

Electronic Theses and Dissertations, 2004-2019

2013

Mathematical And Computational Methods For Freeform Optical Shape Description

Ilhan Kaya
University of Central Florida

 Part of the [Computer Engineering Commons](#)
Find similar works at: <https://stars.library.ucf.edu/etd>
University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Kaya, Ilhan, "Mathematical And Computational Methods For Freeform Optical Shape Description" (2013). *Electronic Theses and Dissertations, 2004-2019*. 2757.
<https://stars.library.ucf.edu/etd/2757>

MATHEMATICAL AND COMPUTATIONAL METHODS FOR FREEFORM OPTICAL
SHAPE DESCRIPTION

by

ILHAN KAYA
B.S. Bilkent University, 2000
M.S. Boğaziçi University, 2004

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2013

Major Professor: Jannick P. Rolland

©2013 Ilhan Kaya

ABSTRACT

Slow-servo single-point diamond turning as well as advances in computer controlled small lap polishing enable the fabrication of freeform optics, specifically, optical surfaces for imaging applications that are not rotationally symmetric. Freeform optical elements will have a profound importance in the future of optical technology. Orthogonal polynomials added onto conic sections have been extensively used to describe optical surface shapes. The optical testing industry has chosen to represent the departure of a wavefront under test from a reference sphere in terms of orthogonal ϕ -polynomials, specifically Zernike polynomials. Various forms of polynomials for describing freeform optical surfaces may be considered, however, both in optical design and in support of fabrication. More recently, radial basis functions were also investigated for optical shape description. In the application of orthogonal ϕ -polynomials to optical freeform shape description, there are important limitations, such as the number of terms required as well as edge-ringing and ill-conditioning in representing the surface with the accuracy demanded by most stringent optics applications. The first part of this dissertation focuses upon describing freeform optical surfaces with ϕ -polynomials and shows their limitations when including higher orders together with possible remedies. We show that a possible remedy is to use edge-clustered-fitting grids. Provided different grid types, we furthermore compared the efficacy of using different types of ϕ -polynomials, namely Zernike and gradient orthogonal Q-polynomials. In the second part of this thesis, a local, efficient and accurate hybrid method is developed in order to greatly reduce the order of polynomial terms required to achieve higher level of accuracy in freeform shape description that were shown to require thousands of terms including many higher order terms under prior art. This comes at the expense of multiple sub-apertures, and as such

computational methods may leverage parallel processing. This new method combines the assets of both radial basis functions and orthogonal phi-polynomials for freeform shape description and is uniquely applicable across any aperture shape due to its locality and stitching principles. Finally in this thesis, in order to comprehend the possible advantages of parallel computing for optical surface descriptions, the benefits of making an effective use of impressive computational power offered by multi-core platforms for the computation of ϕ -polynomials are investigated. The ϕ -polynomials, specifically Zernike and gradient orthogonal Q-polynomials, are implemented with a set of recurrence based parallel algorithms on Graphics Processing Units (GPUs). The results show that more than an order of magnitude speedup is possible in the computation of ϕ -polynomials over a sequential implementation if the recurrence based parallel algorithms are adopted.

Anneme

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor Professor Jannick P. Rolland. We have been working together over a span of seven years on different subjects related to shape descriptions and most recently in the field of optical surface description for freeform optics. I am thankful for her guidance and dedicated support.

I am also deeply thankful to Professor Hassan Foroosh for chairing my committee and for his interest in my topic. I also thank other members of my committee, Dr. Damla Turgut and Dr. Olusegun Illegbusi for their stimulating questions, and Dr. Kevin P. Thompson for sharing his experience in optical system design.

I would like to thank many people for helping me during my graduate studies, including the UCF community, with a special thank to Amy Perry, Rachel Agerton-Franzetta, Diana Camerino, Ayanna Lopez, Traci Freund, Ronda Hill, Terri Rivera-Pons, and Jessica Cheatwood-Alvarez. I am also thankful to Graduate Studies, the College of Engineering and Computer Science, and the College of Optics and Photonics. I would also like to thank the staff at the University of Rochester, where Professor Rolland relocated in 2009, for facilitating my visit there, with a special thank to Gina Kern who was so dedicated to making me feel welcome. I appreciate the help of friends, Ozan Cakmakci, Panomsak Meemon, Kyesung Lee, and many others during my studies.

I would like to thank Professor Ronald F. DeMara and Professor Kalpathy Sundaram.

I am so much indebted to my family, my mother, my brothers and my grandparents.

TABLE OF CONTENTS

LIST OF FIGURES	x
LIST OF TABLES	xv
LIST OF ACRONYMS (or) ABBREVIATIONS	xvi
CHAPTER ONE: INTRODUCTION.....	1
Motivation.....	4
Research Summary	5
Dissertation Outline	6
CHAPTER TWO: POLYNOMIALS AND RADIAL BASIS FUNCTIONS AS OPTICAL SURFACE SHAPE DESCRIPTORS	10
Slope-orthogonal Q-polynomials for Aspheres	10
Zernike Polynomials	23
Gradient Orthogonal Q-polynomials	29
Radial Basis Functions and QR-based Algorithms.....	40
CHAPTER THREE: EDGE-CLUSTERED RAY GRIDS.....	47
Least-squares Data-fitting to Create a Zernike Polynomial Surface	48
The Test Surfaces.....	51
Hexagonal, Chebyshev, Uniform-random and Edge-clustered Grids	53
Results of Efficacy of Fitting the Test Surfaces with Four Different Sampling Grids.....	56

Conclusion	65
CHAPTER FOUR: COMPARISON OF FREEFORM POLYNOMIALS	67
Ray-grids for Data-site Sampling and Test Cases	68
Numerical Simulations for the Efficacy of Zernike versus Q-polynomials	71
The Effect of Irregular Surface Features Height on φ -polynomials Surface Description Efficacy	75
Conclusion	77
CHAPTER FIVE: HYBRID RADIAL BASIS FUNCTIONS AND LOCAL φ -POLYNOMIALS METHOD	80
Hybrid RBF and Local φ -polynomials Method	81
Hybrid RBFs and Local RBFs with Locally Shape Optimization	85
Numerical Experiments for Hybrid RBF and Local φ -polynomials Method	87
Conclusion	95
CHAPTER SIX: ACCELERATION OF COMPUTATION OF φ -POLYNOMIALS	98
General Purpose Computation on GPUs.....	99
Parallelization of Recurrence Relations of φ -polynomials	102
Numerical Results of SIMT Parallelization of φ -polynomials	106
Conclusion	111
CHAPTER SEVEN: CONCLUSION	112

REFERENCES 118

LIST OF FIGURES

Figure 1 Sample fitted with monomials and Chebyshev Polynomials, adapted from [23].	13
Figure 2 The first seven orthogonal Q^{con} polynomials, adapted from [6].....	17
Figure 3 The first seven slope orthogonal Q^{bfs} polynomials, adapted from [6].....	20
Figure 4 (a) The round off errors present in the Zernike polynomial $Z_{25}^0(u^2)$ in explicit computation; (b) The recurrence relation removes the numerical artifacts, adapted from [17]. ..	26
Figure 5 (a) Numerical ill-conditioning associated with $Z_{22}^4(u^2)$; and recurrence relation correctly computes $Z_{22}^4(u^2)$ (b).....	27
Figure 6 The effect of recurrence relations on the accuracy of the dot product of Zernike polynomials for increasing order, n , adapted from Boyd and Yu [29].	28
Figure 7 Cosine version of the gradient orthogonal Q-polynomials (a) $m=1 n=1$ (b) $m=1 n=3$ (c) $m=5 n=0$ (d) $m=5 n=2$, adapted from [12].	33
Figure 8 Gradient fields of the gradient orthogonal Q-polynomials for the given m, n pairs, adapted from [12].	33
Figure 9 (a) 2D Cross-section for fitting with gradient orthogonal Q-polynomials; (b) The sag departure from the best-fit sphere, adapted from [12].	37
Figure 10 (a) 3D view of the off axis section on the parabola and the best-fit sphere intersecting and POI, (b) the sag departure of the parabola off its best-fit sphere over the off-axis section, adapted from [12].	38
Figure 11 Profile of the residual error for the fit with the gradient orthogonal Q-polynomials of the sag shown in Figure 10(b), adapted from [12].	39

Figure 12 RBF optical surfaces in compact HWD design [32, 33].	41
Figure 13 Forming of an RBF surface with Gaussians, $\epsilon=0.19\text{mm}^{-1}$ over a rectangular aperture 40 mm x 80 mm.	41
Figure 14 (a) Ill conditioning of RBF interpolation for $\epsilon =0.2 \text{ mm}^{-1}$; (b) the range of ϵ over which the RBF is ill conditioned, adapted from [37].	42
Figure 15 (a) RBF-QR successfully removes the ill-conditioning for $\epsilon=0.65\text{mm}^{-1}$; (b) RBR-QR is more stable and yielding more accuracy over a range of ϵ , adapted from [37].	45
Figure 16 Zernike coefficients resulting from the least squares fitting of a conventional asphere with 136 Zernike polynomials using the Born and Wolf ordering of terms.	51
Figure 17 The test surfaces described in Eqs. (3.7) -(3.9) (a) A five-term conventional aspheric mirror; (b) A F/1 parabola with 600, 50, and 30 μm bumps; (c) Franke surface. Note that in this illustration the apertures were normalized to 1 in radius.	52
Figure 18 Fitting grids used to demonstrate efficacy of data sampling: (a) hex grid, (b) Cheby-polar grid (c) uni-random grid (Halton points) (d) e_clust-random grid that clusters points towards the boundary over the unit circle.	55
Figure 19 (a) Rotationally symmetric analytic five-term asphere departure from best sphere based on minimum RMS; (b) rotationally symmetric analytic five-term asphere; (c) RMS error in Zernike polynomials approximant performance relative to the analytic function expressed in meters for an increasing number of Zernike coefficients with hex, uni-random, Cheby-polar, and e-clust random sampling grids for the asphere shown in (b).	58
Figure 20 (b) RMS error in Zernike polynomials approximant performance relative to the analytic function expressed in meters for an increasing number of Zernike coefficients with hex,	

uni-random, Cheby-polar, and e-clust-random sampling grids for the F/1 parabola with 3 bumps, shown in (a).....	59
Figure 21 Comparison of the approximants obtained with two different fitting grids for the F/1 parabola with bumps; Top row: Approximant with uni-hex grid sampling with (a) 25 samples, (b) 204 samples, (c) 1990 samples, (d) 4980 samples; Bottom row: Approximant with e_clust-random sampling with (e) 25 samples, (f) 204 samples, (g) 1990 samples, and (h) 4980 samples.	61
Figure 22 (b) Zernike polynomial fit RMS error expressed in meters as a function of coefficient order with hex, uni-random, Cheby-polar, and e_clust-random fitting grids for the freeform Franke surface (Eq. (3.9)) shown in (a).	62
Figure 23 Three regions on the unit circle.	63
Figure 24 Two types of ray-grids used for ϕ -polynomials fitting with about 900 rays in this figure: (a) Hexagonal uniform (b) Edge clustered.	69
Figure 25 Sag departure from the best fit sphere (bfs) : (a) f_1 -bfs, F/1 parabola with the Gaussian bump away from the edge (b) f_2 -bfs, F/1 parabola with the Gaussian bump near the edge of the aperture.	70
Figure 26 Sag fit residual profiles for f_1 ; the F/1 parabola with a Gaussian bump away from the edge of the aperture with $T=80$; (a) fit residual with hexagonal uniform sampling, (b) fit residual with edge clustered sampling. The gradient-orthogonal Q-polynomial and the Zernike polynomial representations give indistinguishable results, so only one is shown.	72
Figure 27 Sag fit residual profiles for f_2 ; the F/1 parabola with Gaussian bump near the aperture edge with $T=80$; (a) fit residual with hexagonal uniform sampling, (b) fit residual with edge	

clustered sampling. Zernike and gradient-orthogonal Q-polynomials perform similarly, so only one is shown.....	73
Figure 28 Comparing Zernike and gradient-orthogonal Q-polynomials as freeform surface representations. The fidelity is investigated with both edge clustered and hexagonal uniform sampling in the case of fitting analytical functions with these surface descriptions; the evolution of the RMS fit residual vs. the number of coefficients for the test case (a) f_1 , F/1 parabola with the bump away from the edge, (b) f_2 , F/1 parabola with the bump at the edge.....	74
Figure 29 Zernike (solid lines) and gradient-orthogonal Q-polynomials (dash lines) surface approximation performance over a range of heights of the rotationally nonsymmetric bump with hexagonal uniform and edge clustered sampling for the test cases (a) f_1 , (b) f_2	76
Figure 30 Domain decomposition with circular subapertures of radius 1.33 mm over a 4 mm x 4 mm square aperture.....	82
Figure 31 Wendland's CSRBFs for weight assignment, adapted [39].	83
Figure 32 Locally optimized shape parameters for hybrid RBFs.	86
Figure 33 The F/1 Parabola where 12.5 μm – 100 μm bumps may be visualized in Figure 34. ..	88
Figure 34 12.5 μm to 100 μm isotropic and anisotropic bumps on F/1 parabola over an 80 mm x 80 mm square aperture.....	88
Figure 35 Decomposition of the aperture of an F/1 parabola into circular subapertures of radii 800 μm along with uniformly distributed sample points, shown only for a -2 mm to 2 mm subregion.....	89

Figure 36 Approximation error profile for an F/1 parabola with several bumps showing the maximum PV errors on the orders of the subnanometer with only 36 local FRINGE Zernike polynomials across an 80 mm x 80 mm aperture.	90
Figure 37 The trade-off in the subaperture radius (a) and subaperture count (b) vs. the number of basis elements within subapertures.	92
Figure 38 Microsoft’s DirectX 11 standard graphics pipeline adapted [54].	100
Figure 39 an example GPU architecture adapted [55].	101
Figure 40 GPU computed low order ϕ -polynomials (a) Zernike, Z_9^3 , (c) gradient orthogonal Q-poly, Q_3^3 ; the difference between the parallel and sequential implementations within 14 significant digits (b) Zernike (d) Q-poly.....	106
Figure 41 (a) Total execution time of the sequential and parallel algorithms of ϕ -polynomials on both CPU and GPU as a function of the grid size (b) speedups of ϕ -polynomials with grid size.	108
Figure 42. Effect of the polynomial order on the computation of ϕ -polynomials (a) computation times on CPU and GPU (b) speedups through parallelization.....	110

LIST OF TABLES

Table 1 The coefficients for the gradient orthogonal Q-polynomials for fitting the sag shown in Figure 10 (b) [12].	39
Table 2 Zernike Polynomials with significant coefficients for a rotationally symmetric conventional asphere.	52
Table 3 Percentage of the number of points per unit area for the two edge clustering grids	64
Table 4 Subnanometer PV errors with a small set of Zernike polynomials (4 th column) or Gaussian RBFs (5 th column) in each subaperture.	91
Table 5 Showing 10 nm PV errors with a small set of Zernike polynomials.	94
Table 6 Effect of the size of the ray grids on the speedup of the computation of ϕ -polynomials.	107
Table 7 Effect of the order of the ϕ -polynomial over the computation time and speedup.	109

LIST OF ACRONYMS (or) ABBREVIATIONS

CPU	Central Processing Unit
CSRBF	Compactly Supported Radial Basis Function
CUDA	Compute Unified Device Architecture
GPU	Graphics Processing Unit
HWD	Head Worn Display
MRF	Magneto-Rheological Finishing
POI	Point Of Intersection
PV	Peak To Valley
RBF	Radial Basis Function
RMS	Root Mean Square
SIMT	Single Instruction Multiple Thread

CHAPTER ONE: INTRODUCTION

Freeform optical components are going to play key roles in the future of optical systems. The ability of these components by definition to depart from rotational symmetry, for the first time, enables truly folded geometries with excellent overall optical correction. These properties enable optical systems with reduced physical sizes through a reduction in element count, with the added property of being lighter weight. In addition to the gain in compactness and reduction in weight, optical designs leveraging freeform components may also yield performance improvements in terms of a gain in étendue, where étendue may be thought of as the product of the field of view and aperture size of the system at a given focal length. In layman terms, as the étendue increases the efficiency of the system may increase together with the resolution or the ability to image a larger field of view or both. One of the early types of optical systems to take advantage of the new fabrication capabilities that enable freeform surfaces is unobstructed all-mirror systems that are being designed enabling ultra-broadband imaging. Some pioneering examples of freeform optical elements have started to emerge in Head Worn Displays (HWDs) [1], projection systems [2], and infrared imagers [3]. As Rolland and Thompson discuss in a recent Optics and Photonics News article, there is a revolution occurring in the field of optical design that is mainly driven by two concurrent but unrelated major developments requiring the optical design community to develop new methods and tools to describe freeform optical surfaces [4].

A first development, prior to even considering freeform optical components but providing some guidance to their development, resulted from discovering that a power series representation

introduced by Abbe [5] for aspheric optics is failing in part because of the lack of orthonormalization. When new polishing methods, small tool polishing, ion beam polishing, and magneto-rheological finishing (MRF) polishing have come to be adopted throughout the industry, this issue has become apparent because rotationally symmetric aspheres were started to be favored in challenged optical designs to insure least number of elements while meeting high performance specification such as for the compact cell phone high resolution cameras and at the other end of the spectrum lithography lenses. New methods to describe optical aspheres were proposed by Forbes in place of historical power series, the Q^{bfs} and Q^{con} polynomials [6], in order to address this issue. Recent work shows that designing with Q-polynomials together with slope constraints that may easily be constrained in the optimization merit function given the unique description of these polynomials, yield optics that is less sensitive to alignment and assembly, a huge gain for higher yield in optical manufacturing that will lower cost while maintaining or improving quality [7, 8].

The second development is the introduction of the slow-servo axis in the diamond turning based manufacturing of optical elements [9]. This development enables the controlled manufacture of optical freeform surfaces, which are not intrinsically rotationally symmetric. The initial impact of this development is to broaden the definition of optical surfaces from a conic surface plus the power series to basis functions that may describe freeform surfaces. Cakmakci et al. proposed and implemented local shape descriptors, in the form of Gaussian basis functions that appear well suited for local shape description [10, 11]. Recently gradient orthogonal Q-polynomials as an addition to a best fit sphere have been proposed to describe optical freeform surfaces in the context of optical manufacturing [12]. Together with the recurrence relations,

they may also provide an efficient and robust optical surface description capability that may also be leveraged in design.

