New modal wave-front sensor: a theoretical analysis

Mark A. A. Neil, Martin J. Booth, and Tony Wilson

Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, United Kingdom

Received August 20, 1999; accepted February 15, 2000

We present a new design of a modal wave-front sensor capable of measuring directly the Zernike components of an aberrated wave front. The sensor shows good linearity for small aberration amplitudes and is particularly suitable for integration in a closed-loop adaptive system. We introduce a sensitivity matrix and show that it is sparse, and we derive conditions specifying which elements are necessarily zero. The sensor may be temporally or spatially multiplexed, the former using a reconfigurable optical element, the latter using a numerically optimized binary optical element. Different optimization schemes are discussed, and their performance is compared. © 2000 Optical Society of America [S0740-3232(00)00206-4]

OCIS codes: 010.7350, 010.1080, 050.1380, 070.2580.

1. INTRODUCTION

Recent developments in wave-front sensors for adaptive optics have mainly concentrated on applications in astronomy for which the most significant source of aberration is atmospheric turbulence.[1,2] The complex temporal and spatial nature of the optical effects of such turbulence coupled with the unavoidably low photon flux with which astronomers must work has led to the predominant use of Shack–Hartmann-type wave-front sensors or, to a lesser extent, zonal curvature sensing and shearing interferometry. Using real-time processing, one then extracts from these data the shape of the aberrated wave front or the magnitude of its modal aberration components, information that is used to drive the adaptive element.

It is generally considered that zonal wave-front sensing is preferable to modal sensing.[1,2] Tyson points out that, theoretically, the required number of zones is roughly equal to the number of modes that must be measured to represent a wave front to the same degree of accuracy; in practice, zonal methods are used, owing to the capabilities of available hardware. Since the correction of complex aberrations (e.g., due to atmospheric turbulence) would require the measurement of a large number of aberration modes, the higher modes would become difficult to measure directly and separation of the effects of, for example, astigmatism and coma would be problematic. Although for atmospheric turbulence it is true that a large number of aberration modes are required to represent the wave-front distortion accurately, in other applications this may not be the case. In confocal microscopy, for example, it has been shown that only the lowest-order Zernike aberrations significantly affect the point-spread function during focusing through a refractive-index mismatch.[3] This is due to a rapid decrease in the magnitude of the Zernike mode with increasing aberration order.[3,4] In such a situation, modal sensing becomes a viable option, especially if used in conjunction with a modal method of correction (e.g., bimorph or membrane mirrors).

In this paper we describe theoretically a novel wave-front sensor that measures directly the size of any chosen Zernike mode present in a wave front. We show that the response of the sensor is linear in mode amplitude, a, over a region around $a = 0$ and is thus ideally suited for incorporation into a closed-loop adaptive system where the wave-front correction occurs within the loop. We also discuss how the sensor can be implemented with a binary optical element and show how several of these sensors can be combined in a multiplexed version that allows simultaneous measurement of multiple different aberration modes.

2. CONCEPTUAL OPERATION OF THE WAVE-FRONT SENSOR

We propose the wave-front sensor shown conceptually in Fig. 1. We wish to measure the amount of a chosen aberration mode that is present in the unknown input wave front. We can describe the input wave front as a uniform intensity but with a phase function, $\Psi(\mathbf{r})$, representing the deviation of the phase of the aberrated wave front from that of an unaberrated plane wave in the aperture A. This input beam is split into two identical beams that then pass through biasing phase plates placed in the back focal plane of a lens. These transmissive phase plates are designed such that the first adds a bias aberration, $\Phi(\mathbf{r})$, to its input wave front and the second subtracts the same aberration from its input wave front. The beams are then focused by the lenses onto detector pinholes placed in the Fourier plane such that the intensity in the detector plane is given by the modulus squared of the Fourier transform (FT) of the total optical field just after the bias plates. In other words, the intensity presented to the first detector may be written as

$$I_1(\mathbf{v}) = |\mathcal{F}\{\exp[j\Psi(\mathbf{r}) + j\Phi(\mathbf{r})]\}|^2 \tag{1}$$

and to the second detector as

$$I_2(\mathbf{v}) = |\mathcal{F}\{\exp[j\Psi(\mathbf{r}) - j\Phi(\mathbf{r})]\}|^2, \tag{2}$$

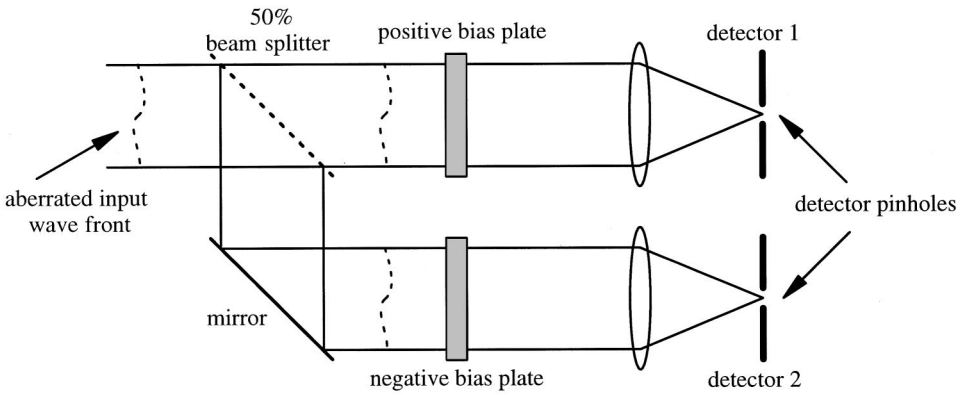

Fig. 1. Schematic description of the aberration sensor that uses biasing elements and Fourier transform lenses.

where $\mathcal{F}$ denotes the FT and $\mathbf{v}$ is the coordinate vector describing the detector plane.

This sensor is a generalization of the curvature (defocus) sensor in which one detector pinhole is placed in front of the nominal focal plane of a lens and one behind it.[1,5] For the special case in which the bias aberration is chosen such that $\Phi(\mathbf{r}) \propto r^2$, the sensor of Fig. 1 is optically identical to the curvature sensor.

