant in form with respect to rotations of axes about the origin $x = y = 0$. By such invariance we mean that when any rotation

$$\left.\begin{aligned} x' &= x\cos\phi + y\sin\phi, \\ y' &= -x\sin\phi + y\cos\phi, \end{aligned}\right\} \tag{2·2}$$

is applied, each polynomial $V(x,y)$ is transformed into a polynomial of the same form, i.e. V satisfies the following relation under the transformation (2·2):

$$V(x,y) = G(\phi)\, V(x',y'), \tag{2·3}$$

$G(\phi)$ being a continuous function with period 2π of the angle of rotation ϕ and $G(0) = 1$.

The following theorem will now be established:

THEOREM 1. *A polynomial $V(x,y)$ of degree n will be invariant in form with respect to rotations of axes about the origin $x = y = 0$, if and only if, when expressed in polar coordinates (r,ϕ), it is of the form*

$$V(r\cos\phi, r\sin\phi) = R(r)\, e^{il\phi}, \tag{2·4}$$

† For example, in the analogous one-dimensional case, the orthogonalization of the sequence

$$1, x, x^2, x^3, \ldots \tag{I}$$

over the range $-1 \leqslant x \leqslant 1$ gives Legendre polynomials. If, however, the terms in (I) are orthogonalized in a different order, other complete sets are in general obtained. The set of Legendre polynomials is distinguished from all other sets which are orthogonal over $-1 \leqslant x \leqslant 1$ by the property that it contains a polynomial of *every* degree.

where l is an integer, positive, negative or zero and $R(r)$ is a polynomial in r of degree n, containing no power of r lower than $|l|$. Moreover, $R(r)$ is an even or an odd polynomial according as l is even or odd.

To prove this theorem, we use the fact that the application of two successive rotations through angles ϕ_1 and ϕ_2 is equivalent to a single rotation through the angle $\phi_1+\phi_2$. It then follows from (2·3) that G must satisfy the functional equation

$$G(\phi_1)\,G(\phi_2) = G(\phi_1+\phi_2). \tag{2·5}$$

The general solution with the period 2π of this equation is (cf. Born (1), p. 153)

$$G(\phi) = e^{il\phi}, \tag{2·6}$$

where l is any integer, positive, negative or zero. On substituting from (2·6) into (2·3), using (2·2) and setting $x'=r$, $y'=0$, it follows that V must be of the form

$$V(r\cos\phi, r\sin\phi) = R(r)\,e^{il\phi}, \tag{2·7}$$

where $R(r) = V(r,0)$ is a function of r alone. Next expanding $e^{il\phi}$ in powers of $\cos\phi$ and $\sin\phi$ and using the fact that V is a polynomial of degree n in $x = r\cos\phi$ and $y = r\sin\phi$, it immediately follows that R must be of the form asserted in Theorem 1.

It will be seen later that there exist an infinity of complete sets of polynomials which are orthogonal for the interior of the unit circle and which contain only such polynomials as are invariant in form with respect to rotations of axes about $x=y=0$. However, the following theorem holds:

THEOREM 2. *There is one and only one*† *set which is*

(*a*) *orthogonal for the interior of the unit circle.*

(*b*) *contains only such polynomials as are invariant in form with respect to rotations of axes about the origin $x=y=0$.*

(*c*) *contains a polynomial for each permissible pair of values of n (degree) and l (angular dependence), i.e. for integral values of n and l such that $n \geqslant 0$, $l \gtreqless 0$, $n \geqslant |l|$ and $n-|l|$ is even.*

To prove this theorem, we choose an integer l' and orthogonalize the sequence of the linearly independent functions

$$r^{|l'|}\,e^{il'\phi},\quad r^{|l'|+2}\,e^{il'\phi},\quad r^{|l'|+4}\,e^{il'\phi},\ldots \tag{2·8}$$

over the interior of the unit circle. That this can be done follows from the Schmidt orthogonalization process. Let us denote the $k+1$th member of the orthogonalized set by $\mathscr{V}^{l'}_{|l'|+2k}$. Then

$$\mathscr{V}^{l'}_{|l'|+2k}(x,y) = \mathscr{R}^{l'}_{|l'|+2k}(r)\,e^{il'\phi}, \tag{2·9}$$

where $\mathscr{R}^{l'}_{|l'|+2k}$ is of the form

$$\mathscr{R}^{l'}_{|l'|+2k}(r) = [\alpha^{l'}_{k,0}\,r^{|l'|} + \alpha^{l'}_{k,2}\,r^{|l'|+2} + \ldots \alpha^{l'}_{k,2k}\,r^{|l'|+2k}], \tag{2·10}$$

and the α's are constants such that

$$\left.\begin{aligned} 2\pi a^{l'}_n\,\delta_{nn'} &= \iint\limits_{x^2+y^2\leqslant 1} \mathscr{V}^{*l'}_n(x,y)\,\mathscr{V}^{l'}_{n'}(x,y)\,dx\,dy \\ &= 2\pi\int_0^1 \mathscr{R}^{*l'}_n(r)\,\mathscr{R}^{l'}_{n'}(r)\,r\,dr, \end{aligned}\right\} \tag{2·11}$$

† Two polynomials which are such that one is a constant multiple of the other are regarded here as identical.

$a_n^{l'}$ being normalization constants to be chosen later. It follows from Schmidt's process that for every value of k, $\alpha_{k,2k}^{l'} \neq 0$, so that $\mathscr{R}_{|l'|+2k}^{l'}$ is a polynomial in r of degree $|l'|+2k$ and consequently $\mathscr{V}_{|l'|+2k}^{l'}$ is a polynomial of degree $|l'|+2k$ in x and y. Moreover, $\mathscr{R}_{|l'|+2k}^{l'}(r)$ is seen to be an even or odd polynomial according as l' is even or odd and contains no powers of r of degree lower than $|l'|$. If now we allow l' to take all possible integral values a set is obtained which satisfies postulates (*b*) and (*c*) of Theorem 2. The set also satisfies the postulate (*a*); for those $\mathscr{V}$ polynomials which have the same upper suffix are orthogonal on account of (2·11) and those which have different upper suffix are orthogonal on account of the orthogonality of the exponential factor over the ϕ domain.

Next we show that there is no other set which satisfies all the requirements of Theorem 2. For let us assume that another set satisfying all the postulates of Theorem 2 exists. Then, by (*b*), each polynomial of this set must be of the form

$$\left.\begin{aligned} \overline{\mathscr{V}}_{|l|+2k}^{l}(x,y) &= \overline{\mathscr{R}}_{|l|+2k}^{l}(r)\,e^{il\phi}, \\ \text{where}\qquad \overline{\mathscr{R}}_{|l|+2k}^{l}(r) &= [\overline{\alpha}_{k,0}^{l}\,r^{|l|} + \overline{\alpha}_{k,2}^{l}\,r^{|l|+2} + \ldots + \overline{\alpha}_{k,2k}^{l}\,r^{|l|+2k}], \end{aligned}\right\} \tag{2·12}$$

and by (*a*), $\overline{\mathscr{V}}$ and $\overline{\mathscr{R}}$ satisfy the orthogonality relations of the form (2·11). Further, according to (*c*), the constant $\overline{\alpha}_{k,2k}^{l} \neq 0$ exists for every integer l' and for every non-negative integer k.