One approach to specifying a freeform surface is to use a base conic surface plus Zernike polynomials to describe the non-rotationally symmetric components [13]. Zernike polynomials are used as a pervasive means of representing optical surface deformations in optical testing, as they are complete and orthogonal over the unit circle. Moreover, the lower order Zernike terms are readily identified with Seidel and H.H. Hopkins aberrations [14, 15, 16] that are used in optical design. However, important limitations in the optical surface descriptions with full aperture Zernike polynomials exist: It may be the case that higher order Zernike terms are required in order to represent optical surfaces with the accuracies required by most stringent optics applications. High order terms possess numerical problems in implementation because of round-off errors. Recurrence relations are adopted as a remedy for this case as in any other orthogonal polynomials [17]. Even when the problems with the numerical round-off errors are bypassed, it is anticipated that the thousands of terms required to describe a freeform surface with subnanometer accuracy is a bottleneck for the optical designers. A second limitation is severe edge ringing associated with ϕ -polynomial surfaces. Edge clustered fitting grids are proposed to overcome the edge-ringing successfully with ϕ -polynomial surfaces [18]. Although for optical design purposes it is suitable to apply effective edge clustered ray grids, for testing purposes, a clustered edge grid may not be easy to implement.

As opposed to full-aperture ϕ -polynomials as optical freeform surface description, local, multi-centric, additive Radial Basis Functions (RBFs) were recently investigated for describing

optical freeform surfaces [10, 11, 19]. While orthogonal φ -polynomials are defined over only specific geometries, such as a circle, RBFs are more general, conforming to any aperture shape. Although the orthonormalization of φ -polynomials over other specific aperture shapes are possible [20], RBFs constitute one basis set that applies to any aperture shape. Forsaking the orthogonality of φ -polynomials, RBFs offer simplicity and geometric flexibility in terms of aperture shapes. However RBFs have their own drawbacks as well. They may suffer from numerical ill-conditioning when their shape is flat or excessive numbers of them are used to describe freeform surfaces.

Motivation

As the optics manufacturing industry is forging ahead in the advancement of their fabrication methods, the mathematical models to describe optical surfaces are required to be retooled and redefined. The major motivation for this work is how to best efficiently describe general optical surface shapes with different aperture geometries and uncommon features. In other words, with the optics manufacturing industry presenting itself with the ability to fabricate most general freeform optical elements, we pose the question of how to best economically and accurately represent general optical shapes. As an impact of these developments, optical designers need to answer the following questions: Does the sampling of the surface have any effect on the accuracy of the description of the optical surface? Is there a way to describe a freeform surface with minimum number of basis elements, as few as 25 terms of φ -polynomials, while at the same time achieving subnanometer accuracy? Do they need to change the basis for the description of freeform elements if they work on an aperture shape different than a circle? Is

it possible to use highly threaded many-core computational platforms to reduce the φ -polynomials computation time through parallel computing? In this thesis, we investigate the possible answers for the above questions while proposing new methods for the description of optical freeform surfaces. We also report on the merits and limitations of working with different basis for freeform optical surfaces under different sampling patterns in addition to devising and implementing parallel algorithms for φ -polynomials computation.

Research Summary

In the first part of this thesis, we show that the ray grids commonly used in sampling a freeform surface, such as a uniform hexagonal sampling grid, to form a database from which to perform a φ -polynomial fit is limiting the efficacy of computation. We present an edge clustered fitting grid that effectively suppresses edge ringing that arises as the φ -polynomials adapt to the fully nonsymmetric features of the optical surface [18]. Secondly, we show that a substantial number of Zernike (φ -polynomial) terms, sometimes thousands, is required in order to achieve subnanometer accuracy. Prior to arriving to the appropriate number of terms, intermediate results with insufficient number of terms exhibit high departure errors at the edge. The impact of this edge-clustered fitting grid on the reduction of edge-ringing and the improvement of surface representation by several orders of magnitude is also compared with uniform hexagonal subgrids centered on rectangular uniform grid, Chebyshev-based radial grids, and polar grids.

As part of an investigation of fitting grids for optical surface description, full aperture φ -polynomials, specifically Zernike and recently introduced gradient orthogonal Q-polynomials, are also investigated in a comparative manner in terms of efficacy of optical surface description

[21]. Results establish the similarity of φ -polynomials for accurately describing freeform surfaces under stringent conditions (i.e. a high departure surface with high local slopes), which is a critical step in the future application of these tools in advanced optical system design and fabrication.

In the second part of this thesis, we developed an efficient, accurate, and localized hybrid method combining assets of both RBFs and φ -polynomials for freeform shape description, which makes it uniquely applicable across any aperture shape due to its domain decomposition and local stitching properties [19]. Results show that the proposed method yields subnanometer accuracy with as few as 25 terms φ -polynomials in each subaperture. Subnanometer accuracy is required for the stringent conditions of lithography and related precision optics applications. Under less stringent conditions, such as for illumination optics, it is shown that the necessary accuracy is achieved using as few as 16 terms of local φ -polynomials in each local partition.

Finally in this dissertation, we have devised and implemented recurrence based parallel algorithms for φ -polynomials in order to take advantage of parallelism on highly threaded computational platforms i.e. Graphical Processing Units (GPUs). The results show that more than an order of magnitude improvement is achieved in computational time over a sequential implementation if recurrence-based parallel algorithms are adopted in the computation of the φ -polynomials [50].

Dissertation Outline

In the next chapter, we present a review of state-of-the-art methods for the description of optical surfaces. The chapter starts with a description of the orthogonal 1D φ -polynomials for

aspheric optical elements. The recurrence relations for the slope orthogonal Q-polynomials as descriptors of rotationally symmetric aspheres are reviewed in this first section. Chapter 2 continues with the presentation of freeform (i.e. 2D and non-rotationally symmetric) ϕ -polynomials. First, the set of Zernike polynomials as a descriptor for freeform elements is explained, followed by the recently introduced gradient-orthogonal Q-polynomials. The recurrence relations for efficient and accurate computation of these polynomials are given in the same section. In the next section, RBFs are introduced along with their numerical properties and QR-based algorithms for this optical surface description.

Chapter 3 focuses upon efficient ray grids for the description of freeform optical elements. The first section describes the least squares fitting process currently used in the preliminary optical design work with optical freeform surfaces and how low order ϕ -polynomials matches the Seidel wavefront aberrations. In the following section, four different ray grids are described: uniform hexagonal subgrids centered on a uniform rectangular grid, a polar grid with Chebyshev-based radial weighing, a uniform random point grid, and an edge-clustered random point grid. In the last section, numerical experiments with different test cases as the examples of highly varying freeform optical surfaces are given.

In Chapter 4, we have compared two different ϕ -polynomials in terms of least squares in order to understand the freeform description capabilities of these two polynomial sets under two different sampling ray grids, uniform hexagonal subgrids centered on the uniform rectangular grid and an edge-clustered fitting grid. In the first section two different ray-grids and test cases are presented. In the numerical experiments section, two different test cases are described with

increasing number of basis elements in order to reveal any similarities or differences between these two methods in terms of least-squares fitting of freeform optical surfaces. The last section is an inquiry of the effect of the height of the surface features on the number of ϕ -polynomials terms required and the impact on the residual peak-to-valley (PV) fitting errors.

In Chapter 5, a local, hybrid, efficient and accurate optical surface description methodology is proposed and shown to have striking significance in the reduction of the order of ϕ -polynomials terms used for freeform surface description. In the first section, the hybrid RBF and local ϕ -polynomials method is presented along with a description of its algorithm. In the next section a variation of the method with local Gaussian RBFs with local shape optimization instead of local ϕ -polynomials is described. Finally in the numerical experiments section a complex freeform surface is represented with the hybrid RBFs and local ϕ -polynomials method. Results show that the surface may be described with as few as 25 terms in each subaperture for subnanometer accuracies. Also in this section the trade-off between the local number of basis elements in the local ϕ -polynomials fits and the size of the subapertures is shown with a brief comparison to its shape optimized local Gaussian RBF counterpart.

In Chapter 6, recurrences based parallel algorithms, devised and implemented on a highly threaded GPU for the acceleration of computation of ϕ -polynomials, are presented. In the first section, general purpose computational methods with GPUs are described along with the brief review of GPU architectures. In the second section, in addition to the pseudo-codes, some parallel algorithms of ϕ -polynomials constructed upon the recurrence relations are shown with a detailed description. In the final section of this chapter, numerical experiments including the

effect of the ray grid size and ϕ -polynomials order on the computation time are carried out as well as the validation of the parallel algorithms and speedups through the parallelism.

Chapter 7 summarizes the main findings and major contributions of this research effort along with possible future directions for the mathematical and computational methods for freeform optical surface description.

CHAPTER TWO: POLYNOMIALS AND RADIAL BASIS FUNCTIONS AS OPTICAL SURFACE SHAPE DESCRIPTORS

In this chapter, we review the state-of-the-art polynomials and radial basis functions (RBFs) for freeform optical surface description. Historically, power series expansions with a conic section of choice are used to describe optical surfaces, which until recently have been dominantly rotationally symmetric, or portions of rotationally symmetric parts, with some limited use of anamorphic aspheres. Failures in this mathematical description model emerged early in 2000 when commercial optical software unwittingly provided optical designers with more aspheric terms than could be support with 32-bit computing. This occurrence then posed the question of how to best describe optical surfaces with high accuracy and minimal cost. The mathematical propositions for this question are reviewed in the next sections. We start with the recently introduced slope orthogonal Q-polynomials [6] for rotationally symmetric optical surfaces. We then review Zernike polynomials as the well-known and currently emerging basis for ϕ -polynomials freeform optical surface descriptions, which are not rotationally symmetric. In the final section of polynomials applied to the description of optical surfaces for optical design, gradient orthogonal Q-polynomials are described along with their recurrence relations. The last section concludes this chapter with a review of RBFs and their stable evaluation with a QR-based approach.

Slope-orthogonal Q-polynomials for Aspheres

The most widely used and conventional method for characterization of optical surface shape, whether that shape is rotationally symmetric or not, is a power series expansion

introduced by Abbe [5] almost a century ago. This power series representation is made more effective with a base conic section of choice as conic sections have some useful optical properties. Hence an optical surface is most generally represented as follows

$$z(\rho) = \frac{c\rho^2}{1 + \sqrt{1 - (1 + \kappa)c^2\rho^2}} + \sum_{m=0}^M a_m \rho^{2m+4}. \quad (2.1)$$

In above Eq. (2.1), ρ and z represents the standard cylindrical coordinates, κ represents the conic constant of choice, c stands for the curvature of the conic. This representation of an optical surface is completed with the aperture radius, ρ_{max} . For aspheric surfaces, because of the rotational symmetry, the sag, $z(\rho)$ has only one independent variable, ρ . For freeform surfaces however, there is also an angular dependence of the sag function, which is represented as $z(\rho, \theta)$.

Although the expression in Eq. (2.1) yields a complete set for approximating the optical surfaces for the required accuracies provided that m is allowed to be large enough, the monomial basis, i.e. ρ^m , is numerically inefficient and provides the surface approximations through heavy cancellation of the terms, which leads to associated least squares approximation and the Gram matrix to become heavily ill-conditioned. One improvement is to apply normalization of the basis such as to adopt $u = \rho/\rho_{max}$ and second is to remove the degeneracies between the basis elements, which is to orthogonalize the basis.

Conditioning is related to the perturbation behavior of a mathematical problem and stability is related to the perturbation behavior of an algorithm [22]. Generally, a well-conditioned problem is the one where a small perturbation in the data causes only negligible

changes in the solution. An ill-conditioned problem is the one where small changes in the data lead to an unacceptable change in the solution. In terms of numerical linear algebra, conditioning of a problem is measured with a condition number. Trefethen defines the condition number as follows [22]:

“Let \mathbf{A} be a nonsingular matrix, consider $\mathbf{Ax}=\mathbf{b}$, the problem of computing \mathbf{b} , given \mathbf{x} , has condition number, $K \leq \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$, with respect to perturbations of \mathbf{x} . The problem of computing \mathbf{x} , given \mathbf{b} , has the condition number, $K \leq \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$ with respect to perturbations of \mathbf{b} . The problem of computing \mathbf{x} , when \mathbf{b} is fixed, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, has the condition number $K = \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$ with respect to perturbations in \mathbf{A} , where $\|\cdot\|$ represents the norm of a matrix.”

\mathbf{A} is ill-conditioned when the condition number, K , is large, and similarly \mathbf{A} is well-conditioned when the condition number is small. It is always expected to lose $\log_{10} K$ digits in the solution of a least square system if the least square matrix is ill-conditioned [22]. An orthogonal basis, since all the basis elements are orthogonal to each other and the associated dot product is zero, contains no degenerate or near-degenerate basis elements, which leads to well-conditioned approximation matrices. The trade-off between the ill-conditioning of a matrix and the accuracy of the solution of the least squares system is best captured through an example. As such, in the following example, the orthogonal Chebyshev polynomials are compared with the monomials (power series) in terms of the least square approximation of a smooth surface [23].

In this section, we present a summary of the example given in [23] to further clarify the differences between the monomials and an orthogonal basis. Let's consider that example [23]: a

1 μm bump over a 60 mm aperture that is fitted with 9 monomials and 9 orthogonal Chebyshev polynomials. The shape of the function is given in Figure 1. The function to be fitted is an exponential of form, $g(\rho) = e^{-\rho^2/81}$. The monomials used for the fit are $\{1, \rho^2, \rho^4, \dots, \rho^{16}\}$. The Chebyshev polynomials are defined to be

$$T_m(\rho) = \cos[m \arccos(\rho)], \quad (2.2)$$

where m is even.

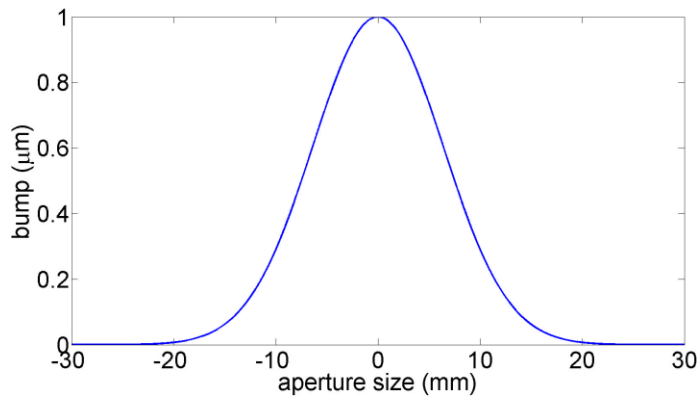


Figure 1 Sample fitted with monomials and Chebyshev Polynomials, adapted from [23].

When we carried out the least squares approximation, coefficients for the monomials are found out to be $\{999.5, -11024.9, 59072.8, -196235.6, 427931.4, -608789.4, 540151.6, -269848.7, 57744.7\}$ nm.

The condition number for this least squares approximation matrix is $5.4711e+5$. So, we expect to lose 6 digits of accuracy because of the ill-conditioning associated with this monomial basis, i.e. $\log(5.4711e5)$ is about 5.74. We see that there is heavy cancellation between the fit coefficients. Even if the fit is required to be within a 1 nm tolerance and the test surface is 1 micron in height, the coefficients are thousands of microns. More importantly, since the approximation matrix has

a non-empty null space, there are infinitely many solutions to this problem. Furthermore, changing the coefficients with a scaled version of the nullspace vectors constitutes more solutions for this problem and does not effectively change the result. For example Forbes [23] mentioned that 539995 could be replaced with 539995-212992, and the fit is still within 1 nm tolerance.

Instead, when doing the least squares fit with orthogonal Chebyshev basis, the fit coefficients are given as {173.6, -314.0, 234.1, -145.5, 76.9, -34.8, 14.0, -4.7, 1.8} nm [23]. The condition number for this orthogonal Chebyshev least squares matrix is just 4. Thus all of the digits in this list are significant, i.e. $\log(4)$ is 0.6, and these coefficients cannot be changed without changing the result of the fit. Since the condition number is very small compared to that of the monomials, this matrix is well-conditioned. This representation of the fit is also more efficient [23]. Furthermore, the coefficients do not change if we include one more basis element to the approximation matrix in the next approximation. If we truncate the number of basis elements at some point, such as 7, then we would expect to have a fit error about the shape and size of the 8th basis element, since this spectrum of coefficients decreases in magnitude. For example, we will expect to lose about 2 nm of accuracy if we truncate the last basis element from the approximation list, since all the digits in this list count. The null space of this approximation matrix is empty (a well-conditioned problem), which means there is a single solution to this approximation problem. The fit coefficients do not change no matter how many basis elements are used in the fit. For example, we carried out the fit with 15 Chebyshev polynomials, and the coefficient list is

{173.6, -314.0, 234.1, -145.5, 76.8, -34.9, 14.0, -4.9, 1.6, -0.5, 0.1, -0.03, 0.006, -0.001, 0.0003}nm.

By examining the above list, we expect to have a subnanometer tolerance in the fit if we had just used 11 basis elements, since the fit coefficient of the 12th basis element is just -0.03 nm.

Thus, although the monomial basis is practically useless after a few terms, for example 6, on the other hand, we can use an orthogonal basis such as Chebyshev basis to arbitrary accuracies significantly set by machine precision. Another useful interpretation of fitting with an orthogonal basis is that the sums of squares of these fit coefficients result in the mean square sag at that point.

After observing that monomial basis totally fail due to ill-conditioning of the associated Gram matrix, and considering the requirements of the optical interferometry testing, Forbes proposed two sets of orthogonal polynomials in [6]. In the following, we will summarize Forbes article [6] for Q-polynomials, namely Q^{con} and Q^{bfs} polynomials.

