Accuracy of Zernike Polynomials in Characterizing Optical Aberrations and the Corneal Surface of the Eye

Luis Alberto Carvalho

Purpose. Zernike polynomials have been successfully used for approximately 70 years in many different fields of optics. Nevertheless, there are some recent discussions regarding the precision and accuracy of these polynomials when applied to surfaces such as the human cornea. The main objective of this work was to investigate the absolute accuracy of Zernike polynomials of different orders when fitting several types of theoretical corneal and wave-front surface data.

Methods. A set of synthetic surfaces resembling several common corneal anomalies was sampled by using cylindrical coordinates to simulate the height output files of commercial videokeratography systems. The same surfaces were used to compute the optical path difference (wave-front [WF] error), by using a simple ray-tracing procedure. Corneal surface and WF error was fit by using a least-squares algorithm and Zernike polynomials of different orders, varying from 1 to 36 OSA-VSIA convention terms.

Results. The root mean square error (RMSE) ranged—from the most symmetric corneal surface (spherical shape) through the most complex shape (after radial keratotomy [RK]) for both the optical path difference and the surface elevation for 1 through 36 Zernike terms—from 421.4 to 0.8 μm and 421.4 to 8.2 μm, respectively. The mean RMSE for the maximum Zernike terms for both surfaces was 4.5 μm.

Conclusions. These results suggest that, for surfaces such as that present after RK, in keratoconus, or after keratoplasty, even more than 36 terms may be necessary to obtain minimum accuracy requirements. The author suggests that the number of Zernike polynomials should not be a global fixed conventional or generally accepted value but rather a number based on specific surface properties and desired accuracy. (*Invest Ophthalmol Vis Sci.* 2005;46:1915–1926) DOI:10.1167/iovs.04-1222

Zernike polynomials (ZPs) are named after Fritz Zernike, who proposed them in 1934.[1] Zernike was a Dutch physicist who worked in several fields of optics and won the Nobel Prize for inventing the phase-contrast microscope in 1935.[2] This instrument is still used today to study biological specimens without the need for dyes. Dyes can help in the visualization process by emphasizing contrast but usually spoil the sample.

After Zernike proposed the ZPs they became rapidly popular among the optical community, perhaps because of certain special properties that allow them to be applied in cases in which the Seidel polynomials (SPs) are not applicable. ZPs have been applied successfully to the field of optics, optical engineering, and astronomy for the past 70 years.[3] Although they have only been applied more recently to the description of the optical aberrations of the human eye,[4,5] they have become a standard in this field also.[6] Nevertheless, there have been some recent discussions (Smolek MK, et al. *IOVS* 1997; 38:ARVO Abstract 4298; Smolek MK, et al. *IOVS* 2002;43: ARVO E-Abstract 3943)[7–10] among colleagues in the eye care community regarding the accuracy and even the usefulness of these polynomials, specifically for application in the visual sciences. There are arguments that ZPs are not sufficient to represent visually significant aberrations,[7,8] and other investigators[9,10] have stated that ZPs should be used carefully when fitting complex surfaces.

The main objective in this work was to conduct a quantitative study of the accuracy of ZPs when applied to typical surfaces in visual optics, ignoring all sources of noise that exist in any real system. The objective was not to redefine what are or are not the visually significant aberrations of the in vivo eye. This subject has been exhaustively discussed in the specialized literature. To conduct a thorough quantitative analysis of the accuracy of ZPs as a method for fitting visual-related surfaces, such as corneal elevation and eye and corneal aberrations, I think it is important not only to conduct experiments using third-party software[11] on corneal elevation and/or wave-front (WF) data from in vivo eyes, but also to implement this method on theoretical surfaces, synthetically generated by computer algorithms, which eliminates the problem of videokeratography measurement errors[12,13] and even avoids the limitations of the molding process of test surfaces on lathes, which are usually limited to spherical, ellipsoidal, or parabolic surfaces. In the analysis conducted in this study I applied ZPs in many different ways and to different theoretical surfaces, avoiding problems such as misalignment, data noise from image processing, and the other just-mentioned technical limitations. This may give an insight into the intrinsic relations between actual surface irregularities and accuracy when fitting them with ZPs.

It should be emphasized that I had no intention of proving that ZPs are the perfect fitting method for general videokeratography and wave-front instrumentation. To do so, it would be necessary to analyze each instrument available in the market, which is quite impossible. What is possible, and was implemented in this study, is to analyze how well these polynomials perform in typical videokeratography and wave-front surfaces. If one commercial instrument has much greater noise than another during the image-processing phase, for example, of course the Zernike fit will provide different accuracy and different propagation of error with this specific instrument.

It should also be stated that, even though actual data from real instruments containing noise were not the point of this study, fitting methods such as ZPs are certainly more advantageous than interpolating methods when one is searching for precise mathematical representations of visual optics surfaces.

The VSIA[6] convention, which has recently also become an ANSI standard,[14] determines, among other aspects, the nomenclature that should be used to refer to each Zernike coefficient,

From the Grupo de Óptica, Instituto de Física de São Carlos, Universidade de São Paulo (IFSC-USP), São Paulo, Brazil; and the Departamento de Oftalmologia, Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil.

Supported in part by Fundação de Amparo à Pesquisa do Estado de São Paulo.

Submitted for publication October 14, 2004; revised November 26, 2004; accepted February 18, 2005.

Disclosure: **L.A. Carvalho**, None

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Luis Alberto Carvalho, Center for Visual Science, Meliora Hall, 262, University of Rochester, New York, NY 14627; lcarvalho@cvs.rochester.edu.

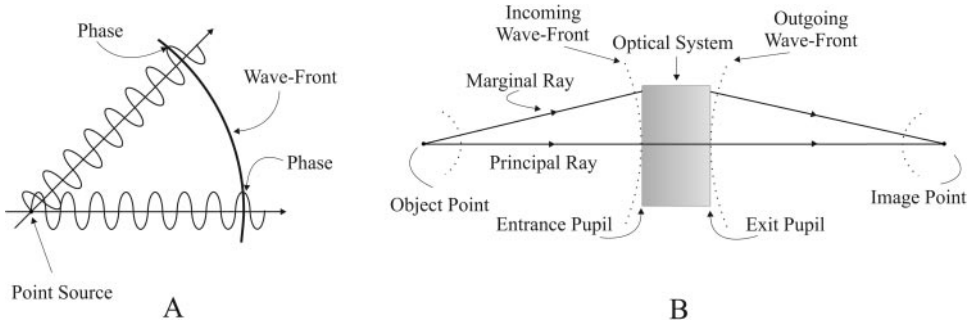

FIGURE 1. Diagrams illustrating the definition of **(A)** a WF and **(B)** a perfect optical system.

to avoid confusion when comparing results from different instrumentation, but does not suggest an upper limit to the number of terms. Nevertheless, many researchers and manufactures throughout the eye care community are currently using, at most, the first 36 terms, regardless of the application. The computations shown herein suggest that the ideal number of Zernike terms should not be a fixed standard for any general surface, but rather one that would allow for a minimum standard error for specific types of surfaces that are being represented by ZPs.

METHODS

Wave Aberration

To understand the importance of ZPs to the field of optics, it is necessary to understand the context in which they are applied and why they may be more efficient than other available methods. To do so, I start by explaining the meaning of WF and *wave aberration* (sometimes also referred to *wave-aberration function* or *wave-front error*—which will be abbreviated herein simply as *W*).

The concept of WF is simple to understand based on the diagram in Figure 1A. The imaginary surface that unites the wave tips of rays of the same phase (or the same optical path) form a virtual surface that is called the wavefront. It may be a perfect sphere, as the WF of any point source of light should ideally be, or it may have a distorted or irregular pattern, which is the case for in vivo eyes and most practical optical systems. A perfect optical system, from a geometric optics point of view, is one that redirects all refracting rays from object point to a single conjugate image point (Fig. 1B). For a perfect optical system, the WF leaving the refracting surface must be centered on the image point, forming a spherical WF. This fact relies on Fermat's principle of least time, which states that the optical path of the principal and marginal rays should be identical.

