DOI 10.1515/aot-2012-0072 — Adv. Opt. Techn. 2013; 2(1): 97–109

Research Article

Christoph Menke* and G.W. Forbes*

Optical design with orthogonal representations of rotationally symmetric and freeform aspheres

Abstract: We review the motivations and intended benefits of orthogonalized representations of aspheric shapes and demonstrate unexpected gains for design purposes. Leaving aside improvements in 'design for manufacturability', we show that local optimization within standard design codes can find solutions with better optical performance when orthogonal descriptions are employed. Our examples include rotationally symmetric systems as well as systems with no global axis of symmetry and even a system of freeform surfaces (i.e., with no individual axes of symmetry).

Keywords: aspheres; freeform optics; optical design; optimization; orthogonal polynomials.

OCIS codes: 080.3620.

***Corresponding authors: Christoph Menke,** Carl Zeiss AG, Carl-Zeiss-Str. 22, 73447 Oberkochen, Germany, e-mail: christoph.menke@zeiss.com; and **G.W. Forbes**, QED Technologies Inc., 1040 University Ave., 14607 Rochester, NY, USA, e-mail: forbes@qedmrf.com

1 Introduction

Production limitations are constantly being tackled to advance the many applications of precision optics. These applications vary from commercial high-volume products such as miniature mobile phone cameras to scientific instruments such as giant monolithic or segmented telescopes and industrial tools such as ultra-high performance lithographic projection systems. Aspheric and freeform optics have come to play an increasingly central role across this domain. Although symmetry typically allows sections of spheres to be fabricated and tested more simply, the successful exploitation of non-spherical optics of growing complexity offers no end of difficulties. Unique challenges are faced by at least four different industries: those that provide optical design software as well as the hardware systems for generating (grinding, diamond turning, etc.), polishing/finishing, and measuring [coordinate measuring machines (CMMs), interferometers, etc.] optical surfaces. Sitting at the foundation of all this is the shared requirement for effective conventions to characterize surface shapes. Among the criteria for the effectiveness of any process for characterizing shape are 'generality', 'efficiency', and 'robustness'. We are primarily concerned here with particular aspects of efficiency.

Regarding 'generality', it is sufficient for many purposes to allow the exclusion of hyper-hemispheres and the like so that surfaces can then be specified by a single-valued sag function. In cylindrical polar coordinates, for example, such a surface can be expressed explicitly as $z=f(\rho,\theta)$. One of the traditional specifications for rotationally symmetric surfaces of this type involves a well-known set of coefficients:

$$z=c\rho^2\Big/\left(1+\sqrt{1-(1+\kappa)\,c^2\rho^2}\right)+\sum_{n=0}A_{2n+4}\,\rho^{2n+4}, \qquad (1)$$

where c and κ are the axial curvature and conic constant and A_4, A_6, A_8,... are monomial degrees of freedom. Only the starting values for the ranges of summation are shown explicitly in this work; options for the upper limits are discussed separately.

'Efficiency' is significant in several ways. Given that a primary task is the communication of shape, it is more efficient if comparable accuracy can be achieved while using fewer degrees of freedom, and also if the associated coefficients require fewer decimal digits. In the context of design, a reduced number of degrees of freedom also makes for faster optimization, but other efficiency-related aspects enter in that context. In particular, ray tracing benefits if the evaluation of the sag function and its derivatives are fast. What is more, it is a decisive advantage across the entire production chain if the representation directly facilitates estimates of manufacturability. The ultimate test in that regard is whether a surface shape is intelligible at a glance, that is, whether it is either 'humanly readable' or requires machine decryption.

'Robustness' is an important numerical consideration in this context, and it is largely tied to cancellation and

round off. This relates in part to being 'future proof'. That is, for a designer using say a 64-bit floating-point code, is it possible to simply keep including more coefficients in order to cope with surfaces of increasing complexity? If not, at what point can nothing be gained by additional terms regardless of the application?

For surfaces that are not rotationally symmetric, Eq. (1) can be generalized by using Cartesian coordinates and writing the sag as:

$$z=(c_x x^2+c_y y^2)\Big/\left(1+\sqrt{1-(1+\kappa_x)c_x^2 x^2-(1+\kappa_y)c_y^2 y^2}\right) + \sum_{j=0}\sum_{k=0} A_{jk}\, x^j y^k, \tag{2}$$

where there are now separate curvatures and conic constants in the x and y directions and a double-indexed set of monomial coefficients. We consider variations on Eqs. (1) and (2) and focus on applications where the peak-to-valley of the sag for each surface is smaller than the semi-diameter of its aperture. Although it is straightforward to avoid this assumption, it means that the associated best-fit sphere is itself not a hyper-hemisphere and that this spherical reference can then be used to boost effectiveness.

Degeneracy causes an infamous lack of robustness when working with the monomial sums in Eqs. (1) and (2). Their troubles arise from the fact that if we subtract two values that differ by less than one part in a million, the result is found to at least six fewer significant digits than is held by the less accurately known of the two original numbers. So, for example, unless there are more than six digits in our floating-point system, such an operation yields no accuracy at all. This sort of cancellation within monomial sums not only limits the number of terms that can effectively be used but also means that a surprisingly large number of digits must often be retained in the associated coefficient values. These weaknesses can be addressed by orthogonalizing the additive polynomials and using recurrence relations for their evaluation [1]. Impressively, such a step also effectively decouples these degrees of freedom so that terms can be dropped or added as judged to be appropriate at a glance. As we discuss in Section 2, by tailoring the orthogonalization to this context, various aspects of efficiency can be improved significantly at the same time.

Our main objective in what follows, however, is to demonstrate the varying success of standard local optimization routines when designing systems where the aspheric and freeform surface shapes are characterized in different representations. We are interested not only in the optical performance of the end results but also the convergence of the optimization process. The gains to be won by applying simple manufacturability-related design constraints have been impressively demonstrated elsewhere (e.g., [2, 3]). All these considerations are higher level measures of efficiency for the surface characterization. The design examples that we consider, which include tilted and even freeform elements, are presented in Section 3.

2 Orthogonal bases for optical surface shape

Rather than tailor a surface description to be well suited to particular shapes that have value in special circumstances (e.g., conic sections and more general Cartesian ovals for systems with small fields of view), it is more effective in general to anticipate the predominance of manufacturability as a driving consideration. Accordingly, the departure of a surface from a best-fit sphere immediately takes center stage. More specifically, manufacturability is typically related to variations in, and differences between, the local principal curvatures across the surface. In turn, these are driven directly by the rates of change of the displacement to the surface measured along the normal to the best-fit sphere. It is natural then to seek a characterization of shape in which the weighted mean square gradient of this normal departure from best-fit sphere is just a sum of the squares of the associated coefficients. In fact, this requirement is found to fully determine the characterization [4, 5].