Instead of using the monomials that are given in Eq. (2.1), we could have replaced the monomials with a set of orthogonal polynomials, Q^{con} 's. Then, a surface sag can be represented with a conic base plus the departure from the conic, such as given as Eq. (2.3), also in [6]. In this way, ill-conditioning of the Gram matrix is removed since the orthogonality will not allow it to be ill-conditioned.

$$z(\rho) = \frac{c\rho^2}{1 + \sqrt{1 - (1 + \kappa)c^2\rho^2}} + D_{con} \left(\frac{\rho}{\rho_{max}} \right). \quad (2.3)$$

In Eq. (2.3), the first part of the equation represents the base conic, and $D_{con}(u)$ represents the sag departure from the conic. All other variables are as the same as Eq. (2.1). The departure from the conic is represented in [6] as

$$D_{con}(u) = u^4 \sum_{m=0}^M a_m Q_m^{con}(u^2). \quad (2.4)$$

Q^{con} polynomials are related to the Jacobi polynomials such that the associated Gram matrix G is diagonal. Under a unity weight function, the dot product between two basis elements forms the contents of the Gram matrix shown in [6] as follows

$$G_{nm} = \left\langle u^8 Q_m^{con}(u^2) Q_n^{con}(u^2) \right\rangle = 2 \int_0^1 Q_m^{con}(x) Q_n^{con}(x) x^4 dx, \quad (2.5)$$

where angle brackets denote a weighted average, and the dot product under the unit weight reduces the integral form given in Eq. (2.5). Since these Q^{con} polynomials are orthogonal, the associated Gram matrix is diagonal. The relationship between the orthogonal Q^{con} polynomials and Jacobi polynomials, P , is given in [6] as follows

$$Q_m^{con}(x) = P_m^{(0,4)}(2x-1). \quad (2.6)$$

A few initial polynomials are $\{1, 6x-5, 28x^2-42x+15 \dots\}$. In Figure 2, we show the first 7 polynomials from this list.

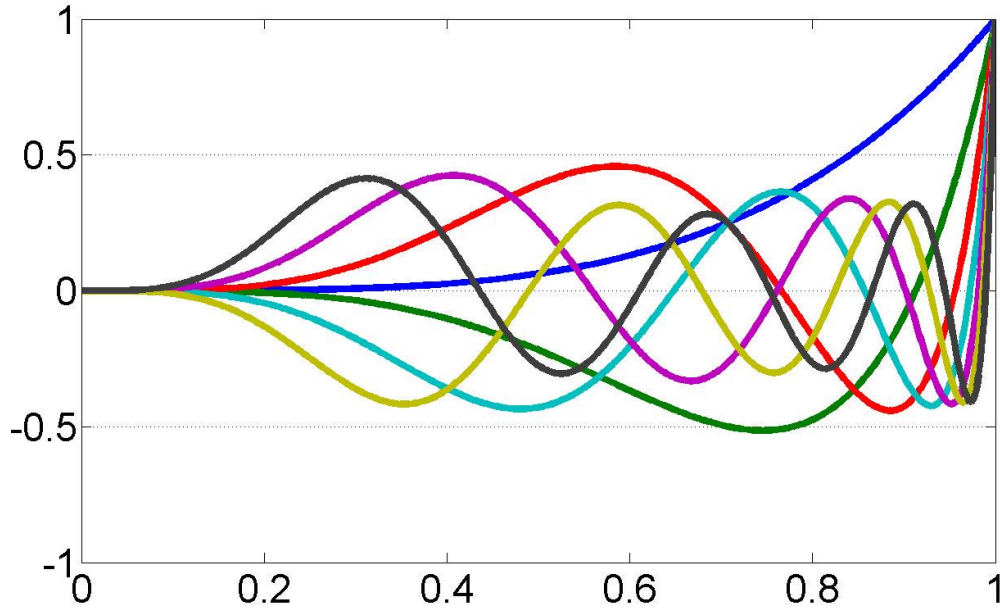


Figure 2 The first seven orthogonal Q^{con} polynomials, adapted from [6].

Similar to the generation of Q^{con} orthogonal polynomials, Forbes derived orthogonal Q^{bfs} polynomials. Two main significant differences between Q^{bfs} and Q^{con} are the use of a best-fit sphere as the base surface for Q^{bfs} (as opposed to a conic for Q^{con}) and the orthogonalization in slope for Q^{bfs} as opposed to sag for Q^{con} , motivated thereafter. First, aspheric surfaces are most cost effective when their deviation from a best fit sphere is restrained to meet the needs of metrology and fabrication, thus the choice of a sphere for the base surface. Moreover significantly, limiting the absolute maximum slope of the departure leads to enhancements in manufacturability of aspheres as it extends the slope range over which metrology can be successfully performed and reduces the sensitivity to alignment. Thus a representation such as Q^{bfs} , where the square root of the sum of the coefficients squared represents the Root Mean Square (RMS) slope error is most convenient as this sum may be computed on the fly during

optimization of a surface in lens design. As such, the maximum slope can also be simultaneously constrained as the RMS and max slope errors are intimately related.

Most fabrication shops use for the definition of the best-fit sphere the one that touches the surface at its axial point and around its perimeter. The best-fit sphere curvature is effectively calculated in [6] as

$$c_{bfs} = \frac{2f(\rho_{max})}{(\rho_{max}^2 + f(\rho_{max})^2)}, \quad (2.7)$$

where $f(\rho_{max})$ is the sag at the perimeter, and ρ_{max} is the aperture radius. The sag can then be written [6] as

$$z(\rho) = \frac{c_{bfs}\rho^2}{(1 + \sqrt{1 - c_{bfs}^2\rho^2})} + D\left(\frac{\rho}{\rho_{max}}\right), \quad (2.8)$$

where the departure from the best fit sphere is defined in [6] as

$$D(u) = \frac{u^2(1-u^2)}{\sqrt{1 - c_{bfs}^2\rho^2}} \sum_{m=0}^M a_m Q_m^{bfs}(u^2). \quad (2.9)$$

In Eq. (2.9), u is the normalized radial coordinate. Note that by having the term $u^2(1-u^2)$ appear in the numerator, the departure from the best fit sphere is as required zero at the edge and its axial point, the denominator is the cosine of the angle between the normal of the best-fit sphere and the optical axis. In order to construct the RMS slope of the departure along the normal from

the sum of squares of the coefficients, a_m , the slope functions, $S_m(u)$ must be orthogonal. The slope functions are defined in [6] to be

$$S_m(u) = \frac{d}{dx} \left\{ u^2 (1-u^2) Q_m^{bfs}(u^2) \right\}. \quad (2.10)$$

A dot product with a weighted function is defined such that the orthogonal polynomials do not grow unboundedly towards the ends of the interval,

$$\frac{\int_0^1 S_m(u) S_n(u) w(u) u du}{\int_0^1 w(u) u du}, \quad (2.11)$$

where S_n and S_m are orthogonal slope functions and $w(u)$ is the weight function, $(u^2(1-u^2))^{-0.5}$ [6]. With this dot product, the first function can be taken to be a constant and normalized. Then the new members of the orthogonal Q^{bfs} polynomials can be made orthogonal to all the previously computed Q^{bfs} polynomials. An appropriate procedure for this orthogonalization is to use a modified Gram-Schmidt algorithm. The first few of the polynomials are

$$\left\{ 1, \frac{1}{\sqrt{19}}(13-16x), \sqrt{\frac{2}{95}} [29-4x(25-19x)], \dots \right\}. \quad (2.12)$$

In Figure 3, we have shown seven of the slope orthogonal Q^{bfs} polynomials. The advantage of using this set of orthogonal polynomials as compared to that of monomials are described with examples in [6, 23, 24]. As an application example, Ma et al. recently showed that the design of a 28 element lithographic lens and an optimization integrated RMS slope constraint resulted in

an order of magnitude decrease in overall sensitivity to tilts and decenters with Q-polynomials [7]. Ma et al. also reported similar findings in the investigation of a high-resolution cell phone camera [8].

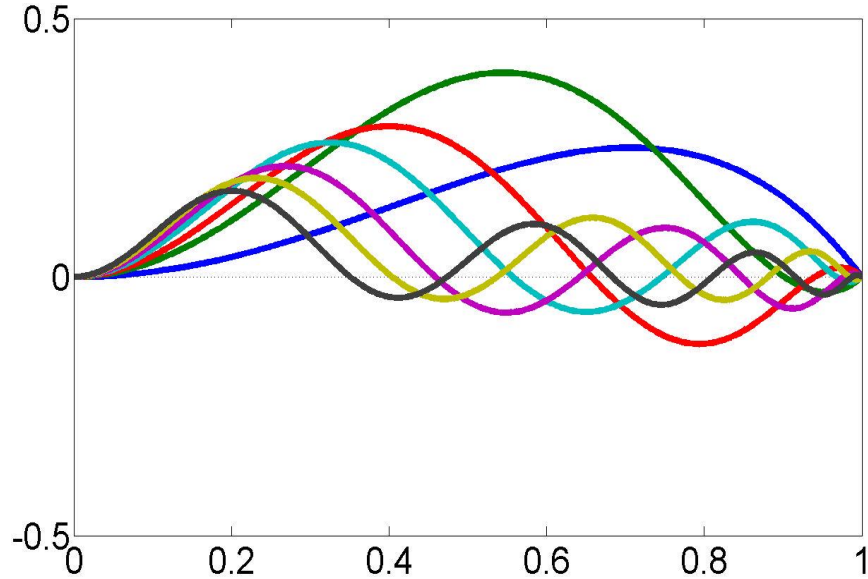


Figure 3 The first seven slope orthogonal Q^{bfs} polynomials, adapted from [6].

In order to efficiently calculate the Q-polynomials, Forbes used recurrence relations. Often used with orthogonal polynomials, recurrence relations provide simplicity and stability for the numerical calculations that would otherwise be affected by numerical cancellation and round-off errors leading to an ill-conditioned system of equations. For a Q^{con} , a standard 3-term recurrence relation, defined in [17], is given as

$$Q_{m+1}^{con}(u^2) = (rv_{1m} + rv_{2m}u^2)Q_m^{con}(u^2) - rv_{3m}Q_{m-1}^{con}(u^2). \quad (2.13)$$

In Eq. (2.13), u is the normalized radial coordinate, $u = \rho/\rho_{\max}$, rv_{1m} , rv_{2m} , and rv_{3m} are recurrence variables defined below in Eqs. (2.14)-(2.16), and m starts at 1. The recurrence relation is initialized with the first two polynomials, which are $Q_0^{con}(u^2)$ and $Q_1^{con}(u^2)$, 1 and $6u^2 - 5$ respectively. After initialization, any Q^{con} polynomial of order m can be computed with the recurrence relation whose variables are defined in [17] as

$$rv_{1m} = -\frac{(2m+5)(m^2+5m+10)}{(m+1)(m+2)(m+5)}, \quad (2.14)$$

$$rv_{2m} = \frac{2(m+3)(2m+5)}{(m+1)(m+5)}, \quad (2.15)$$

$$rv_{3m} = \frac{(m+3)(m+4)m}{(m+1)(m+2)(m+5)}. \quad (2.16)$$

For the Q^{bfs} polynomials, however, there is no standard 3-term recurrence relation. Instead they satisfy an unconventional 3-term recurrence relation with a set of auxiliary polynomials, $P_m(u^2)$, [24]. The unconventional 3-term recurrence relation for Q^{bfs} polynomials is defined in [24] as

$$Q_{m+1}^{bfs}(u^2) = \frac{[P_{m+1}(u^2) - g_m Q_m^{bfs}(u^2) - h_{m-1} Q_{m-1}^{bfs}(u^2)]}{f_{m+1}}. \quad (2.17)$$

The auxiliary polynomials, $P_m(u^2)$, are a special form of Jacobi polynomials which satisfy a conventional 3-term recurrence relation given in [24] as

$$P_{m+1}(u^2) = (2 - 4u^2)P_m(u^2) - P_{m-1}(u^2). \quad (2.18)$$

The recurrence relation in Eq. (2.18) is initialized with first two auxiliary polynomials, $P_0(u^2)$ and $P_1(u^2)$, which are 2 and $6-8u^2$, respectively. After initialization, any $P_m(u^2)$ of order m can be computed.

The unconventional 3-term recurrence relation given in Eq. (2.17), contains recurrence variables g_m , h_{m-1} , and f_{m+1} . These variables can be found for each iteration of the recurrence relation progressively starting with $m=2$, $f_0=2$, $f_1=19^{0.5}/2$, and $g_0=-0.5$ and using the recursions given in [24] as

$$h_{m-2} = -\frac{m(m-1)}{2f_{m-2}}, \quad (2.19)$$

$$g_{m-1} = -\frac{(1 + g_{m-2}h_{m-2})}{f_{m-1}}, \quad (2.20)$$

$$f_m = \sqrt{(m(m+1) + 3 - g_{m-1}^2 - h_{m-2}^2)}. \quad (2.21)$$

Once the variables and auxiliary polynomials defined above are computed, they can be iterated through the unconventional 3-term recurrence relation defined in Eq. (2.17) by first initializing the recurrence with the first two polynomials $Q_0^{bfs}(u^2)$ and $Q_1^{bfs}(u^2)$, which are 1 and $19^{-0.5}(13-16u^2)$, respectively. All of the Q^{bfs} and Q^{con} polynomials illustrated in Figure 2 and Figure 3 are computed with the recurrence relations shown in this section.

Zernike Polynomials

Zernike polynomials are orthogonal polynomials over the unit circle. Since their introduction by F. Zernike while developing the theory of phase-contrast microscopy in the 1930s [13], Zernike polynomials have emerged as a pervasive means of describing as-fabricated optical surface deformations. More recently, Zernike polynomials have further emerged to illustrate the field dependence of the polynomial coefficients in rotationally symmetric optical systems [25]. In optical design and manufacturing, Zernike polynomial representations of surface departure, placed as an added layer on top of a conic surface, form an enabling fundamental basis as they are complete and orthogonal over the unit circle and, in addition, the lower-order terms are readily identified with the Seidel aberrations. Moreover, H.H. Hopkins wavefront aberration function may also be described in terms of Zernike polynomials [15]. The forms of the lower order Zernike polynomials and the associated optical wavefront aberrations are shown in detail in [26]. The Zernike polynomials provide a mapping between an optical surface under consideration and wavefront aberrations, central to optical system design. Overall, Zernike polynomials are one of the major tools in optical applications ranging from modeling optical surfaces to representing wavefront test data and defining residual error profiles.

The Zernike polynomials are defined in standard form in Born and Wolf [27] as follows

$$Z_n^m(\rho, \theta) = R_n^{\pm m}(\rho) \begin{cases} \cos m\theta \\ \sin m\theta \end{cases}, \quad (2.22)$$

where $n \geq m$ and $m-n$ is even. Instead of the radial variable, ρ , a normalized variable $u = \rho/\rho_{\max}$ may be adopted. This representation shows that Zernike polynomials are composed of Fourier series in angular direction. The radial polynomial in explicit form is given in [27] as

$$R_n^{\pm m}(\rho) = \sum_{q=0}^{\frac{n-m}{2}} (-1)^q \frac{(n-q)!}{q! \left(\frac{n+m}{2} - q\right)! \left(\frac{n-m}{2} - q\right)!} \rho^{n-2q}. \quad (2.23)$$

The radial polynomial shown above in Eq. (2.23) comprises even powers of the radial variable scaled with factorial coefficients. Radial polynomial is of power n , which contains no powers of ρ less than m . Forbes in [17] presented another useful representation of the radial polynomial, which is first given in [28]

$$R_n^m(\rho) = \rho^m Z_{nf}^m(\rho^2), \quad (2.24)$$

where Z_{nf}^m is an orthogonal polynomial, which is of power $nf = (n-m)/2$. The Zernike polynomials are strongly related to orthogonal Jacobi polynomials to the extent that the radial polynomial is sometimes referred as one-sided Jacobi polynomial. Authors in [17, 28] depicted this relationship as

$$Z_{nf}^m(\rho^2) = P_{nf}^{(0,m)}(2\rho^2 - 1), \quad (2.25)$$

where $P_{nf}^{(0,m)}$ is the Jacobi polynomial. This form of the radial polynomial is more concise compared to the explicit form given in Eq. (2.23).

In Boyd and Yu [29], where seven spectral methods were compared for approximations of surfaces, each method's virtues and drawbacks were listed; Zernike basis is listed as one of the best spectral methods due to its spectral convergence and fewer number of basis elements for the same accuracy as compared to that of the Chebyshev-Fourier basis. Although the Zernike polynomials are one of the best tools for representing wavefront data and optical surfaces, which may both be rotational symmetric or not, high-order terms become necessary for their representation. A representation based upon the explicit form of the Zernike polynomials given above in Eq.(2.23), especially for the higher-order terms, suffers from the round-off errors produced by numerical cancelation. This most often leads to ill-conditioned system of equations for the least squares procedures for surface approximations. Author of [17] summarizes this situation as “*It has not been generally appreciated that, in practice, this is a road to grief.*”

Thanks to the relationship with the Jacobi polynomials given in Eq.(2.25), Zernike polynomials satisfy a conventional 3-term recurrence relation. The standard 3-term recurrence relation for Zernike polynomials is given in [17] as

$$Z_{nf+1}^m(u^2) = (rv_{1nf} + rv_{2nf}u^2)Z_{nf}^m(u^2) - rv_{3nf}Z_{nf-1}^m(u^2), \quad (2.26)$$

where the u represents the normalized radial coordinate as before, and rv_{1nf} , rv_{2nf} , and rv_{3nf} are the recurrence variables. For each recurrence relation iteration, the recurrence variables need to be computed. They are defined in [17] as

$$rv_{1nf} = -\frac{(s+1)\left[(s-nf)^2 + nf^2 + s\right]}{(nf+1)(s-nf+1)s}, \quad (2.27)$$

$$rv_{2nf} = \frac{(s+2)(s+1)}{(nf+1)(s-nf+1)}, \quad (2.28)$$

$$rv_{3nf} = \frac{(s+2)(s-nf)nf}{(nf+1)(s-nf+1)s}, \quad (2.29)$$

with $s=m+2nf$. For each azimuthal order m , this recurrence relation is initialized with $Z_0^m(u^2)$ and $Z_1^m(u^2)$, which are 1 and $[(m+2)u^2-(m+1)]$, respectively. The recurrence relation then can be iterated for any order of Zernike polynomials. Forbes states that the recurrence relations not only remove the round-off errors in the computation of the polynomials in explicit form, thus the ill-conditioning of the least squares and Gram matrix, but also they provide computational advantages by reducing the computational cost from a $O(M^2)$ process to a $O(M)$ process [17]. In Figure 4, a high-order Zernike term $Z_{25}^0(u^2)$ is shown with its associated round-off errors if the explicit form is followed, and the remedy for round-off errors, the recurrence relation.