The performance of the sensor depends on the aperture shape, the chosen bias aberration, and the size and shape of the pinhole, all of which we may choose. In general, our aim is to obtain the magnitude of the various modes in an orthogonal modal expansion of the input phase aberration $\Psi(\mathbf{r})$. We now proceed to show that if we choose our bias aberration to be exactly one of those orthogonal modes, then, to first order, the sensor will be sensitive to that mode in the input wave front while rejecting all others. Let us consider the case in which the detection pinholes are infinitely small and positioned on the optical axis. We take our sensor output signal to be the difference between the intensities at the two detectors. Let $\Phi(\mathbf{r}) = af_i(\mathbf{r})$ and $\Psi(\mathbf{r}) = bf_k(\mathbf{r})$, where f_i and f_k are orthogonal functions representing the modal expansion of the phase aberration and a and b are representing the magnitude of the input aberration and the applied bias. The intensities at the two detectors are then given by

$$I_{1,2}(\mathbf{0}) = \left| \int\int_A \exp(jaf_k \pm jbf_i)\mathrm{d}A \right|^2. \tag{3}$$

The sensor output is

$$\Delta I = I_1(\mathbf{0}) - I_2(\mathbf{0}), \tag{4}$$

and we define the sensitivity, S, of the sensor as the gradient of ΔI with respect to a, the size of the input aberration, at $a = 0$:

$$S = \left. \frac{\partial \Delta I}{\partial a} \right|_{a=0}. \tag{5}$$

The first derivative of the intensities with respect to a can be shown to be

$$\frac{\partial I_{1,2}}{\partial a} = 2\,\mathrm{Im}\left[\int\int_A \exp(jaf_k \pm jbf_i)\mathrm{d}A \int\int_A f_k \times \exp(jaf_k \mp jbf_i)\mathrm{d}A \right], \tag{6}$$

and it follows that the sensitivity is given by

$$S = 4\,\mathrm{Im}\left[\int\int_A \exp(jbf_i)\mathrm{d}A \int\int_A f_k \exp(-jbf_i)\mathrm{d}A \right]. \tag{7}$$

The two exponential terms can be expanded as a Maclaurin series to separate the imaginary parts, giving

$$S = 4\left(b \int\int_A f_i \mathrm{d}A \int\int_A f_k \mathrm{d}A - b \int\int_A f_i f_k \mathrm{d}A \times \int\int_A \mathrm{d}A + \cdots \right). \tag{8}$$

So, if the bias b is sufficiently small that only first-order terms are significant and the aberration modes f_i and f_k have zero mean across the aperture, only the second term remains on the right-hand side of Eq. (8). It follows that if f_i and f_k are members of a set of functions orthogonal over the aperture, A, then the sensitivity will be given by

$$S \approx -4bA \int\int_A f_i f_k \mathrm{d}A = -4bAC\,\delta_{ik}, \tag{9}$$

where δ_{ik} is a Kronecker delta and C is the constant of orthogonality. S will be nonzero only if f_i and f_k are the same function. The sensor would therefore respond only to an input aberration mode that is identical to its bias mode. This result is valid for any aperture shape and any set of orthogonal functions with zero mean defined over that aperture.

For the more general case we would like to use detector pinholes of finite size and a bias, b, of finite value in order to maximize the sensitivity. Under these circumstances there will in general be some cross sensitivity between modes. One particularly useful modal expansion is that which uses Zernike polynomials to describe the phase aberration over a circular aperture. We now proceed to analyze this system more fully, using Zernike modes as our orthogonal basis and finite-sized circularly symmetric detector pinholes. We show that the sensor still predominantly responds to its design mode but that there exists some cross-sensitivity between certain modes.

3. THEORETICAL TREATMENT OF A WAVE-FRONT SENSOR BASED ON ZERNIKE POLYNOMIALS

For our analysis of the sensor, we consider circular pinholes and apertures and use the Zernike circle polynomials as our orthonormal basis functions. We choose to use the definition of Noll[6] and the indexing scheme used by Mahajan.[7] This is described in Appendix A.

If we assume that only phase and not amplitude aberrations are present, any circular input wave front can be represented by the exponential of its phase function, $\Psi(r,\theta)$, which can be written as a series of Zernike polynomials.[6,8] The wave front may therefore be described by

$$\exp[j\Psi(r,\theta)]\mathrm{circ}(r) = \exp\left[j\sum_{k=0}^{\infty} a_k Z_k(r,\theta)\right]\mathrm{circ}(r), \tag{10}$$

where

$$\mathrm{circ}(r) = \begin{cases} 1 & \text{for } r \leqslant 1 \\ 0 & \text{for } r > 1, \end{cases} \tag{11}$$

and $Z_k(r,\theta)$ is a Zernike polynomial. For simplicity, from now on we shall omit $\mathrm{circ}(r)$ in our descriptions.

Let us consider the case in which the input wave front contains only one Zernike term, which is denoted by the subscript i, and the sensor is designed to detect the Zernike mode denoted by the subscript k. Let the amount of input aberration be a and the size of the bias be b. Thus the total wave-front distortion in the first path is given by $aZ_k(r,\theta) + bZ_i(r,\theta)$, whereas in the second path, where the bias is negative, the wave front is given by $aZ_k(r,\theta) - bZ_i(r,\theta)$. If we now introduce polar coordinates (ν, ξ) in the detector plane, we can write the intensities incident on the detector plane more generally from Eq. (3) as

$$I_{1,2}(\nu,\xi) = \frac{1}{8\pi^3}\left|\int_0^{2\pi}\int_0^1 \exp[jaZ_k(r,\theta) \pm jbZ_i(r,\theta)] \times \exp[j\nu r\cos(\theta - \xi)]r\,\mathrm{d}r\,\mathrm{d}\theta\right|^2. \tag{12}$$

The factor of $1/8\pi^3$ ensures that the total summed power in the two detector planes is equal to unity. For finite-sized circular detector pinholes of radius ν_p, we integrate the intensity in the detector plane and obtain the detected difference signal (the signal from detector 1 minus the signal from detector 2) as

$$\Delta W_{ik} = \int_0^{2\pi}\int_0^{\nu_p} I_1(\nu,\xi)\nu\,\mathrm{d}\nu\,\mathrm{d}\xi - \int_0^{2\pi}\int_0^{\nu_p} I_2(\nu,\xi)\nu\,\mathrm{d}\nu\,\mathrm{d}\xi, \tag{13}$$

and again as above we define the sensitivity of a mode i sensor to an input mode k as

$$S_{ik} = \left.\frac{\partial\Delta W_{ik}}{\partial a}\right|_{a=0} = \int_0^{\nu_p}\int_0^{2\pi}\left.\frac{\partial I_1}{\partial a}\right|_{a=0} - \left.\frac{\partial I_2}{\partial a}\right|_{a=0}\mathrm{d}\xi\,\nu\,\mathrm{d}\nu. \tag{14}$$