2.1 Rotationally symmetric case

For a rotationally symmetric surface of semi-diameter $\rho_{\max}$, an alternative to Eq. (1) takes the form:

$$z=c\rho^2\Big/\left(1+\sqrt{1-c^2\rho^2}\right)+\frac{1}{\sqrt{1-c^2\rho^2}}N_{\text{bfs}}(\rho/\rho_{\max}), \tag{3}$$

where the pre-factor on N_{bfs} is just division by the cosine that effectively converts a displacement measured in the z direction to departure along the local normal vector of the sphere. Note that the argument of N_{bfs} varies from zero to unity, and that the sphere in Eq. (3) can be designated as a best-fit sphere if this departure vanishes at both limits. Accordingly, in terms of $u=\rho/\rho_{\max}$, the normal departure from best-fit sphere is expressed as:

$$N_{\text{bfs}}(u) = u^2(1-u^2)\sum_{n=0} a_n^0\, Q_n^0(u^2), \tag{4}$$

where Q_n^0 is a polynomial of order n. (In keeping with the next subsection, the superscripted zero signifies rotational symmetry.) Given a weighting to be used in the average over the aperture, these polynomials are uniquely determined by requiring that the mean square slope of $N_{\text{bfs}}(u)$, that is, the mean square of $\frac{d}{du}N_{\text{bfs}}$, is just the sum over n of $(a_n^0)^2$. For one well-suited weighting, the resulting polynomials are discussed in [4], and a sample of them is plotted in Figure 1.

A glimpse of some of the strengths of this orthogonal characterization can be had upon taking representative aspheres from the patent literature and expressing their shape by using Eqs. (1) and (4). Such a result is presented in Figure 2 where these two characterizations match down to the nanometer level. The shape of this surface is not at all apparent from the traditional characterization. By contrast, as the zeroth basis element in Figure 1 attains a peak value of 0.25, the new coefficients in Figure 2 make it immediately clear that this asphere has approximately 2300/4≈600 microns of departure from a best-fit sphere. This orthogonal characterization also reveals at a glance that there are no significant high-order components in the shape. What is more, only approximately one-third of the digits is now required because the cancellation between the different monomial components is avoided. It was also seen at a glance that the last two terms could be dropped while leaving the others unchanged and retaining a nanometer level of agreement. These are all indications of significant efficiency gains.

2.2 Freeform optics

In many applications of freeforms, it is sufficient to choose the domain for orthogonalization to be a circular cylinder that tightly encloses the aperture. In this case, $\rho_{\max}$ is taken to be the radius of the enclosing cylinder and N_{bfs} becomes a function of the polar angle mentioned leading into Eq. (1). It turns out that all the monomial terms of Eq. (2) are accounted for if we now express the normal departure of Eq. (3) in polar coordinates as:

$$N_{\text{bfs}}(u,\theta)=u^2(1-u^2)\sum_{n=0}a_n^0\,Q_n^0(u^2)+\sum_{m=1}u^m\sum_{n=0}[a_n^m\cos(m\theta)+b_n^m\sin(m\theta)]\,Q_n^m(u^2). \tag{5}$$

Again, only the lower limits of the indices of summation are shown; much as for the sums in Eq. (2), their upper limits can be chosen in a number of ways. The first line of Eq. (5) contains the m=0 terms that are precisely those of Eq. (4).

The equivalence of Eq. (5) and a Cartesian monomial sum like that in Eq. (2) can be appreciated by expanding the right-hand side and then equating real and imaginary parts of the identity:

$$u^m(\cos m\theta+i\sin m\theta)=u^m e^{im\theta}=(u\cos\theta+iu\sin\theta)^m=(X+iY)^m. \tag{6}$$

With this, it can be seen that all terms up to, say, order T are included if Eq. (2) is summed over positive indices satisfying $j+k\le T$, or Eq. (5) is summed over:

$$\begin{cases}2n+4\le T, & m=0,\\ m+2n\le T, & m>0.\end{cases} \tag{7}$$

Note that the T+1 degrees of freedom at order T are b_n^0 with j=0,1,...T and k=T-j for the sum in Eq. (2), whereas, for Eq. (5), they are a_n^m and b_n^m where m has the same parity (even/odd) as T with $0\le m\le T$ and

$$n=\begin{cases}\frac{1}{2}(T-4), & m=0,\\ \frac{1}{2}(T-m), & m>0.\end{cases} \tag{8}$$

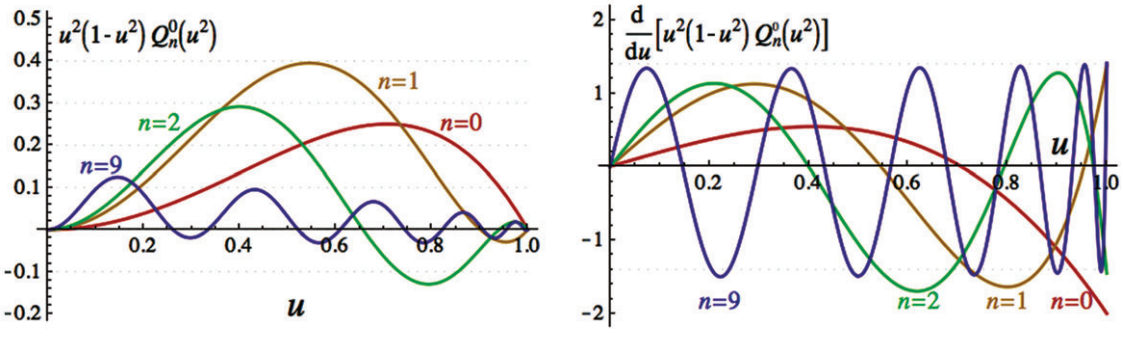

Figure 1 Low-order slope-orthogonal elements of Eq. (4). The derivatives plotted (right) reveal their Fourier-like orthogonality.

n	A_{2n+4} (mm^{-2n-3})	a_n (nm)
0	-2.83553693E-08	2,252,133
1	-1.12122261E-11	-177,695
2	-2.05192812E-16	-32,347
3	-1.55525080E-20	2,293
4	-4.77093112E-24	-211
5	8.39331135E-28	dropped
6	-8.97313681E-32	dropped

Figure 2 Alternative prescriptions for L616 of US patent no. 6,646,718 [6]. The clear aperture (CA) is 143.652 mm. This surface has κ=0 and the inverse of the axial curvature for Eq. (1) is 101.25424 mm, whereas that of the best-fit sphere for Eq. (3) is 108.02985 mm.

[When counting to identify the T+1 terms for any T>2, note that b_n^0 does not enter Eq. (5), and this drops one of the terms when T is even.] If the rotationally symmetric components of Eq. (5) are supplemented by u^0 and u^2 (i.e., piston and power), all terms up to order T in either Eq. (5) or in a Cartesian monomial sum have equal numbers of degrees of freedom and are precisely interchangeable linear combinations of each other.

For any given weighting function, the polynomials in Eq. (5) can be determined uniquely by requiring that the weighted mean square gradient of $N_{\mathrm{bfs}}(u,\theta)$ is just the sum of the squares of all the coefficients, see [5]. Although – as shown in the plots in Appendix 1 – the results differ significantly from Zernike polynomials, these basis elements can be labeled in terms that are familiar from the Zernike domain. The four basis members that typically dominate the spectra are plotted in Figure 3 and are referred to in what follows with familiar labels: astigmatism, coma, trefoil, and spherical aberration. Also, note that, although the tilt terms associated with a_0^1 and b_0^1 may sometimes be significant, they are of minor importance as far as manufacturability is concerned. In fact, they can be taken to be zero without loss of generality provided the part has tip and tilt freedoms when it is configured in the system during optimization.