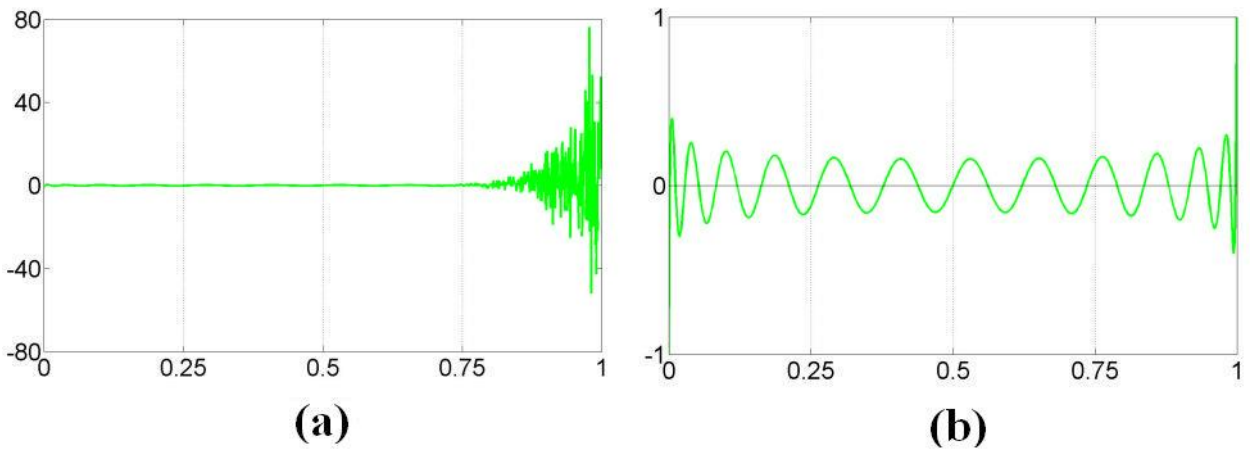


Figure 4 (a) The round off errors present in the Zernike polynomial $Z_{25}^0(u^2)$ in explicit computation; (b) The recurrence relation removes the numerical artifacts, adapted from [17].

By looking at Figure 4 above, we can clearly see that large numerical cancellations lead to the round off errors making the polynomial towards the edge unstable, off almost two orders of magnitude, and of chaotic sign. In Figure 4, we also observe that the recurrence relation computes the polynomial with the exact magnitude and correct oscillations.

To make matters more explicit, we present another example in two-dimensional form shown in Figure 5. In Figure 5(a) a high-order Zernike with its round-off errors produced by numerical cancellations is shown and the accuracy in the computation is evidently off by a full order of magnitude. Fine scale details are not observed if the explicit form of the polynomial is used in the computation. However when the recurrence relation defined in Eq. (2.26) is used, the polynomial peaks at one at the edge of the normalized aperture and clear sine-like details are present in the computation of the polynomial.

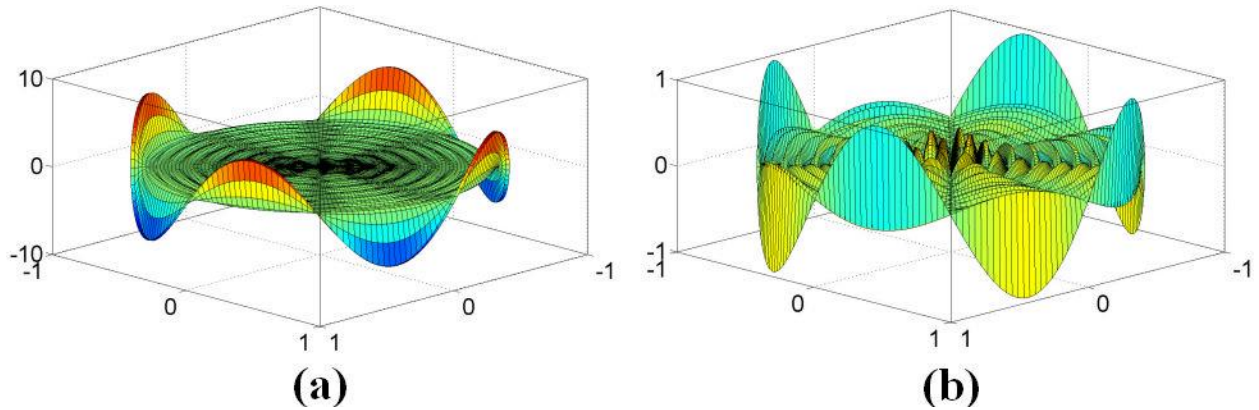


Figure 5 (a) Numerical ill-conditioning associated with $Z_{22}^4(u^2)$; and recurrence relation correctly computes $Z_{22}^4(u^2)$ (b).

The ill-conditioning associated with the explicit form of Zernike polynomials are also investigated by Boyd and Yu [29]. They have compared the dot product of the radial polynomial with itself for both the explicit power series representation and a 3-term recurrence relation. We have shown Figure 3 of their paper [29] for illustration of the ill-conditioning of the Zernike polynomials in explicit form in Figure 6.

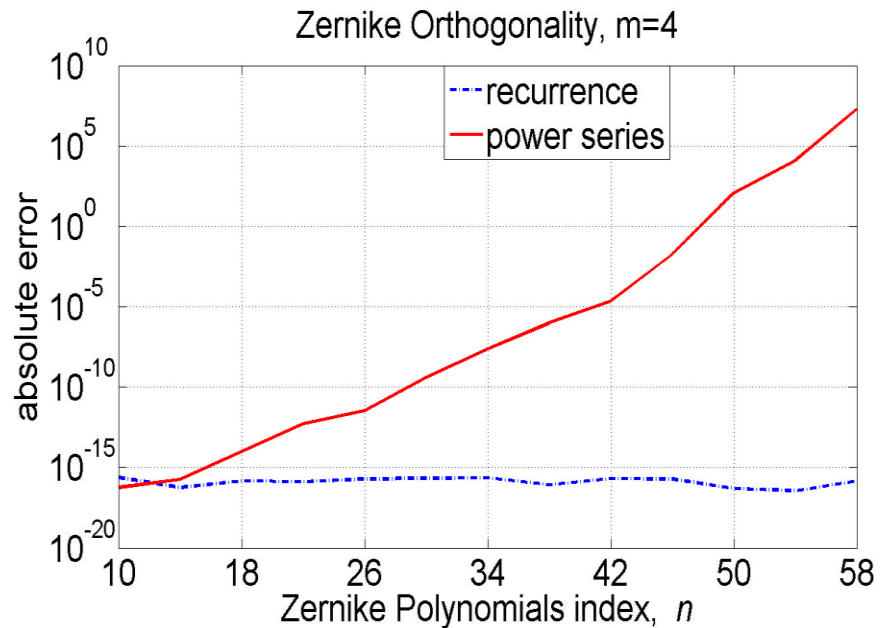


Figure 6 The effect of recurrence relations on the accuracy of the dot product of Zernike polynomials for increasing order, n , adapted from Boyd and Yu [29].

In Figure 6, authors presented the errors in evaluating the dot product of a radial component of the Zernike polynomial with itself for the increasing powers of n , while keeping the azimuthal variable, $m=4$ for both explicit and recursive evaluations. The accuracy of the dot product is lost as the degree of the polynomial is increased with the explicit power series computation, which is highly ill-conditioned and unstable for large n . On the other hand, the recurrence relation

provides a stable computation and preserves its accuracy even for the larger degrees of the polynomials. Concisely, the orthogonality of Zernike polynomials is maintained with the recurrence relation even for the higher degrees because of the stability of the recursion.

Similar to the slope orthogonal polynomials, an optical surface characterization based upon Zernike polynomials with the help of a best-fit sphere is represented as

$$z(\rho, \theta) = \frac{c_{bfs}\rho^2}{\left(1 + \sqrt{1 - c_{bfs}\rho^2}\right)} + \sum_{m=0}^M u^m \sum_{n=0}^N \left[a_n^m \cos(m\theta) + b_n^m \sin(m\theta) \right] Z_n^m(u^2), \quad (2.30)$$

where $z(\rho, \theta)$ represents the sag of the surface, as the surface is not necessarily symmetric, c_{bfs} represents the curvature of the best fit sphere, $u^m Z_n^m(u^2)$ represents the standard Born and Wolf Zernike polynomials of order n [27], and u is the normalized radial coordinate.

Gradient Orthogonal Q-polynomials

Recently a new set of orthogonal polynomials over a circular aperture has been developed by Forbes, orthogonalized with respect to the mean square gradient over an enclosing circular aperture with the goal of facilitating measures of manufacturability, e.g. optical testing, pad polishing [12]. These polynomials will be referred to in this text as gradient-orthogonal Q-polynomials following from the Q-polynomial form developed earlier for rotationally symmetric aspheric surfaces. Since the common method to express an optical surface is to define the departure of the surface from its best fitting conic with an orthogonal set of polynomials, Forbes decided to conform to this methodology in the definition of gradient orthogonal Q-polynomials in order to facilitate estimates for manufacturability of these surfaces and to

integrate with optical design environments. Orthogonal polynomials have the advantage of expressing an optical surface as a spectrum of coefficients in decreasing order, which helps in interpreting the frequency content of an optical surface. In terms of optical manufacturing and testing of an optical surface, the shapes closer to a sphere are easier to produce. Thus the rate of change of departure of a surface along the local normal from its best fitting sphere must be specified and considered because the local principal curvatures are related to the derivatives of the departure.

Similar to slope orthogonal Q^{bfs} polynomials, a two-dimensional freeform optical surface with gradient orthogonal Q-polynomials is represented in [12] as follows

$$z(\rho, \theta) = \frac{c_{bfs}\rho^2}{\left(1 + \sqrt{1 - c_{bfs}\rho^2}\right)} + D\left(\frac{\rho}{\rho_{max}}, \theta\right), \quad (2.31)$$

where the departure from the best fit sphere is specified in [12] as

$$D(u, \theta) = \frac{1}{\sqrt{1 - c_{bfs}^2\rho^2}} \left\{ \begin{array}{l} u^2(1 - u^2) \sum_{n=0}^N a_n Q_n^{bfs}(u^2) \\ + \sum_{m=1}^M u^m \sum_{n=1}^N [b_n^m \cos(m\theta) + c_n^m \sin(m\theta)] Q_n^m(u^2) \end{array} \right\}. \quad (2.32)$$

In Eq. (2.32), u represents the normalized radial coordinate as before, ρ_{max} is the radius of the enclosing circular aperture, $Q_n^{bfs}(u^2)$ represents the slope orthogonal polynomials, $Q_n^m(u^2)$ represents the gradient orthogonal Q-polynomials, c_{bfs} represents the curvature of the best-fit sphere. The entity within braces corresponds to the departure of the optical surface from its best-fit sphere along the local normals of that sphere. The first line on top in Eq. (2.32) accounts for

the rotationally symmetric Q^{bfs} polynomials contributions to the departure along the normal, whereas the nonsymmetric contributions are defined with the gradient orthogonal Q-polynomials in the second line of the Eq. (2.32). The departure along the local normals of the best fitting sphere is converted to a sag deviation along the principal axis of interest by dividing it with the cosine of the angle between the principal axis and the local normal of the best-fit sphere, which is the square root in the denominator in Eq. (2.32). For a surface description, the truncation of the sums of the polynomials in Eq. (2.32) is carried out by selecting a truncation point, T , which constrains the highest degree of the polynomials, $n+2m$.

In order to construct the gradient orthogonal Q-polynomials, Forbes made use of the fact that the mean square gradient of the normal departure from the best-fit sphere is given by the sum of the squares of the coefficients of the surface description in Eq. (2.32) [12],

$$\left\langle |\nabla D(u, \theta)|^2 \right\rangle = \left\langle \left(\frac{\partial D}{\partial u} \right)^2 + \frac{1}{u^2} \left(\frac{\partial D}{\partial \theta} \right)^2 \right\rangle = \sum_{n,m} \left[(a_n)^2 + (b_n^m)^2 + (c_n^m)^2 \right], \quad (2.33)$$

where angle brackets define the mean of the entity over the aperture. The average of a function over the aperture is usually found by taking a double integral of the function with an appropriate weight. Forbes used the following function in [12] for defining the weights,

$$w(u) = \frac{1}{u\sqrt{1-u^2}}. \quad (2.34)$$

Then the average over the aperture of a function, $g(u, \theta)$ is given in [12] by

$$\langle g(u, \theta) \rangle = \frac{1}{\pi^2} \int_0^1 \int_{-\pi}^{\pi} g(u, \theta) (1-u^2)^{-1/2} d\theta du. \quad (2.35)$$

Since the trigonometric modes for different azimuthal orders m and m' are orthogonal by definition, the radial parts of the polynomials need to be orthogonalized with Gram-Schmidt orthogonalization, with the constraint given in Eq. (2.33). Some of the orthogonal polynomials are given in [12]. We present here a couple of the first gradient orthogonal Q-polynomials for each azimuthal order $m=1, 2, 3$, and $n=0, 1$ as below:

$$\begin{aligned} Q_n^1(u^2) &= \left\{ 1, \frac{4}{\sqrt{14}}(1-u^2) \dots \right\} \\ Q_n^2(u^2) &= \left\{ \frac{1}{\sqrt{2}}, \frac{(9-8u^2)}{\sqrt{38}} \dots \right\} \\ Q_n^3(u^2) &= \left\{ \frac{4}{3\sqrt{6}}, \frac{48(10-9u^2)}{9\sqrt{1110}} \dots \right\} \end{aligned} \quad (2.36)$$

In Figure 7, we have illustrated the cosine version of the gradient orthogonal Q-polynomials for different m and n values. We have plotted two polynomials from the sequence for each azimuthal order m . It is important to note that these polynomials are generated with the constraint that their gradients fields are orthogonal to each other. Their gradient fields are shown in Figure 8. The gradient is a vector. So the gradient for different points in the aperture forms a vector field that is shown in Figure 8. For each azimuthal order these gradients are orthogonal to each other and for Figure 8, it can be verified that the dot products of the gradients shown in each row are zero. Also when we examine the gradient orthogonal Q-polynomials shown in Figure 7 and their respective gradients in Figure 8, we can observe that when there are steep slopes in the Q-polynomial, the gradient field has a peak, and when there are flat regions, gradients are zero.

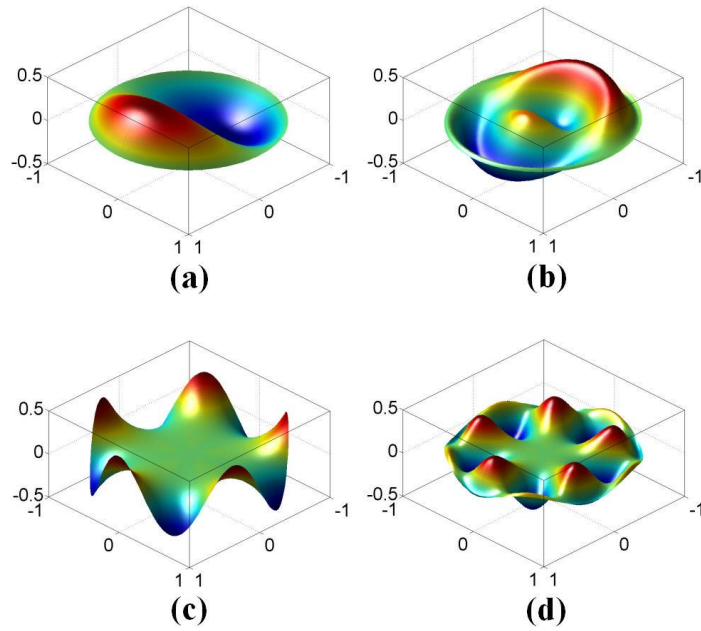


Figure 7 Cosine version of the gradient orthogonal Q-polynomials (a) $m=1$ $n=1$ (b) $m=1$ $n=3$ (c) $m=5$ $n=0$ (d) $m=5$ $n=2$, adapted from [12].

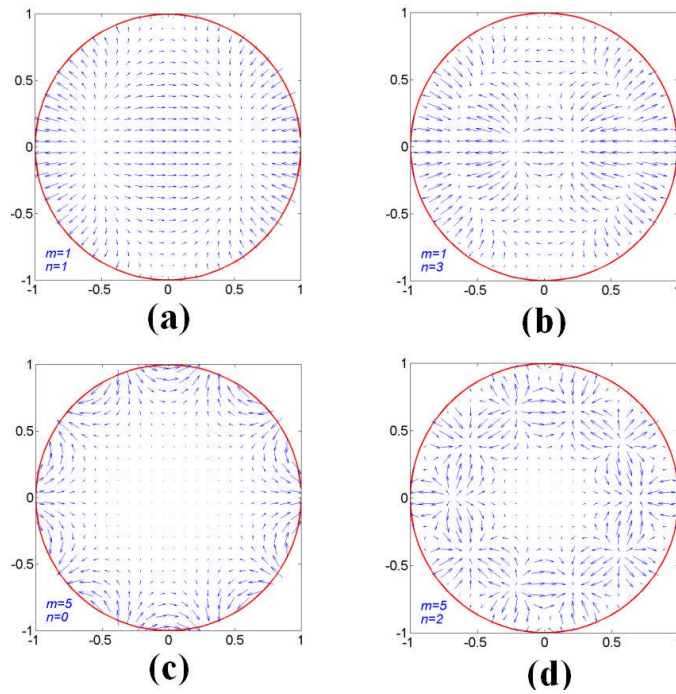


Figure 8 Gradient fields of the gradient orthogonal Q-polynomials for the given m, n pairs, adapted from [12].

In the generation of Figures 7 and 8, we have made use of the recurrence relations defined for gradient orthogonal Q-polynomials in [12]. Similarly to the slope orthogonal Q^{bfs} polynomials, gradient orthogonal Q-polynomials satisfy an unconventional 3-term recurrence relation with the help of a set of auxiliary orthogonal polynomials.