Let l' be a particular value of l. On multiplying $r\overline{\mathscr{R}}_{|l'|+2k}^{*l'}$ successively by $\overline{\mathscr{R}}_{|l'|}^{l'}, \overline{\mathscr{R}}_{|l'|+2}^{l'}, \overline{\mathscr{R}}_{|l'|+4}^{l'} \ldots \overline{\mathscr{R}}_{|l'|+2k-2}^{l'}$ and integrating over $0 \leqslant r \leqslant 1$ it follows on account of (2·11) that when l' is even (odd), $\overline{\mathscr{R}}_{|l'|+2k}^{l'}$ and similarly $\mathscr{R}_{|l'|+2k}^{l'}$ are orthogonal with a weighting factor r to every even (odd) power r^j such that

$$|l'| \leqslant j \leqslant |l'| + 2k - 2. \tag{2·13}$$

Consequently any linear combination of $\mathscr{R}_{|l'|+2k}^{l'}$ and $\overline{\mathscr{R}}_{|l'|+2k}^{l'}$, and in particular the function

$$\mathscr{F}(r) = \alpha_{k,2k}^{l'}\,\overline{\mathscr{R}}_{|l'|+2k}^{l'}(r) - \overline{\alpha}_{k,2k}^{l'}\,\mathscr{R}_{|l'|+2k}^{l'}(r) \tag{2·14}$$

is orthogonal to any polynomial $\mathscr{P}(r)$ which contains only powers r^j of r such that the inequalities (2·13) hold and which is an even or odd polynomial according as l' is even or odd. Now $\mathscr{F}(r)$ itself satisfies the conditions imposed on $\mathscr{P}(r)$ and hence must be orthogonal to itself,

$$\int_0^1 \mathscr{F}^*(r)\,\mathscr{F}(r)\,r\,dr = 0.$$

Since the integrand is positive or zero throughout the range of integration it follows that $\mathscr{F}(r) \equiv 0$, whence

$$\alpha_{k,2k}^{l'}\,\mathscr{R}_{|l'|+2k}^{l'}(r) = \alpha_{k,2k}^{l'}\,\overline{\mathscr{R}}_{|l'|+2k}^{l'}, \tag{2·15}$$

showing that for each value l' of l, $\overline{\mathscr{R}}_{|l'|+2k}^{l'}$ is a constant multiple of $\mathscr{R}_{|l'|+2k}^{l'}$ so that the two sets are identical. This concludes the proof of Theorem 2.

The set (2·9) contains $\frac{1}{2}(n+1)(n+2)$ linearly independent polynomials of degree $\leqslant n$. Hence every monomial $x^i y^j$ ($i \geqslant 0, j \geqslant 0$ integers) and consequently every polynomial may be expressed as a linear combination of a finite number of the $\mathscr{V}$'s. By Weierstrass's theorem on approximations by polynomials it then follows that the $\mathscr{V}$ set is complete.†

† We are indebted to Dr E. Ruch for some helpful discussions concerning the completeness of orthogonal sets.

It is of interest to note that there exist other complete orthogonal sets for the interior of the unit circle which consist of polynomials invariant in form with respect to rotations of axes about the origin. This immediately follows from similar considerations as outlined in the footnote on p. 41 in connexion with the uniqueness of the Legendre polynomials. Such sets, however (clearly infinite in number), do not satisfy requirement (*c*) of our uniqueness theorem, namely, that they contain a polynomial for each pair of the permissible values of n and l.

2·2. From the manner in which the $\mathscr{V}$ polynomials were defined it is seen that the *radial polynomials* $\mathscr{R}^l_{|l|}$, $\mathscr{R}^l_{|l|+2}$, $\mathscr{R}^l_{|l|+4} \dots \mathscr{R}^l_{|l|+2k} \dots$ *may be regarded as functions obtained by orthogonalizing the sequence*

$$r^{|l|}, \quad r^{|l|+2}, \quad r^{|l|+4}, \quad \dots, \quad r^{|l|+2k}, \quad \dots \tag{2·16}$$

with the weighting factor r *over the range* $0 \leqslant r \leqslant 1$. From Schmidt's orthogonalization process it then immediately follows that the ratios of the constants α which occur in each polynomial are real. Since polynomials which differ from each other only by a constant multiplicative factor are regarded here as identical, we may take the α's (and consequently the $\mathscr{R}$'s) to be real. Moreover, since in (2·16) only the absolute value of l enters, it follows that

$$\mathscr{R}^{-l}_n(r) = \beta^l_n \mathscr{R}^l_n, \tag{2·17}$$

where the β's are constants depending only on the normalization of the two polynomials $\mathscr{R}^{-l}_n$ and $\mathscr{R}^l_n$. Hence, if the normalization is chosen so that $\beta^l_n = 1$ for all l and n, then $\mathscr{R}^{-l}_n(r) = \mathscr{R}^l_n(r)$ and consequently

$$\mathscr{V}^{\pm m}_n(x, y) = \mathscr{R}^m_n e^{\pm im\phi}, \tag{2·18}$$

$m = |l|$ denoting non-negative integers.

Now the circle polynomials of Zernike (Zernike (7), Nijboer (5)) which, as already mentioned, play an important part in certain optical problems, are real polynomials of the form†

$$U^{\pm m}_n(x, y) = R^m_n \begin{matrix}\cos m\phi \\ \sin m\phi\end{matrix}. \tag{2·19}$$

Moreover the complex combinations

$$U^m_n(x, y) \pm i U^{-m}_n(x, y) = R^m_n(r)\, e^{\pm im\phi} \tag{2·20}$$

are easily seen to satisfy all conditions of Theorem 2. *Hence the Zernike circle polynomials are simply the real polynomials* $\mathscr{R}^m_n(r)\cos m\phi$ *and* $\mathscr{R}^m_n(r)\sin m\phi$ *associated with* (2·18).

It will now be shown that with the help of our results and from certain well-known results concerning Jacobi polynomials explicit expressions for the radial polynomials $\mathscr{R}^m_n(r)$ may almost immediately be written down.

3. *Explicit expression for the radial polynomials.* Since $\mathscr{R}^{\pm m}_n(r)$ is a polynomial in r of degree n which contains no power of r lower than m and is an even or odd polynomial according as n is even or odd, it follows that each $\mathscr{R}$ can be expressed in the form

$$\mathscr{R}^{\pm m}_n(r) = t^{\frac{1}{2}m} \mathscr{Q}_{\frac{1}{2}(n-m)}(t), \tag{3·1}$$

† Zernike actually used the same symbol $U^m_n(x, y)$ for $R^m_n(r)\cos m\phi$ and $R^m_n(r)\sin m\phi$.

where $t = r^2$ and $\mathcal{Q}_{\frac{1}{2}(n-m)}(t)$ denotes a polynomial in t of degree $\frac{1}{2}(n-m)$. The orthogonality condition (2·11) becomes, remembering that $\mathcal{R}$ and consequently $\mathcal{Q}$ are now real,

$$\tfrac{1}{2}\int_0^1 t^m \mathcal{Q}_k(t)\,\mathcal{Q}_{k'}(t)\,dt = a_n^{\pm m}\delta_{kk'}, \tag{3·2}$$

where
$$k = \tfrac{1}{2}(n-m), \quad k' = \tfrac{1}{2}(n'-m).$$

Since the radial polynomials $\mathcal{R}_n^{\pm m}(r)$ are the functions obtained by orthogonalizing the sequence (2·16) (with $|l| = m$) with the weighting factor r over $0 \leqslant r \leqslant 1$, it is seen from (3·1) and (3·2) that polynomials $\mathcal{Q}_0(t), \mathcal{Q}_1(t), \mathcal{Q}_2(t), \ldots, \mathcal{Q}_k(t), \ldots$ may be obtained by orthogonalizing the sequence

$$1, \quad t, \quad t^2, \quad \ldots, \quad t^k, \quad \ldots, \tag{3·3}$$

with the weighting factor $w(t) = t^m$ over the range $0 \leqslant t \leqslant 1$.

Now functions $G_k(p,q,t)$ obtained by orthogonalizing (3·3) with a more general weighting factor
$$w(t) = t^{q-1}(1-t)^{p-q} \quad (q>0, p-q>-1)$$
over the range $0 \leqslant t \leqslant 1$ are well known. They are the so-called *Jacobi* (or hypergeometric) *polynomials*.†

$$G_k(p,q,t) = \frac{k!\,(q-1)!}{(p+k-1)!}\sum_{s=0}^{k}(-1)^s\frac{(p+k+s-1)!}{(k-s)!\,s!\,(q+s-1)!}t^s, \tag{3·4}$$

whose orthogonality and normalization properties are given by

$$\int_0^1 t^{q-1}(1-t)^{p-q}G_k(p,q,t)\,G_{k'}(p,q,t)\,dt = b_k(p,q)\,\delta_{kk'}, \tag{3·5}$$

where
$$b_k(p,q) = \frac{k!\,[(q-1)!]^2\,(p-q+k)!}{(q+k-1)!\,(p+k-1)!\,(p+2k)}. \tag{3·6}$$

(With this choice of b_k, $G_k(p,q,0) = 1$ for all values of k, p and q for which G is defined.) Hence on comparing (3·2) and (3·5) it follows that

$$\mathcal{Q}_k(t) = \sqrt{\{2a_n^{\pm m}/b_k(m+1,m+1)\}}\,G_k(m+1,m+1,t). \tag{3·7}$$