We can represent the sensitivities by a matrix, **S**, in which each row represents a sensor designed for a particular Zernike mode and each column represents the input wave-front mode. If we take a general input wave front containing several Zernike modes of amplitude a_k and express it as a vector **a**, then the matrix vector product $\mathbf{w} = \mathbf{Sa}$ gives us the signals detected on the output of the sensor (to a linear approximation). In general use, however, we detect the signals **w** and would like to know which Zernike modes are present in the input wave front. This can be done simply by computing the inverse of **S**, and then the amplitudes of the Zernike modes are given by $\mathbf{a} = \mathbf{S}^{-1}\mathbf{w}$. In the ideal case, **S** would be diagonal with all of the off-diagonal elements identically zero. In practice, this does not hold, but we will show that the matrix **S** is indeed sparse. Most elements will be shown to be necessarily zero, and general conditions will be derived to determine which elements are nonzero. First, we take Eq. (12) and write it as

$$\begin{aligned} I_{1,2}(\nu,\xi) &= \frac{1}{8\pi^3}|\mathcal{F}[\exp(jaZ_k \pm jbZ_i)]|^2 \\ &= \frac{1}{8\pi^3}\mathcal{F}[\exp(jaZ_k \pm jbZ_i)] \times \mathcal{F}[\exp(jaZ_k \pm jbZ_i)]^*, \end{aligned} \tag{15}$$

where $*$ denotes the complex conjugate. Differentiation by parts with respect to a gives the derivative at $a = 0$ as

$$\left.\frac{\partial I_{1,2}}{\partial a}\right|_{a=0} = \frac{1}{4\pi^3}\mathrm{Im}\{\mathcal{F}[\exp(\pm jbZ_i)]\mathcal{F}[\tilde{Z}_k\exp(\mp jb\tilde{Z}_i)]\}, \tag{16}$$

where we define

$$\tilde{Z}_i(r,\theta) = Z_i(r,\theta + \pi). \tag{17}$$

Let us now take the case in which $Z_i(r,\theta) = R_i(r)\cos(m_i\theta)$ and use the following expansion of the exponential in terms of Bessel functions,[9]

$$\exp(jz\cos\phi) = \sum_{p=-\infty}^{\infty} j^p J_p(z)\cos(p\phi), \tag{18}$$

where $J_p(\cdot)$ is the pth-order Bessel function of the first kind. The first FT in Eq. (16) becomes

$$\mathcal{F}[\exp(\pm jbZ_i)] = \sum_{p=-\infty}^{\infty} j^p\mathcal{F}[J_p(\pm bR_i)\cos(pm_i\theta)]. \tag{19}$$

In this summation the functions are of the form $f(r)\cos(\gamma\theta)$. The properties of the FT's of such functions are summarized in Appendix B and further allow us to write,

$$\mathcal{F}[\exp(\pm jbZ_i)] = 2\pi\sum_{p=-\infty}^{\infty} j^{p(m_i+1)}(\pm)^p\cos(pm_i\xi) \times \int_0^1 J_p(bR_i)J_{pm_i}(\nu r)r\,\mathrm{d}r, \tag{20}$$

and, similarly choosing $Z_k(r,\theta) = R_k(r)\cos(m_k\theta)$, we can write the second FT in Eq. (16) as

$$\begin{aligned}\mathcal{F}[\tilde{Z}_k \exp(\mp jb\tilde{Z}_i)] &= \pi \sum_{q=-\infty}^{\infty} j^{q(\mp)^p}\Big\{ j^{qm_i+m_k}\cos[(qm_i+m_k)\xi'] \\ &\times \int_0^1 R_k J_q(bR_i) J_{qm_i+m_k}(\nu r) r\,\mathrm{d}r \\ &+ j^{qm_i-m_k}\cos[(qm_i - m_k)\xi'] \\ &\times \int_0^1 R_k J_q(bR_i) J_{qm_i-m_k}(\nu r) r\,\mathrm{d}r\Big\}, \quad (21)\end{aligned}$$

where $\xi' = \xi + \pi$. The differential intensity signal at the detector plane is therefore given by

$$\begin{aligned}\left.\frac{\partial I_1}{\partial a}\right|_{a=0} - \left.\frac{\partial I_2}{\partial a}\right|_{a=0} &= \frac{1}{\pi}\mathrm{Im}\Bigg(\sum_{\substack{p=-\infty \\ (p+q)\ \text{odd}}}^{\infty}\sum_{q=-\infty}^{\infty} (-1)^q j^{(p+q)(m_i+1)} \\ &\times \Big\{ j^{m_k}\cos(pm_i\xi)\cos[(qm_i+m_k)\xi'] \\ &\times \int_0^1 J_p(bR_i) J_{pm_i}(\nu r) r\,\mathrm{d}r \\ &\times \int_0^1 R_k J_q(bR_i) J_{qm_i+m_k}(\nu r) r\,\mathrm{d}r \\ &+ j^{-m_k}\cos(pm_i\xi)\cos[(qm_i - m_k)\xi'] \\ &\times \int_0^1 J_p(bR_i) J_{pm_i}(\nu r) r\,\mathrm{d}r \\ &\times \int_0^1 R_k J_q(bR_i) J_{qm_i - m_k}(\nu r) r\,\mathrm{d}r\Big\}\Bigg), \quad (22)\end{aligned}$$

where the condition that $(p + q)$ must be odd is applied to the summations. To calculate the detected power sensitivity S_{ik}, we must now integrate over the detector pinholes according to Eq. (14). At this point we can use the standard orthogonality relationships between sines and cosines for the integral over ξ. In particular, we now note that if we had chosen Z_k to have a negative azimuthal index, then the terms in Eq. (22) would all be of the form $\cos(A).\sin(B)$ and when integrated over ξ would give exactly zero sensitivity. This implies that a sensor for a mode of positive azimuthal order is insensitive to modes of negative azimuthal order. Likewise, by a simple coordinate transform (rotating by $\pi/2$ about the origin) we can similarly deduce that a sensor for a mode of negative azimuthal order is insensitive to modes of positive azimuthal order.