For each azimuthal order m , the auxiliary polynomials $P_n^m(u^2)$ satisfy a standard 3-term recurrence relation given in [12] as

$$P_{n+1}^m(u^2) = [A_n^m + B_n^m u^2] P_n^m(u^2) - C_n^m P_{n-1}^m, \quad (2.37)$$

where the recurrence variables are defined in [12] as

$$A_n^m = \frac{(2n-1)(m+2n-2)[4n(m+n-2) + (m-3)(2m-1)]}{D_n^m}, \quad (2.38)$$

$$B_n^m = -\frac{2(2n-1)(m+2n-3)(m+2n-2)(m+2n-1)}{D_n^m}, \quad (2.39)$$

$$C_n^m = \frac{n(2n-3)(m+2n-1)(2m+2n-3)}{D_n^m}, \quad (2.40)$$

$$D_n^m = (4n^2 - 1)(m+n-2)(m+2n-3). \quad (2.41)$$

The recurrence relation shown in Eq. (2.37) is initialized with $P_0^m(u^2) = 1/2$ and $n=1$, and the first polynomial in the set, $P_1^m(u^2)$. Special handling is required for when $m=1$, and $P_1^m(u^2)$ is defined to account for the special case given in [12] as follows

$$P_1^m(u^2) = \begin{cases} 1 - \frac{u^2}{2} & m = 1 \\ m - \frac{1}{2} - (m-1)u^2 & m > 1 \end{cases}. \quad (2.42)$$

By iterating through n , any order n of the auxiliary polynomials can be generated for each azimuthal order m . Once auxiliary polynomials are computed then, they can be used for the unconventional recurrence relation for the gradient orthogonal Q-polynomials. The unconventional recurrence relation for the gradient orthogonal Q-polynomials is given in [12] as

$$Q_n^m(u^2) = \frac{[P_n^m(u^2) - g_{n-1}^m Q_{n-1}^m(u^2)]}{f_n^m}, \quad (2.43)$$

where the recurrence variables f_n^m and g_n^m are defined in [12] as

$$g_{n-1}^m = \frac{G_{n-1}^m}{f_{n-1}^m}, \quad (2.44)$$

$$f_n^m = \sqrt{F_n^m - g_{n-1}^m g_{n-1}^m}. \quad (2.45)$$

The f_n^m and g_n^m variables can be computed to any order n for a fixed m through iteration over n , starting at $n=1$. The unconventional recurrence relation shown in Eq. (2.43) is initialized with $Q_0^m(u^2) = 1/2f_0^m$. After f_n^m and g_n^m are computed for a fixed m , and up until the desired order n , then the recurrence relation defined in Eq. (2.43) is iterated over n to find the gradient orthogonal Q-polynomial for the fixed azimuthal order m . The details of the recurrence relations can be found in the appendix A of [12].

In order to illustrate a description of a surface with the gradient orthogonal Q-polynomials, an example is given by Forbes [12]. An implementation of the gradient orthogonal Q-polynomials is carried out in what follows in order to validate and explain in detail a characterization of an optical surface in terms of gradient orthogonal Q-polynomials through the step by step implementation of the example presented in [12] by Forbes. An off-axis section of a simple parabolic surface is fitted with gradient orthogonal Q-polynomials. The paraxial radius of curvature of the parabola is $\frac{1}{c} = 20$ mm, and the center of the off-axis section of interest is offset 20 mm away from the optical axis (z-axis). The radius of curvature of the best-fit sphere is $\frac{1}{c_{bfs}} = 37.405$ mm [12]. The best-fit sphere is the one that touches the surface at its axial point. The best-fit sphere curvature is calculated by taking the mean value of the sag around the perimeter,

$$c_{bfs} = \frac{2\langle f(\rho_{\max}, \theta) \rangle}{\rho_{\max}^2 + \langle f(\rho_{\max}, \theta) \rangle^2}, \quad (2.46)$$

where the angle brackets denote the average of the sag around the perimeter over θ . The off-axis section of interest has a diameter of 20 mm. In Figure 9 (a) a two dimensional cross section of the parabola and its best-fit sphere intersecting at the point of intersection (POI) along the local normal are shown. The sag departure from the best-fit sphere along the local normal is presented in Figure 9 (b).

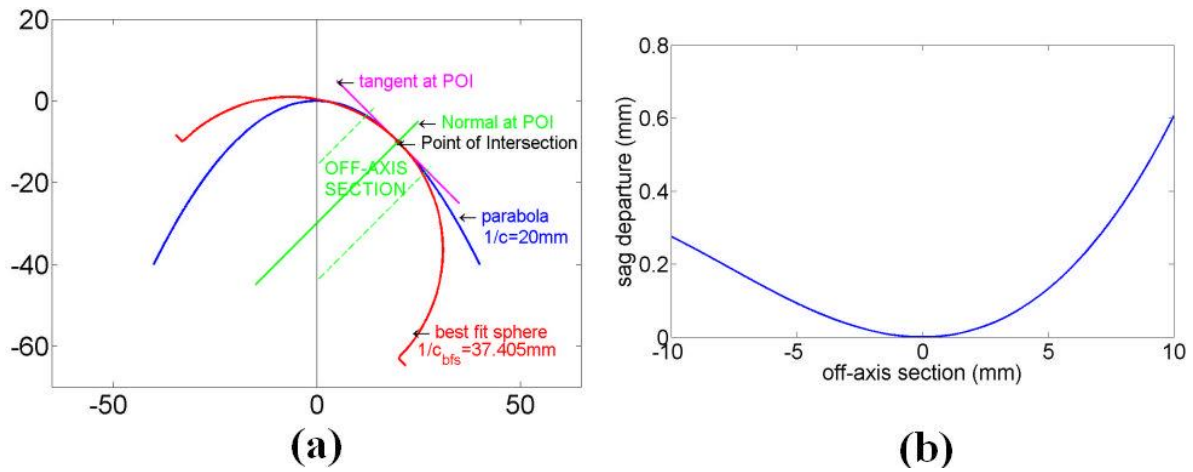


Figure 9 (a) 2D Cross-section for fitting with gradient orthogonal Q-polynomials; (b) The sag departure from the best-fit sphere, adapted from [12].

In three dimensions, the positions of the best-fit sphere and parabola are shown in Figure 10 (a). The red grid shows the best-fit sphere that touches the parabola at the POI, and the green section shows the off-axis section of interest. The green line is the normal at the POI, $(20, 0, -10)$ mm. In Figure 10 (b) the sag departure from the best-fit sphere for the off-axis section of interest is shown. Note the similarity between Figure 10 (b) and Figure 9 (b), which is just the central line of the former.

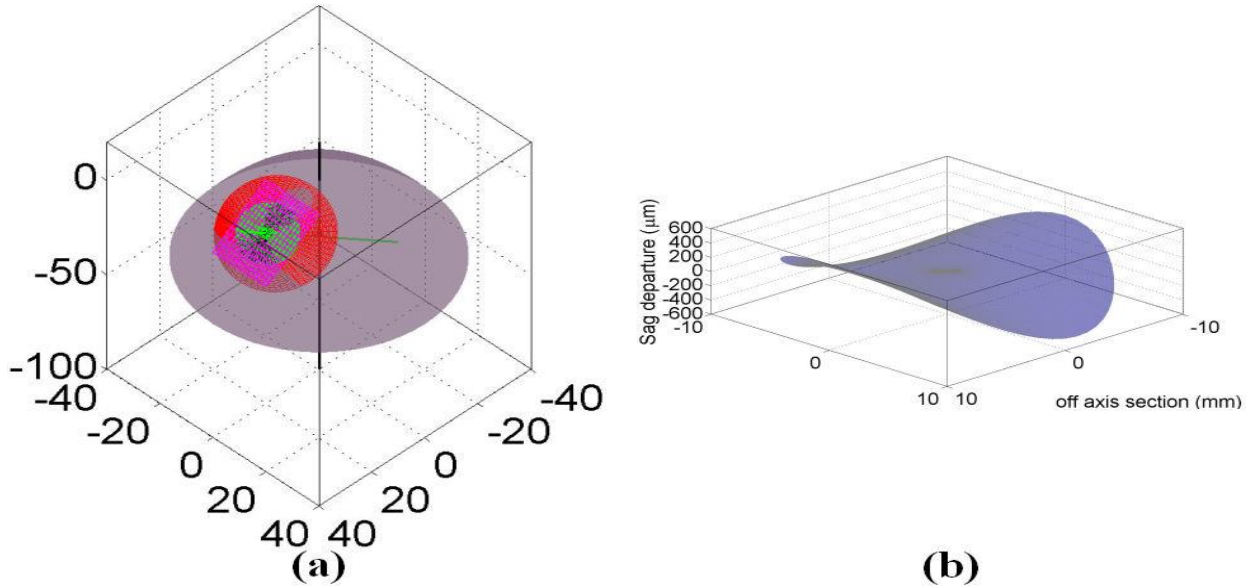


Figure 10 (a) 3D view of the off axis section on the parabola and the best-fit sphere intersecting and POI, (b) the sag departure of the parabola off its best-fit sphere over the off-axis section, adapted from [12].

After performing a least square fit of the sag shown in Figure 10 (b), we arrive at the coefficients for the fit. The tolerance for the fit is 1 nm. The truncation of the series expansion shown in Eq. (2.32) is $T=8$, which is $m+2n=8$. The slope orthogonal and gradient orthogonal Q-polynomials are entirely computed with the recurrence relations in order to achieve robustness and stability. For each azimuthal order m , the coefficients of the fitting Q-polynomials are given in Table 1. In Table 1, we can see that the coefficients decrease in magnitude as their order n decreases, and the smallest coefficient is 1 nm, which is the tolerance we have for the error profile.

Table 1 The coefficients for the gradient orthogonal Q-polynomials for fitting the sag shown in Figure 10 (b) [12].

b_n^m (nm)	m										
		0	1	2	3	4	5	6	7	8	
n	0	11509	199278	592756	-72134	6311	-274	-27	8	-1	
	1	-218	-187945	16062	115	-145	17	-1			
	2	6	1353	-243	5	2					
	3		-35	5							

The residual error profile for this least square fit with the gradient orthogonal Q-polynomials is shown in Figure 11. It is clear from Figure 11 that the Peak to Valley (PV) error never reaches the tolerance level of 1nm that is set for the fit.

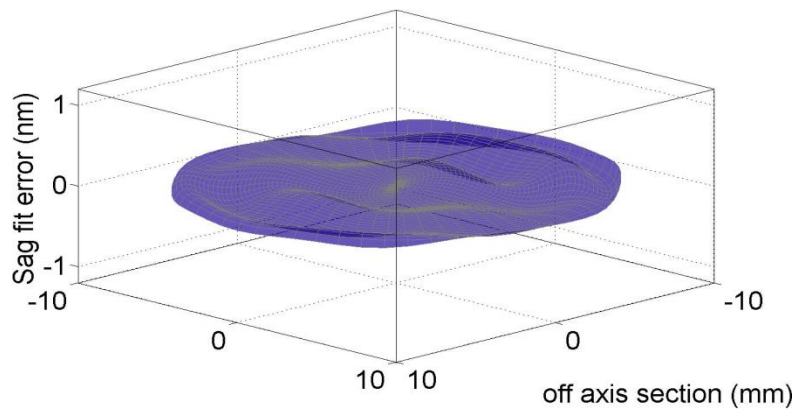


Figure 11 Profile of the residual error for the fit with the gradient orthogonal Q-polynomials of the sag shown in Figure 10(b), adapted from [12].

Radial Basis Functions and QR-based Algorithms

RBFs can be seen as a general optical surface description methodology forsaking the orthogonality of the polynomials in exchange for much improved simplicity and geometric flexibility in terms of aperture shapes. As opposed to the polynomials being orthogonal only over specific aperture shapes, such as Zernike and Q-polynomials over the unit circle, RBFs can be used over any aperture shape. They are simple to implement, they do not require any mesh or triangulation or polygonizations. The RBF description of a surface that is developed here is based upon a summation of a basic function translated across the aperture of the optical element. Linear combinations of the translation of the basic function form the foundation of this surface description methodology. It is the implementation of translation as a fundamental parameter that is innovative and powerful for application to optical systems without rotational symmetry. RBFs provide comparable accuracy to polynomials, and spectral convergence might be achieved [30]. Furthermore, Fornberg and Zuev reported that 10 node RBF interpolation of an arctangent function results in the same accuracy (10^{-5}) as the interpolation with the 170 Chebyshev polynomials provided the node locations and RBF basis functions are properly optimized [31].

Cakmakci et al. made use of Gaussians RBFs centered uniformly over an aperture for designing of HWDs freeform surfaces [10]. Cakmakci and Rolland designed and implemented some of the first pioneering examples of compact and lightweight Head Worn Displays (HWDs) employing RBF freeform surfaces [32, 33, 34]. Figure 12 shows sample HWDs designed by Rolland and Cakmakci with RBFs. A RBF freeform surface is described as:

$$z(\mathbf{x}) = \sum_{n=1}^N \phi(\varepsilon^2 \|\mathbf{x} - \mathbf{x}_n\|_2) a_n, \mathbf{x} \in R^s, \quad (2.47)$$

where a_n represents the weights in the combination, \mathbf{x}_n represents the center, \mathbf{x} is a point in the aperture, ε is the shape factor, and ϕ are the basis functions.



Figure 12 RBF optical surfaces in compact HWD design [32, 33].

We can observe that the basis functions are radial with the distance from the center, $\|\mathbf{x} - \mathbf{x}_n\|_2$ with the definition shown in Eq. (2.47). An example of a freeform RBF surface formation is shown in Figure 13.

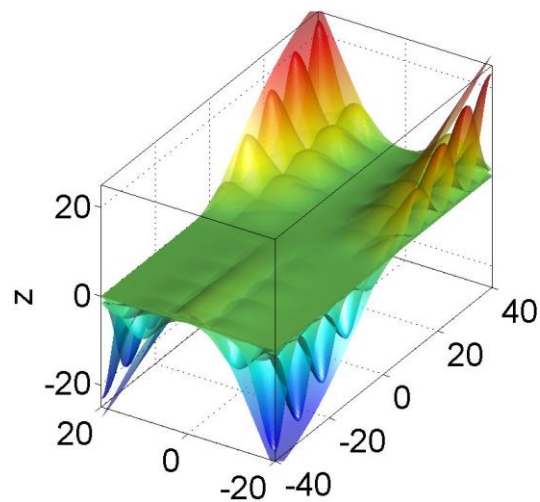


Figure 13 Forming of an RBF surface with Gaussians, $\varepsilon=0.19\text{mm}^{-1}$ over a rectangular aperture 40 mm x 80 mm.

In generating Figure 13, we have made use of 100 uniformly distributed Gaussians with the shape parameter 0.19 mm^{-1} . The coefficients a_n become the scaling weights for each of the Gaussian basis functions to form the overall approximated surface shown. In Figure 13, we can clearly see the individual Gaussians scaled with the weights. The overall RBF surface touches all of the Gaussian peaks at their center. A zero height surface in solid green is also shown in the center.

Unfortunately, giving up the orthogonality constraint does not come without a price with RBFs. As a consequence, severe ill-conditioning may occur, especially when the shape parameter ϵ is small, which corresponds to a flattening of the basis functions. In the flat basis function limit, Driscoll and Fornberg showed that limiting interpolants exist and converge to the form of polynomials [35]. In Figure 14 (a), we show a severe ill-conditioning of the RBF interpolation for a shape parameter corresponding to the flat basis functions. The range of shape parameters for which the RBF is ill-conditioned for this surface is shown in Figure 14 (b).

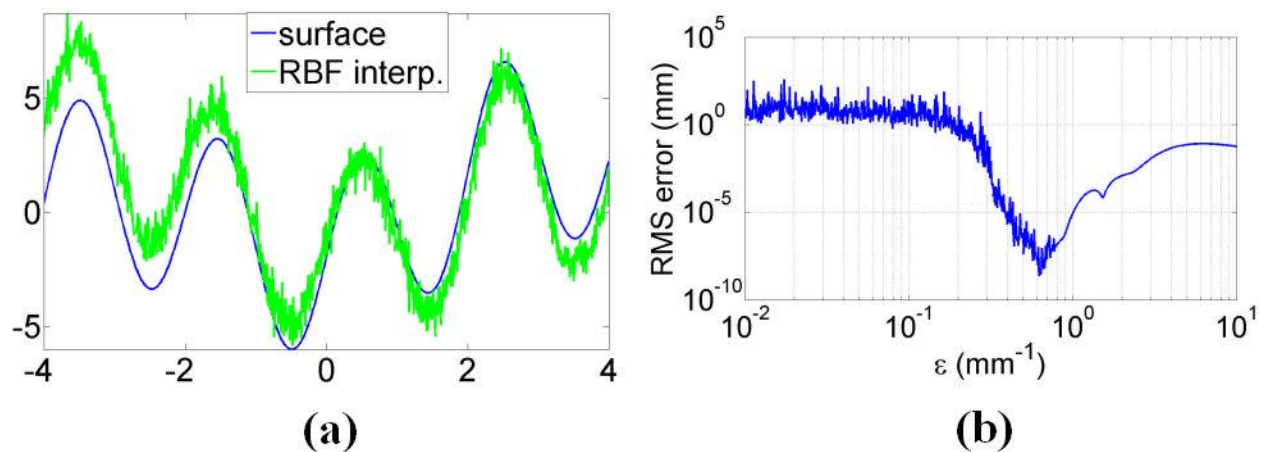


Figure 14 (a) Ill conditioning of RBF interpolation for $\epsilon = 0.2 \text{ mm}^{-1}$; (b) the range of ϵ over which the RBF is ill conditioned, adapted from [37].