Just as for the rotationally symmetric case, it is instructive to express freeform surfaces from the patent literature in both the representations considered here. One such example is shown in Figure 4 where the listed coefficients from the patent include all terms up to order 10 in Eq. (2) and are in that upper table. Both conic constants vanish for this mirror and the radii in mm are $1/c_x$=-452.62713 and $1/c_y$=-443.43539. This mirror has a plane of symmetry and the coordinates are aligned so that A_{jk}=0 whenever j is odd. Note that two terms were dropped from the specification,

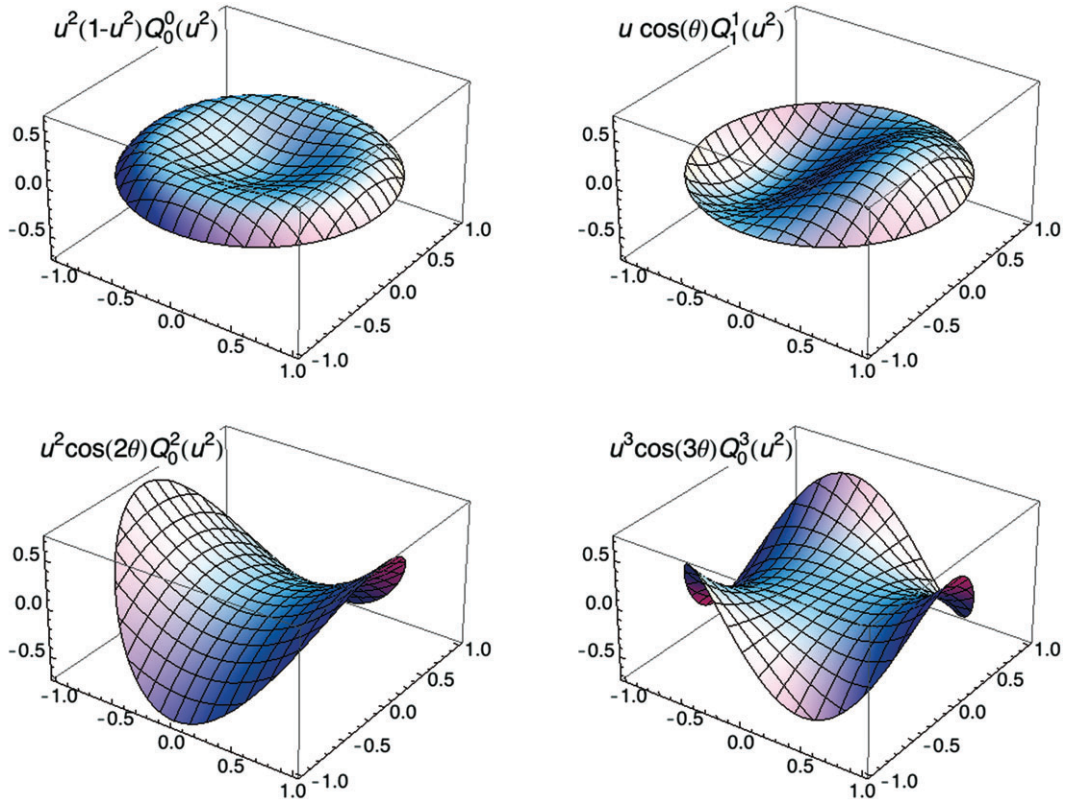

Figure 3 Plots of the lower-order basis elements for m=0, 1, 2, and 3 (spherical, coma, astigmatism, and trefoil, respectively). These are all plotted on the same scale and are of 4th, 3rd, 2nd, and 3rd order with peak values of 0.25, 0.41, 0.71, and 0.54, respectively.

j

k	0	2	4	6	8	10
0		6.086975E-05	1.345443E-10	3.310262E-16	1.020537E-21	1.653756E-28
1		-9.657166E-08	-4.424293E-13	-1.749391E-18	-6.739667E-24	
2	8.446692E-05	3.881972E-10	1.672236E-15	7.515349E-21	-3.800397E-27	
3	-1.773111E-08	-5.340721E-13	-4.286471E-18	-2.305324E-23		
4	2.103339E-10	1.962740E-15	1.613314E-20	1.939290E-26		
5	-4.450410E-14	-3.972902E-18	-4.548523E-23			
6	1.204820E-15	2.276418E-20	9.938038E-26			
7	-3.751270E-18	-6.515923E-23				
8	1.243271E-20	1.259617E-25				
9	-1.671689E-23					
10	2.740074E-26					

A_{jk} (mm^{1-j-k})

m

n	0	1	2	3	4	5	6	7	8	9	10
0	70,531	-225,291	25,894	199,399	-3,583	-2,651	1,886	-339	-55	41	-5
1	-43	223,995	-11,377	-198	-2,604	801	-46	-37	5		
2	-82	-12,916	3,592	-994	158	10	-5				
3	-10	1,568	-256	10	2						
4		-20	-3								

a_n^m (nm)

Figure 4 Alternative representations of M6 from 'projection optics 37' of US patent no. 2012/0069315 A1 [7]. The Cartesian coefficients from the patent are listed along with the new spectrum of coefficients (to T=10) for this shape.

namely (j,k)=(0,0) and (0,1), which makes the polynomial and its first derivatives vanish at the origin.

When using Eq. (5), the plane of symmetry means that b_n^m vanishes for all m and n when the coordinates are related by (x,y)=(u sin θ, u cos θ). (This choice makes changing the sign on θ equivalent to changing the sign on x and, just as for the conventional description, all the odd terms then drop out.) With $\rho_{\max}$=174.2 and 1/c=-478.12597, the resulting values of a_n^m are given in the lower table of Figure 4. The illuminated aperture for this part is roughly elliptical with an aspect ratio of 93%, and hence is slightly smaller than the enclosing circle used here. These two characterizations match to better than 1 nm over the enclosing circle, but the second of them requires approximately one-third the number of digits and is plainspoken: this shape is evidently dominated by a couple of hundred microns of each of coma (a_1^1) and trefoil (a_0^3). Keep in mind that, as demonstrated in [5], the value of the mean tilt (a_0^1) is largely irrelevant to the shape; it relates more to the orientation of the mirror and, much as was done with A_{01} in the original specification, it can be dropped from the spectrum provided tip and tilt freedoms are used during optimization. The issue of how easily such an asphere can be produced is discussed briefly in the next subsection.

2.3 Manufacturability

Different stages of production have their own limitations, but manufacturability is oftentimes coupled to entities such as local differences and/or global variations in the principal curvatures of the surface. By using the methods developed in [8], the accessible domain in such cases can be estimated in terms of an ellipsoid centered at the origin in parameter space of the coefficients of Eqs. (4) or (5). A unique strength of those representations is that manufacturability generally improves as their coefficients become smaller. Because the four elements plotted in Figure 3 tend to dominate the spectra, an idea of the accessible domains can be gleaned by determining the extent of the ellipsoid in just these few dimensions.