From (3·7) and (3·1) we obtain, remembering that $k = \frac{1}{2}(n-m)$,

$$\mathcal{R}_n^{\pm m}(r) = \sqrt{\{2a_n^{\pm m}/b_k(m+1,m+1)\}}\,r^m G_k(m+1,m+1,r^2). \tag{3·8}$$

Following Zernike we shall choose the normalization so that

$$\mathcal{R}_n^{\pm m}(1) = 1. \tag{3·9}$$

The corresponding values of the normalization constants $a_n^{\pm m}$ can easily be determined. We have from (3·8) and (3·9)

$$\sqrt{\{b_k(m+1,m+1)/2a_n^{\pm m}\}} = G_k(m+1,m+1,1). \tag{3·10}$$

The value $G_k(m+1,m+1,1)$ can be obtained from the generating function (cf. Courant-Hilbert (2), p. 77)

$$\frac{[z-1+\sqrt{\{1-2z(1-2r^2)+z^2\}}]^m}{(2zr^2)^m\sqrt{\{1-2z(1-2r^2)+z^2\}}} = \sum_{s=0}^{\infty}\binom{m+s}{s}G_s(m+1,m+1,r^2)\,z^s. \tag{3·11}$$

† See, for example, Madelung ((4), p. 57), Courant and Hilbert ((2), p. 76) and Kemble ((3), p. 594). In the account of Courant and Hilbert there appear two misprints, the weighting factor being incorrectly given as $x^q(1-x)^{p-q}$ instead of $x^{q-1}(1-x)^{p-q}$, and the range for q is incorrectly given as $q > -1$ instead of $q > 0$.

For $r = 1$ the left-hand side reduces to $(1+z)^{-1}$ and we obtain, on expanding this expression into a power series and on comparing it with the right-hand side of (3·11):

$$G_s(m+1, m+1, 1) = (-1)^s \Big/ \binom{m+s}{s}. \tag{3·12}$$

Hence from (3·10) and (3·12),

$$\sqrt{\{2a_n^{\pm m}/b_k(m+1, m+1)\}} = (-1)^{\frac{1}{2}(n-m)} \binom{\frac{1}{2}(n+m)}{\frac{1}{2}(n-m)}. \tag{3·13}$$

Substituting from (3·6) into (3·13) with $k = \frac{1}{2}(n-m)$, we obtain

$$a_n^m = a_n^{-m} = \frac{1}{2n+2}. \tag{3·14}$$

With this choice of normalization we have from (3·8) and (3·13)

$$\mathscr{R}_n^{\pm m}(r) = (-1)^{\frac{1}{2}(n-m)} \binom{\frac{1}{2}(n+m)}{\frac{1}{2}(n-m)} r^m G_{\frac{1}{2}(n-m)}(m+1, m+1, r^2). \tag{3·15}$$

On substituting from (3·4) for the Jacobi polynomials we finally obtain the required explicit expressions for the radial polynomials:

$$\mathscr{R}_n^{\pm m}(r) = \sum_{s=0}^{\frac{1}{2}(n-m)} (-1)^s \frac{(n-s)!}{s!\,\{\frac{1}{2}(n+m)-s\}!\,\{\frac{1}{2}(n-m)-s\}!} r^{n-2s}. \tag{3·16}$$

We can also immediately write down the generating function for $\mathscr{R}_n^{\pm m}(r)$ by setting $n = m+2s$ in (3·15) and then substituting into (3·11). This gives

$$\frac{[1+z-\sqrt{\{1-2z(1-2r^2)+z^2\}}]^m}{(2zr)^m \sqrt{\{1+2z(1-2r^2)+z^2\}}} = \sum_{s=0}^{\infty} z^s \mathscr{R}_{m+2s}^{\pm m}(r). \tag{3·17}$$

The expressions (3·15) and (3·16) for the radial parts of the circle polynomials are in agreement with those obtained by Zernike (7).

4. *A related orthogonal set.* The methods of the previous sections may also be used to derive explicit expressions for other sets of functions orthogonal for the interior of the unit circle. We shall discuss briefly only one other set which has analogous properties to that of the Zernike polynomials.

Instead of polynomials in x and y alone, we now consider polynomials W in x, y, and $r = \sqrt{(x^2+y^2)}$ which are invariant in form with respect to rotations of axes about the origin. By an argument similar to that used to establish Theorem 1, one finds that if W is such a polynomial of degree n in x, y and r it must be of the form

$$W(r\cos\phi, r\sin\phi, r) = S(r)\, e^{il\phi}, \tag{4·1}$$

where l is an integer, positive, negative or zero and $S(r)$ is a polynomial in r of degree n such that it contains no powers of r lower than $|l|$.

Again, an infinity of orthogonal sets of this type can be shown to exist. It may, however, be proved by an argument similar to that used in § 2·2 that there is only one W set which has the additional property that it contains a member for each permissible pair of values of the suffixes n and l, i.e. for n, l integers and $n \geqslant |l|$. We denote the polynomials of this special set by $\mathscr{W}_n^l$ and the corresponding radial parts by $\mathscr{S}_n^l$. Then $\mathscr{W}_n^{l'}$ may be obtained by orthogonalizing in place of (2·8) the sequence

$$r^{|l'|} e^{il'\phi}, \quad r^{|l'|+1} e^{il'\phi}, \quad r^{|l'|+2} e^{il'\phi}, \quad \ldots, \quad r^{|l'|+k} e^{il'\phi}, \quad \ldots \tag{4·2}$$

over the interior of the unit circle. If we normalize the polynomials so that

$$\mathscr{S}^l_n(1) = 1$$

for all n and l, then we find, following the procedure of § 3, that

$$\mathscr{S}^{\pm m}_n(r) = (-1)^{n-m}\binom{m+n+1}{n-m} r^m G_{n-m}(2m+2, 2m+2, r), \tag{4·3}$$

and that

$$\int_0^1 \mathscr{S}^{\pm m}_n(r)\,\mathscr{S}^{\pm m}_{n'}(r)\, r\,dr = \frac{1}{2n+2}\delta_{nn'}, \tag{4·4}$$

m again denoting a non-negative integer. From (4·3) and (3·4) we obtain the following explicit expressions for $\mathscr{S}^{\pm m}_n(r)$:

$$\mathscr{S}^{\pm m}_n(r) = \sum_{s=0}^{n-m} (-1)^s \frac{(2n+1-s)!}{s!\,(n-m-s)!\,(n+m+1-s)!} r^{n-s}. \tag{4·5}$$

The first few of the radial polynomials $\mathscr{S}$ are exhibited in Table 1. Using (4·3) and the formula (cf. Courant and Hilbert, (2), p. 77)

$$G_k(p,q,r) = \frac{(q-1)!}{(q+k-1)!} r^{1-q}(1-r)^{p-q} \frac{d^k}{dr^k}[r^{q+k-1}(1-r)^{p-q+k}],$$

Table 1. *First few of the polynomials* $\mathscr{S}^{\pm m}_n(r)$

m \ n	0	1	2	3
0	1	$3r-2$	$10r^2-12r+3$	$35r^3-60r^2+30r-4$
1	—	r	$5r^2-4r$	$21r^3-30r^2+10r$
2	—	—	r^2	$7r^3-6r^2$
3	—	—	—	r^3

it follows that $\mathscr{S}^{\pm m}_n(r)$ may be expressed in the form

$$\mathscr{S}^{\pm m}_n(r) = \frac{r^{-(m+1)}}{(n-m)!}\left(\frac{d}{dr}\right)^{n-m}[r^{m+n+1}(r-1)^{n-m}]. \tag{4·6}$$

The generating function for $\mathscr{S}^{\pm m}_n(r)$ is obtained by replacing in (3·11) $m+1$ by $2m+2$, r^2 by r and using (4·3) with $n = m+s$. This gives

$$\frac{[1+z-\sqrt{\{1+2z(1-2r)+z^2\}}]^{2m+1}}{r^{m+1}(2z)^{2m+1}\sqrt{\{1+2z(1-2r)+z^2\}}} = \sum_{s=0}^{\infty} z^s \mathscr{S}^{\pm m}_{m+s}(r). \tag{4·7}$$

A comparison of (4·7) and (3·17) shows that the following relation exists between the $\mathscr{S}$ set and some members of the $\mathscr{R}$ set:

$$r\mathscr{S}^{\pm m}_n(r^2) = \mathscr{R}^{\pm(2m+1)}_{2n+1}(r). \tag{4·8}$$

By an argument similar to that used in § 2, it may be shown that the set of the $\mathscr{W}$ functions is complete. Since the $\mathscr{W}$ functions are polynomials in x, y and $r = \sqrt{(x^2+y^2)}$. whereas the Zernike U functions are polynomials in x and y alone, the $\mathscr{W}$ functions may be expected in some applications to be more appropriate than the polynomials of Zernike.