In this sense, the objective of a lens designer is to produce an optical system that forms a perfect image at image space. The problem is that there is also an economic factor involved. Usually optical materials of uniform refractive indexes and lenses with spherical surfaces are much more cost effective. Paraxial rays usually form a good-quality image for these systems, but marginal rays degrade them. The departure from a perfect image is called wave aberration and is illustrated in Figure 2. Because spherical surfaces intrinsically generate undesired differences in the optical path for different rays, the lens designer usually has to group several lenses with different parameters to obtain the best possible image. There are arguments toward a similar solution regarding the human eye[15,16] (although we do not know whether nature did this also for economic reasons!). Fortunately, today there are several very sophisticated optical design software programs that make this task much less empiric than it was years ago.[3]

A WF with aberrations may be described by comparing it with a reference WF, which is usually chosen to be the spherical WF that leaves the exit pupil (Fig. 2). The reference WF has its vertex tangent

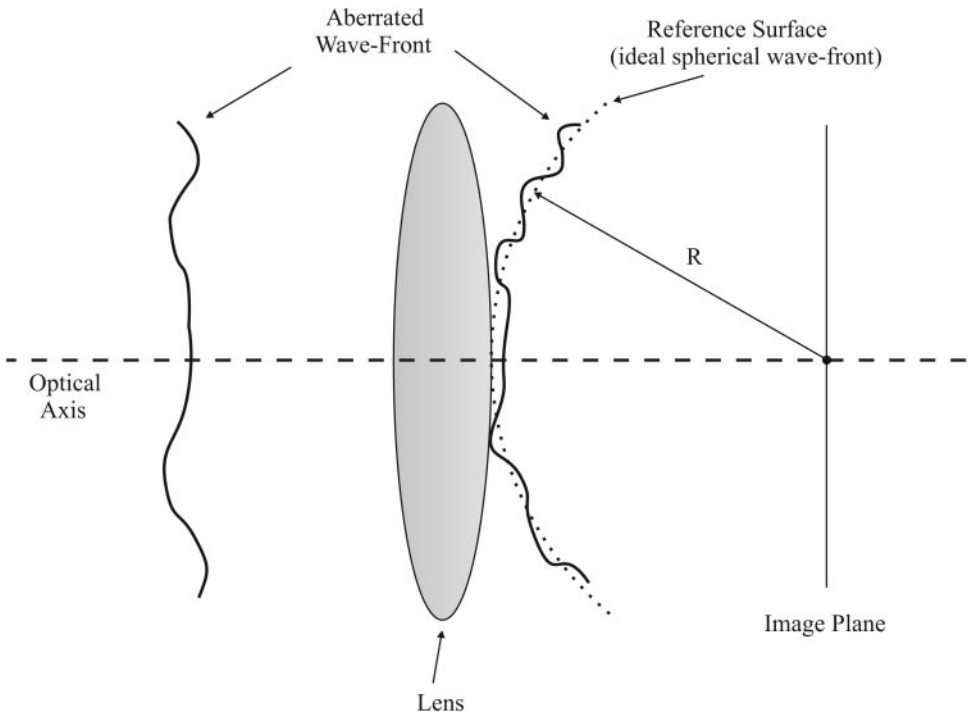

FIGURE 2. Definition of the WF function, also called the optical path difference or wave-aberration (*W*) function. The WF aberration function *W* is determined by comparing the actual WF leaving the exit pupil with a reference surface, which is an ideal spherical WF. The actual exit pupil is not shown, but may be assumed to be tangential to the apex of the back surface of the lens and to have the same diameter as the lens.

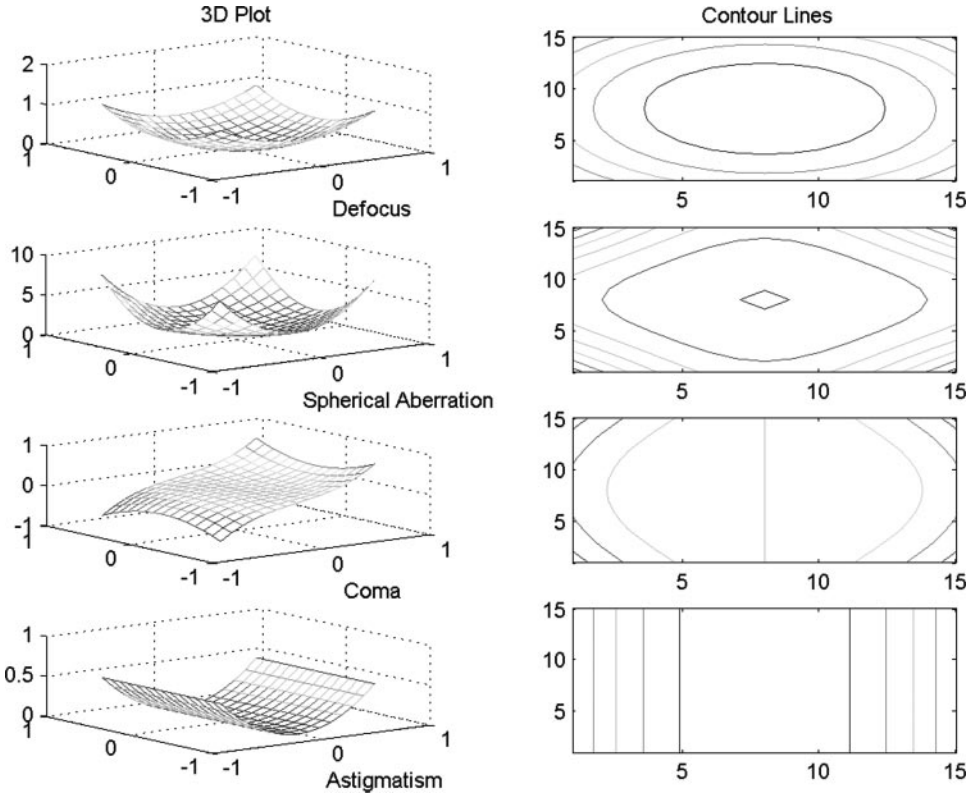

FIGURE 3. A plot (MatLab; the MathWorks, Natic, MA) of some of the most common Seidel aberrations, represented in two- and three-dimensional format. The third-order approximation of aberration theory leads to what are called the primary aberrations. (In the United Kingdom, these aberrations are viewed as the first-order corrections to the paraxial theory, whereas in the United States they are viewed as the third-order correction to the first-order theory.)

to the exit pupil and is centered on the image point with a radius of curvature *R.* For each point on the exit pupil, the optical path difference (OPD), represented by the wave-aberration function (W), is measured between the spherical reference surface and the aberrated WF along the radius of the reference surface. The wave aberration is usually described as a function of Cartesian or Polar coordinates ($W(x,y)$ or $W(\rho,\theta)$) obtained over the exit pupil, which is now used to describe the wave aberration in a continuous format.

There are several methods of representing the aberration function W. The usual procedure is to represent it as a polynomial expansion. This is very useful because each term of the expansion may itself represent a specific type of aberration and also may determine how much of it is present on the entire WF. There are two sets of such polynomials that traditionally have been used for the description of the aberration function. In optical design the *Seidel polynomials* are typically used. In optical testing and measurement, W usually has to be deciphered (and not chosen or minimized, as happens in optical design), and the procedure that became common is to fit W with a set of ZPs. I will briefly describe the SPs and some of their terms, to show why they are different from ZPs and why ZPs are a better solution for more complex applications.

Seidel Polynomials

SPs are usually represented in a polar coordinate system (ρ,θ) at the exit pupil where ρ is the radial distance and θ is the polar angle. The wave-aberration function W can be represented mathematically by a set of SPs by

$$W(\rho,\theta) = \sum_{i,j,k} C_{ijk}\,\overline{H}^{i}\,\rho^{j}\cos^{k}\theta, \qquad (1)$$

where C_{ijk} is the WF aberration coefficient, H is the fractional image/object height for the chief ray, and the other terms are the radial order and polar frequency. The use of normalized pupil coordinates is a matter of convenience, and dimensionality is maintained by the C_{ijk} coefficient.

The SPs are defined as the five lower-order terms where $i + j = 4$ in the expansion given by equation 1. The most familiar aberrations are spherical aberration, coma, astigmatism, field of curvature, and distortion. Figure 3 shows two- and three-dimensional representations of some of the Seidel aberrations.

Although the SPs may be used to represent most of the common aberrations present in optical systems, there are certain restrictions. These restrictions arise as soon as the optical system becomes nonrotationally symmetric, which happens when an optical component is tilted or decentralized, relative to the optic axis. This is certainly a major problem in visual optics.[17] When decentralization and tilt are present, there is no term in the SP expansion that can account for the wave aberration induced, because SPs are based on rotationally symmetric terms. When this problem arises, ZPs make a difference. As will be shown in the next section, they have all the invariance, normalization, and other interesting properties of the Seidel expansion, but they can also account for a more complete set of possible optical imperfections.