For cases where m_i is zero or positive, then the only way that terms in the summation integrate to nonzero results are when

$$pm_i = \pm(qm_i + m_k) \quad \textbf{or} \quad pm_i = \pm(qm_i - m_k). \quad (23)$$

If $m_i = 0$ then nonzero sensitivity is obtained only when $m_k = 0$, thus a sensor for a circularly symmetric mode is only sensitive to other circularly symmetric modes and likewise a sensor for a noncircularly symmetric mode is insensitive to circularly symmetric modes. Applying this condition of $m_i = m_k = 0$ to Eq. (22) and integrating over ξ according to Eq. (14) gives

$$\begin{aligned}S_{ik}|_{m_i=m_k=0} &= 4\,\mathrm{Im}\Bigg\{ \sum_{\substack{p=-\infty \\ (p+q)\ \text{odd}}}^{\infty}\sum_{q=-\infty}^{\infty} (-1)^q (j)^{(p+q)} \\ &\times \int_0^{\nu_p}\Big[\int_0^1 J_p(bR_i) J_0(\nu r) r\,\mathrm{d}r \\ &\times \int_0^1 R_k J_q(bR_i) J_0(\nu r) r\,\mathrm{d}r\Big]\nu\,\mathrm{d}\nu\Bigg\} \\ &= 4\,\mathrm{Im}\Bigg\{\int_0^{\nu_p}\Big[\int_0^1 \exp(jbR_i) J_0(\nu r) r\,\mathrm{d}r \\ &\times \int_0^1 R_k \exp(-jbR_i) J_0(\nu r) r\,\mathrm{d}r\Big]\nu\,\mathrm{d}\nu\Bigg\}. \quad (24)\end{aligned}$$

Turning to now to the case of modes of positive azimuthal order, we note that the conditions set by Eq. (23) along with the condition that $(p + q)$ must be odd results in the fact that m_k/m_i must itself be an odd integer if nonzero terms are to exist. Thus a sensor for a mode of a given azimuthal order is insensitive to other modes whose azimuthal order is not an odd multiple of that sensor azimuthal order. Applying this condition to Eq. (22) and again integrating according to Eq. (14) therefore gives

$$\begin{aligned}S_{ik}|_{m_k/m_i\ =\ \text{odd}} &= 2(-1)^{\frac{m_k+m_i}{2m_i}} \\ &\times \int_0^{\nu_p}\Bigg\{\sum_{p=-\infty}^{\infty}\int_0^1 J_p(bR_i) J_{pm_i}(\nu r) r\,\mathrm{d}r \\ &\times \int_0^1 R_k\Big[J_{p+\frac{m_k}{m_i}}(bR_i) \\ &- J_{p-\frac{m_k}{m_i}}(bR_i)\Big] J_{pm_i}(\nu r) r\,\mathrm{d}r\Bigg\}\nu\,\mathrm{d}\nu. \quad (25)\end{aligned}$$

Equation (25) can be further simplified as terms in the summation for negative values of p are identical to those for positive values of p. This fact, along with the fact the summation converges very quickly for low and moderate values of b and ν_p, makes Eq. (25) particularly suitable for numerical computation of sensitivity matrix coefficients. We should also note that the same result would be obtained here had we chosen $Z_i(r,\theta) = R_i(r)\sin(-m_i\theta)$ and $Z_k(r,\theta) = R_k(r)\sin(-m_k\theta)$, since this simply corresponds to a change of coordinate system.

These results show that most elements of the sensitivity matrix, **S**, are identically zero and that the matrix is always sparse. Moreover, this result is independent of the nature of the radial components of Z_i and Z_k and would apply to other choices of radial functions. Therefore this result also holds for annular apertures and pinholes. In this case the Zernike circle polynomials are replaced by Zernike annular polynomials,[10] which are closely related to the circle polynomials as given in Eq. (A1), but the radial function is replaced by

$$R_n{}^m\left[\left(\frac{r^2-\epsilon^2}{1-\epsilon^2}\right)^{1/2}\right], \tag{26}$$

where ϵ is the inner radius of the annulus. Since the azimuthal component retains the same form, we again find that the elements of **S** are nonzero only if m_k is an odd multiple of m_i.

There are two parameters with which we can tune the performance of the sensor: the magnitude of the bias, b, and the size of the pinhole radius, ν_p. Ideally, we require a sensor for which S_{ii} is as large as possible while all S_{ik} are zero or very small when $i \neq k$. Consider the sensitivity when the pinhole is very small, $\nu_p \to 0$. This can be derived from Eq. (24) for the case when $m_i = m_k = 0$ as

$$S_{ik}|_{m_i=m_k=0} = 2\nu_p{}^2\,\mathrm{Im}\left[\int_0^1 \exp(jbR_i)r\,\mathrm{d}r \times \int_0^1 R_k \exp(-jbR_i)r\,\mathrm{d}r\right]. \tag{27}$$

It can also be seen that for sufficiently small bias, b, such that b^2 and higher terms are insignificant, the sensitivity becomes

$$S_{ik}|_{m_i=m_k=0} = 2b\nu_p{}^2\left(\int_0^1 R_i\,r\,\mathrm{d}r\int_0^1 R_k\,r\,\mathrm{d}r - \int_0^1 r\,\mathrm{d}r\int_0^1 R_kR_i\,r\,\mathrm{d}r\right); \tag{28}$$

however, in this case the first two integrals are zero and the last integral is equal to $\delta_{ik}/2$ from the orthogonality relationship [Eq. (A4)], so the sensitivity becomes

$$S_{ik} = -\frac{b}{2}\nu_p{}^2\delta_{ik}. \tag{29}$$

For the remaining cases, when $m_i = m_k \neq 0$, we can use the fact that for a small argument a Bessel function is given approximately by the first term in its Maclaurin power series expansion[11] such that as $\nu \to 0$:

$$J_{pm_i}(\nu r) \approx \frac{1}{(pm_i)!}\left(\frac{\nu r}{2}\right)^{pm_i}. \tag{30}$$

It follows that the terms in the summation of Eq. (25) are vanishingly small unless $p = 0$, leaving

$$S_{ik}|_{m_k/m_i\,=\,\mathrm{odd}} \approx (-1)^{\frac{m_k+m_i}{2m_k}}\nu_p{}^2\int_0^1 J_0(bR_i)r\,\mathrm{d}r \times \int_0^1 R_k[J_{\frac{m_k}{m_i}}(bR_i) - J_{-\frac{m_k}{m_i}}(bR_i)]r\,\mathrm{d}r. \tag{31}$$

Using the relationship $J_{-n}(.) = (-)^nJ_n(.)$ [Ref. 11] and Eq. (30), we find that for small b then terms in b exist only when $m_i = m_k$ and we obtain

$$S_{ik}|_{m_i/m_k\,=\,1} \approx -b\nu_p{}^2\int_0^1 r\,\mathrm{d}r\int_0^1 R_kR_i r\,\mathrm{d}r = -\frac{b}{2}\nu_p{}^2\delta_{ik}. \tag{32}$$

In addition to the above, we find that for the cases of off-diagonal sensitivity that are not identically equal to zero (i.e., when $i \neq k$ and m_k/m_i is odd), then the residual sensitivity from Eqs. (24) and (25) is at most $O(b\nu_p^4) + O(b^3\nu_p^2)$. Thus, to summarize, we find that the terms in the sensitivity matrix as the bias and detector pinhole radius tend to zero are given by

$$S_{ik} = \begin{cases} -\frac{b}{2}\nu_p{}^2 & i = k \\ \leqslant O(b^3\nu_p{}^2) + O(b\nu_p{}^4) & i \neq k,\ \frac{m_k}{m_i}\ \text{is odd} \\ 0 & \text{otherwise} \end{cases}. \tag{33}$$

4. SENSITIVITY MATRIX

The above analysis shows that the sensitivity matrix produced for this type of sensor is indeed sparse but that there is also a trade-off between maximizing the size of the on-diagonal elements while minimizing the size of the off-diagonal elements. In particular, Eq. (33) shows that to make the on-diagonal elements as large as possible, the pinhole size and the bias should be increased, whereas to minimize the relative size of the off-diagonal elements, the pinhole size and bias should be kept small. A further complication arises in that beyond the limits of validity of the approximations made in Eqs. (29) and (32) the on-diagonal sensitivity itself saturates and reaches a maximum before declining at higher values of bias and pinhole size.