In conclusion, it is a pleasure to thank Professor Max Born, F.R.S., and Dr H. Schneider for their interest in this work and for many stimulating and helpful discussions. One of us (A. B. B.)† is also indebted to the University of Edinburgh for the award of an I.C.I. Fellowship during the tenure of which this work was carried out.

REFERENCES

(1) BORN, M. *Natural philosophy of cause and chance.* (Oxford, 1949).
(2) COURANT, R. and HILBERT, D. *Methoden der Mathematischen Physik*, vol. 1, 2nd ed. (Berlin, 1931).
(3) KEMBLE, E. *The fundamental principles of quantum mechanics*, 1st ed. (New York, 1937).
(4) MADELUNG, E. *Die Mathematischen Hilfsmittel des Physikers* (Berlin, 1936).
(5) NIJBOER, B. R. A. Thesis, University of Groningen (1942).
(6) NIJBOER, B. R. A. *Physica, 's Grav.*, 13 (1947), 605.
(7) ZERNIKE, F. *Physica, 's Grav.*, 1 (1934), 689.
(8) ZERNIKE, F. and BRINKMAN, H. C. *Proc. Akad. Sci. Amst.* 38, no. 2 (1935).
(9) ZERNIKE, F. and NIJBOER, B. R. A. Contribution in *La Théorie des Images Optiques* (Paris, 1949. Éditions de la Revue d'Optique), p. 227.

DEPARTMENT OF MATHEMATICAL PHYSICS
UNIVERSITY OF EDINBURGH

† Present address: Physics Division, National Research Council, Ottawa.

4.2

Reprinted from JOURNAL OF THE OPTICAL SOCIETY OF AMERICA, Vol. 43, No. 3, 218, March, 1953
Printed in U. S. A.

The X_n and Y_n Functions of Hopkins, Occurring in the Theory of Diffraction

E. WOLF
Department of Mathematical Physics, The University, Drummond Street, Edinburgh, 8, Great Britain
(Received November 20, 1952)

IN a paper which concerns the disturbance near focus of waves of radially nonuniform amplitude, and which forms an extension of Lommel's classical analysis,[1] Hopkins[2] was led to the introduction of two new functions,

$$\left.\begin{aligned} X_n(y,z)&=y\frac{\partial U_n}{\partial y}=\sum_{s=0}^{\infty}(-1)^s(n+2s)\left(\frac{y}{z}\right)^{n+2s}J_{n+2s}(z),\\ Y_n(y,z)&=-y\frac{\partial V_n}{\partial y}=\sum_{s=0}^{\infty}(-1)^s(n+2s)\left(\frac{z}{y}\right)^{n+2s}J_{n+2s}(z).\end{aligned}\right\}\quad(1)$$

Here U_n and V_n denote the Lommel functions and J_n is a Bessel function of the first kind. Recently Boivin[3] derived interesting "multiplication" theorems which express $U_n(\alpha^2y,\alpha z)$, $V_n(\alpha^2y,\alpha z)$, $X_n(\alpha^2y,\alpha z)$, and $Y_n(\alpha^2y,\alpha z)$ in series involving these functions with successive orders but with arguments (y, z). These formulas are of importance when one studies diffraction by concentric arrays of ring-shaped apertures.

I pointed out elsewhere[4] that the Y_n functions can be expressed in terms of the Lommel V_n functions by means of the simple relation

$$Y_n(y,z)=\tfrac{1}{2}\left[\frac{z^2}{y}V_{n-1}(y,z)+yV_{n+1}(y,z)\right],$$

and similarly

$$X_n(y,z)=\tfrac{1}{2}\left[\frac{z^2}{y}U_{n+1}(y,z)+yU_{n-1}(y,z)\right].\quad(2)$$

It is therefore possible to express the new formulas of Boivin for $X_n(\alpha^2y,\alpha z)$ and $Y_n(\alpha^2y,\alpha z)$ in terms of the ordinary Lommel functions with arguments (y, z), giving*

$$\left.\begin{aligned} X_n(\alpha^2y,\alpha z)&=\sum_{s=0}^{\infty}\frac{(\alpha^2-1)^sy^s}{2^{s+1}s!}\left[\frac{z^2}{y}U_{n-s+1}(y,z)\right.\\ &\qquad\left.+2sU_{n-s}(y,z)+yU_{n-s-1}(y,z)\right],\\ Y_n(\alpha^2y,\alpha z)&=\sum_{s=0}^{\infty}(-1)^s\frac{(\alpha^2-1)^sy^s}{2^{s+1}s!}\left[\frac{z^2}{y}V_{n+s-1}(y,z)\right.\\ &\qquad\left.-2sV_{n+s}(y,z)+yV_{n+s+1}(y,z)\right].\end{aligned}\right\}\quad(3)$$

Alternatively, one can replace y by α^2y and z by αz in (2) and apply Boivin's formulas for $U_n(\alpha^2y,\alpha z)$ and $V_n(\alpha^2y,\alpha z)$. One then obtains the simpler relations

$$\left.\begin{aligned} X_n(\alpha^2y,\alpha z)&=\sum_{s=0}^{\infty}\frac{(\alpha^2-1)^sy^s}{2^{s+1}s!}\\ &\qquad\times\left[\frac{z^2}{y}U_{n-s+1}(y,z)+\alpha^2yU_{n-s-1}(y,z)\right],\\ Y_n(\alpha^2y,\alpha z)&=\sum_{s=0}^{\infty}(-1)^s\frac{(\alpha^2-1)^sy^s}{2^{s+1}s!}\\ &\qquad\times\left[\frac{z^2}{y}V_{n+s+1}(y,z)+\alpha^2yV_{n+s+1}(y,z)\right].\end{aligned}\right\}\quad(4)$$

The U and V functions of the higher orders which occur in (3) and (4) can be easily calculated with the help of the recurrence relations

$$\left.\begin{aligned} U_n(y,z)+U_{n+2}(y,z)&=\left(\frac{y}{z}\right)^nJ_n(z),\\ V_n(y,z)+V_{n+2}(y,z)&=\left(\frac{z}{y}\right)^nJ_n(z).\end{aligned}\right\}\quad(5)$$

Equations (3) or (4), together with (5), make it possible to evaluate $X_n(\alpha^2y,\alpha z)$ and $Y_n(\alpha^2y,\alpha z)$ from tables of the four basic functions $U_1(y,z)$, $U_2(y,z)$, $V_0(y,z)$, $V_1(y,z)$ of Lommel's original problem. This will considerably simplify calculations on the basis of Boivin's theory of the effect of waves presenting a radial variation of amplitude.

[1] E. Lommel, Abh. Bayer. Akad. Wiss. **15**, 229 (1885).
[2] H. H. Hopkins, Proc. Phys. Soc. (London) **B62**, 22 (1949).
[3] A. Boivin, J. Opt. Soc. Am. **42**, 60 (1952).
[4] E. Wolf, Proc. Roy. Soc. (London) **A204**, 535 (1951).
* An error in sign of the term $sU_{n-s}(y,z)$ in Boivin's formula for $X_n(\alpha^2y,\alpha z)$ is corrected here.

Reprinted from JOURNAL OF THE OPTICAL SOCIETY OF AMERICA, Vol. 39, No. 9, 752–756, September, 1949
Printed in U. S. A.

43

On the Corrector Plates of Schmidt Cameras

E. H. LINFOOT AND E. WOLF
The Observatory, Cambridge, England

(Received May 2, 1949)

The paper discusses the problem of designing the aspheric surface of a Schmidt camera so as to obtain optimum performance, in an agreed sense, over the field taken as a whole. A solution is obtained in a form which allows the optimum plate-profile to be quickly determined for a Schmidt camera of given aperture-ratio and field-size, working over a given spectral range. It is shown that at apertures near $f/3$ the optimum plate can be obtained by slightly decreasing the strength of an ordinary "color-minimised" Schmidt plate, while in wide-field systems working at apertures near $f/1$, it is better to use a plate with neutral zone at the edge of the aperture.