Special Properties of ZPs

ZPs have certain special properties that make them an interesting expansion set for the description of general surfaces in the fields of optical engineering and in physiological optics. First, some of these properties will be depicted, and later, simulations will be conducted to confirm this affirmation. The ZPs themselves will not be formally introduced nor will the theory behind their recursive equations for generating individual terms, given that most of the vision science community should be acquainted by now with these fundamental concepts and also because this is a thoroughly covered topic in the specialized literature.[1,5,12,13]

ZP properties may be summarized by stating that they form a "complete set of orthonormal polynomials" in the three-dimensional space inside the unit circle domain.[1] This guarantees that one may fit any piece-wise continuous surface in space, given that a sufficiently large number of terms are used. This is certainly the case for most corneas[18] that have not undergone physical trauma, early post-kerato-

plasty, in keratitis, or in any other severe disease that may cause abrupt changes in curvature.

To demonstrate this fact, I will revise some of the theory behind the properties of these polynomials and also provide examples with a well-known power series. I will demonstrate that for a limited number of terms in a power series or a polynomial series there should also be limited accuracy in data fitting. This may seem quite an obvious affirmation and should be valid for any fitting method applied to any typical problem, but what is not so obvious is how this fitting accuracy varies for specific types of surfaces and at what cost/benefits in terms of computational time, which is one of the objectives of this study. Because Fourier series (FS) have analogous properties when compared with ZPs, I will begin to demonstrate this fact by using FS for a simple signal and then for a more complex signal. Afterward ZPs will be applied to both simple and more complex surfaces.

A FS may be defined as the expansion of a function $f(x)$ as an infinite summation of sines and cosines with different coefficients and arguments. Its general expression is given by

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx) \quad (2)$$

where the coefficients in the equation are given by

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)dx, \quad (3)$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)dx, \quad (4)$$

and

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx. \quad (5)$$

FS are an elegant and simple method to break up a periodic function into a set of simple terms to fit a solution to whatever the desired accuracy. One of the main applications of a FS is to allow the segmentation of complex signals (whether they are in image-processing applications or electric engineering, for example) into its basic frequencies—so that a noisy signal may be represented by a set of well-behaved periodic functions, each with its own frequency and amplitude. Figure 4 shows two completely different types of signals: (A) a simple sinusoidal wave given by a sine function and (B) a square wave composed of several different frequencies. Superimposed on these signals are several successive attempts to fit them with FS with a different number of terms. It can be easily seen that for type 1, the increase in the number of terms (frequencies) in the FS has little consequence after a certain number of terms. (Actually, after the second term is added, the accuracy changes very little, because this is the original number of terms used to generate $f(x)$.) For signal type 2, the story is different. Because the signal is much more complex (has several frequencies associated with it) the truncation of the FS at different frequencies has significant influence on accuracy.

To demonstrate the errors associated with a different number of terms when fitting a function $f(x)$ with an FS we calculated the root mean square error (RMSE) associated with each function, given by

$$\text{RMSE} = \sqrt{(f - IF_n)^2}, \quad (6)$$

where IF_n represents the FS up to the *nth* order. Results for a successive number of terms for the square wave and sinusoidal functions are provided in Table 1.

As shown in Table 1 and the graph in Figure 5, as the number of terms increase, RMSE decreases, as expected. This decline is in agreement with the fact that I have mentioned, but notice that, for a less complex function (one that contains less components), the RMSE decreases more rapidly after a given term. Why does this happen? What are the fundamental mathematical properties of the FS (and as will be shown, also of ZPs) that make them suitable for such fitting procedures, and why are the errors associated in the fitting process so closely related to the number of terms? These properties lay behind the concepts of "othogonality" and "completeness"; the formal definition of these terms follows.

Two functions (the terms "functions" and "polynomials" are used interchangeably), $f(x)$ and $g(x)$, are said to be orthogonal over the interval $a \le x \le b$ with weighting function $w(x)$ if

$$[f(x)\,|\,g(x)] = \int_a^b f(x)g(x)w(x)dx = 0, \quad (7)$$

and, in addition, if also

$$\int_a^b [f(x)]^2\, w(x)dx = 1 \quad (8)$$

and

$$\int_a^b [g(x)]^2\, w(x)dx = 1, \quad (9)$$

the functions are also said to be "normalized." When functions are both orthogonal and normalized, they are called "orthonormal." If the set of functions and polynomials has more than two terms, the generalized form of these properties, if each term is represented by ϕ, is

$$\int_a^b w(x)\phi_n(x)\phi_m(x)dx = \delta_{mn}c_n, \quad (10)$$

where n and m represent the indexes of each polynomial, C_n is a constant and, when it assumes a value of 1, the polynomials are also normalized, and δ represents the Kronecker delta[19] and assumes a value of 0 if $m = n$ and 1 if $m \neq n$. The general properties given by equations (7–10) are also applicable to functions or polynomials defined in larger domains (such as the x–y plane). The only difference is that a double integral should be implemented. The orthonormal functions are also said to be "complete" in the closed interval $x \in (a,b)$ if, for every piece-wise continuous function $f(x)$ in this interval, the squared error

$$E_n = [f - (c_1\phi_1 + c_2\phi_2 + c_3\phi_3 + \ldots c_n\phi_n)]^2 \quad (11)$$

converges to 0 as $n \to \infty$. Symbolically, a set of functions is complete if

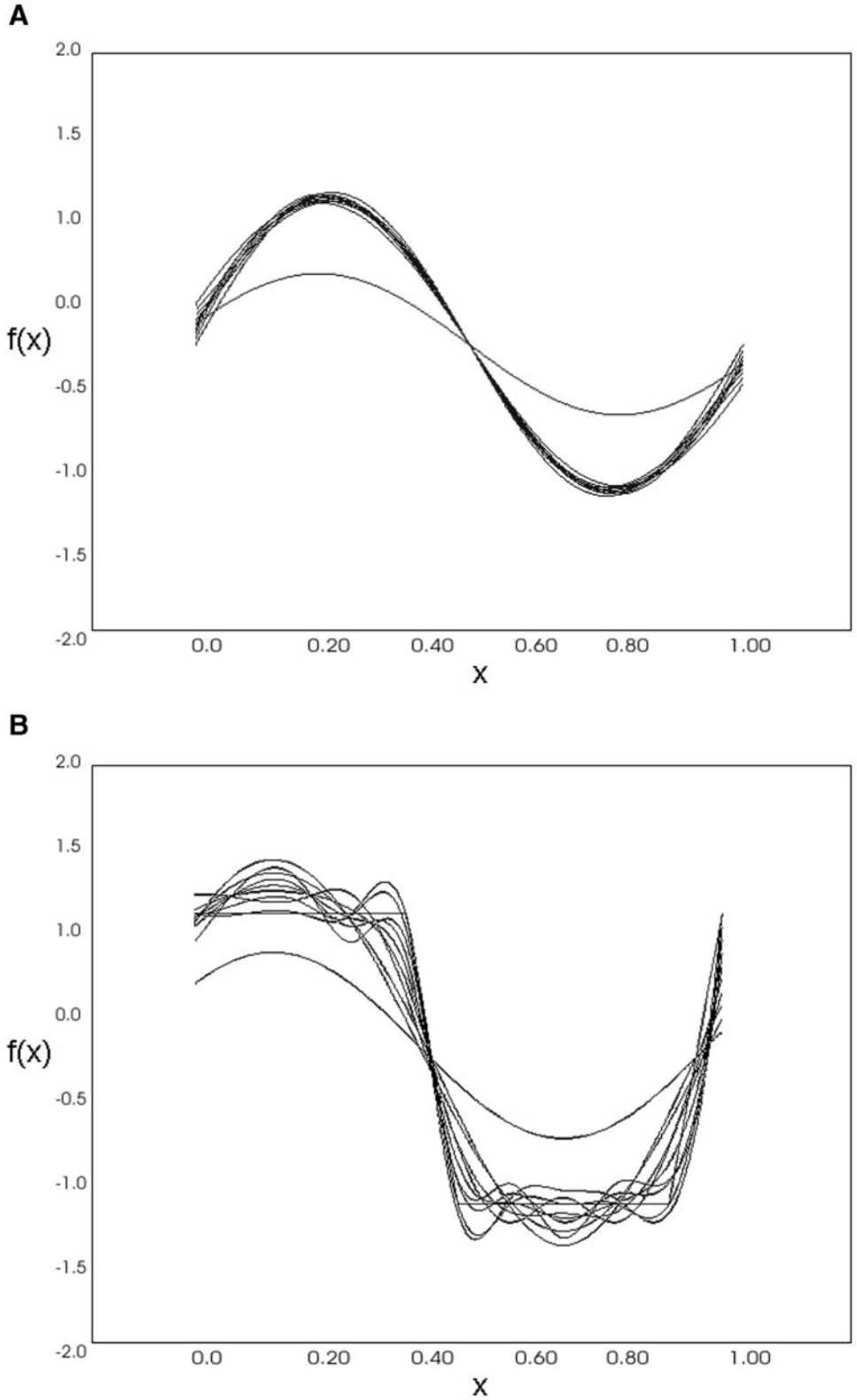

FIGURE 4. (**A**) Sinusoidal and (**B**) square-wave functions, with different-order Fourier fittings.