The case of stitched interferometric testing is used here as an example. It was shown in [8] that, provided the CA is not too small, the size of that ellipsoid is independent of the part size. It does depend on part speed, however, which is defined here as η=$c\rho_{\max}$. The results for a representative test configuration are plotted in Figure 5. It can be seen that, provided a transmission sphere of appropriate size is available, a freeform should be able to be stitched when these low-order coefficients do not exceed a few hundred microns. (Note that the part in Figure 4 falls within this domain.) By contrast, the domain for full-aperture testability is a spheroid of approximately ten times smaller semi-diameter, as indicated by the dotted line in Figure 5. Such rules of thumb can be determined for other production processes, of course. They are of enormous value and can be refined as needed for progressively more sophisticated and accurate estimates of manufacturability.

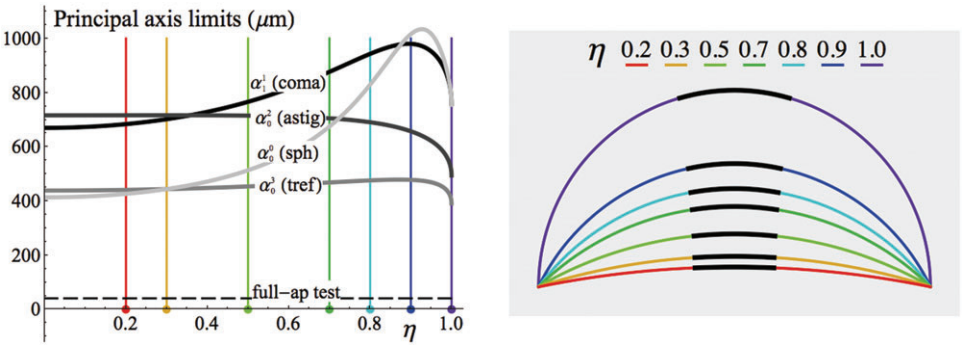

Figure 5 The semi-diameters of the principal axes of the 'ellipsoid of stitchability' vary with speed of the best-fit sphere. The plotted limits are for a 1 K camera at HeNe wavelength at up to a root mean square (RMS) of half Nyquist with subapertures larger on average than the black arcs on the cross-sections of best-fit spheres indicated (right).

3 Design case studies

We now present several design case studies and compare the optimization behavior of various surface types. All designs share a common feature: a large number of aspheric coefficients, either on a single surface or on multiple surfaces. With the exception of the last example, the surfaces are rotationally symmetric. In addition to the standard aspheric descriptions of Eqs. (1) and (2), with or without a normalization radius, we use several orthogonal representations, namely Zernike polynomials, the orthogonal polynomials introduced in [9] and called Q_{con} in [4], and the gradient-orthogonal polynomials Q_n^m of Eqs. (3), (4), and (5) (which, for m=0, are precisely Q_{bfs} as discussed in [5]).

The starting systems are simple spherical or low-order aspheric systems. Complexity is increased by adding more aspheric coefficients and re-optimizing the systems without changing the error function. To explore the generality of our findings, we use two commercial optical design codes, namely ZEMAX® [10] and CodeV® [11], as well as the Carl Zeiss custom optical design software.

3.1 Cassegrain telescope

A simplistic Cassegrain-type telescope was used in [12] to demonstrate some parametrically defined surfaces when the standard aspheric description fails. The telescope is shown in Figure 6. It consists of a concave spherical primary mirror and an aspheric convex secondary located near the ray caustic. Table 1 lists the design parameters.

We used ZEMAX® to compare various aspheric surface types: the standard aspheric description called 'Even Asphere' in ZEMAX® with a polynomial expansion up to the 16th order; the 'Extended Asphere' which is similar to the Even Asphere but can support higher orders and uses a normalized radial coordinate; the 'Zernike Standard Sag' surface which supports the standard Zernike polynomials up to 20th order (using only the rotationally symmetric terms for the Cassegrain); and the ZEMAX® implementations 'Qcon_recurr' and 'Qbfs_recurr' of the sets of orthogonal polynomials [10] that use numerically stable recurrence formulas for the associated computations. All surface types except the Even Asphere use a normalization

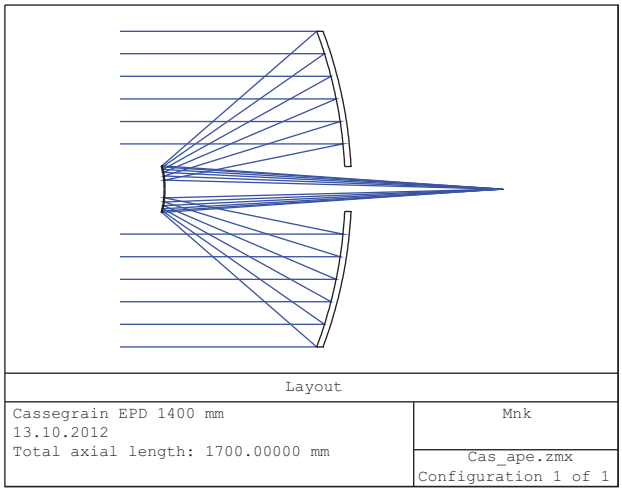

Figure 6 Cassegrain-type telescope with a spherical primary mirror and an aspheric secondary mirror near the caustic.

Table 1 Parameters of the Cassegrain-type telescope.

Primary mirror radius of curvature	2000 mm
Primary mirror diameter of aperture	1400 mm
Vertex separation between mirrors	800 mm
Distance from secondary mirror to image plane	1500 mm

radius. We fixed this radius at 105 mm to ensure that all rays intercept the secondary mirror within this domain.

For simplicity, only the wavefront of the on-axis field point was optimized and we started with a purely spherical system. The fourth-order aspheric coefficient of the secondary mirror was released and the system optimized until convergence. This was repeated for the higher-order aspheric coefficients, varying one additional coefficient at a time while keeping the lower-order coefficients variable and each time starting with the last optimized system.

The optimization results are shown in Figure 7. The RMS wavefront error is measured at a wavelength of 633 nm. The results for the Even Asphere and the Extended Asphere surface types were identical, thus only the latter are shown. All descriptions yield nearly the same wavefront error when up to eight aspheric coefficients are used. Releasing additional coefficients slightly improves the results for the Extended Asphere. However, the wavefront errors are significantly larger than those of the Q polynomials when more than eight coefficients are used (and of the Zernike surface with nine coefficients). This indicates numerical instabilities during optimization for the non-orthogonal surface type. These instabilities are likely to be associated with the degeneracy of the monomial sums, which was discussed in Section 1. The final Q-based systems (14 coefficients=30th order) can be converted to within insignificant errors into the standard asphere representation after the fact, but the optimizer seems unable to find this solution if the surface is described with a non-orthogonal monomial sum during design.

It is interesting to note that, when all aspheric coefficients are released at once, the orthogonal representations (Zernike, Q_{con}, Q_{bfs}) yield exactly the same results as with the previously described procedure. The wavefront error for the Extended Asphere improves slightly (0.17 waves instead of 0.19), but is still more than a factor of 5 larger than the wavefront error with the same number of coefficients for the orthogonal representations. Again, this points out that numerical instabilities can be avoided by the use of orthogonal polynomials.

Of course, this example is artificial and only intended to test the ability of the optimizer to find a certain aspheric shape with a specific surface representation. It demonstrates that orthogonal surface types can have advantages when more than eight or so aspheric coefficients are used.