1. INTRODUCTION

As a rule the corrector plates of Schmidt cameras are designed without paying much attention to the off-axis images. The aspheric profile is calculated so that the system gives an error-free axial image in light of one particular wave-length, and at the same time the paraxial convexity of the plate is chosen so as to make the axial color-error as small as possible.

This simple procedure gives satisfactory results in medium-sized cameras of not too short focal ratio. However, it is evident from the theoretical point of view that some improvement could be obtained by a more careful choice of plate-profile. For the sizes and shapes of the off-axis images depend on the monochromatic off-axis aberrations of the system as well as on its chromatism, and the correct choice of plate-profile would take account of these two aberrations simultaneously, combining them with axial undercorrection to the best advantage over the field taken as a whole.

The present paper shows how a choice of plate-profile can be made which, in a certain sense, optimises the optical performance of the camera over a given field. In Section 2 the aberration theory of the Schmidt camera is summarised in a form suitable for the later discussion. In Section 3 a definition of effective image-radius is set up, and in Section 4 this image-radius is evaluated, for a Schmidt of given aperture-ratio and field, in terms of two parameters a and p which describe the plate-profile. Lastly, in Section 5, the results obtained are used to design optimum plate-profiles in two numerical cases and to draw one or two general conclusions. In particular, it appears that in wide-field systems of apertures approaching $f/1$ the optimal plate-profile is one which is nearly flat at the edge; while at apertures near $f/3$ an optimised plate can be obtained, with neutral zone at or near 0.866 of the full radius, simply by shortening the wave-length at which the plate is designed to give axial stigmatism.

2. THE ABERRATIONS OF THE SCHMIDT CAMERA

We wish to see what choice of plate-profile will give optimum imaging (in some agreed sense) over the whole of a given field.

Let R denote the radius of curvature of the mirror, H the radius of the effective plate-aperture, and set

$$\mu = H/R. \tag{1}$$

Since the focal length f of the system is nearly equal to $\frac{1}{2}R$, it follows that 2μ is very nearly equal to the numerical aperture of the system, so that in an $f/2$-camera $\mu=\frac{1}{8}$, approximately.

It can be shown that, to the order of approximation adopted in treating μ^2 as negligible in comparison with unity, there is no loss of generality in supposing that the corrector plate has its aspheric surface towards the mirror.*

If we use ordinary Cartesian coordinates in the plane of the aperture-stop, with origin on the axis of the system, the edge of the aperture stop is the circle $x^2+y^2=H^2$. It is more convenient to introduce new coordinates u, v, connected with x, y by the scale-relation,

$$x = Hu, \quad y = Hv; \tag{2}$$

then the edge of the aperture stop is the circle $u^2+v^2=1$ (see Fig. 1).

To simplify the formulas, we now choose R as the unit of length, setting $R=1$. Then $H=\mu$, and Eqs. (2) become

$$x = \mu u, \quad y = \mu v. \tag{3}$$

The plate-profile which gives axial stigmatism in light for which the plate has refractive index N is then given by the equation

$$(N-1)s = \tfrac{1}{4}\mu^4(r^2-r_0^2)^2(1+(5/2)r_0^2\mu^2) + \tfrac{3}{8}\mu^6(r^2-r_0^2)^3 + \text{constant} + O(\mu^8), \tag{4}$$

where s denotes the 'figuring depth,'** $r^2=u^2+v^2$ and r_0 is the radius of the neutral zone of the plate. The actual focal length of the system is then

$$f = \tfrac{1}{2}(1-\mu^2 r_0^2)^{-\frac{1}{2}} \tag{5}$$

* See Linfoot (2) for a proof of this statement and also of Eqs. (4), (9), and (11) below.

** i.e., the depth of glass which, laid on to a plane surface, would produce the desired profile. It is convenient to allow s to take both positive and negative values and to suppose that $s=0$ at the center of the aspheric surface.

and its numerical aperture

$$2\lambda = 2\mu(1-\mu^2 r_0^2)^{\frac{1}{2}}. \tag{6}$$

If terms of order μ^6 may be neglected, (4) can be written in the simpler form

$$(N-1)s = \tfrac{1}{4}\mu^4(r^4 - ar^2) + O(\mu^6), \tag{7}$$

where $a = 2r_0^2$, or in the equivalent form

$$(N-1)s = \tfrac{1}{4}\lambda^4(r^4 - ar^2) + O(\lambda^6), \tag{8}$$

in virtue of (6). Whether or not terms of order μ^6 are small enough to be negligible in the equation of the plate-surface, the off-axis aberrations δX, δY of the system, measured in seconds of arc, are given at the point of angular off-axis displacement $\varphi = O(\mu)$ in the u-direction on the spherical field surface by the equation†

$$\delta X + i\delta Y = \tfrac{1}{4}K\mu^3\varphi^2\left[\frac{N+1}{2N}\frac{\partial}{\partial u} + \tfrac{1}{2}u\frac{\partial^2}{\partial u^2} + \frac{i}{2N}\frac{\partial}{\partial v} + \tfrac{1}{2}iu\frac{\partial^2}{\partial u \partial v}\right](r^4 - ar^2) + O(K\mu^7), \tag{9}$$

where $K = 648{,}000/\pi$. Since $\mu = \lambda[1 + O(\lambda^2)]$, this equation remains valid if μ is replaced by λ. In an $f/2$ system, λ^2 and μ^2 are approximately 1/64 and (9) gives δX and δY accurately to within a few percent of their biggest values.

In light of a different wave-length, for which the refractive index of the plate is n, the rays of the axial parallel pencil are no longer brought to a sharp focus. If

$$n - N = O(\mu^2), \tag{10}$$

the aberrations are now given by the equation

$$\delta X + i\delta Y = \tfrac{1}{4}K\mu^3\varphi^2\frac{n-1}{N-1}\left[\frac{n+1}{2n}\frac{\partial}{\partial u} + \tfrac{1}{2}u\frac{\partial^2}{\partial u^2} + \frac{i}{2n}\frac{\partial}{\partial v} + \tfrac{1}{2}iu\frac{\partial^2}{\partial u \partial v}\right](r^4 - ar^2) + \tfrac{1}{4}K\mu^3\frac{n-N}{N-1}\left(\frac{\partial}{\partial u} + i\frac{\partial}{\partial v}\right) \times (r^4 - ar^2) + O(K\mu^7). \tag{11}$$

Let n_0 be the value of the refractive index of the plate at some selected point of the spectrum and suppose, what is true in a practical case, that $n - n_0$ and $N - n_0$ are both $O(\mu^2)$. Then if we use, in place of δX, δY, the quantities δX^*, δY^* defined by the equation

$$\delta X^* + i\delta Y^* = \tfrac{1}{4}K\mu^3\varphi^2\left[\frac{n_0+1}{2n_0}\frac{\partial}{\partial u} + \tfrac{1}{2}u\frac{\partial^2}{\partial u^2} + \frac{i}{2n_0}\frac{\partial}{\partial v} + \tfrac{1}{2}iu\frac{\partial^2}{\partial u \partial v}\right](r^4 - ar^2) + \tfrac{1}{4}K\mu^3\frac{n-N}{n_0-1}\left(\frac{\partial}{\partial u} + i\frac{\partial}{\partial v}\right)(r^4 - ar^2), \tag{12}$$

the error committed is only $O(K\mu^7)$ in each case.

From (12) it follows that the aberrations δX, δY of the system depend, to within an accuracy $O(K\mu^7)$, only on $n - N$, a, μ and φ; to this accuracy, they are effectively independent of the choice of n_0 because the value chosen for n_0 can only be varied by $O(\mu^2)$.

The two fifth-order aberrations of the Schmidt camera are lateral spherical aberration and a species of higher astigmatism. We can balance the lateral spherical aberration by undercorrecting the system on axis, and can compensate the higher astigmatism to some extent by increasing the 'central bulge' of the plate, i.e., by increasing a. However, too big a departure from the value which minimizes the greatest slope of the plate-profile (namely $a = \frac{3}{2}$) will give excessive chromatism and so spoil the performance of the system.