One solution to remedy the ill-conditioning of RBFs is to use a QR-based algorithm. Fornberg et al. devised a QR approach [36] based upon the polynomial expansions of Gaussians in order to overcome the numerical ill-conditioning associated with RBFs. Their method expands the Gaussians over Chebyshev polynomials for the radial component, with a Gaussian weighting function along that dimension, and trigonometric functions for the angular components. A Gaussian over the polar coordinates is expanded by Fornberg et al. [36] as follows

$$\phi(r, \theta, r_n, \theta_n) = \sum_{j=0}^{\infty} \left\{ \sum_{m=0}^{\frac{(j-p)}{2}} d_{j,m} c_{j,m} T_{j,m}^c(r, \theta) + \sum_{m=1-p}^{\frac{(j-p)}{2}} d_{j,m} s_{j,m} T_{j,m}^s(r, \theta) \right\}, \quad (2.48)$$

where r_n, θ_n represents the Gaussian center, $d_{j,m} c_{j,m}$, and $s_{j,m}$ are expansion coefficients which depends on the shape parameter ε , and $T_{j,m}^{c,s}(r, \theta)$ are the final expansion functions. The cosine version of the expansion functions are given in [36] as

$$T_{j,m}^c(r, \theta) = e^{-\varepsilon^2 r^2} r^{2m} T_{j-2m}(r) \cos((2m + p)\theta), \quad (2.49)$$

where $T_{(j-2m)}(r)$ are the standard Chebyshev polynomials, and p is 0 or 1 depending on the index variable j . The sine version of the expansion function is exactly the same as Eq. (2.49), except the cosine function is replaced with a sine function. When we analyze the expansion functions given in Eq. (2.49), we observe that for the small shape parameter, first term is a Gaussian that becomes unity, therefore leaving us with the orthogonal Chebyshev polynomials and a monomial term in the radial component and orthogonal trigonometric functions in the angular direction.

The method then applies the QR decomposition on the resulting expansion matrix in order to yield a well-conditioned basis. When considering the application to large shape factors as well, the RBF-QR method may suffer from numerical overflow as the expansion coefficients, $d_{j,m}$ start to diverge quickly depending upon ε and the index variable. In Figure 15 (b) such an overflow of expansion coefficients is seen for the range of shape parameters larger than 3. In this case the RBF method shown in Eq.(2.47) may be used instead, given that in this case RBFs mostly do not suffer from ill-conditioning.

More recently, Fasshauer and McCourt devised another RBF-QR approach [37] similar to Fornberg et al.'s QR algorithm. This method works by deploying eigenfunctions of Gaussians that are related to Hermite polynomials. The eigenfunctions of the Gaussians are given in [37] as

$$\varphi_n(x) = \gamma_n e^{-\delta^2 x^2} H_{n-1}(\alpha \beta x), \quad (2.50)$$

where α is the global scale parameter, β , γ_n , and δ are defined in [37] as auxiliary parameters, and $H_{n-1}(x)$ is the Hermite polynomials of order $n-1$. The multivariate eigenfunctions are obtained with a tensor product of the eigenfunctions over the different dimensions. With these eigenfunctions at hand, a Gaussian can be represented as a summation of eigenfunctions

$$\phi(x - x_n) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x) \varphi_k(x_n), \quad (2.51)$$

where λ_k is the corresponding eigenvalue [37]. The multivariate case is obtained through a tensor product of eigenfunctions and a multiplication of the eigenvalues on each dimension. The well-conditioned basis is acquired though a QR decomposition on the resulting expansion

matrix. In Figure 15, we have shown a RBF-QR application for the surface shown in Figure 14 to successfully remove the ill-conditioning of RBF interpolation.

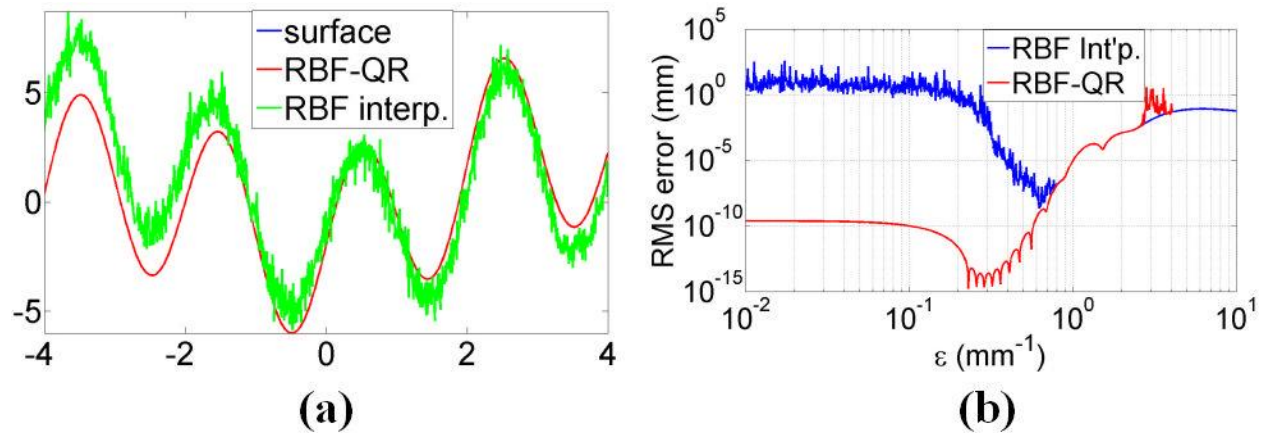


Figure 15 (a) RBF-QR successfully removes the ill-conditioning for $\epsilon=0.65\text{mm}^{-1}$; (b) RBR-QR is more stable and yielding more accuracy over a range of ϵ , adapted from [37].

Fasshauer and McCourt also suggest using a regression method with RBF-QR, which consists of internally truncating the data to lower rank approximations to maintain accurate approximants [37]. This large data reduction requires that the original surface is greatly oversampled. Whenever high orders of orthogonal polynomials are necessary within the use of the RBF-QR method, recurrence relations can be used to remove ill conditioning associated with these polynomials [38]. For a more detailed description of RBF-based methods, the book by Fasshauer contains an in-depth analysis of the RBF methods [39].

The RBF-QR method, while correctly removing ill conditioning associated with RBFs of small shape factors, has prominent aspects notable for optical designers. As the method itself expands the Gaussians onto polynomials, it makes use of a large number of terms in the expansion in order to represent the Gaussians with desired accuracies. Moreover, the Chebyshev

polynomials with trigonometric modes are only orthogonal over circular apertures. The expansion coefficients suffer overflows when the shape parameter, ε , becomes larger; as a consequence, standard RBF methods appear as a viable alternative to compensate for the deficiency in QR-based algorithms.

CHAPTER THREE: EDGE-CLUSTERED RAY GRIDS

In this chapter, we focus on some of the important limitations to φ -polynomials, specifically Zernike polynomial representation of optical surfaces in describing the evolving freeform surface descriptions. There are two major bottlenecks in representing an optical surface sag or wavefront aberration function over the unit circle with φ -polynomials to the accuracies demanded for most stringent optics applications. One is the numerical ill-conditioning associated with computation of the higher-order polynomials that might be required to achieve an accurate representation of an as-fabricated optical surface. Forbes addressed this bottleneck with the development of three-term recurrence relations for Zernike polynomials that were described in chapter 2 [17]. Second is the substantial number of Zernike terms required, sometimes thousands. Prior to arriving at the proper number of terms in the representation of the optical surface, intermediate results with an insufficient number of coefficients exhibit high-departure errors at the edges. Importantly, as it is shown in this chapter, the rate of convergence to an adequate number of terms in the representation is sensitive to the surface sampling. The content of this chapter was published in our article [18].

Specifically in this chapter, we show that the ray grids commonly used in sampling a freeform surface to form a database from which to perform a φ -polynomial fit is limiting the efficacy of computation. *We show an edge clustered fitting grid that effectively suppresses that edge ringing that arises as the polynomial adapts to the fully nonsymmetric features of the surface.* The impact of this fitting grid on the reduction of edge ringing and the associated improvement of surface-fit quality by several orders of magnitude compared to the current fitting

grids are demonstrated in this chapter. We compare the effectiveness of fit with commonly used grid types such as 1) a set of uniform size hexagonal subgrids centered on a uniform rectangular grid, 2) a polar grid with a Chebyshev-based radial sample points, and 3) a uniformly random point fitting grid to show the significance of the edge-clustered random fitting grid on the accuracy of surface approximation.

This chapter is organized as follows; in the next section a detailed explanation of the least squares procedure to map the freeform surface based on the sampled database created on the fitting grids to a Zernike polynomial representation is presented. In the succeeding section, the surfaces that were created to be the benchmark cases for freeform surfaces are described. In the following section, four fitting grids that supply the database for the least-square fitting process are detailed. Finally, the numerical results that show Root Mean Square (RMS) errors as a function of the number of Zernike polynomial coefficients for each test surface with different fitting grids are reported, before concluding this chapter.

Least-squares Data-fitting to Create a Zernike Polynomial Surface

A freeform surface over a circular aperture may be represented as a function $f(\rho, \theta)$ that is a weighted sum of Zernike basis functions that form a complete and orthogonal set over the unit disk or circle as

$$f(\rho, \theta) = \sum_{k=1}^N c_k Z_n^m(\rho, \theta), \quad (3.1)$$

where c_k represents the coefficient associated with each Zernike polynomial. If the function is known beforehand, the coefficients c_k could be determined by taking a double integral of the function multiplied with the corresponding Zernike polynomial over the circle. In optical system design, these surfaces are to be determined from optimization of the user selected variable coefficients, based on wavefronts that are initiated at different field points, by a real ray trace of a grid (typically uniform, rectangular) of rays in the pupil at one wavelength through a model of a complete optical system. In this scenario, the least squares approach given in Eq.(3.2) is used to find the coefficients c_k associated with each Zernike polynomial at each wavelength and field point. Here, the goal is to represent the sag of a surface to be used as an exact representation of a freeform surface, to the extent possible. In Eq. (3.2) a least squares system of equations is shown as

$$\mathbf{A}\mathbf{c} = \mathbf{f}. \quad (3.2)$$

The matrix A is M by N , where N is the number of Zernike polynomial coefficients to be fit, and M is the number of sample points throughout the aperture, and $M > N$. Each row corresponds to a sample point, where each Zernike polynomial is computed. Each column corresponds to a Zernike polynomial evaluated over all the points in the circular aperture. The coefficients c_k are the weights multiplying the columns of the matrix A to match the surface sag vector, f . The coefficients vector, c , comprises the coefficients c_k . Once the coefficients are determined, the approximant can be evaluated at any point across the aperture as would routinely occur in evaluating the impact of the surface in an overall optical system design.

In general a system consisting of M equations and N unknowns, where $M > N$, does not have an exact solution. This is an *overdetermined* system. The residual, \mathbf{r} , can be made very small by a suitable choice of coefficients \mathbf{c} , as

$$\mathbf{r} = \mathbf{f} - \mathbf{A}\mathbf{c}. \quad (3.3)$$

This residual is minimal if and only if it lies perpendicular to the range of matrix \mathbf{A} . It then satisfies

$$\mathbf{A}^* \mathbf{r} = 0, \quad (3.4)$$

where $*$ represents the transpose of the matrix \mathbf{A} . Considering Eqs. (3.3) and (3.4), the set of coefficients that gives the minimal residual is then given by

$$\mathbf{c} = (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* \mathbf{f}. \quad (3.5)$$

Instead of Eq. (3.5), which provides the normal set of equations for the general least squares that may be unstable, a QR decomposition of \mathbf{A} is taken leading to the coefficients

$$\mathbf{c} = \mathbf{R}^{-1} (\mathbf{Q}^* \mathbf{f}), \quad (3.6)$$

where \mathbf{Q} is the orthogonal column matrix, and \mathbf{R} is the upper triangular matrix of the QR decomposition. The QR based algorithm given in Eq. (3.6) to compute the solution of the least squares system is very stable. It is used throughout this dissertation for the solution of any least squares problem. In Figure 16, as an example, we have shown the 136 Zernike polynomials coefficients that are computed as a result of least squares fitting of a conventional rotationally

symmetric asphere, which is shown in Figure 17 (a) and whose description is given in Eq. (3.7). As can be seen in Figure 16, most of the coefficients are almost zero up to the working double precision limit, as expected. The non-zero valued coefficients are aligned with increasing orders of the Zernike polynomial terms that are affiliated with spherical aberration. Table 2 provides a list of the nonzero Zernike polynomial coefficients of Figure 16.

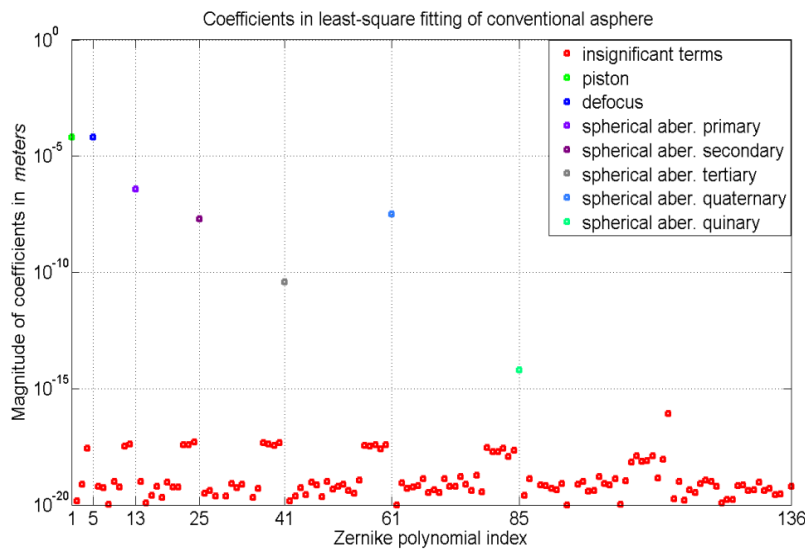


Figure 16 Zernike coefficients resulting from the least squares fitting of a conventional asphere with 136 Zernike polynomials using the Born and Wolf ordering of terms.

The Test Surfaces

In order to investigate the effectiveness of different sampling distributions over the unit circle, we chose three different test cases for analytic surface shapes as a benchmark suite whose analytical expressions in mm units are given in Eqs. (3.7)-(3.9). The second two test surfaces are designed to exercise features of a next generation freeform surface shape that includes not only

conic and polynomial terms but also multi-centric additive or subtractive functions. The first test surface is a rotationally symmetric five-term aspheric mirror surface shown in Figure 17 (a).

Table 2 Zernike Polynomials with significant coefficients for a rotationally symmetric conventional asphere

Index	n	m	Born & Wolf Zernike Polynomial	Aberration type
1	0	0	1	Piston
5	2	0	$2\rho^2 - 1$	Defocus
13	4	0	$6\rho^4 - 6\rho^2 + 1$	4th spherical aberration
25	6	0	$20\rho^6 - 30\rho^4 + 12\rho^2 - 1$	6 th spherical aberration
41	8	0	$70\rho^8 - 140\rho^6 + 90\rho^4 - 20\rho^2 + 1$	8 th spherical aberration
61	10	0	$252\rho^{10} - 630\rho^8 + 560\rho^6 - 210\rho^4 + 30\rho^2 - 1$	10 th spherical aberration
85	12	0	$924\rho^{12} - 2772\rho^{10} + 3150\rho^8 - 1680\rho^6 + 420\rho^4 - 42\rho^2 + 1$	12 th spherical aberration

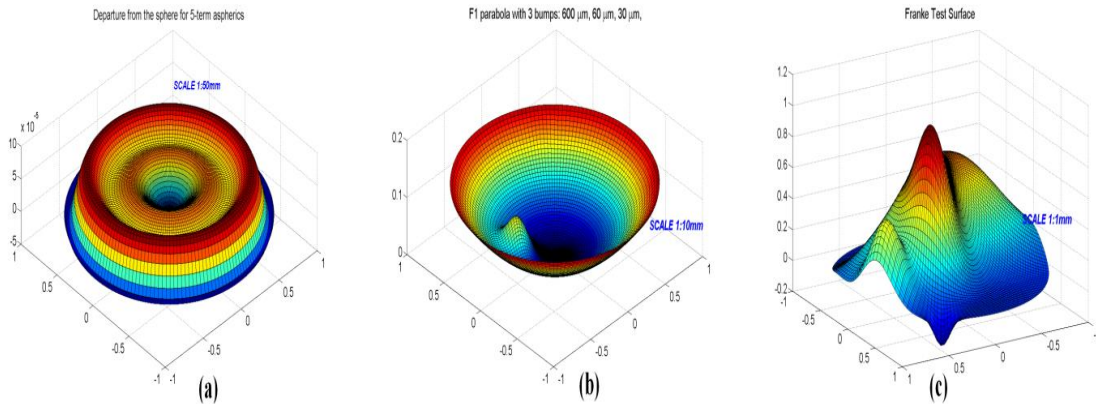


Figure 17 The test surfaces described in Eqs. (3.7) -(3.9) (a) A five-term conventional aspheric mirror; (b) A F/1 parabola with 600, 50, and 30 μm bumps; (c) Franke surface. Note that in this illustration the apertures were normalized to 1 in radius.

$$f_{\text{aspherics}}(\rho) = \frac{(\rho^2) / 200}{1 + \sqrt{1 + 1 / (200)^2 (\rho^2)}} + 1.876 \times 10^{-5} \rho^2 - 0.5 \times 10^{-7} \rho^4 + 0.545 \times 10^{-10} \rho^6 - 0.25 \times 10^{-13} \rho^8 + 0.4 \times 10^{-17} \rho^{10}. \quad (3.7)$$

$$f_{\text{three-bumps}}(x, y) = \frac{(x^2 + y^2)}{80} + 0.05e^{-0.25[(x-7)^2 + (y+6)^2]} + 0.6e^{-0.49[(x+3)^2 + (y-2)^2]} + 0.03e^{-0.81[(x-5)^2 + (y-7)^2]}. \quad (3.8)$$

$$f_{\text{Franke}}(x, y) = 0.75e^{-0.25[(9x-2)^2 + (9y-2)^2]} + 0.75e^{-[(9x+1)^2 / 49 + (9y+1)^2 / 10]} + 0.5e^{-0.25[(9x-7)^2 + (9y-3)^2]} - 0.2e^{-[(9x-4)^2 + (9y-7)^2]}. \quad (3.9)$$

The second test surface is an F/1 parabolic surface with three Gaussian bumps of different heights added and standard deviations (as shown in Eq. (3.8)), aiming to be a general example of a multiple bump surface that might be encountered in optical surface manufacturing and illustrated in Figure 17 (b). The third test surface is a Franke [39] test function (description is given in Eq. (3.9)), shown in Figure 17 (c) that comes from the scattered-data approximation literature that is widely used to assess the approximation capabilities of different mathematical bases.