3.2 Three-mirror anastigmat

The second design study is a three-mirror anastigmatic objective. We used the CodeV® sample lens 'threemir.len' as a starting system, increased the entrance pupil diameter from 100 mm (f/2.5) to 125 mm (f/2), and reduced the full-field angle from 5°×1° to 0.5°×0.5°. Figure 8 shows the layout of the starting system. All mirrors are rotationally symmetric aspheres, but they do not share a common axis.

In the starting system, three parameters per surface are varied: the curvature, the conic constant, and the fourth-order aspheric coefficient. Twelve more parameters control the distances between the mirrors, the positions of the mirror vertices, and of the image surface, adding up to a total of 21 variable parameters.

The rectangular field was sampled with six equally distributed field points. We included constraints in the error function to ensure that no mirror blocked the light between other mirrors and that the image is accessible. Distortion was controlled to be smaller than 0.5%. We applied the following optimization strategy: aspheric coefficients were always released simultaneously on all three mirrors, but only one coefficient per mirror at a time. The last optimized system was used as the starting system

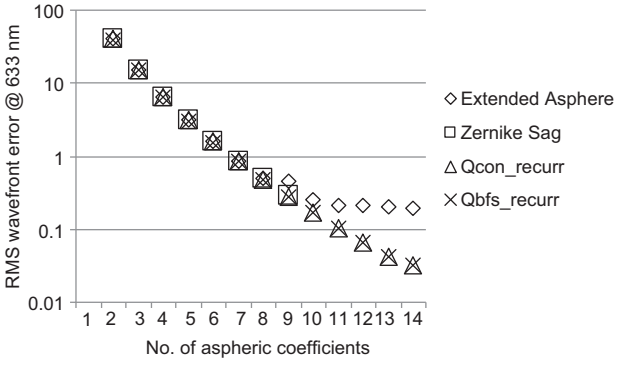

Figure 7 Optimization results for the Cassegrain-type telescope with increasing number of aspheric coefficients on the secondary mirror.

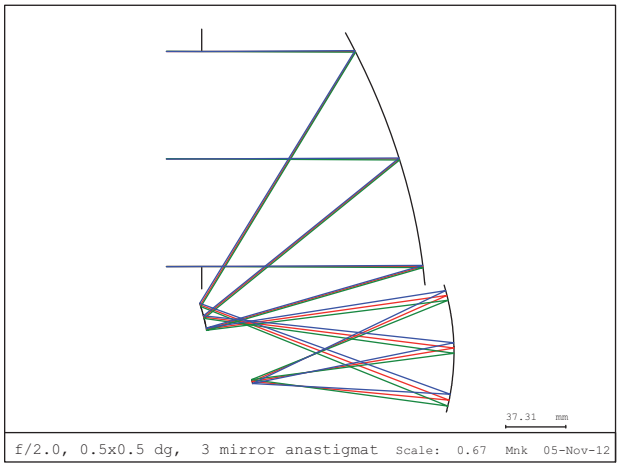

Figure 8 f/2, 0.5° × 0.5°, three-mirror anastigmat.

for 100 cycles of optimization before the next aspheric coefficients were turned into variables. The resulting RMS wavefront errors for four different surface types (standard asphere, Fringe Zernike, Q_{con}, Q_{bfs}) are shown in Figure 9.

In the starting system, the mirrors are described as standard aspheres. We used the built-in conversion with an automatic choice of the normalization radius to generate the starting systems for the Fringe Zernike (using only the rotationally symmetric terms), Q_{con}, and Q_{bfs} surface types. Whereas the conversion to Zernike and Q_{con} is exact, the conversion to Q_{bfs} is an approximate fit. The Q_{bfs} surface type in CodeV® does not yet have a conic term as an option. (Such an option has been introduced [8], but because it breaks the close coupling to manufacturability in most applications, it is recommended only for exceptional cases, e.g., extremely fast parts.) Instead of conic plus 4th-order coefficient, we started with the 4th- and 6th-order coefficients for the Q_{bfs} surfaces to retain the same number of variables, and then re-optimized the system. The resulting Q_{bfs} system showed much worse performance than the other systems: 0.17 waves of composite RMS wavefront error compared with 0.03 waves.

However, when two more aspheric coefficients are added (radius plus four aspheric coefficients per mirror, thus 27 variables in total), the performance of the Q_{bfs} system is superior to all other systems: 0.014 waves compared with 0.016 for Q_{con} and Zernike and 0.018 for the standard asphere. Varying more coefficients does not help to improve the Zernike system at all, but gives a slight improvement for the other surface types.

In this design example, we obtain a 25% reduction in RMS wavefront error simply by using an orthogonal representation with the same number of parameters. The optimizer finds a deeper minimum for the same error function. This seems unlikely to be caused by lack of numerical robustness because the number of parameters per surface is small: only four aspheric coefficients. There now appears to be a more complex interaction between the three aspheric mirrors. By changing the surface representation, the dependence of the error function on the basic parameters changes even though the essential definition of the error function remains unchanged. Transformation of the parameter space evidently enables the optimizer to find a deeper minimum in this case.

Importantly, changing the surface type does not affect the family of surface shapes that is covered by the representations: When the final Q-based systems are converted to standard aspheres, the associated wavefront error remains the same. That is, the minima also exist in the standard asphere representation, but the optimization algorithm was unable to find them.

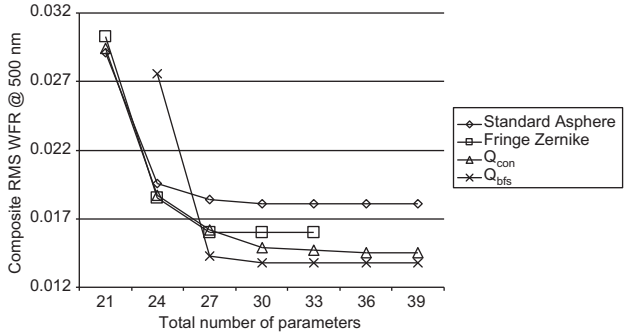

Figure 9 Optimization results for the three-mirror anastigmat with increasing number of aspheric coefficients.

3.3 Four-mirror off-axis system

Our next design example is a four-mirror off-axis system based on US patent no. 6,577,443 [13] with an image-side numerical aperture of 0.12, a reduction ratio of 0.25, and a radial field of 4 mm. The basic layout is shown in Figure 10. All mirrors are rotationally symmetric with respect to a common optical axis. The outermost field point is positioned at 200 mm from the optical axis. The stop is on the second mirror from the image side, producing a telecentric image.

Once again, we performed this design example with ZEMAX®. The error function consisted of operands controlling the wavefront error, distortion, and telecentricity for five equally distributed and weighted field points. Operands were added to prevent the blocking of rays by the mirrors. The overall length was restricted to be smaller than 1600 mm. Starting with a purely spherical system, we used the same optimization strategy as for the three-mirror system: release one aspheric coefficient simultaneously on all mirrors and optimize 100 cycles before varying the next higher-order aspheric coefficients. Figure 11 illustrates the optimization behavior that strongly depends on the chosen surface type.

When more than two aspheric terms per surface are employed, the system with rotationally symmetric Zernike

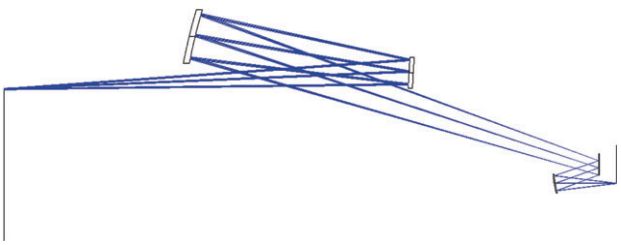

Figure 10 Layout of the four-mirror off-axis system.