$$\lim_{m\to\infty}\int_a^b [f(x) - \sum_{n=0}^{m} c_n\phi_n(x)]^2 w(x)dx = 0 \qquad (12)$$

for every value of x in the considered interval. If this same concept is applied to a set of functions or polynomials for which the domain now is the x,y plane, the symbolic representation becomes

$$\lim_{m\to\infty}\int_a^b\int_c^d [f(x,y) - \sum_{n=0}^{m} c_n\phi_n(x,y)]^2 w(x,y)dxdy = 0, \qquad (13)$$

where the closed interval is defined both in the x-axis (a,b) and the y-axis (c,d). Also, if these polynomials are in cylindrical coordinates—that is, $\phi(\rho,\theta)$—then equation 13 can be written in terms of these new

Table 1. RMSE for the Sine and Square-Wave Functions

n	Sine Function	Square-Wave Function
1	0.5671	0.8800
2	0.3108	0.5373
3	0.2962	0.5035
4	0.2965	0.4962
5	0.2854	0.4400
6	0.2854	0.4400
7	0.2817	0.3838
8	0.2897	0.3765
9	0.2849	0.3427

The abrupt error minimization after the second term is included in the sine approximation, and the low variation after that. On the contrary, for the square wave, there was a gradual error minimization as the number of terms successively increased. Truncation of the RMSE at the fourth significant algorithm was purely arbitrary and has no relation to the intrinsic floating-point precision for inverse Fourier transform calculations in the software (MatLab; The MathWorks, Natick, MA).

coordinates, and the limiting interval will also be determined by (ρ,θ). More specifically, it is useful to write these polynomials in cylindrical coordinates when the domain has polar symmetry, which often happens in the case of optical apertures and also the human pupil. This is why the VSIA standards for ZPs are given in cylindrical coordinates. ZPs also obey all the properties discussed thus far—a fact that is not proven herein, given that there are very good references that demonstrate these properties and also for the sake of brevity. The reader is directed to Born[1] for a thorough discussion and demonstration of these properties and also an explanation of the recursive formulas for generating ZPs of any order. In this way, it can also be affirmed that ZPs form a "complete set of orthonormal polynomials" inside the unit circle domain $(0 - 1, 0 - 2\pi)$.

To determine the efficiency of ZPs in fitting simple and complex surfaces, the same techniques applied in the above example using FS were applied to synthetic corneas and synthetic OPD aberrations calculated using these corneas. The following section explains how these surfaces were generated.

Synthetic Corneal Surfaces

Several surfaces that resemble typical corneal shapes were used, such as spherical, ellipsoid, ellipsotoric, keratoconic, and post-radial keratectomy (post-RK). Based on the coordinate system shown in Figure 6 shapes were as follows.

Spheres, Ellipsoids, and Ellipsotorics. For the ellipsotoric family the equation

$$\rho = \sqrt{2z_c r_a - pz_c^2} \tag{14}$$

was used, where r_a is the apical radius, $p = 1 - e^2$ is the "shape factor," and e is the eccentricity. When e is set to 0, the shape factor is 1, and equation 14 becomes a sphere with radius r_a. Another typical parameter is the "asphericity" (Q) of the surface, which is equal to $-e^2$, so that the shape factor may also be written as $p = 1 + Q$. To model an astigmatic cornea, an ellipsotoric surface[20] was used, where the apical radius is a function of the polar angle (θ)

$$r_a(\theta) = \frac{1}{\left[\frac{1}{r_h} + \left(\frac{1}{r_v} - \frac{1}{r_h}\right)\sin^2\theta\right]}, \tag{15}$$

where r_v and r_h are the vertical and horizontal apical radii, respectively.

Varying these parameters allows simple spherical surfaces to astigmatic surfaces to be generated with elliptic profiles of different eccentricities.

Keratoconus. For the keratoconic surface the following parametric equation was applied

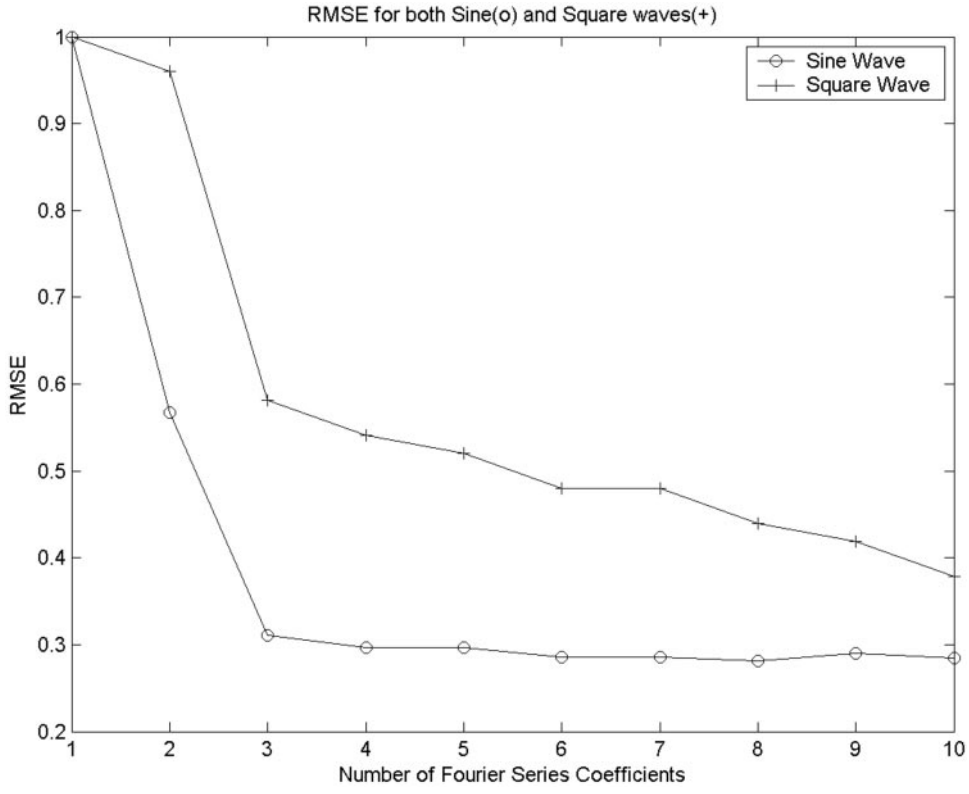

Figure 5. The RMSE for the FS fit for the sine and square-wave functions. The series associated with the simpler symmetry (fewer frequencies) converges to 0 more rapidly than that associated with the square wave. Also, it requires fewer terms to converge to a lower RMSE.