Hexagonal, Chebyshev, Uniform-random and Edge-clustered Grids

In this section, we present four different fitting grid patterns that form the database content for the least squares fits to yield the Zernike polynomial coefficients that are then used to describe the surface sag at any point on the optical surface. The resulting Zernike polynomial definition of the surface is re-sampled during either an analysis of the optical system performance or during optical system optimization. When optical surfaces are specified by a polynomial function in an optical system design and analysis simulation environment, the

predicted performance of the optical system is computed by an often sparse sample of the wavefront obtained by ray tracing. Typically, in a commercial raytrace code, uniform rectangular grids are used to form a sampled database for the wavefront at an individual wavelength and at an individual field point where the grid density is often in a user specified range between 64×64 and 1024×1024 .

In this chapter, four different sampling grids are considered to create the database that is applied to transform the specially selected set of analytic test surfaces defined by a small number of coefficients to an equivalent surface, but now computed from a Zernike polynomial description in all cases. The four types of grids applied to the transformation process are 1) a set of uniform size hexagonal subgrids centered on a uniform rectangular grid (hex grid), 2) a polar grid with Chebyshev-based radial weighting (Cheby-polar grid), 3) a uniformly random-point grid (uni-random grid), and 4) an edge-clustered, random-point grid (e_clust-random grid). Figure 18 displays each of these grid types using approximately 450 points over the unit circle.

The purpose of compiling and illustrating the evolution of the fit accuracy is to highlight the relationship between the spatial-frequency content of the test surface and the number of Zernike terms required. Moving forward, one of the features of the broader class of freeform surfaces will be an ability to introduce surface features with higher spatial-frequency content. With the hex grid setting, the unit circle is divided into regular hexagonal cells. The center point for each hexagonal cell is also included as a sample point. We have chosen a hexagonal grid structure for the uniform setting since a circular aperture is more uniformly covered with hexagons rather than rectangular cells.

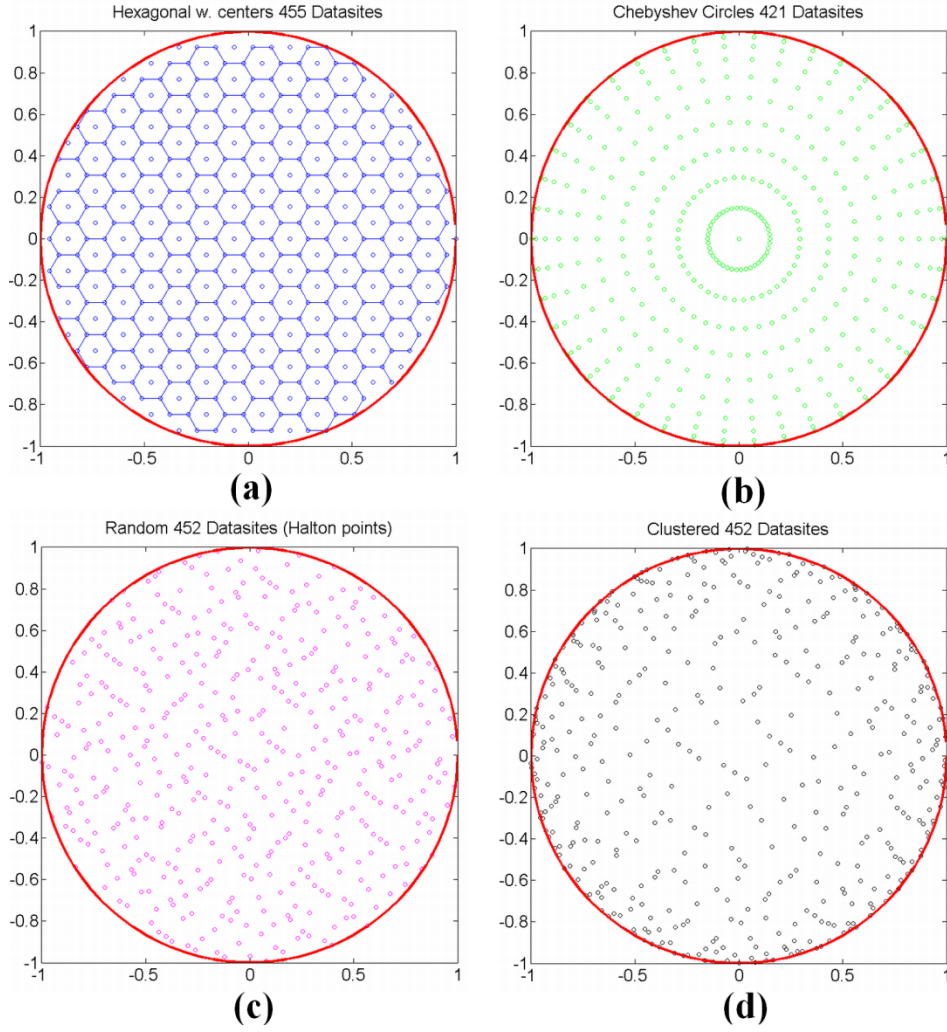


Figure 18 Fitting grids used to demonstrate efficacy of data sampling: (a) hex grid, (b) Cheby-polar grid (c) uni-random grid (Halton points) (d) e_clust-random grid that clusters points towards the boundary over the unit circle.

In the Cheby-polar grid, the sample points are placed along an expanding set of circles at the roots of the Chebyshev polynomial of degree n in the ρ direction. The points along the ρ direction are called Chebyshev abscissas. Therefore we call these circles Chebyshev circles. The radii of the circles are given as

$$\rho_k = \cos\left(\frac{(2k+1)\pi}{2n}\right), k = 0, 1, \dots, n-1. \quad (3.10)$$

The third grid type, uni-random, consists of simply randomly placed sample points. A point sampling function generates uniformly distributed random points across the unit square (see [39], Halton points). The fourth sampling method is termed e_clust-random point sampling. Here the radial coordinate of the random Halton points is modified with a sine function weighting to move them towards the boundary of the unit circle. If a point has coordinates (ρ, θ) then the corresponding clustered point has $(\sin(\pi\rho/2), \theta)$ as its coordinates. This type of edge-clustered sampling was previously used in the context of an RBF-QR method by Fornberg et al. [36]. As will be shown in the next section, it is the fourth fitting grid that has clearly demonstrated the best efficacy.

Results of Efficacy of Fitting the Test Surfaces with Four Different Sampling Grids

To study the efficacy of each grid type for determining the coefficients of a Zernike polynomial fit to each of the test surfaces, over an increasing set of coefficients, we have sampled each surface with the four different fitting grids, carried out the least squares approximation to create the Zernike polynomial fit and then evaluated the approximant at around 7000 points that are on a uniform grid as would be the case in any commercial lens design software. Then, RMS errors in the resulting approximant to the test surface, which quantifies the difference between the height of the test surface computed using Eqs. (3.7)-(3.9) and the Zernike polynomial fit determined for each fitting grid type, are recorded in meters in order for 10^{-9} on the vertical scale to correspond to nanometers. For each approximation, the size of the approximation matrix is M by N , where N is the number of Zernike polynomials coefficients, and M is the number of samples over the circular aperture, where $M \approx 1.5N$. We have chosen this ratio

such that the computational cost is not too high in the approximation; meanwhile the number of samples is sufficient to avoid working within an interpolation setting (i.e. $M=N$). In Chebyshev-based clustering methods, this ratio has been studied in 1D, and optimum results are obtained when this ratio is around 1.5 for different bases [40]. We have analyzed if the results in our experiments are affected with this ratio. For example, if we increase this ratio to 2.3, and kept the number of samples between 25 to 4980, the observed impact is a slower rate of convergence when using the Cheby-polar grid and the e-clust random grid, which produce effectively the coincident results however with 2 or 3 orders of magnitude less accuracy, which is still much better than the uniform hexagonal grid or random point outcomes. In another set-up, where we decided to increase the number of samples across the circular aperture to an interval between 37 to 7639 while keeping the number of basis elements the same in order to increase this ratio to 2.3, similar to the previous experiment, we noticed that all major trends remained unchanged, except that the Cheby-polar grid and the e-clust random grid results coincided at the expense of needing extra samples compared to the results presented in what follows. For instance, we have reached $2e-14$ level of accuracy for the Franke test case both with 7639 samples and the ratio being 2.3, and with 4980 samples and the ratio being 1.5. Since in our work, we are also looking for the most economical solutions to optical surface description, we are pleased with a ratio of surface samples to basis elements kept at 1.5 so the edge clustered grid has a faster spectral rate of convergence and the number of samples needed to reach a level of accuracy is lower than if using a ratio of 2.3. It is important to note that the computational time that it takes to solve a least square problem grows with $O(M^3)$, where M is the dimension of the approximation matrix, i.e. total number of samples.

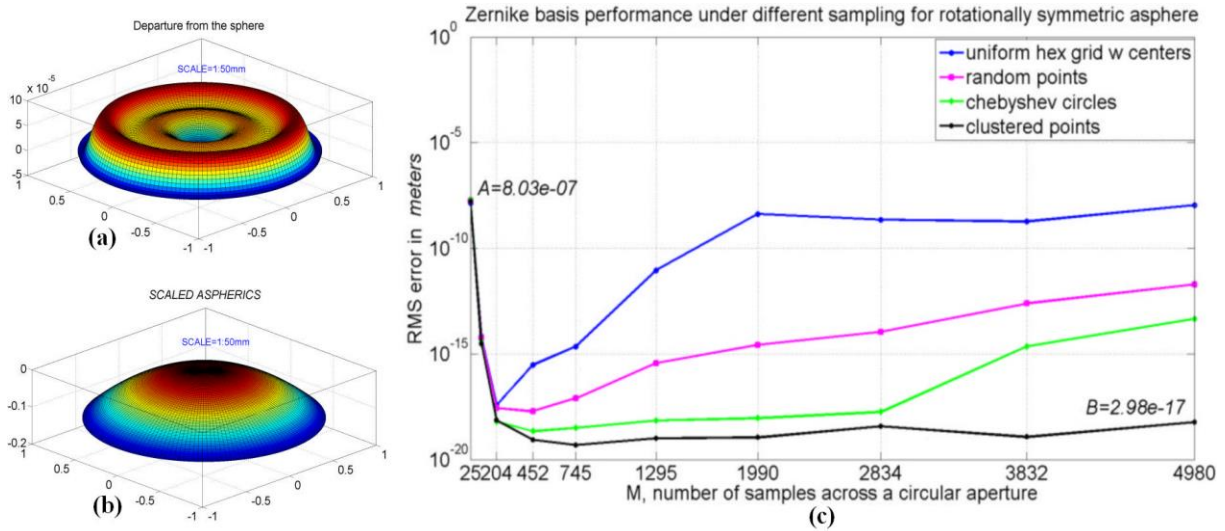


Figure 19 (a) Rotationally symmetric analytic five-term asphere departure from best sphere based on minimum RMS; (b) rotationally symmetric analytic five-term asphere; (c) RMS error in Zernike polynomials approximant performance relative to the analytic function expressed in meters for an increasing number of Zernike coefficients with hex, uni-random, Cheby-polar, and e-clust random sampling grids for the asphere shown in (b).

In Figure 19 (c), we have plotted the RMS errors in meters versus the number of sample points on a uniform grid for the four different fitting grids for the analytic conventional rotationally symmetric test surface described in Eq. (3.7), and shown in Figure 6 (a)-(b). This test surface is well approximated with all fitting strategies by using around 200 points. We see that initially, all the fitting grids achieve good RMS errors, and the geometry of the fitting grid has little influence on the quality of the approximation. As shown in Figure 19(c), all fitting grids quickly achieved sub-nanometer accuracies with less than 100 points. However, as the number of Zernike coefficients increases, the RMS error grows with hex grid, uni-random, and Cheby-polar grids, whereas the edge clustered random point set produced the most stable results with errors kept on the order of machine precision. As this surface is a conventional rotationally symmetric surface rather than a freeform surface, the number of Zernike polynomials is minimal, i.e. 66

terms, to reach the high accuracy levels demanded by precision or lithography applications (also see Figure 16).

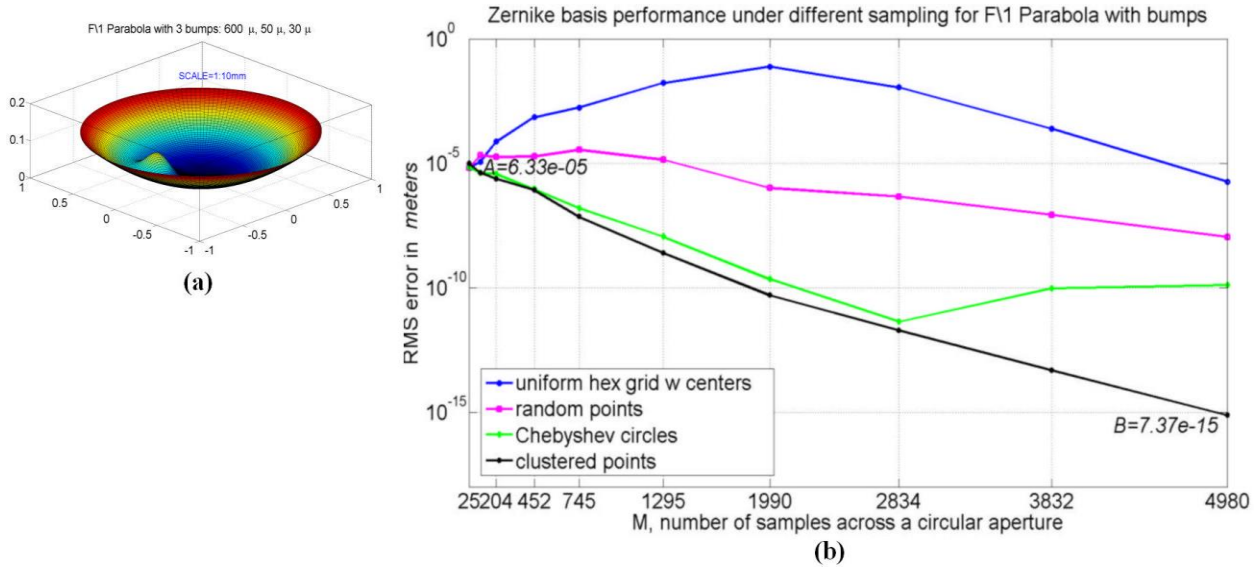


Figure 20 (b) RMS error in Zernike polynomials approximant performance relative to the analytic function expressed in meters for an increasing number of Zernike coefficients with hex, uni-random, Cheby-polar, and e-clust-random sampling grids for the F/1 parabola with 3 bumps, shown in (a).

In Figure 20 (b), we illustrate the RMS errors in meters versus the total number of sample points on a uniform grid for the four fitting grids applied to fitting a Zernike polynomial to the F/1 parabola with three bumps described in Eq. (3.8). Results show that the hex grid and the uni-random grid produce large errors in the order of 10^{-5} m to 1m, which is huge compared to the sizes of the bumps on the F/1 parabola spanning 30 to 600 μ m. This is caused by the dominant effect of the oscillations on the circular boundary when there is no edge weighting. On the other hand, applying either of the edge clustered sampling grids, Cheby-polar or e_clust-random, produces *exponentially decaying errors* as the number of sampling points increases. However, it is only with the e_clust-random fitting grid that we can obtain an approximant described with

machine precision accuracy, as illustrated in Figure 20 (b). The Cheby-polar grid produces a similar RMS error trend as that of the e_clust-random grid until the number of point samples reaches around 3000, after which it yields less accuracy. Only the edge clustered sampling grids achieved sub-nanometer accuracies, while the unclustered fitting grids could not even describe the surface with micron accuracies. Hex grid and uni-random grid point sampling produced consistently poor surface approximants when fitting Zernike polynomials to intrinsically nonsymmetric surfaces.

The reason that all of the fitting grids initially produce RMS errors around 5×10^{-5} m for the F/1 parabola with bumps is best captured in Figure 21, where we show the resulting approximants with the hex grid fitting (top row) and e_clust-random fitting (bottom row) grids while increasing the number of sample points and in proportion the number of Zernike polynomial coefficients for the F/1 parabola with bumps. Results show that initially the number of sample points used in the evaluation of the approximation, is so small that the bumps on the F/1 parabola are under-sampled. However, as the number of Zernike polynomial coefficients increases to better fit the surface (and in conjunction the number of sampling points used in the evaluation), the least squares process tries to match the sag values at more points with a higher number of Zernike polynomial terms, causing severe oscillations at the edges when distributions that are not edge clustered are used as seen in the upper row of Figure 21. As the lower row of displays in Figure 21 shows, the e_clust-random fitting grid successfully describes the surface without significant edge ringing, therefore producing much better approximants than a sampling without edge clustering.

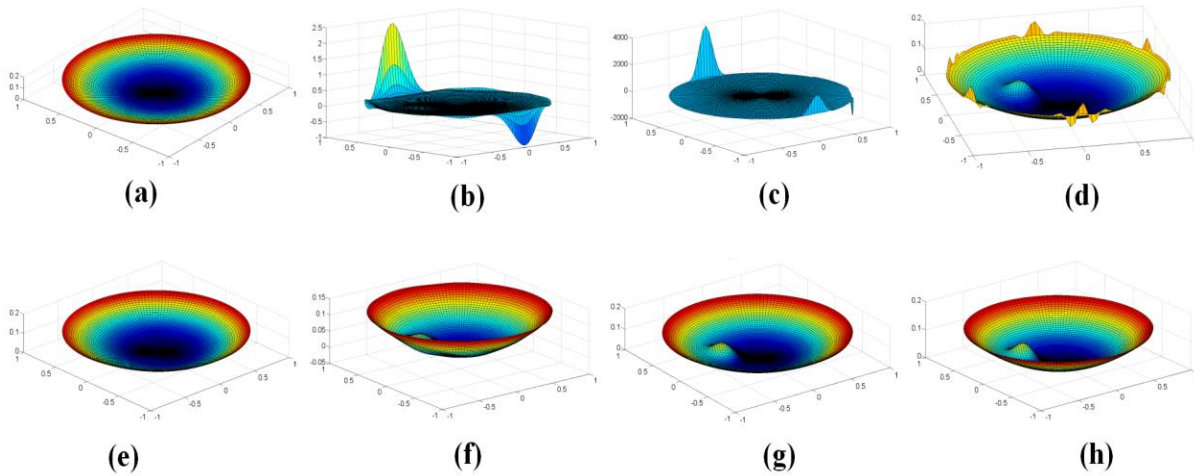


Figure 21 Comparison of the approximants obtained with two different fitting grids for the F/1 parabola with bumps; Top row: Approximant with uni-hex grid sampling with (a) 25 samples, (b) 204 samples, (c) 1990 samples, (d) 4980 samples; Bottom row: Approximant with e_clust-random sampling with (e) 25 samples, (f) 204 samples, (g) 1990 samples, and (h) 4980 samples.