As an example, we present results for a sensor designed to operate with near-maximum on-diagonal sensitivity alone. We find that a fairly broad maximum sensitivity peak is achieved for all modes in the region of $b = 0.7$ and $\nu_p = \pi$. Table 1 shows the matrix, **S**, for modes $i,k = 4$ to 19, obtained with these values of bias and pinhole size. Each row represents a sensor designed for mode i; each column represents an input mode k. We do not include tip, tilt, and piston, since piston has no effect and we consider it better to use other methods to detect tip and tilt. It can be seen that elements of **S** are zero unless m_k is an odd multiple of m_i and the matrix is

Table 1. Sensitivity Matrix, S

Input Mode	Sensor Designed for Mode															
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
4	−0.664	0	0	0	0	0	0	0.227	0	0	0	0	0	0	0	0
5	0	−0.492	0	0	0	0	0	0	0.157	0	0	0	0	0	0	0
6	0	0	−0.492	0	0	0	0	0	0	0.157	0	0	0	0	0	0
7	0	0	0	−0.685	0	0.035	0	0	0	0	0	0	0.243	0	0.052	0
8	0	0	0	0	−0.685	0	−0.035	0	0	0	0	0	0	0.243	0	−0.052
9	0	0	0	0	0	−0.506	0	0	0	0	0	0	0	0	0.188	0
10	0	0	0	0	0	0	−0.506	0	0	0	0	0	0	0	0	0.188
11	0.178	0	0	0	0	0	0	−0.754	0	0	0	0	0	0	0	0
12	0	0.147	0	0	0	0	0	0	−0.687	0	0	0	0	0	0	0
13	0	0	0.147	0	0	0	0	0	0	−0.687	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	−0.469	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	−0.469	0	0	0	0
16	0	0	0	0.206	0	0.008	0	0	0	0	0	0	−0.712	0	0.030	0
17	0	0	0	0	0.206	0	−0.008	0	0	0	0	0	0	−0.712	0	−0.030
18	0	0	0	0	0	0.148	0	0	0	0	0	0	0	0	−0.652	0
19	0	0	0	0	0	0	0.148	0	0	0	0	0	0	0	0	−0.652

Table 2. Inverse of the Sensitivity Matrix, $\mathbf{S}^{-1}$

Input Mode	Sensor Designed for Mode															
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
4	−1.637	0	0	0	0	0	0	−0.493	0	0	0	0	0	0	0	0
5	0	−2.180	0	0	0	0	0	0	−0.499	0	0	0	0	0	0	0
6	0	0	−2.180	0	0	0	0	0	0	−0.499	0	0	0	0	0	0
7	0	0	0	−1.627	0	−0.183	0	0	0	0	0	0	−0.556	0	−0.207	0
8	0	0	0	0	−1.627	0	0.183	0	0	0	0	0	0	−0.556	0	0.207
9	0	0	0	0	0	−2.157	0	0	0	0	0	0	0	0	−0.622	0
10	0	0	0	0	0	0	−2.157	0	0	0	0	0	0	0	0	−0.622
11	−0.386	0	0	0	0	0	0	−1.442	0	0	0	0	0	0	0	0
12	0	−0.465	0	0	0	0	0	0	−1.563	0	0	0	0	0	0	0
13	0	0	−0.465	0	0	0	0	0	0	−1.563	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	−2.133	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	−2.133	0	0	0	0
16	0	0	0	−0.470	0	−0.099	0	0	0	0	0	0	−1.565	0	−0.137	0
17	0	0	0	0	−0.470	0	0.099	0	0	0	0	0	0	−1.565	0	0.137
18	0	0	0	0	0	−0.491	0	0	0	0	0	0	0	0	−1.676	0
19	0	0	0	0	0	0	−0.491	0	0	0	0	0	0	0	0	−1.676

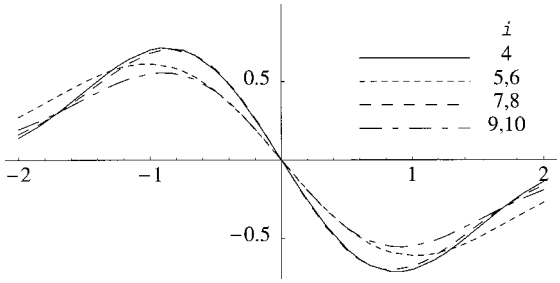

Fig. 2. Output signal from wave-front sensors for Zernike modes 4–10. The bias $b = 0.7$, and the pinhole radius $\nu_p = \pi$.

dominated by the diagonal terms. However, for each aberration sensor there is significant sensitivity to other modes; for example, the defocus sensor will also detect spherical aberration. We also note that sensitivity between modes of different azimuthal order ($m_i \neq m_k$), although not actually zero, is significantly less than that between modes of the same azimuthal order ($m_i = m_k$). The inverse matrix $\mathbf{S}^{-1}$ is given in Table 2.

Figure 2 shows the output signal of the sensor calculated with Eq. (13) when the input and the bias aberrations are the same, for Zernike modes 4–10. Each mode shows a similar response curve, which is linear in a re-

gion around $a = 0$. The curves that overlap exactly are where the aberrations that are being sensed have the same azimuthal order but opposite sign. Even then the similarities between the different curves are notable.