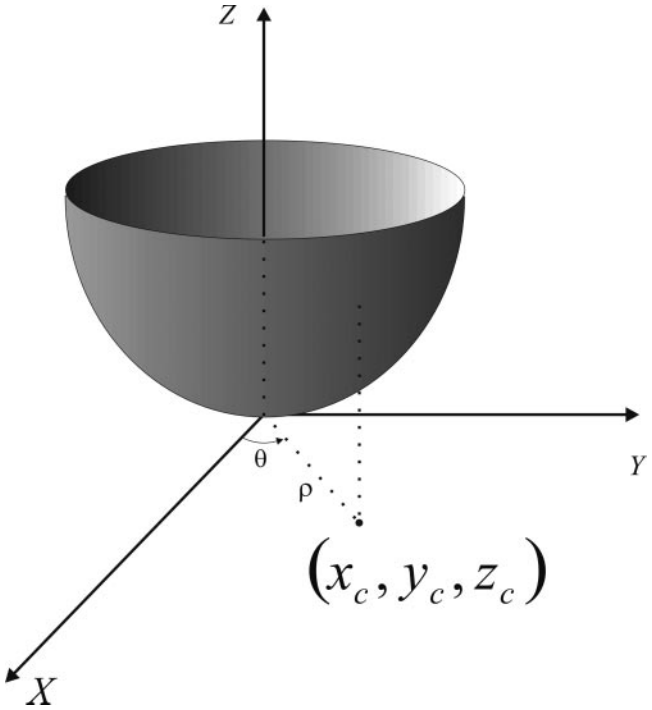

FIGURE 6. Cylindrical coordinate system for representation of corneal and WF elevation data.

$$z_c = \begin{Bmatrix} r_0 - \sqrt{r_0^2 - \rho^2} \\ r_0 - \sqrt{r_0^2 - \rho^2} + \dfrac{a}{2}\left[1 - \cos \pi\left(\dfrac{\rho - \rho_1}{\rho_2 - \rho_1}\right)\right] \\ r_0 - \sqrt{r_0^2 - \rho^2} \end{Bmatrix} \begin{matrix} \text{for} & 0 < \rho \le \rho_1 \\ \text{for} & \rho_1 < \rho \le \rho_2. \\ \text{for} & \rho_2 < \rho \le \rho_3 \end{matrix} \quad (16)$$

Parameter a is keratoconus intensity. This surface has continuous curvature (proportional to the second derivative), an important feature to mimic in vivo corneal surfaces.[18] Although decentralized keratoconus is also common in in vivo eyes, this surface mimics a centralized keratoconus—that is, one for which the apex coincides with the intersection of the optic axis with the anterior cornea. This simplifies the computational efforts of synthetic Placido image simulation. There is also the possibility of tilting this surface by using rotation and translation transformation matrices commonly used in computer graphics,[21,22] to obtain decentralized keratoconus. This was not implemented in the present work for the sake of brevity, although it is certainly an important analysis to undertake in further work.

Post-Radial Keratectomy. An equation suggested by Klein,[18] a generalized version of a surface proposed by Rand et al.,[23] was used for the post-RK cornea

$$z_c = r_0 - \sqrt{r_0^2 - \rho^2} + a \cos(8\theta) R(\rho), \quad (17)$$

where the corneal height (z) is now a function of the axial distance (ρ) and the polar angle (θ). R is a parametric factor that depends on the axial distance, so that

$$R(\rho) = \begin{Bmatrix} 0 \\ 1 - \cos\left[\pi \dfrac{(\rho - \rho_1)}{\rho_2 - \rho_1}\right] \\ 0 \end{Bmatrix} \begin{matrix} \text{for} & \rho \le \rho_1 \\ \text{for} & \rho_1 < \rho \le \rho_2 \\ \text{for} & \rho > \rho_2 \end{matrix} \quad (18)$$

where a here is the "wound intensity," proportional to the depth of the scalped incisions. This surface is a sphere of radius r_0 with an added sinusoidal corrugation. Table 2 is a list of surfaces that were used in the simulations and their respective parameters, according to the description just provided.

Examples of in-focus simulated Placido images for surfaces A through I are shown in Figure 7. The ZP fit of the surface profiles were compared to the original surface points for each Placido disc (polar coordinates). The RMSE for the entire surface was computed as

$$\text{RMSE} = \frac{1}{5760} \sum_{n=1}^{16} \sum_{\theta=1}^{360} [\{z(\rho(\theta,n),\theta) - z_c(\rho(\theta,n),\theta)\}^2]^{1/2}, \quad (19)$$

where z is the ZP fit of the corneal height, z_c is the theoretical surface height computed from equations 12, 13, 15, 16, and 17, and the ρ values are presented in their parameterized form—that is, as a function of polar angle (θ) and Placido disc (n). An analogous procedure was implemented to compute the errors associated with the WF error fit.

Generating the Aberration Function from Corneal Height Data

The second type of surface in which the ZP expansion was tested is the OPD computed from the synthetic corneas given in the previous section. This section explains the method used to generate OPD points. The diagram shown in Figure 8 illustrates the method used for optical aberration calculations.

Figure 8 illustrates the image formation by a refracting surface separating two media of different refractive indexes (n and n'). An object point at position (P) localized at the object plane (O) has its image formed at point (P′) on the image plane (I). A marginal ray intersects the exit pupil at some point (ρ_c,θ_c) on the corneal surface. W

TABLE 2. Surfaces Used in the Simulation

Surface	Description	Equation	Parameters
A	Sphere	14	$p = 1$, $r_a = 7.80$
B	Ellipsoid	14	$p = 0.50$, $r_a = 7.80$
C	Ellipsotoric	14,15	$p = 0.75$, $r_h = 7.50$, $r_v = 8.00$
D	Ellipsotoric	14,15	$p = 0.50$, $r_h = 8.00$, $r_v = 7.50$
E	Ellipsotoric	14,15	$p = 0.30$, $r_h = 9.00$, $r_v = 7.00$
F	Keratoconic	16	$a = 0.0125$, $\rho_1 = 1.00$, $\rho_2 = 3.00$, $r_0 = 7.00$
G	Keratoconic	16	$a = 0.0250$, $\rho_1 = 1.50$, $\rho_2 = 3.00$, $r_0 = 8.00$
H	Keratoconic	16	$a = 0.0500$, $\rho_1 = 2.00$, $\rho_2 = 3.00$, $r_0 = 9.00$
I	Post-RK	17,18	$a = 0.0250$, $\rho_1 = 2.00$, $\rho_2 = 4.00$, $r_0 = 7.00$
J	Post-RK	17,18	$a = 0.0500$, $\rho_1 = 2.00$, $\rho_2 = 5.00$, $r_0 = 8.00$

Parameters were chosen in accordance with two principles: (1) published values of typical corneal shape factors (p) and (2) values that generated severe surface curvature changes resembling cases of astigmatism, keratoconus, and post-RK. Surfaces are separated into three families: ellipsotoric (A–E), keratoconic (F–H), and post-RK (I, J).

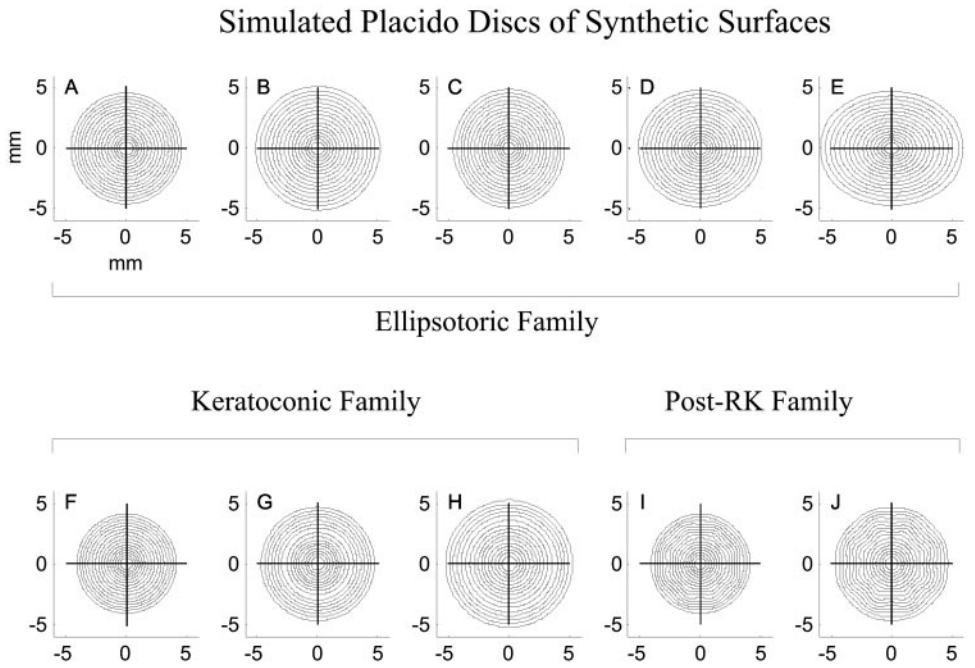

FIGURE 7. Simulated Placido images obtained for surfaces listed in Table 1 using the ray-tracing procedures described elsewhere.[12] The central crosses are 5 × 5 mm and may be used as qualitative reference to show the changes in Placido image size as the surface becomes more or less prolate, more or less astigmatic, and so on.

along the marginal ray is calculated as the difference in optical path length from the chief ray. For a single ray

$$\partial W = nl + nl' - ns - ns'. \quad (20)$$

The object distance (s) is chosen to be at infinity (>6 m), and the image distance s' is calculated by approximating the cornea to a lens, with radius (r_m) calculated from the mean value of all axial radii of curvature of the surface

$$r_m = \frac{1}{360} \sum_{n=1,\theta=1}^{16,360} \frac{\rho(\theta,n)}{\sin \alpha(\theta,n)} \quad (21)$$

where

$$\alpha(\theta,n) = \arctan\left(\frac{\rho(\theta,n)}{r_a - z_c(\theta,n)}\right), \quad (22)$$

where r_a is the apical radius (computed in the corneal elevation simulation phase). Alternatively, but with little loss in precision and generalization, the paraxial approximation may be applied and the r_m is then replaced by the apical radius of curvature. When the surface is astigmatic, the mean value of the vertical plus horizontal radii may be used [$(r_v + r_h)/2$].