Axial undercorrection in n-light can be obtained by choosing a plate-profile (4) with $N \neq n$. The quantity N here takes on the role of a parameter which, together with a, specifies the coefficients of r^2 and r^4 in the geometrical plate-profile. (N and a still possess physical interpretations, of course; N is the refractive index at which the plate would give a stigmatic axial image, while a is $2r_0^2$—twice the squared radius of the neutral zone in our present units.) Changes in N and a will, in accordance with (4), alter the sixth-power term in the plate-profile, as well as the squared and fourth-power terms, but only the changes in the squared and fourth-power terms affect the form of the expressions (11) and (12) for the aberrations of the system.

Since n and N enter (12) only through their difference $n - N$, the optimal choice of N depends strongly on the

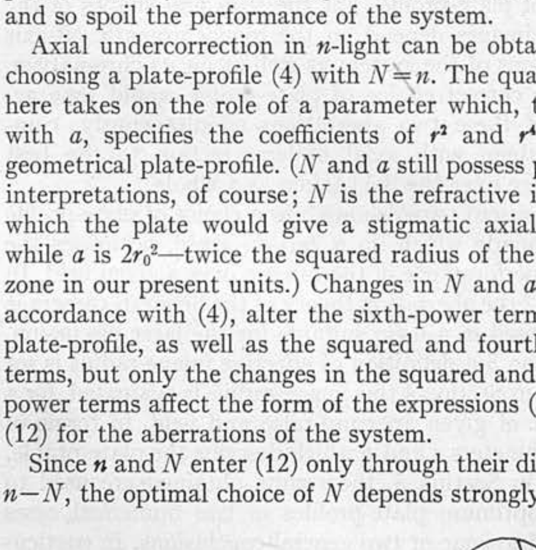


FIG. 1. The Schmidt camera.

† In a system of central obstruction-ratio $\leq \frac{1}{2}$, we have $\sin\varphi_0 \leq \mu$, so that the condition $\varphi = O(\mu)$ is always satisfied.

spectral characteristics of the light to be transmitted by the system.

In wide-field systems, $(n-N/n_0-1)\ll\varphi_0^2$ and it appears from (12) that chromatism is not very important. But in large astronomical Schmidt cameras, working over a 5- or 6-degree field at a focal ratio of $f/2.5$ or longer, considerations of chromatism play an essential part in determining the optimum plate-profile.

3.

The notion of optimising the performance of the system over a given field $\varphi\leqq\varphi_0$ can be interpreted in various ways. The most primitive interpretation, and for some purposes the best, is to take the greatest image-diameter for all object points in the given field and for all wave-lengths in the adopted spectral range, and to say that the performance is optimised when this diameter is made as small as possible. But even if diffraction effects are disregarded, this procedure does not correspond very well to the requirements of the physical situation when, for example, the effective intensity of the light rises to a peak at a certain point in its spectrum and tails off to very small values at either end.

Moreover, in cases where the size of the image is comparable with the resolving power of the photographic film, the performance is affected by photographic image-spread and there is no very close correspondence, even in monochromatic light, between image-diameter as defined above and the diameter of the star-images actually formed on a photographic plate.

In such cases it seems more appropriate physically, as well as more convenient analytically, to define the effective radius of a single monochromatic image as the square root of the expression

$$\frac{1}{\pi H^2}\iint_{x^2+y^2\leqq H^2}[(\delta X)^2+(\delta Y)^2]dxdy \tag{13}$$

and the effective monochromatic image-radius for the system working over the field $\varphi\leqq\varphi_0$ as the square root of

$$E=\frac{2}{\pi H^2\varphi_0^2}\int_0^{\varphi_0}\varphi d\varphi\iint_{x^2+y^2\leqq H^2}[(\delta X)^2+(\delta Y)^2]dxdy$$

$$=\frac{2}{\pi\varphi_0^2}\int_0^{\varphi_0}\varphi d\varphi\iint_{u^2+v^2\leqq 1}[(\delta X)^2+(\delta Y)^2]dudv. \tag{14}$$

This definition will therefore be adopted in the present discussion. It is equivalent, if we imagine the ray-density distribution in each stellar image to be replaced by a mass-distribution, to defining the effective radius of the image as its radius of gyration about its center $(\delta X,\ \delta Y)=(0,0)$ and the effective monochromatic image-radius over the whole field as the mean-square average of these effective radii over the field-circle $\varphi\leqq\varphi_0$.

The value of $E=E(\varphi_0;a,N-n)$ is different for different wave-lengths, and in assessing the performance of the system it is necessary to take account of the fact that the images are not monochromatic.

4. EVALUATION OF E

The Eq. (12) can be written in the form

$$\left.\begin{aligned}\delta X&=K\mu^3[Pu(r^2-\tfrac{1}{2}a)+\varphi^2u^3]+O(K\mu^7)\\ \delta Y&=K\mu^3[Qv(r^2-\tfrac{1}{2}a)+\varphi^2u^2v]+O(K\mu^7)\end{aligned}\right\}, \tag{15}$$

where $\varphi=O(\mu)$ and the quantities

$$P=\frac{2n_0+1}{2n_0}\varphi^2+\frac{n-N}{n_0-1},\quad Q=P-\varphi^2 \tag{16}$$

are each $O(\mu^2)$.

By (11), (12) and (14) we can write

$$E=E^*[1+O(\mu^2)], \tag{17}$$

where

$$E^*=\frac{2}{\pi\varphi_0^2}\int_0^{\varphi_0}\varphi d\varphi\iint_{u^2+v^2\leqq 1}[(\delta X^*)^2+(\delta Y^*)^2]dudv \tag{18}$$

$$\left.\begin{aligned}\delta X^*&=K\mu^3[Pu(r^2-\tfrac{1}{2}a)+\varphi^2u^3]\\ \delta Y^*&=K\mu^3[Qv(r^2-\tfrac{1}{2}a)+\varphi^2u^2v]\end{aligned}\right\}. \tag{19}$$

From (19)

$$\begin{aligned}(\delta X^*)^2+(\delta Y^*)^2=K^2\mu^6[(P^2u^2+Q^2v^2)(r^2-\tfrac{1}{2}a)^2\\ +2\varphi^2u^2(Pu^2+Qv^2)(r^2-\tfrac{1}{2}a)+\varphi^4u^4r^2].\end{aligned} \tag{20}$$

Now

$$\begin{aligned}\iint_{u^2+v^2\leqq 1}(P^2u^2+Q^2v^2)(r^2-\tfrac{1}{2}a)^2dudv\\ =\tfrac{1}{2}(P^2+Q^2)\iint_{u^2+v^2\leqq 1}r^2(r^2-\tfrac{1}{2}a)^2dudv\\ =\frac{\pi}{16}(P^2+Q^2)(a^2-\tfrac{8}{3}a+2);\end{aligned} \tag{21}$$

$$\iint_{u^2+v^2\leqq 1}(Pu^2+Qv^2)u^2(r^2-\tfrac{1}{2}a)dudv=\frac{\pi}{16}(3P+Q)(\tfrac{1}{2}-\tfrac{1}{3}a); \tag{22}$$

$$\iint_{u^2+v^2\leqq 1}u^4r^2dudv=\int_0^{2\pi}\cos^4\theta d\theta\int_0^1 r^7dr=\frac{3\pi}{32}. \tag{23}$$

Hence,

$$\iint\limits_{u^2+v^2\le 1} [(\delta X^*)^2+(\delta Y^*)^2]dudv$$

$$=\frac{\pi}{16}K^2\mu^6[(P^2+Q^2)(a^2-\tfrac{8}{3}a+2)$$

$$+2(3P+Q)(\tfrac{1}{2}-\tfrac{1}{3}a)\varphi^2+\tfrac{3}{2}\varphi^4]. \quad (24)$$