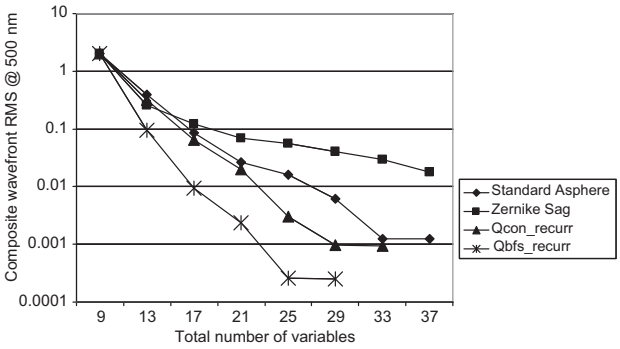

Figure 11 Optimization results for the four-mirror off-axis system.

surfaces exhibits the worst performance. At first glance this is surprising, but it may be due to the ring field which has the effect that all mirrors except the pupil mirror are hit by rays only in a small portion of their full aperture. Although the Zernike polynomials are orthogonal over the unit circle, they appear to be poorly suited to this case because of their strong gradients in the outer annulus.

The standard asphere representation leads to better results than the Zernike polynomials, but the Q_{con} surface type is superior to both of them. Q_{con} reaches the final RMS wavefront error level with a smaller number of variables (a total of 29 compared with 33, corresponding to five and six aspheric coefficients on each mirror, respectively), and additionally the level is approximately 20% lower. Nothing is gained by varying higher-order coefficients from that point.

The system with Q_{bfs} surfaces reaches by far the best RMS wavefront error: a factor of 10 lower than Q_{con} – the second best – with 25 variables (four aspheric coefficients per mirror). The final level still is approximately 75% lower with Q_{bfs} than with Q_{con} – a dramatic four times reduction in wavefront error with a smaller number of variables. As shown in Figure 12, the final configurations differ in the positions and sizes of the first and second mirror. Once again, the Q_{bfs} system can be converted into the standard asphere representation without loss of performance.

It is interesting to speculate why the Q polynomials do so much better than the Zernikes in this off-axis system. Both sets of polynomials are orthogonal over the unit circle. However, in contrast to the Zernike polynomials, the slopes of the Q_{bfs} polynomials remain modest even near the edge of the aperture (see the discussion of the figures in Appendix 1). We have no explanation for the superior behavior of the Q_{con} over the Zernike polynomials in this design example. As an aside, note that both the Zernikes and Q_{bfs} can be scaled to be orthogonal over annular apertures, as discussed in [8], but this option is currently unavailable in the design code.

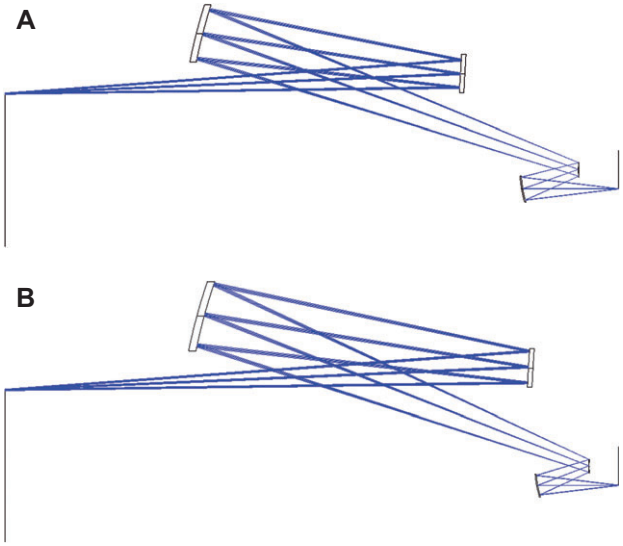

Figure 12 Layout of the four-mirror off-axis optimized system. (A) Standard aspheres. (B) Q_{bfs} surfaces.

3.4 Freeform prism

Our last design study is an example with surfaces that are not rotationally symmetric. The system is a wedge-shaped freeform prism for a head-mounted display based on US patent no. 5,959,780 [14]. The prism magnifies the image of a microdisplay into the eye pupil of the observer. Figure 13 shows the basic layout. The effective focal length of the patent system is scaled to 33 mm, the diameter of the pupil is 15 mm, and the full field of view is 32°×24°. The system is symmetric about the *yz*-plane (the paper plane).

The original prism consists of two freeform surfaces and a planar surface. This surface, next to the pupil, is used

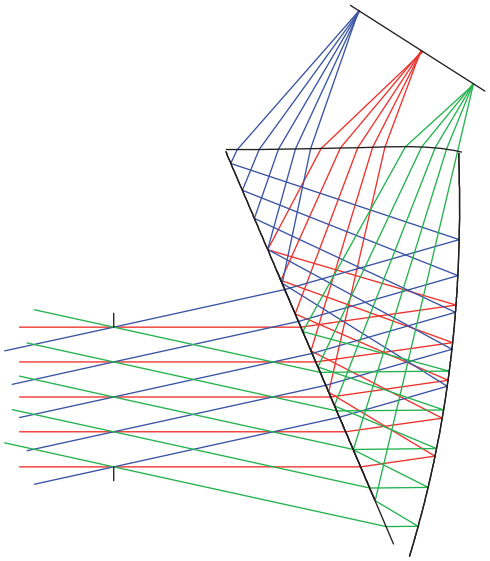

Figure 13 Freeform prism for a head-mounted display.

twice: first in total internal reflection, then in refraction. The condition for total internal reflection at this surface is controlled for all rays during optimization. Additional structural constraints ensure that the prism remains physically realizable and we limited the distortion to be <8%.

To investigate the optimization behavior of the Q polynomials of Eq. (5) for this case, we implemented the new representation in the Carl Zeiss custom optical design software. The prism surfaces were represented either by standard XY polynomials, by Fringe Zernike polynomials, or by this new set of Q polynomials. The originally planar surface was also converted to a freeform surface. For each of the surfaces, 34 parameters were varied during the optimization leading to a total number of variables >100. The variables were chosen so that the Q and the Zernike polynomials covered the same function space as the XY polynomials (as described in Section 2.2 and displayed in Figure 4).

Figure 14 gives the error function values that are reached at the end of each optimization cycle. When the prism surfaces are described by XY polynomials, the error function falls until the optimization stagnates after 30 cycles at a value of approximately 33% of the initial level. With Q polynomials, however, the error function falls much more rapidly and achieves superior performance in the first four or five optimization cycles, followed by a slower reduction. After 25 cycles, the error function did not improve further. The error function is then at 17% of the initial level – a reduction of the XY polynomial result by a factor of two. This reduction in error function corresponds to an average 15% improvement in modulation transfer function (MTF) at 30 lp/mm over the full field, see Figure 15. Except for the first few cycles, the result with the Zernike surfaces is falling in between the XY and the Q polynomials, but the end result offers less than half the improvement seen with Q.