For the $s \rightarrow \infty$ assumption, the focal distance of the lens is equal to the image distance ($s' = f$) and the *Lens Maker* equation with a single refracting surface of radius r_m may be applied

$$\frac{1}{s} = \frac{1}{f} = \frac{n' - n}{r_m}. \quad (23)$$

After s' is determined, the distances of the marginal and chief rays can be computed, and equation 20 can be applied over the entire Placido image domain

$$W = n\sqrt{(z_c - s)^2 + x_c^2 + y_c^2} + n'\sqrt{(s' - z_c)^2 + x_c^2 + y_c^2} - ns - n's', \quad (24)$$

where W, z_c, x_c, and y_c are all parameterized functions of ($\rho(\theta,n)$, θ). From this equation, the optical aberration (ρ,θ) for each corneal surface point $z_c(\rho,\theta)$ obtained from the theoretical surface can be calculated.

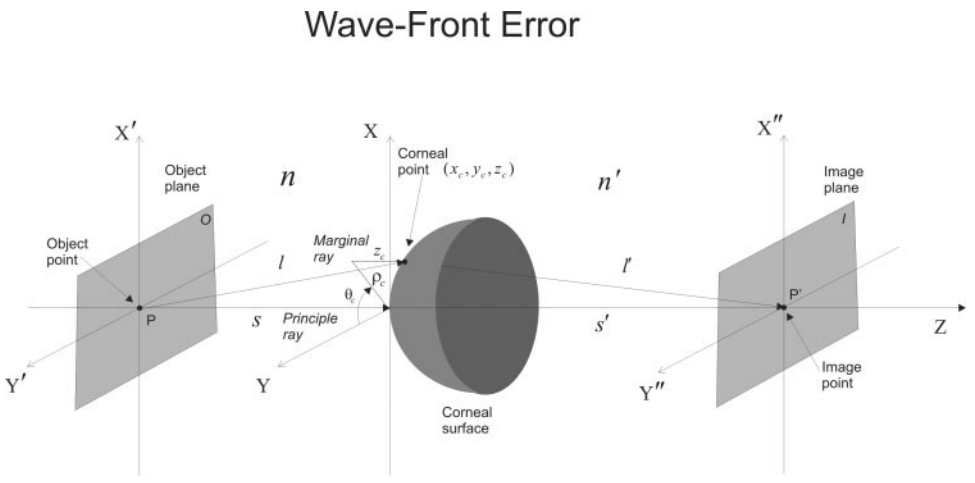

FIGURE 8. The ray-tracing procedure for wave-aberration calculation based on the synthetic corneal data. The WF error function W was calculated for all surfaces shown in Table 2 and an exit pupil of 8 mm in diameter was used in all cases.

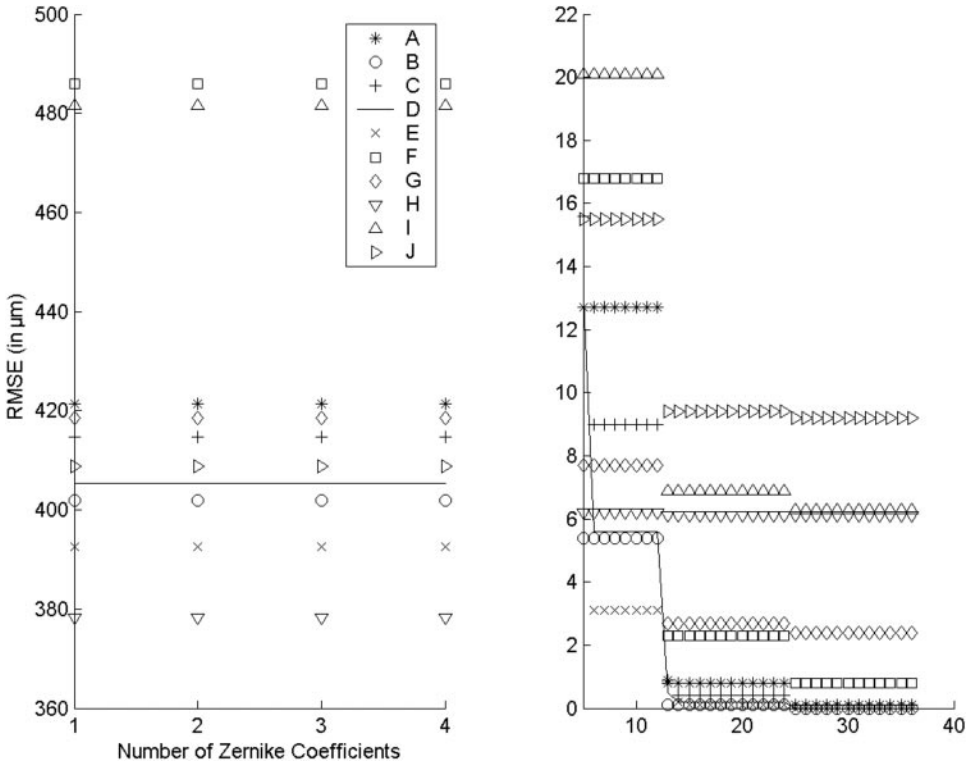

FIGURE 9. The RMSE of the corneal surface as a function of the number of ZP coefficients. The Zernike terms were grouped into two categories because the error range is ~500 to 0 μm, which makes it difficult to visualize small errors in a single graph. The two groups included coefficients 1 to 4 and coefficients 5 to 36. Surfaces are as listed in Table 2, and correspond to: A–E, the ellipsotoric family; F–H, the keratoconic family; and I, J, the post-RK family. A break in the *right* panel *y*-axis was inserted after five coefficients, to avoid superposition of most of the lower errors, due to an error of approximately 44 μm on surface E.

Figure 7 shows that the Placido image dimensions vary over the Cartesian planes of different synthetic surfaces. This happens because in the simulation algorithms presented in the previous section, the surface minimum and maximum height were the limiting parameters for computation of Cartesian positions of discs, and not the contrary (for more implementation details please refer to Ref. 12). Because of this implementation detail, all Placido image domains and exit pupil domains of all the synthetic images were limited to a total radial distance of 4 mm when computing the corneal elevation, optical aberrations, and errors. This radial distance limit guarantees that all Zernike fit errors for different surfaces are compared for domains having the same area. Aberrations for smaller pupils were not computed, because an 8-mm pupil maximizes aberration and at the same time serves as a reasonable approximation of maximum in vivo pupil sizes.[24]

The aberration data are fit to a set of ZPs, and the errors involved are computed in a procedure analogous to the one used for the corneal height, described in the previous section.

Fitting Corneal Elevation and WF with Different-Order ZPs

To perform the fitting routines, I represented corneal elevation data or wave-front aberration data as a parameterized function of polar coordinates (ρ,θ). This representation is illustrated in Figure 6. In this manner, corneal or WF surface elevation can be approximated by the series

$$z_c(\rho(n,\theta),\theta) \approx \sum_{j=0}^{36} C_j Z_j(\rho(n,\theta),\theta), \quad (25)$$

where ρ is a parametric function of the Placido disc number and angle (n,θ), C_j are the Zernike coefficients, and Z_j are the ZPs.