5. IMPLEMENTATION IN BINARY OPTICS

An alternative implementation that is more versatile than that of Fig. 1 is to use binary phase optics to generate the positive and negative bias by use of a single diffractive element. The method, which originated in computer-generated holography,[12] allows the generation of focal spots containing arbitrary amounts of any aberration.[13]

Suppose we take a wave front $f(x, y) = \exp\{j[\phi(x, y) + \tau(x, y)]\}$, including a desired bias aberration, $\phi(x, y)$ and a linear phase tilt, $\tau(x, y)$. It can be shown that binarizing this wave front by taking the sign of the real part of $f(x, y)$ produces a binary wave front $g(x, y)$ with amplitude $+1$ or -1 that can be described by the Fourier series expansion as

$$g(x, y) = \frac{2}{\pi}\left\{\exp[j(\phi + \tau)] + \exp[-j(\phi + \tau)] - \frac{1}{3}\exp[j3(\phi + \tau)] - \frac{1}{3}\exp[-j3(\phi + \tau)] + \cdots\right\}. \quad (34)$$

In the Fourier plane of a lens the individual components in the expansion are now diffracted to spatially separated positions because each generated component has a different overall tilt. This results in diffracted orders positioned at relative displacements of $\{+1, -1, +3, -3, +5, -5 \ldots\}$ from the axis with relative powers of $\{1, 1, 1/9, 1/9, 1/25, 1/25, \ldots\}$. Each order itself carries a correspondingly scaled analogue phase modulation $\{+1,-1,+3,-3,+5,-5 \ldots\} \times \phi(x, y)$. Clearly, the $+1$ and -1 diffraction orders here correspond to our positive bias and negative bias, respectively. In the configuration of Fig. 3, the aberrated input wave is incident on such a binary phase plate that lies in the back focal plane of the lens. The intensity distribution in the Fourier plane of the lens then contains all of the above-mentioned diffraction orders separated spatially. The $+1$ and -1 orders are identical to the intensity distributions at detectors 1 and 2 in Fig. 1 but reduced in magnitude by a factor $8/\pi^2$, or ~81%. Suitably placed pinhole detectors obscure the light from the other orders. This method of using a binary phase element is thus not only convenient in that it multiplexes both bias elements in one but also in that it provides a highly accurate method of defining the required bias patterns with a digital binary element.

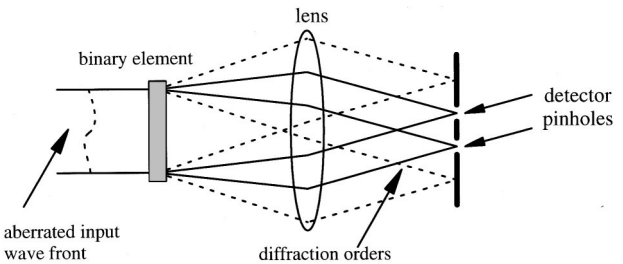

Fig. 3. Schematic description of the aberration sensor that uses a binary phase optical element to produce the two "biased" spots.

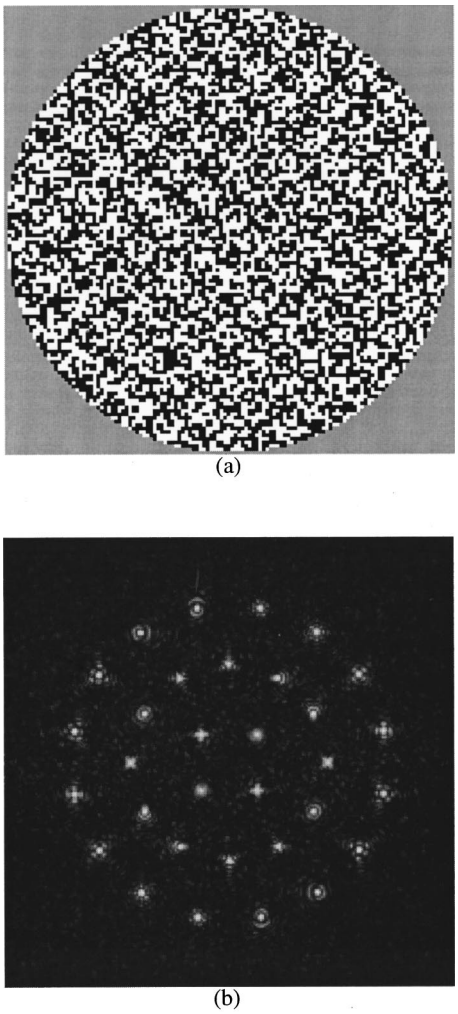

Fig. 4. (a) Optimized multiplexed wave-front sensing mask (transmission: white = 1, black = −1, and gray = 0), (b) simulated output intensity in focal plane for plane wave-front input (logarithmic scale over 20-dB range).

If the binary element were reconfigurable, for example, if it were a spatial light modulator, it would be possible to change the bias aberration so that the sensor could sequentially measure each Zernike mode. It would, however, be more desirable to have a single fixed element that could allow us to measure the different modes simultaneously. For a multiplexed filter of this kind the binary optical element produces a number of symmetrical pairs of spots in different positions in the focal plane, each pair corresponding to a different Zernike mode. A suitably designed array of pinhole detectors can then be used to derive the output signal for each mode.

Such a mask can be designed with standard optimization techniques (e.g., direct binary search) where the binary phase element starts as a grid of random binary pixels and we cycle through each pixel in a pseudorandom sequence, choosing the optimum value that maximizes

some cost function associated with the intensity sensitivity to each mode at the center of each spot according to the equation

$$\left.\frac{\partial I(\nu,\xi)}{\partial a_k}\right|_{a_k=0} = -\frac{1}{4\pi^3}\mathrm{Re}\left\{\int_0^1\int_0^{2\pi} U(r,\theta) \times \exp[jr\nu\cos(\theta-\xi)]\,\mathrm{d}\theta r\,\mathrm{d}r \times \int_0^1\int_0^{2\pi} jZ_k U^*(r,\theta) \times \exp[-jr\nu\cos(\theta-\xi)]\,\mathrm{d}\theta r\,\mathrm{d}r\right\}, \tag{35}$$

where $U(r,\theta)$ is the binary optical element transmission and Z_k is the aberration to be sensed. One of these binary phase masks of size 128×128 pixels and its corresponding detector plane intensity pattern are shown in Fig. 4. The characteristic point-spread function corresponding to the aberration to be sensed can clearly be seen in each spot pair.

Calculation of the detected power is best achieved by noting that the power falling onto a detector of a fixed size positioned anywhere in the output plane is given by the convolution of the intensity pattern with that detector pattern. Thus we can get the differential power sensitivity by performing a convolution of the differential intensity pattern with the detector pattern, which can be achieved by simply using standard fast Fourier transform techniques. Figure 5 shows diagrammatic representations of the sensitivity matrices shown for (a) the previously described time-multiplexed sensor (Table 1) and (b) the optimized spatially multiplexed sensor. For the latter we again use the value of pinhole size $\nu_p = \pi$ but rely on the optimization procedure to maximize the on-diagonal sensitivity to a uniform level and simultaneously to suppress off-diagonal sensitivity where its value is known to be exactly zero in the nonmultiplexed case. The impulse height in these diagrams represents the sensitivity matrix element values, with the top corner being the (0, 0) element of each matrix.