It turns out that the surface sag and the gradients are at similar levels for the final systems. In this case study, one observes faster convergence to a deeper minimum with the new Q polynomials. This starkly illustrates a potential impact of an appropriate surface representation.

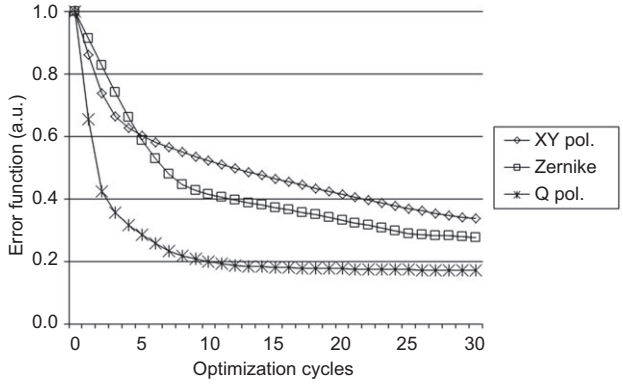

Figure 14 Comparison of optimization results for the freeform prism with surfaces described by XY, Zernike, and Q polynomials.

4 Concluding remarks

Creating a more convenient standard for communicating shape was a primary goal of the orthogonal bases discussed here. In particular, offering fewer digits and human intelligibility as well as a closer link to manufacturability were central considerations. Benefits were also expected in terms of design for manufacturability, where the simple constraints enabled by the new characterization could better define regions to be searched for both local and global optimizers. Such benefits have already been demonstrated impressively in terms of maintaining performance while eliminating inflection points and significantly reducing tolerance requirements for assembly (e.g., [2, 3]). Our results build on this by establishing that, as an unexpected bonus, gains in raw optical performance and convergence may also be won.

Although no single characterization of shape is expected to be optimally suited to all surfaces in the

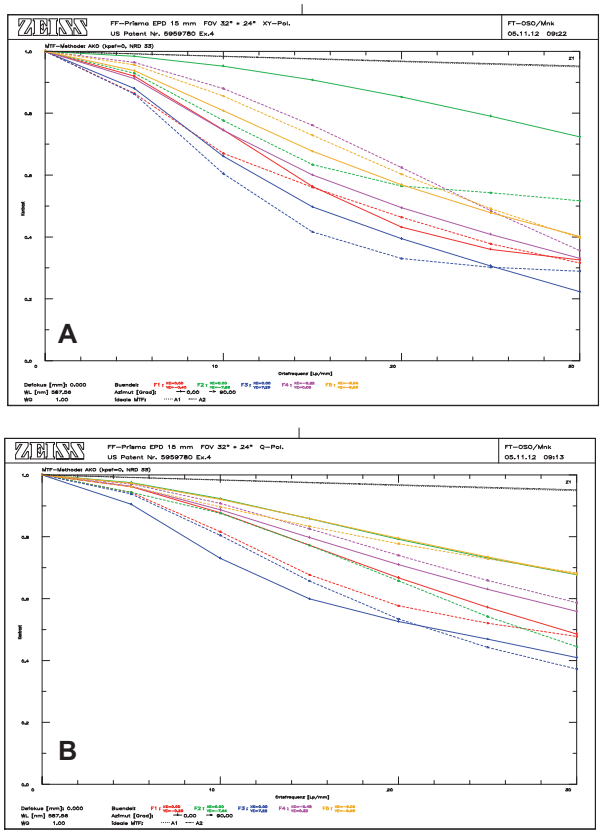

Figure 15 MTF results for the freeform prism. (A) XY polynomials. (B) Q polynomials.

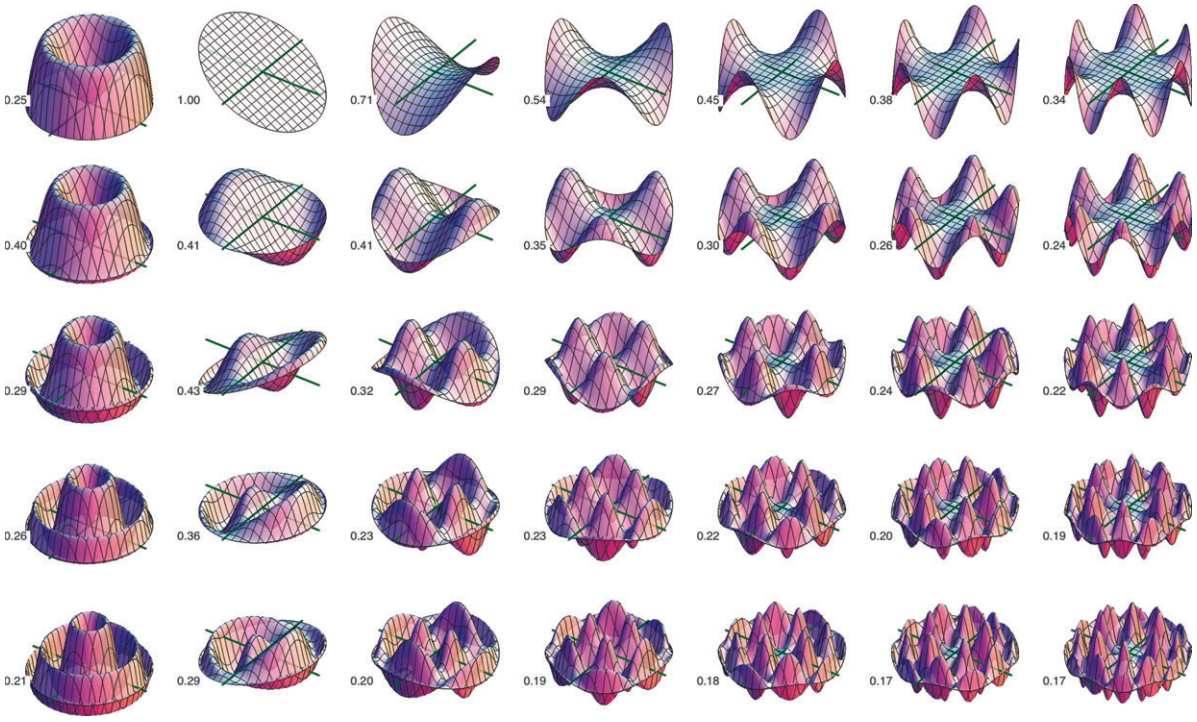

Figure 16 Plots of the lower-order Q basis elements for m=0, 1, 2,...6 running horizontally and n=0, 1, 2,...4 running vertically. The peak absolute value attained by each of the elements is given as an inset.