To find the Zernike coefficients for a specific corneal height, a minimum square fit is performed for all N data points. This procedure consists of minimizing the sum

$$S = \sum_{\theta=1,n=1}^{360,16} \{z_c(\rho(n,\theta),\theta) - \sum_{j=1}^{36} C_j Z_j(\rho(n,\theta),\theta)\}^2 \quad (26)$$

relative to each Zernike coefficient; $dS/dC_t = 0$ for $t = 1,\ldots,k$, must be found, where k is the total number of coefficients, so that

$$\frac{dS}{dC_t} = \sum_{\theta=1,n=1}^{360,16} z_c(\rho(n,\theta),\theta) Z_t(\rho(n,\theta),\theta) - \sum_{j=1}^{36} C_j \sum_{\theta=1,n=1}^{360,16} Z_j(\rho(n,\theta),\theta) Z_t(\rho(n,\theta),\theta) = 0, \quad (27)$$

from which we extract a linear system $AC = b$ with 36 equations and 36 unknown values of C. By solving this linear system through conventional procedures, such as the *Gaussian elimination method*, the 36 Zernike coefficients are found for each surface.

RESULTS

Figure 9 is a plot of the RMSEs for all corneal surfaces in Table 2 versus the number of Zernike terms used in the fitting process. In Figure 10, the analogous results for the WF errors calculated for each of these surfaces are shown.

DISCUSSION

As shown in Figure 9, errors for ZPs up to the third order (from the first to the fourth terms) correspond to unacceptable errors that range from approximately 360 to 460 μm. This is obviously the result of using a restricted number of terms that cannot account for surface irregularities; moreover, they are also not efficient, even in simple surfaces such as spheres (surface A). At this stage, the RMSE has apparently no strict relation to the complexity of the sur-

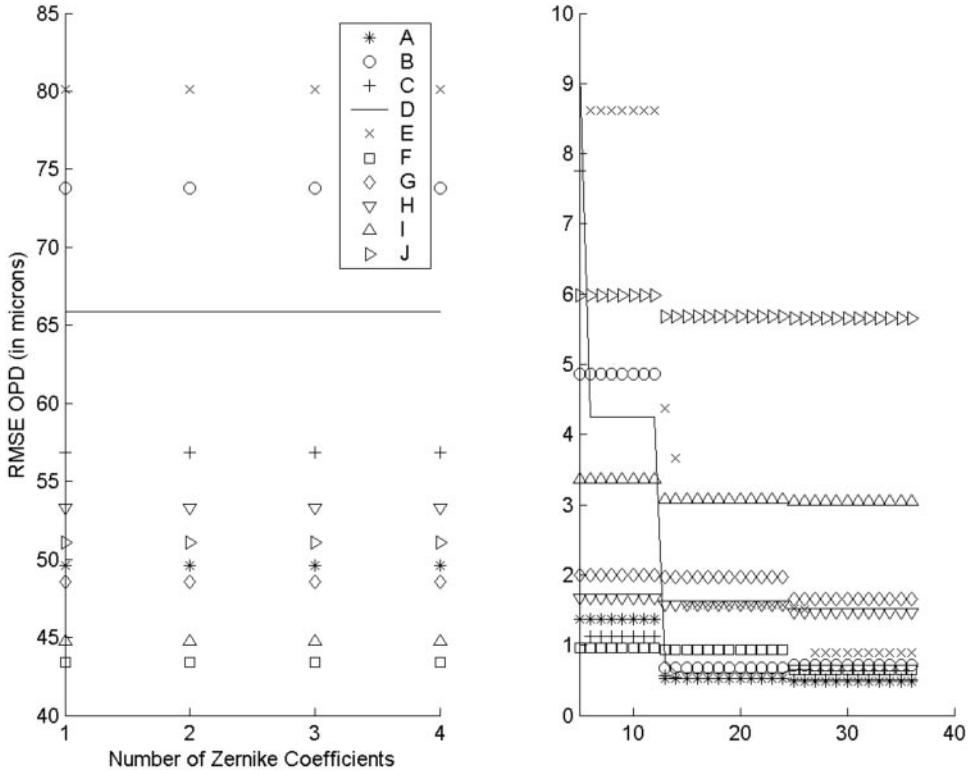

FIGURE 10. The RMSE for the WF error (OPD) as a function of the number of ZP coefficients. ZPs and surfaces are grouped in the same manner as described in Figure 9. A break in the *right* panel *y*-axis was inserted after five coefficients, to avoid superposition of most of the lower errors, due to an error of approximately 36 μm on surface E.

face, given that lowest errors are associated with a keratoconic surface (H). The absence of such a relation is obviously a coincidence. The same behavior was observed in the case of the WF error (Fig. 10), although errors were in a lower range for third-order polynomials (from 45 to 80 μm, approximately). This finding is an interesting fact, but it is not difficult to understand why the errors were so much lower for the WF error than they were for the corneal surface. It is because, although corneal anomalies do propagate to the WF aberration, the deviations involved in the OPD are much more subtle than those of the corneal surface, which is a positive factor that contributes to visual acuity.

For the next sequence of coefficients (4–12; Fig. 9) errors were in a much smaller range, from approximately 2.5 to 20 μm for the cornea and 2 to 10 μm for the WF error. In the case of corneal surfaces in this second group, the order in which the surfaces appeared is more coherent with the symmetry complexity. But there are still more complex surfaces that result in a lower RMSE than other simpler ones. The latter sequences are where the errors are more coherent, in that they correlate with what is expected, for both corneas and WF errors. The RMSE diminishes as the surface shape becomes simpler—as expected, since this behavior was also observed in the FS examples.

Zernike coefficients are a reliable and well-established method for representing optical aberrations in many fields of optics, and they will probably continue to be a standard in most fields of optics, including visual optics. Nevertheless, as results of this work indicate, certain specific considerations should be made when applying ZPs to different problems. Depending on the complexity of the corneal and WF surfaces, accuracy varies, from 421.4 to 0.8 μm, to 421.4 to 8.2 μm, respectively, and the mean RMSE for a maximum of 36 Zernike terms for both surfaces was 4.5 μm—a high error for most applications in visual optics today involving wave-front and videokeratography measurements. Klein[18] has obtained accuracy of up to 0.1 μm for elevation of synthetic surfaces with an algorithm that avoids the skew ray problem, and Guirao and Artal[13] have obtained accuracy of 0.2 μm in practical measurements on test surfaces, using a commercial videokeratograph and synthetic ellipsoidal surfaces.

In contrast, in regard to specific surfaces, examined one at a time for the spherical surface, for example, after coefficient number 10, errors were already very close to 2.0 μm. This means that there are two ways in which ZPs could be successfully used: (1) If there were prior information or a parameter that told in advance the complexity of the surface, certain assumptions could be made regarding the necessary or sufficient number of elements in the Zernike expansion; and (2) a securely high number of terms could be used in all cases, providing more reliable results both for simple and complex surfaces. The great disadvantage of this second option is the computational cost. The least-squares method used for calculating the ZP coefficients involves the inversion of a square matrix. When the number of coefficients is doubled, the matrix becomes four times greater, and so does the computing time. This means that using 36 coefficients instead of 18, for example, takes four times more computational time. In some cases, this is not acceptable, depending on the type of application. For laboratory instrumentation, computational time is, in most cases, not such a relevant issue because most of what is done is experimental in nature, and algorithms should be tested and retested.[25] On the contrary, for commercial diagnostic instrumentation used by eye care professionals, processing time should be a major factor. With most commercial videokeratography instruments available today, data processing takes only a few seconds,[26–29] which means that algorithms that take on the order of minutes would be unacceptable for professional clinical use. I used commercial programming language (MatLab; The MathWorks, Natick, MA), input files which have the same number of data points as the Eyesys videokeratograph (5760), and three different IBM-compatible computers with processors and RAM, respectively, of 1.6 GHz with 1 GB, 1.7 GHz with 1 GB, and 1.7 GHz with 0.5 GB. Processing times for each number of Zernike coefficients are illustrated in Figure 11.