Set

$$\alpha=\frac{1}{2n_0},\quad p=\frac{n-N}{n_0-1}\frac{1}{\varphi_0{}^2}; \quad (25)$$

then, by (16),

$$P=(1+\alpha)\varphi^2+p\varphi_0{}^2,\quad Q=\alpha\varphi^2+p\varphi_0{}^2;$$

$$\int_0^{\varphi_0}\varphi d\varphi(P^2+Q^2)=\varphi_0{}^6[\tfrac{1}{6}(1+2\alpha+2\alpha^2)+\tfrac{1}{2}p(1+2\alpha)+p^2],$$

$$\int_0^{\varphi_0}\varphi d\varphi\cdot\varphi^2(3P+Q)=\varphi_0{}^6[\tfrac{1}{6}(3+4\alpha)+p],$$

$$\int_0^{\varphi_0}\varphi d\varphi\cdot\varphi^4=\tfrac{1}{6}\varphi_0{}^6.$$

Substituting from (24) into (18) and using the last three equations, we finally obtain

$$E^*=\tfrac{1}{8}K^2\mu^6\varphi_0{}^4\{(a^2-8a/3+2)[\tfrac{1}{6}(1+2\alpha+2\alpha^2)$$
$$+\tfrac{1}{2}p(1+2\alpha)+p^2]$$
$$+2(\tfrac{1}{2}-\tfrac{1}{3}a)[\tfrac{1}{6}(3+4\alpha)+p]+\tfrac{1}{4}\}, \quad (26)$$

and the effective image-radius in n-light is

$$E^{\frac{1}{2}}=(E^*)^{\frac{1}{2}}\{1+O(\mu^2)\}.$$

If we define

$$e^*=(a^2-8a/3+2)[\tfrac{1}{6}(1+2\alpha+2\alpha^2)$$
$$+\tfrac{1}{2}p(1+2\alpha)+p^2]$$
$$+2(\tfrac{1}{2}-\tfrac{1}{3}a)[\tfrac{1}{6}(3+4\alpha)+p]+\tfrac{1}{4}, \quad (27)$$

where α, p are given as before by (25), then the effective image-radius in n-light is $\frac{1}{2}(e^*/2)^{\frac{1}{2}}K\mu^3\varphi_0{}^2$.

Figure 2 shows $(e^*)^{\frac{1}{2}}$ as a function of a and p. $(e^*)^{\frac{1}{2}}$ is effectively independent of the choice of n_0 since, as already remarked, this is true of $\delta X^*+i\delta Y^*$. In the actual computation of Fig. 2, n_0 was taken as 1.55.

5.

Conclusions of some practical interest can be drawn from Fig. 2 about the optimum choice of plate-profile in Schmidt cameras of given focal ratio and field-size, working in light of known wave-length range.

As an example, suppose that the camera works at $f/3.5$ over a field of 8° diameter, i.e., with obstruction ratio $\frac{1}{2}$. Then $\varphi_0=1/14$ and if $n_0=1.55$ we obtain from (25)

$$\alpha=\frac{1}{3.10}=0.32,\quad p=\frac{n-N}{0.55}\cdot(14)^2=356(n-N). \quad (28)$$

Suppose that at the two ends of the adopted wave-length-range n has the values 1.53123 and 1.54011. This would be the case with a plate of Chance glass HC525588

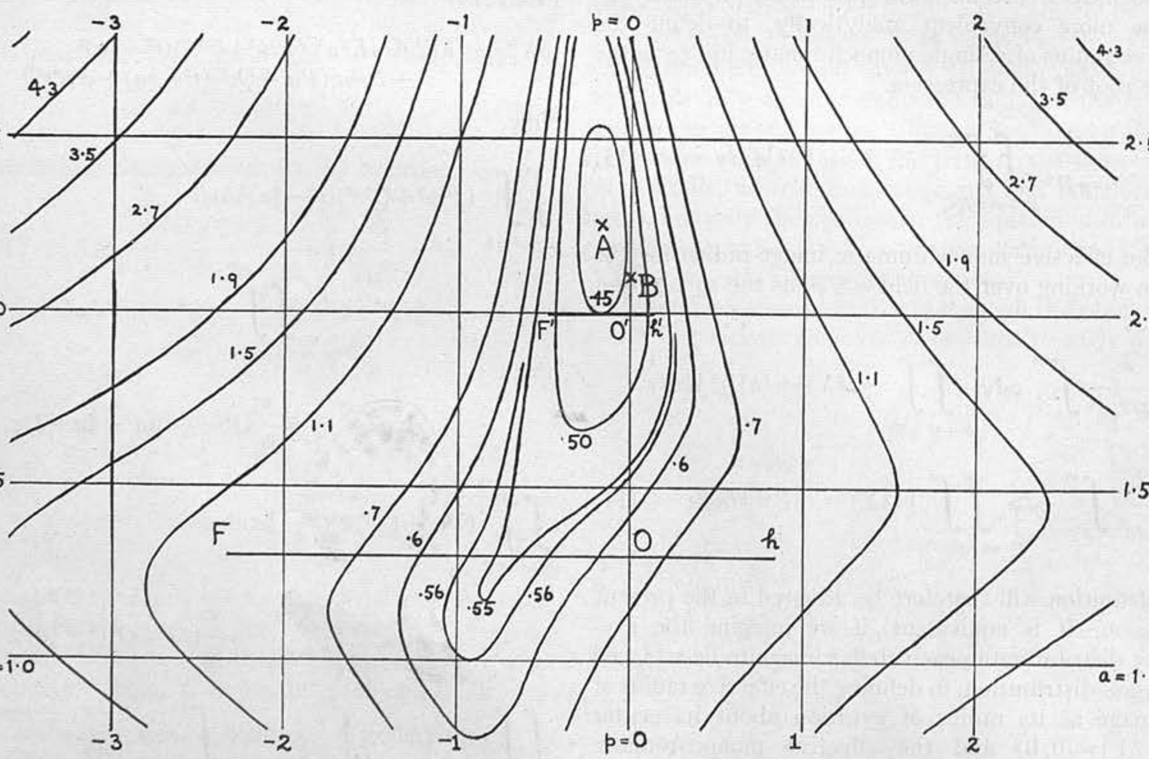


FIG. 2. Level curves for $(e^*)^{\frac{1}{2}}$ as a function of a and p.

and an adopted spectral range $F-h$, which is the most commonly used range of the astrographic spectrum. Then, by (28), the variation in p corresponding to this spectral range is

$$356(1.54011-1.53123)=3.16. \tag{29}$$

If a horizontal linear segment, of fixed length equal to 3.16 units, is moved into a position on Fig. 2 which makes $(e^*)^{\frac{1}{2}}$ as small as possible throughout the segment, then the common ordinate a of its points gives the optimum radius r_0 of the neutral zone on the plate by means of the equation

$$a=2r_0^2/R^2, \tag{30}$$

while the optimum value of N can be read off in the following way. To each point of the movable segment corresponds a particular value of n, these values running linearly from $n_F=1.53123$ at the left-hand end of the segment to $n_h=1.54011$ at its right-hand end. By (12) and (25), the value of n which falls on the line $p=0$ in Fig. 2 corresponds to the wave-length for which the system is axially stigmatic; the appropriate choice of N is therefore equal to this value.

To make $(e^*)^{\frac{1}{2}}$ as small as possible over the whole of the linear segment, it is sufficient to move the segment horizontally until the value of $(e^*)^{\frac{1}{2}}$ is balanced at its two end-points, and then, keeping this balance, to slide the segment up or down until the value of $(e^*)^{\frac{1}{2}}$ at the two end points is a minimum. The selected position of the segment is shown as FOh on the graph.

This method of balancing the aberrations is best suited to systems intended for use as spectrograph cameras. With systems intended solely for direct photography, it would be possible to proceed in a more refined way by taking into account the varying density of the photographic spectrum along the adopted wave-length range and trying to minimise the weighted mean of e^* over this range. But so much elaboration hardly seems appropriate in a discussion of the present character.

From Fig. 2 it is seen that, for this system, the 'best' values of a are those between 1.2 and 1.5, and that there is little to choose between different values in this range, provided that N is suitably chosen in each case. The effective image radius in F and h light is then

$$\tfrac{1}{2}(e^*/2)^{\frac{1}{2}}K\mu^3\varphi_0^2=0.136(e^*)^{\frac{1}{2}}=0.12 \text{ second of arc.}$$

As a second example, consider an $f/1$ camera with obstruction ratio $\frac{1}{3}$, working over a 19° field. Here we have

$$\varphi_0=\tfrac{1}{6}, \quad p=(n-N/0.55)\cdot 6^2=65.44(n-N),$$

and the length of the p-segment corresponding to a variation 0.00888 in the value of n is now only 0.6 unit. This shortening of the segment, by a factor 5 compared with the first example, is the graphical expression of the much smaller importance, in the second system, of chromatism relative to the monochromatic aberrations.