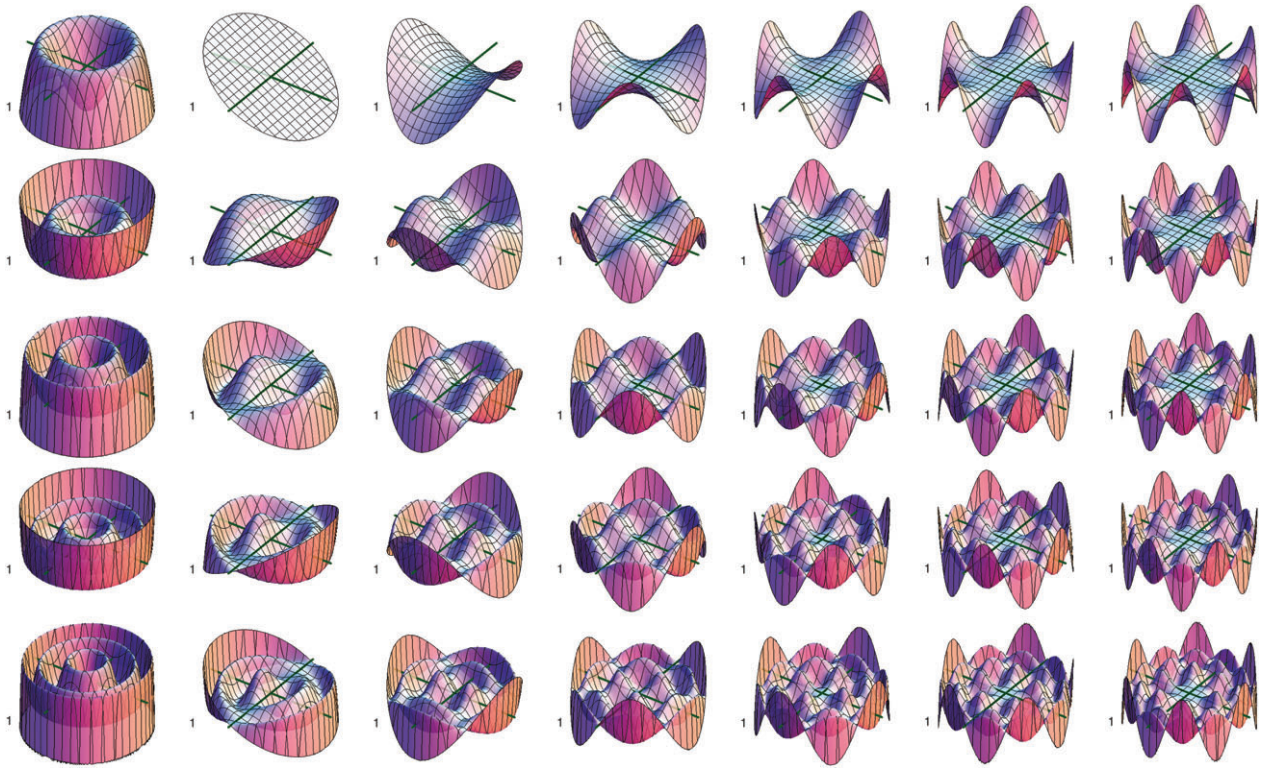

Figure 17 Plots of the Zernike basis elements for comparison with the elements of Q in Figure 16. The signs have been chosen to match the topology of the corresponding elements of Q, and piston and power have been dropped from the first column for maximal consistency between these two figures. By convention, they are all scaled to have a peak absolute value of unity.

current span of optical systems, our design studies demonstrate that orthogonal polynomials sometimes offer surprising advantages. It is helpful that orthogonal bases benefit from robust and efficient algorithms, but such internal mathematics is unimportant from a user's perspective; adopting and understanding the degrees of freedom – just a simple spectrum – are all that is required. In our opinion, the sooner we all become familiar with such spectra, the better for our whole industry.

Acknowledgments: Part of the work of C.M. has been financially supported from the German Ministry of Science and Education (BMBF) in the frame of the project 'FREE' (FKZ: 13N10827).

Received November 16, 2012; accepted January 11, 2013

Appendix 1

To give a better appreciation of the gradient-orthonormal basis elements of Eq. (5), we have created some additional graphics. As in Figure 3, only the reflection-symmetric elements (involving cosine) are drawn in Figure 16; the antisymmetric elements (involving sine) follow simply upon rotating these by $\pi/(2m)$ about the z axis. Note that the plots in the first column (where m=0) are just the curves of Figure 1 spun about the z axis. This grid of plots is arranged to match the layout of coefficients in the lower table of Figure 4. The plotted shapes and the inset amplitudes therefore give an intuitive meaning to such spectra of coefficients.

For comparison with the Zernike basis elements, an analogous set of plots is given in Figure 17. Note that, as indicated by the green lines of the x and y axes, the elements in the first column are now non-zero at center and edge. This means that the associated coefficients in this case change axial thicknesses as well as the curvature of the best-fit sphere. As they are built from zeroth-order polynomials, that is, just constants, the remainder of the first row (n=0) is identical in shape to their counterparts in Figure 16. However, the strong peaks and high gradients near the aperture edge for all n>0 are significant distinctions that presumably lead to the different optimization results reported in Section 3. For example, the peak gradient value divided by the RMS of the function value is more than an order of magnitude higher for the Zernikes in the last row (i.e., n=4) than for the corresponding elements of Q. The general differences are perhaps easiest to appreciate initially by a pairwise comparison of the plots in the second columns (i.e., m=1).

References

[1] G. W. Forbes, Opt. Exp. 18, 13851–13862 (2010).
[2] D. Stephenson, Proc. SPIE OptiFab TD07–38 (2011).
[3] B. Ma, K. P. Thompson, K. Sharma and J. P. Rolland. Front. Optics Tech. Digest FTh3E.3 (2012).
[4] G. W. Forbes, Opt. Exp. 15, 5218–5226 (2007).
[5] G. W. Forbes, Opt. Exp. 20, 2483–2499 (2012).
[6] K.-H. Schuster, D. R. Shafer, W. Ulrich, H. Beierl and W. Singer. US patent 6,646,718 B2 (2003).
[7] H.-J. Mann, J. Zellner, A. Dodoc, C. Zahlten, C. Menke et al., US patent 2012/0069315 A1 (2012).
[8] G. W. Forbes, Opt. Exp. 19, 9923–9941 (2011).
[9] J. Kross and R. Schuhmann, Optik 70, 76–85 (1985).
[10] ZEMAX®, in 'ZEMAX® Optical Design Program Users' Manual' (2011).
[11] Synopsys, Inc., in 'CodeV® Reference Manual' (Synopsys, Inc., Mountain View, CA, 2011).
[12] S. A. Lerner and J. M. Sasian, Appl. Opt. 39, 5205–5213 (2000).
[13] U. Dinger and H.-J. Mann, US patent 6,577,443 (2003).
[14] T. Togino and J. Takahashi, US patent 5,959,780 (1999).

Christoph Menke is a Staff Scientist at Carl Zeiss, Oberkochen. In 1998, he joined the Optical Design Department of the Carl Zeiss Corporate Research and Technology. His main work areas include the design of photographic and lithographic lenses, freeform systems, and optimization algorithms in optical design software. He holds a PhD in Mathematics from the University of Ulm, Germany. Since 2007, he has been a lecturer at the University of Stuttgart on Optical System Design.

Greg Forbes has been Senior Scientist at QED Technologies since 2000. Although the company is based in Rochester (NY, USA), Greg lives in Sydney (Australia). He developed the algorithms that underpin and drive QED's subaperture polishing and stitched-interferometry systems that have helped transform the commercial production of a wide range of high-precision optics. Following his PhD in Theoretical Physics at the Australian National University, he was a Fulbright Fellow at the Optical Sciences Center (Tucson, 1984), a tenured faculty member of The Institute of Optics (Rochester, 1985–1994), and a Research Professor at Macquarie University (Sydney, 1994–2000). Throughout his career, optical modeling has remained one of his primary interests.