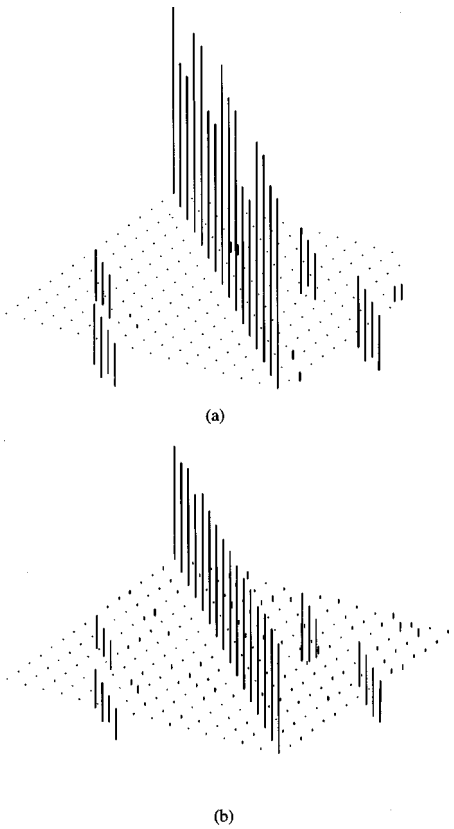

Fig. 5. Sensitivity matrices: (a) time multiplexed, (b) spatially multiplexed, optimized for sensitivity with suppression of certain off-diagonal elements and weighting of diagonal elements for better uniformity.

Figure 5(a) has been scaled to that of Fig. 5(b), assuming a theoretical 81% efficiency obtained by use of a binary optical element, and a factor of 16 because the sensor must be time multiplexed 16 times. Even with this normalization, the on-diagonal elements of the spatially multiplexed device are still lower than those of the temporally multiplexed device. Apart from the inevitable penalty of producing a multiplexed element, this can in part be explained by the fact that the spatially multiplexed device uses a significant part of the space bandwidth available on the 128-pixel-diameter element: Full use of the space bandwidth would incur a reduction of efficiency to 60% (owing to the pixelation, the output intensity pattern is effectively apodized by a sinc function). Increasing the number of pixels in the element would reduce this effect significantly. One advantage of the optimized element is that additional flexibility can be introduced, for example, to equalize the height of on-diagonal terms or possibly to further suppress off-diagonal terms.

6. CONCLUSIONS

In this paper we have presented a theoretical analysis of a new modal wave-front sensor intended for use in a closed-loop adaptive system. Conceptually, the sensor consists of two beam paths: In the first path a bias aberration is added to the input wave front; in the second path the same bias aberration is subtracted from the input wave front. Aberrated focal spots are produced by focusing these beams onto two pinhole photodetectors. The output signal from the sensor is taken as the difference between the two photodetector signals. When Zernike circle polynomials are used as the bias modes, the response of the sensor can be summarized by a sparse sensitivity matrix that is easily inverted to allow calculation of the magnitude of each Zernike component in the input wave front. We have shown that most elements of the matrix are necessarily zero. In the limiting case, when the detector pinholes and the bias are small, the sensitivity matrix is diagonal. We have also shown that other choices of radial polynomials and detector geometries will yield similar results.

The individual sensors can be temporally or spatially multiplexed to allow sequential or simultaneous measurement, respectively, of each aberration mode. A numerically optimized binary optical element may be used to generate the focal spot pattern required for the spatially multiplexed sensor. We have shown how the optimiza-

tion process may be used to tailor the properties of the sensitivity matrix, for example, to suppress cross sensitivity between modes or to equalize the on-diagonal elements of the matrix.

In a future paper we shall consider the application of this new sensor to confocal microscopy and discuss theoretically its operation for different imaging modes.

APPENDIX A: ZERNIKE CIRCLE POLYNOMIALS

We have chosen to use a single index as described by Mahajan[7] to refer to the Zernike circle polynomials. Many authors (e.g., Ref. 6) use two indices n and m that relate to the order of the radial and azimuthal variations, respectively. This method, although less compact, is more intuitive than a single index, so it is useful to be able to switch between the two systems. A small alteration of Mahajan's system allows each polynomial to be uniquely defined by either the single index, i, or the dual index (n, m). We refer to those polynomials containing a cosine term as having a positive value of m, whereas those with a sine term take negative m. n may take the value of any positive integer or zero; m is restricted to values given by the conditions that $n \geqslant |m|$ and $n - |m|$ is even. The single index i starts at 1 for $n = 0$ and rises in integer steps, first with rising n, then with rising magnitude of m, and finally with the sign of m. The polynomials may then be defined as

$$Z_n^m(r,\theta) = \begin{cases} \sqrt{2(n+1)}R_n^{-m}(r)\sin(-m\theta) & m < 0 \\ \sqrt{n+1}R_n^0(r) & m = 0, \\ \sqrt{2(n+1)}R_n^m(r)\cos(m\theta) & m > 0 \end{cases} \tag{A1}$$

with

$$R_n^m(r) = \sum_{s=0}^{(n-|m|)/2} \left[\frac{(-1)^s (n-s)!}{s![(n+m)/2-s]![(n-m)/2-s]!} r^{n-2s} \right] \mathrm{circ}(r). \tag{A2}$$

Table 3. First 22 Zernike Circle Polynomials

i	n	m	$Z_i(r,\theta)$	Aberration Term
1	0	0	1	Piston
2	1	1	$2r\cos(\theta)$	Tip
3	1	−1	$2r\sin(\theta)$	Tilt
4	2	0	$\sqrt{3}(2r^2-1)$	Defocus
5	2	2	$2\sqrt{3}r^2\cos(2\theta)$	Astigmatism
6	2	−2	$2\sqrt{3}r^2\sin(2\theta)$	Astigmatism
7	3	1	$2\sqrt{2}(3r^3-2r)\cos(\theta)$	Coma
8	3	−1	$2\sqrt{2}(3r^3-2r)\sin(\theta)$	Coma
9	3	3	$2\sqrt{2}r^3\cos(3\theta)$	
10	3	−3	$2\sqrt{2}r^3\sin(3\theta)$	
11	4	0	$\sqrt{5}(6r^4-6r^2+1)$	Spherical (1st)
12	4	2	$\sqrt{10}(4r^4-3r^2)\cos(2\theta)$	
13	4	−2	$\sqrt{10}(4r^4-3r^2)\sin(2\theta)$	
14	4	4	$\sqrt{10}r^4\cos(4\theta)$	
15	4	−4	$\sqrt{10}r^4\sin(4\theta)$	
16	5	1	$2\sqrt{3}(10r^5-12r^3+3r)\cos(\theta)$	
17	5	−1	$2\sqrt{3}(10r^5-12r^3+3r)\sin(\theta)$	
18	5	3	$2\sqrt{3}(5r^5-4r^3)\cos(3\theta)$	
19	5	−3	$2\sqrt{3}(5r^5-4r^3)\sin(3\theta)$	
20	5	5	$2\sqrt{3}r^5\cos(5\theta)$	
21	5	−5	$2\sqrt{3}r^5\sin(5\theta)$	
22	6	0	$\sqrt{7}(20r^6-30r^4+12r^2-1)$	Spherical (2nd)

Table 3 shows the first 22 Zernike modes thus defined, and the terms used to describe these modes as common aberrations where appropriate.