The three curves were fit with second-order polynomials of time (t) as a function of number of Zernike coefficients (n)

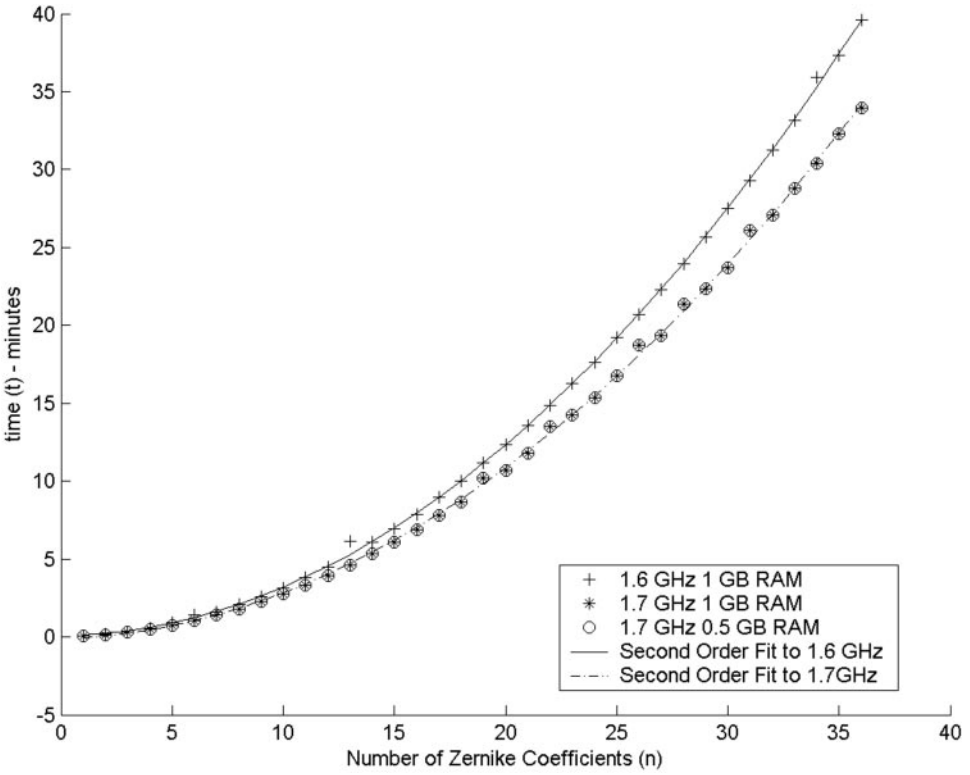

FIGURE 11. Processing times for 1 through 36 coefficients for three computer setups shown in the key. Data for the three computers were fit with second-order polynomials (parabolas) and, as shown by the bottom curve, there was practically no difference in processing time for the 1.7-GHz computers with different amounts of RAM.

$$t(n) = 0.0303n^2 + 0.0057n + 0.1113 \qquad (28)$$

and

$$t(n) = 0.0253n^2 + 0.0449n - 0.1415 \qquad (29)$$

where equation 29 refers to the 1.6-GHz computer data and equation 29 refers to both 1.7-GHz computers, since there was practically no difference in processing time for the latter computers, regardless of the quantity of RAM. From the graphs shown in Figure 11, the meaning of "acceptable processing times," at least for commercially available systems, should become clearer. Of course, there is no universally accepted standard for the least or the most time an eye care instrument should take to process patient data. Nevertheless, common sense tells us that an instrument that would take more than half an hour to process and print the results of patient examinations would probably have very little success among eye care professionals in today's fast-paced world. Moreover, extrapolating both graphs shown in Figure 11, using equations 28 and, 29 for a total of 72 coefficients, for example, shows that times would be 157.5901 and 134.0723 minutes, respectively, for the 1.6- and 1.7-GHz processors. These processing times are more than 100 times greater than are obtained with typical commercially available videokeratographs.[27]

For more quantitative values regarding times needed for phases before the specific Zernike coefficient processing, such as the image-processing phase, for both videokeratography systems and WF instrumentation, there are several resources (see Refs. 25–29).

Future work should be undertaken to make other practical conclusions regarding the most efficient procedures for application of ZPs in visual optics. A study should be conducted to recalculate the ZP coefficients of surfaces such as those presented herein for more than 36 terms, and an analysis should be made of the computational efficiency in terms of the desired accuracy when fitting complex surfaces. A more ideal definition of the optimum number of coefficients should be adopted as a standard in visual optics. The greatest challenge is the accuracy to which the corneal elevations and eye aberrations can be measured and the precision with which they should be represented, not the mathematical tools used to fit the surfaces. The most important decision factor today is computational time, given that, as has been shown in this work, ZPs are a sophisticated method and are not a limiting factor, since one can generate as many terms as desired.

References

1. Born M. *Principles of Optics.* New York: Pergamon Press; 1975; 464–466.
2. Guenther R. *Modern Optics.* New York: John Wiley & Sons; 1990.
3. Geary JM. *Introduction to Lens Design with Practical Zemax Examples.* Richmond, VA: Willmann-Bell; 2002.
4. Howland HC, Howland B. A subjective method for the measurement of monochromatic aberrations of the eye. *J Opt Soc Am.* 1977;67:1508–1518.
5. Liang J, Grimm B, Goelz S, Bille JF. Objective measurement of wave aberrations of the human eye with the use of a Hartmann-Shack WF sensor. *J Opt Soc Am A.*1994;11:1949–1957.
6. Thibos LN, Applegate RA, Schwiegerling JT, Webb R, and the VSIA Standards Taskforce Members. *Standards for Reporting the Optical Aberrations of Eyes.* Washington, DC: Optical Society of America, Visual Science and Applications; 1999.
7. Smolek MK, Klyce SD. Zernike polynomial fitting fails to represent all visually significant corneal aberrations. *Invest Ophthalmol Vis Sci.* 2003;44:4676–4681.
8. Klyce SD, Karon MD, Smolek MK. Advantages and disadvantages of the Zernike expansion for representing wave aberration of the normal and aberrated eye. *J Refractive Surg.* 2004;20:S537–S541.
9. E-Letter. *Invest Ophthalmol Vis Sci*. available at http://www.iovs.org/cgi/eletters/44/11/4676#114.
10. Carvalho LA. Preliminary results of neural networks and Zernike polynomials for classification of videokeratography maps. *Optom Vis Sci.* 2005;82:151–158.
11. Sarver EJ, Applegate RA. Modeling and predicting visual outcomes with VOL-3D. *J Refract Surg.* 2000;16:S611–S616.

12. Carvalho LA. Absolute accuracy of Placido-based videokeratographs to measure the optical aberrations of the cornea. *Optom Vis Sci.* 2004;81:616–528.
13. Guirao A, Artal P. Corneal wave aberration from videokeratography: accuracy and limitations of the procedure. *J Opt Soc Am.* 2000;17:955–965.
14. American National Standards Institute. Optical Laboratories Association, ANSI standard ASC Z80; February 20, 2004:7.
15. Artal P, Guirao A, Berrio E, Williams DR. Compensation of corneal aberrations by the internal optics of the eye. *J Vision.* 2001;1:1–8.
16. Artal P, Guirao A. Contributions of the cornea and the lens to the aberrations of the human eye. *Opt Lett.* 1998;23:1713–1715.
17. Liou HL, Brennan NA, Anatomically accurate, finite model eye for optical modeling. *J Opt Soc Am A*. 1997;14:1684–1695.
18. Klein SA. Corneal topography reconstruction algorithm that avoids the skew ray ambiguity and the skew ray error. *Optom Vis Sci.* 1997;74:945–962.
19. Butkov E. *Mathematical Physics.* Reading, MA: Addison Wesley; Facsimile edition; 1968.
20. Burek H, Douthwaite WA. Mathematical models of the general corneal surface. *Ophthalmic Physiol Opt*. 1993;13:68–72.
21. Foley JD, van Dam A, Feiner SK, Hughes JF, Phillips RL, eds. *Introduction to Computer Graphics.* Menlo Park, CA; Addison-Wesley; 1993.
22. Zhigang X, Plastock RA. *Schaum's Outline of Computer Graphics.* 2nd ed. New York: McGraw-Hill; 2000.
23. Rand RH, Howland HC, Applegate RA. Mathematical model of a Placido disk keratometer and its implications for recovery of corneal topography. *Optom Vis Sci*. 1997;74:926–930.
24. Iskander DR, Collins MJ, Mioschek S, Trunk M. Automatic pupillometry from digital images. *IEEE Trans Biomed Eng.* 200451: 1619–1627.
25. Carvalho LA. A simple and effective algorithm for detection of arbitrary Hartmann-Shack patterns. *J Biomed Inform.* 2004;37: 1–9.
26. Igarashi H, Kojima M, Igarashi S, Yoshida A, Cheng HM. A simple and effective video keratometric system. *Acta Ophthalmol Scand.* 1995;73:336–339.
27. Carvalho LA, Stefani M, Romao AC, et al. Videokeratoscopes for dioptric power measurement during surgery. *J Cataract Refract Surg.* 2002;28:2006–2016.
28. Carvalho LA, Tonissi SA, Castro JC. Preliminary tests and construction of a computerized quantitative surgical keratometer. *J Cataract Refract Surg.* 1999;25:821–826.
29. Carvalho LA, Castro JC, Carvalho LAV. Measuring higher order optical aberrations of the human eye: techniques and applications. *Braz J Med Biol Res.* 2002;35:1395–1406.