In this case the selected position of the segment, shown as $F'O'h'$ on the graph, gives

$$a=2.0, \quad N=1.5383$$

as the optimum values; then $(e^*)^{\frac{1}{2}}=0.52$ and the effective image radius is $31.66(e^*)^{\frac{1}{2}}=16.5$ seconds of arc. The value $a=2.0$ corresponds to a plate with its neutral zone at the edge of the aperture.

It will be seen that in the $f/3.5$ system of the first example the effect of the monochromatic errors is merely to displace the best choice of 'stigmatic wave-length' some way from the center towards the violet end of the effective range, and that even for the unusually wide field of 8° the best value of a is substantially the same as that which optimises the axial image.***

Besides indicating optimum values of a and N in numerical cases, Fig. 2 provides answers to one or two questions of a more general character. For example, the 'best monochromatic plate,' represented on the graph by the point A at which $(e^*)^{\frac{1}{2}}$ attains its least value 0.43, is the plate with

$$a=2.25, \quad p=\frac{n-N}{n_0-1}\frac{1}{\varphi_0^2}=-0.18,$$

that is,

$$a=2.25, \quad N=n+0.10\varphi_0^2.$$

The term $0.10\varphi_0^2$ indicates the axial undercorrection required to balance the monochromatic errors of the system over the given field. For a monochromatic Schmidt in which the corrector plate is designed to give axial stigmatism, the optimum value of a is 2.1; this system is represented on the graph by the point B at which $(e^*)^{\frac{1}{2}}$ takes its least value along the line $p=0$, namely 0.46.

More generally still, Fig. 2 shows that in astronomical Schmidt cameras of focal ratio $f/3$ or $f/3.5$ and with obstruction ratio not exceeding $\frac{1}{2}$, the monochromatic aberrations are of small importance compared with the color-error, while the opposite is true for large-field systems working at focal ratios near to $f/1$. An $f/2$ system occupies an intermediate position, in which neither type of aberration can be said to dominate the other.

Bibliography

(1) B. Strömgren, Vierteljahresschrift der Astr. Ges. **70**, 65–89 (1935).

(2) E. H. Linfoot, M.N.R.A.S. **109** (in press).

*** The effective radius of the axial image is least when $a=4/3$. This follows from (26) on substituting for p its value (25) and setting $\varphi_0=0$, when we obtain $E^*=\frac{1}{8}K^2\mu^6(n-N)^2/(n_0-1)^2 \times(a^2-8a/3+2)$. The diameters of the color-confusion circles are least when $a=3/2$ (Strömgren (1), p. 81).

4.4

(*Reprinted from Nature, Vol. 172, p. 615, October 3, 1953*)

MICROWAVE OPTICS

WITHIN the past two decades much research has been carried out on electromagnetic radiation of wave-lengths which belong to the transition region between the ordinary radio region and the optical region of the electromagnetic spectrum. Because of certain properties characteristic of this range, and also because the wave-lengths in this transition range are of the order of magnitude of most mechanical devices handled in laboratories, this new branch of electromagnetic theory, often called microwave optics, offers to physicists many interesting new opportunities. These are being well explored, for example, in connexion with communications and radar.

During the four days June 22–25, a symposium on microwave optics was held at McGill University, Montreal, under the joint sponsorship of the Eaton Electronic Research Laboratory, of Commission VI of the International Scientific Radio Union (U.R.S.I.) in Canada and the United States, and of the Electronic Research Directorate, Air Force Cambridge Research Center (U.S.A.). The symposium marked the scientific opening of the Eaton Electronic Research Laboratory, which now, in its third year of existence, is already playing a prominent part in this field of research, under the directorship of Prof. G. A. Woonton.

The main aim of the symposium was to bring together scientists who work not only on microwave optics but also on related subjects, such as light optics, information theory, operational methods, etc., thus enabling exchange of information and a closer collaboration between research workers in these fields.

The symposium was attended by about 140 scientists mainly from Canada and the United States and by seven guest speakers from Europe. During the four days, altogether sixty-four papers dealing with a wide range of subjects were presented. These numbers themselves indicate the great interest now shown in microwave optics. There were altogether eight half-day sessions on the following subjects: scattering (two sessions); microwave optical systems and aberrations; electromagnetic diffraction (two sessions); electromagnetism and diffraction; Fourier transforms and information theory; radio lenses, and other topics of special interest. In addition, there was a round-table conference on Fourier transforms, under the chairmanship of Dr. Roy C. Spencer, of the Air Force Cambridge Research Center. It is impossible in a short article to analyse the very extensive material presented at the symposium, and consequently only a brief account of the main subjects discussed will be given.

Papers presented in the session on scattering were mainly concerned with power series solutions of scattering problems, and with calculations and measurements of scattering cross-sections and radar cross-sections. Some integral-equation techniques were also described and applied to problems of diffraction by spheres, cones and wedges, subject to simple boundary conditions.

The session on microwave optical systems and aberrations began with a survey of some of the research carried out at the Naval Research Laboratory in Washington, and described investigations on microwave optical systems for use in radar scanning antennæ. Other papers were concerned with wide-angle focusing properties of paraboloidal mirrors, with applications of Rayleigh's tolerance criteria to microwave systems, and with the behaviour of the geometrical field near caustics. The ideal lens of Luneberg and its variants were discussed by several speakers in this session and also in the session on radio lenses. An interesting contribution to aberration theory was made in a paper concerned with characteristic curves in the image space, where a method was proposed for the derivation of functions orthogonal over parabolic and other domains; these should prove useful in certain theoretical investigations of diffraction.

In the session on radio lenses and other topics of special interest, an interesting account was given of some of the current researches carried out at the Ministry of Supply Telecommunication Research Establishment in Great Malvern; it dealt mainly with the use of optical techniques at millimetre wavelengths. Several papers discussed the behaviour of various artificial dielectrics.

The papers presented in the two sessions just mentioned indicated the similarity as well as the differences between systems employed in microwave optics and in light optics: microwave systems work as a rule with a smaller number of surfaces and larger numerical apertures; in microwave optics aspherical surfaces and asymmetrical combinations are the rule; in light optics they are the exception. In systems working with ordinary light the illumination is, as a rule, uniform; in microwave systems it is usually tapered.

Electromagnetism and especially electromagnetic diffraction were particularly well represented. Many of the contributions dealt with various aspects of diffraction by apertures in plane screens, and were followed by lively discussions. The theoretical papers were mainly concerned with approximate methods of solution, while those dealing with the experimental aspects described methods and results of measurements of diffraction fields. The literature of diffraction by apertures in plane screens is very extensive, and the comments from the audience indicated the difficulties faced by all but the experts on this class of problems when trying to assess the merits of the various methods of solution and the present state of knowledge in this field. The present writer shares the view expressed by some of the scientists taking part in the discussion, that a formal solution to a diffraction problem is often of relatively little value unless supplemented by computational data, and that a good deal more effort ought to be devoted to this end. (In this connexion it may be of some interest to mention that although Sommerfeld's classical solution of the half-plane problem was obtained as early as 1896 and numerous discussions of this problem have been published since, it was not until 1952 that detailed diagrams were given, showing the behaviour of the amplitude and phase and of the energy flow (Poynting vector) in the neighbourhood of the diffracting edge (W. Braunbek and G. Laukien, *Optik*, 9, 174 (1952); the diagrams bring out interesting features not previously suspected.)

Other papers dealt with asymptotic solutions (for large wave-numbers) of linear partial differential equations, with a new method (called 'cliff method') for approximate integration, and with the application of operational techniques to the development of diffraction integrals. Another group of papers was concerned with scalar treatment of diffraction problems and with the analogies between the field of geometrical optics and certain diffraction fields. A generalization of Fermat's Principle was suggested in a paper entitled "The Geometrical Theory of Diffraction". A good deal of attention was given to a paper concerned with the 'creeping wave' in the theory of diffraction. It outlined a method for problems concerned with the diffraction by objects not larger than about a hundred wave-lengths; the asymptotic expansions can then no longer be used, as there may be a considerable contribution from a wave creeping around the back of the object.

As already mentioned, one session was devoted to Fourier transforms and information theory. Accounts of useful Fourier and operational techniques were given and various applications described. Some analogies between optics and information theory were