The orthogonality relationship is given by

$$\frac{1}{\pi}\int_0^1\int_0^{2\pi} Z_n^m(r,\theta)Z_{n'}^{m'}(r,\theta)\,\mathrm{d}\theta r\,\mathrm{d}r = \delta_{nn'}\delta_{mm'}, \tag{A3}$$

where δ_{ij} denotes the Kronecker delta. We have chosen to use the definition of $Z_n^m(r,\theta)$ given by Noll[6]; the normalization is such that, for $n \neq 0$, the mean of each polynomial $Z_n^m(r,\theta)$ over the unit circle is zero and the variance is π (for $n = 0$, the mean is 1 and the variance is 0).

Using the single index scheme, we refer to a Zernike polynomial as $Z_i(r,\theta) = R_i(r)\cos(m_i\theta)$ or $Z_i(r,\theta) = R_i(r)\sin(-m_i\theta)$ where appropriate. In Eq. (A3), the integral in θ, as a consequence of the orthogonality of the trigonometric functions, gives rise to the second Kronecker delta, $\delta_{mm'}$, which is zero unless the azimuthal orders are identical, in which case the integral in r gives rise to $\delta_{nn'}$. The following properties of the radial polynomials can thus be determined:

$$\int_0^1 R_i(r)R_k(r)r\,\mathrm{d}r = \begin{cases} \delta_{ik}/2 & m_i = m_k = 0 \\ \delta_{ik} & m_i = m_k \neq 0 \end{cases}. \tag{A4}$$

Moreover, it can also be shown that

$$\int_0^1 R_i(r)r\,\mathrm{d}r = 0 \qquad m_i = 0. \tag{A5}$$

APPENDIX B: FOURIER TRANSFORMS OF POLAR FUNCTIONS

Consider the FT of a function $Z(r,\theta) = f(r)\cos(m\theta)$. This is defined as

$$\mathcal{F}\{Z(r,\theta)\} = \int_0^{2\pi}\int_0^1 f(r)\cos(m\theta) \times \exp[j\nu r\cos(\theta-\xi)]r\,\mathrm{d}r\,\mathrm{d}\theta. \tag{B1}$$

If we let $\phi = \theta - \xi$ and rearrange the integration we find that

$$\mathcal{F}\{Z(r,\theta)\} = \int_0^1 f(r)\Bigg[\cos(m\xi)\int_0^{2\pi}\cos(m\phi) \times \exp(j\nu r\cos\phi)\mathrm{d}\phi - \sin(m\xi) \times \int_0^{2\pi}\sin(m\phi)\exp(j\nu r\cos\phi)\mathrm{d}\phi\Bigg]r\mathrm{d}r. \quad \text{(B2)}$$

Using symmetry considerations, the second integral in ϕ is found to be equal to zero and the first integral can be written analytically with Ref. 11,

$$J_m(\nu r) = \frac{j^{-m}}{2\pi}\int_0^{2\pi}\cos(m\phi)\exp(j\nu r\cos\phi)\mathrm{d}\phi, \quad \text{(B3)}$$

to give the result

$$\mathcal{F}\{Z(r,\theta)\} = 2\pi j^m\cos(m\xi)\int_0^1 f(r)J_m(\nu r)r\mathrm{d}r. \quad \text{(B4)}$$

The equivalent process for a function $Z(r,\theta) = f(r)\sin(m\theta)$ yields

$$\mathcal{F}\{Z(r,\theta)\} = 2\pi j^m\sin(m\xi)\int_0^1 f(r)J_m(\nu r)r\mathrm{d}r. \quad \text{(B5)}$$

REFERENCES

1. J. W. Hardy, *Adaptive Optics for Astronomical Telescopes* (Oxford U. Press, Oxford, UK, 1998).
2. R. K. Tyson, *Principles of Adaptive Optics* (Academic, London, 1991).
3. M. J. Booth, M. A. A. Neil, and T. Wilson, "Aberration correction for confocal imaging in refractive index mismatched media," J. Microsc. **192**, 90–98 (1998).
4. P. Török, P. Varga, and G. Nemeth, "Analytical solution of the diffraction integrals and interpretation of wave-front distortion when light is focused through a planar interface between materials of mismatched refractive indices," J. Opt. Soc. Am. A **12**, 2660–2671 (1995).
5. F. Roddier, "Curvature sensing and compensation: a new concept in adaptive optics," Appl. Opt. **27**, 1223–1225 (1988).
6. R. J. Noll, "Zernike polynomials and atmospheric turbulence," J. Opt. Soc. Am. **66**, 207–211 (1976).
7. V. N. Mahajan, "Zernike circle polynomials and optical aberrations of systems with circular pupils," Eng. Lab. Notes in Opt. Photon. News, Aug. 1994 [Appl. Opt. **33**, 8121–8124 (1994)].
8. M. Born and E. Wolf, *Principles of Optics* (Pergamon, Oxford, UK, 1975).
9. A. Gray and G. B. Matthews, *A Treatise on Bessel Functions and Their Applications to Physics* (Dover, New York, 1966).
10. V. N. Mahajan, "Zernike annular polynomials and optical aberrations of systems with annular pupils," Eng. Lab. Notes in Opt. Photon. News, Nov. 1994 [Appl. Opt. **33**, 8125–8127 (1994)].
11. M. Abramovitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
12. W. H. Lee, "Computer-generated holograms: techniques and applications," in *Progress in Optics*, E. Wolf, ed. (Elsevier, New York, 1978), Chap. 3.
13. M. A. A. Neil, M. J. Booth, and T. Wilson, "Dynamic wavefront generation for the characterization and testing of optical systems," Opt. Lett. **23**, 1849–1851 (1998).