The reason the edge clustered fitting grids produce better surface approximants can be explained by their success in reducing the boundary errors over the unit circle. In the same way that adjusting an equispaced grid to Chebyshev points is the remedy for the Runge type oscillations that result in using equispaced points in 1D, the edge clustered fitting grids act as a remedy for ϕ -polynomials edge ringing for surface shapes with offset localized structured such as seen with multi-centric RBFs [18].

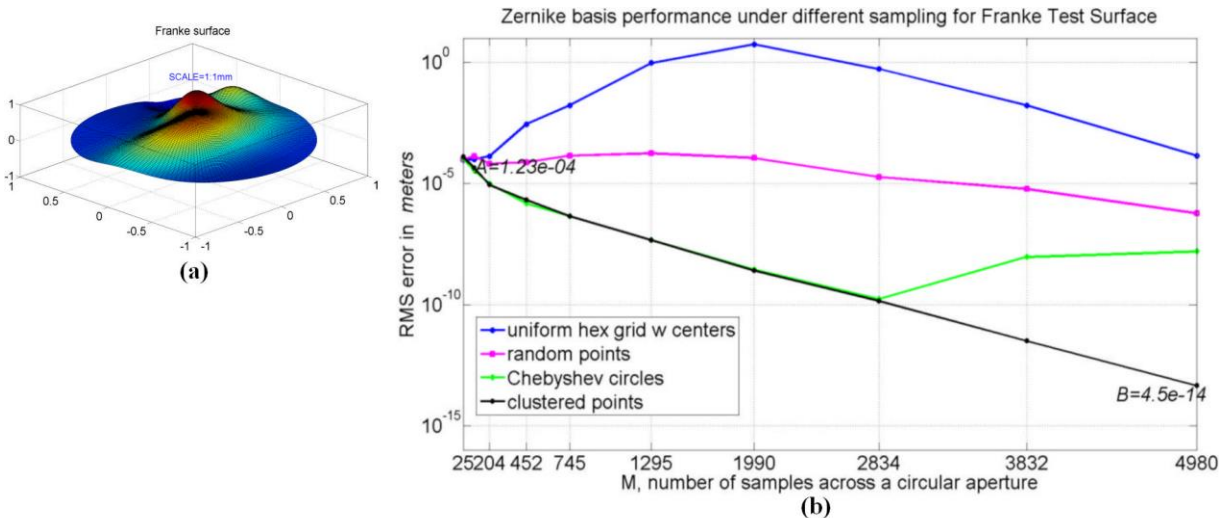


Figure 22 (b) Zernike polynomial fit RMS error expressed in meters as a function of coefficient order with hex, uni-random, Cheby-polar, and e_clust-random fitting grids for the freeform Franke surface (Eq. (3.9)) shown in (a).

In Figure 22(b), we have plotted the RMS errors in Zernike approximants compared to the analytic equation as the number of Zernike coefficients increases for the Franke surface (Eq. (3.9)), which is selected to be a stressing example of the important characteristics of a next generation optical freeform surface. Similar to the results presented for the F/1 parabola with bumps, edge clustered fitting grids produced *excellent approximants with errors decaying exponentially* with respect to the number of sample points used in the least squares approximation. Two to nine orders of magnitude improvement over the unweighted fitted grids was achieved. The reason the unweighted fitting grids produce large RMS errors as the number of Zernike polynomial coefficients is increased is because the edge oscillations must be controlled with edge weighting or clustering. Only the edge clustered fitting grids achieve sub-nanometer accuracy levels (i.e. 10^{-10} m) demanded by lithography applications. Cheby-polar fitting grids produced similar results to e_clust-random fitting grids until around 2800 points.

After this point, Cheby-polar fitting grids produce approximants with growing RMS errors. The stable and exponentially decaying errors produced by the edge clustered sampling make this method the method of choice for fitting Zernike polynomials to freeform optics surfaces and is the key result of this chapter.

In order to analyze the behavior change between the two different edge clustering grids (namely Cheby-polar and e-clust-random grids) that is observed in both Figures 20 (b) and 22(b), (see green and black lines at and after 2834 sample points), we have divided the unit circle into 3 parts. In each region in Figure 23, we determined the percentage of the number of points per unit area and reported that metric for two edge clustering grids types in Table 3.

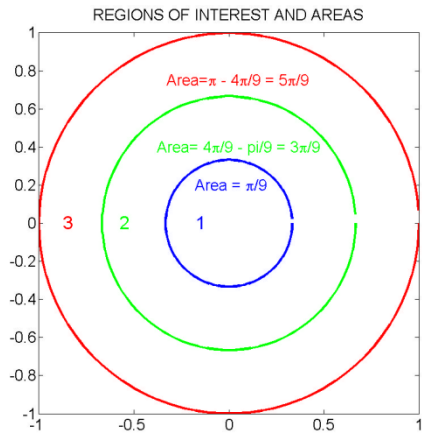


Figure 23 Three regions on the unit circle.

As the degree of the Zernike polynomial increases in order to better fit the surface shown in Figure 22 (a), the variations on the form of the polynomial are increased towards the edge, see for example Figure 4 (b) in chapter 2. In order to account for these variations for the Zernike polynomials towards the edge, and decouple the basis elements - especially higher orders - increased sampling need to occur in the outermost region, R3, not in R1. Only e_clust-random

grid provide this requirement and therefore stably reduce the error as the polynomial order increases, see Figures 20 (b) and 22 (b).

Table 3 Percentage of the number of points per unit area for the two edge clustering grids

Samples	Chebyshev polar grid				Edge-clustered random grid			
	R1 (%)	R2 (%)	R3 (%)	RMS (m)	R1 (%)	R2 (%)	R3 (%)	RMS (m)
745	15.5	10.2	10.7	4.4e-7	4.7	5.6	15.7	4.6e-7
1990	18.2	9.1	10.9	2.8e-9	4.7	5.7	15.7	2.5e-9
2834	19.2	8.9	10.7	1.7e-10	4.7	5.6	15.6	1.4e-10
3832	22.6	7.5	10.9	8.7e-9	4.7	5.7	15.7	3.0e-12
4980	20.0	8.6	10.9	1.8e-8	4.7	5.6	15.6	4.5e-14

In the process of optical design, due to the number of parameters, fields, wavelengths, etc. the number of sample points within any one ray set used in evaluating the metric for optimization is minimal, often less than 100. However, once a solution is established, the analysis of the performance is typically conducted with hundreds of thousands if not millions of ray samples per surface. In the context of the process being pursued here, the Zernike coefficients that characterize the surface are computed with a relatively sparse optimization grid. The result is that there is a spatial frequency content on the wavefront that will not typically be sampled during the optimization. This gap may result in a poor match to the wavefront by the surface model. It is necessary in fact, for the surfaces shown here, especially the parabola with a

series of three bumps and the Franke surface, to greatly exceed the number of samples that can be applied in a typical lens design software if the distinguishing features are to be picked up in the surface models as revealed by the high resolution analysis that is readily conducted in a ray tracing code. Additionally, as compiled here using a standalone code, the number of Zernike terms that needs to be available in the optimization procedure exceeds by thousands the parameterization sets available during an optical design, which currently rarely exceeds 100.

Conclusion

In this chapter, we have investigated the effect of edge clustering points towards the boundary of the unit circle when fitting Zernike polynomials to a general surface shape that represents a family of freeform optical surfaces. We demonstrated that these grids are effective and that the edge-clustered random-fitting grid is particularly effective. We have also compared this fitting grid with a hexagonal sub-grid spaced on uniform centers and a simple random fitting grid for optical surface approximation with Zernike polynomials. We have observed that edge-clustered fitting grids produced very good approximants, and improved the approximation performance by several orders of magnitude compared to that of fitting grids without edge clustering. For rotationally symmetric aspheres, it turns out that sampling grid do not have a particular significance, as all the fitting grids produced very good approximants with a small number of Zernike terms and samples. For highly varying freeform surfaces (e.g. surfaces with no specific symmetry) like Franke or surfaces with mid spatial frequencies, only the edge-clustered sampling method achieved sub-nanometer accuracies. Since this work was completed, Forbes used this finding to fit mid spatial frequency defects on optical surfaces with a Cheby-

polar like fitting grid, however points were displaced from the original “spoke structure” to simulate some equivalence to the randomness similar to edge clustered fitting grid shown in this work [41]. The RMS errors produced by edge clustered fitting grids stably and exponentially decreased as the number of point samples was gradually increased for freeform optical surfaces. In the test cases considered, edge clustered random sampling has achieved in 2D what a Chebyshev-radial spaced sample achieves in 1D in removing the impact of edge ringing.

CHAPTER FOUR: COMPARISON OF FREEFORM POLYNOMIALS

In this chapter, we have comparatively assessed optical surface characterization with two different ϕ -polynomials, namely the widely used Zernike polynomials and the recently introduced gradient orthogonal Q-polynomials in terms of the least squares approximation. Various forms of polynomials for describing freeform optical surfaces exist in optical design and to support fabrication. Among the several forms of optical surface description, a popular method is to add orthogonal polynomials onto a conic section; the latter can equally be a sphere if the conicity is zero. In order to achieve numerical robustness when higher-order polynomials are utilized to describe freeform surfaces, recurrence relations are a key enabler. A detailed review of ϕ -polynomials (both Zernike and gradient orthogonal Q-polynomials) and their associated recurrence relations are given in chapter 2 of this dissertation. Results shown in this chapter establish the equivalence of both sets of ϕ -polynomials in accurately describing freeform surfaces with clustered sampling grids in terms of least squares. Quantifying the accuracy of these two freeform surface descriptions is a critical step in the future application of these tools in both advanced optical system design and optical fabrication.

This chapter is organized as follows: In the next section, two benchmark test cases based on an F/1 parabolic surface with generic asymmetric features are described along with two different ray-based sampling strategies. In the succeeding section, we show the results of performing least-squares fits of these analytical surface test cases by utilizing the two sets of ϕ -polynomials with two different sampling grids. Prior to concluding this chapter, the effect of

heights of the nonsymmetric features on the Root Mean Square (RMS) fit residual is investigated. The major content of this chapter is published in our article [21].

Ray-grids for Data-site Sampling and Test Cases

In chapter 2, we have shown some examples of sag representation with the recently introduced gradient orthogonal Q-polynomials and Zernike polynomials (see Eq. (2.30), Eq. (2.31), and Eq. (2.32)). In order to achieve numerical robustness and for the removal of the ill-conditioning associated with the higher order ϕ -polynomials, throughout this chapter the recurrence relations are utilized, which are given in chapter 2. A brief review of surface descriptions with ϕ -polynomials is provided in chapter 2.

We showed previously that, in the context of fitting a set of polynomials to a continuous analytical surface, edge clustered fitting grids demonstrate the best efficacy in fitting optical surfaces in a least-squares sense [18]. Thus, we will make use of an edge clustered fitting grid in our performance evaluations. Edge clustered sampling is created by first generating random Halton points and then applying a sine function on the radial coordinate to move these points towards the boundary of the aperture, as described in chapter 3.

To continue to illustrate the effectiveness of edge clustered ray grids and to enable a comparison with earlier evaluations, we also provide results using hexagonal subgrids centered on a uniform rectangular grid. We have sampled the optical surface with hexagonal subgrids rather than rectangular subgrids as the circular aperture is more uniformly covered with this strategy. In Figure 24, we have illustrated two examples of the sample grids that will be used in the ϕ -polynomials comparisons. Throughout this dissertation, when comparisons are made

between two sets of ϕ -polynomials with two different sampling grids, we make sure that we use about the same number of samples on each of these grids. The number of samples on each grid is empirically determined as approximately $9 \cdot k^2$, where k is the highest order of the ϕ -polynomials in the least square fit.

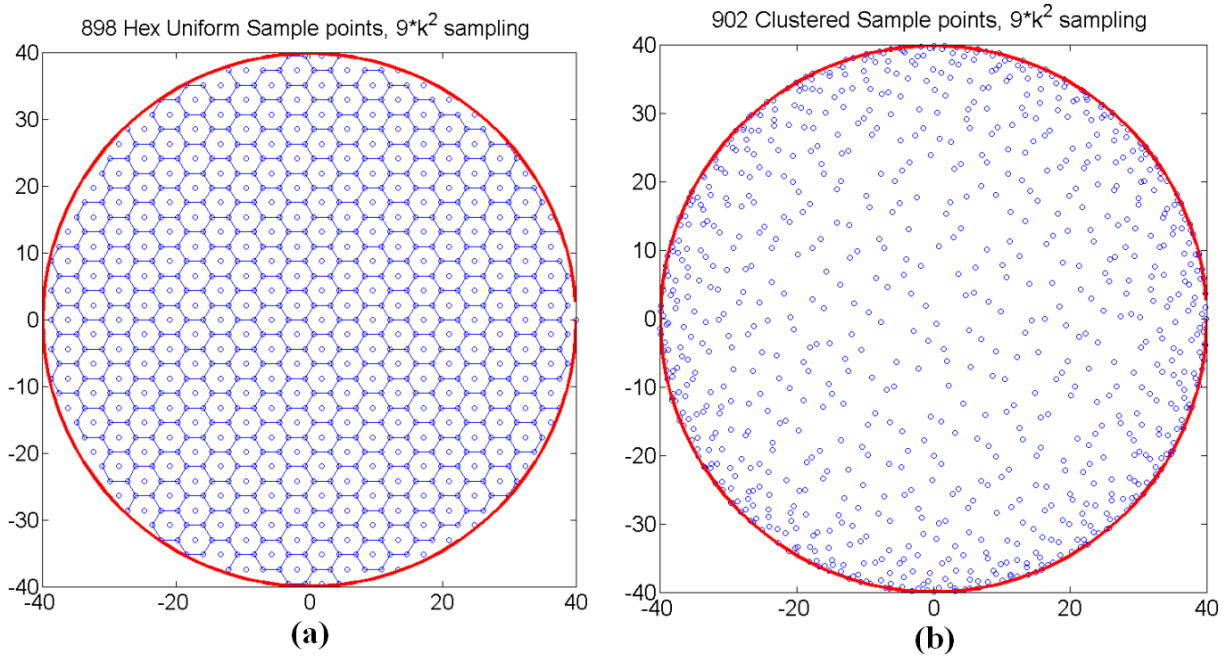


Figure 24 Two types of ray-grids used for ϕ -polynomials fitting with about 900 rays in this figure: (a) Hexagonal uniform (b) Edge clustered.

In order to investigate the effectiveness of the gradient-orthogonal Q-polynomials versus the Zernike polynomials using the ray grids given in Figure 24, we have formed a benchmark test suite consisting of analytical functions. The first test case is an F/1 parabola with a Gaussian bump away from the edge of the aperture. The second test case is again the same F/1 parabola with the Gaussian bump placed now at the edge of the parabola. The aperture diameter for the F/1 parabola is chosen to be 80 mm. The Gaussian bump is 12.5 μm in height and has a

2.357 mm standard deviation. The analytical definitions for the test cases are given in Eq. (4.1) and Eq. (4.2) as

$$f_1(x, y) = \frac{x^2 + y^2}{320} + 0.0125e^{-0.09[(x-20)^2 + (y-5.5)^2]}, \quad (4.1)$$

$$f_2(x, y) = \frac{x^2 + y^2}{320} + 0.0125e^{-0.09[(x+26)^2 + (y-26)^2]}, \quad (4.2)$$

where f_1 represents the F/1 parabola with the Gaussian bump away from the edge, and f_2 represents the F/1 parabola with the Gaussian bump near the edge. To illustrate these two test cases, we plotted the sag departure from a best-fit sphere in Figure 25.

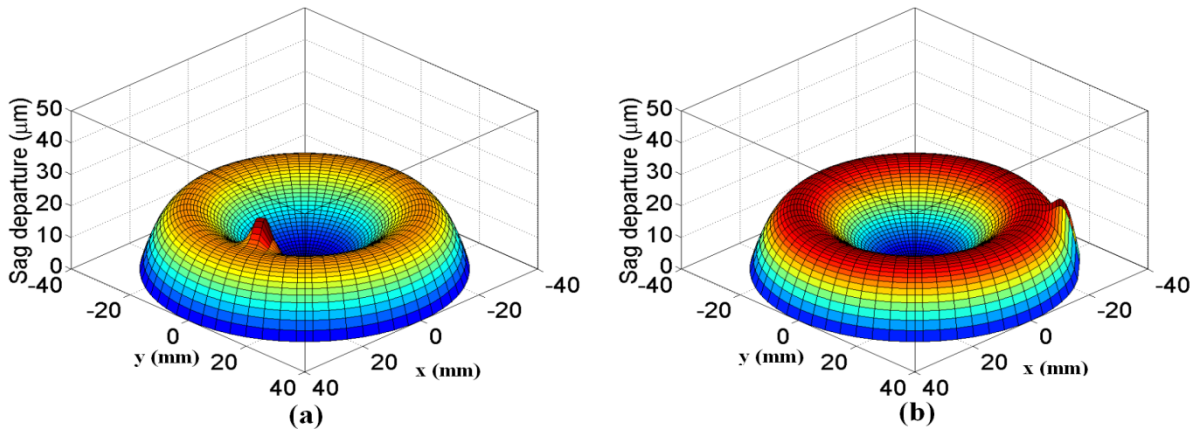


Figure 25 Sag departure from the best fit sphere (bfs) : (a) f_1 -bfs, F/1 parabola with the Gaussian bump away from the edge (b) f_2 -bfs, F/1 parabola with the Gaussian bump near the edge of the aperture.

Figure 25 (a) illustrates the sag departure of the first test case that is an 80 mm diameter aperture F/1 parabola with the Gaussian bump away from the edge of the aperture. Figure 25 (b) shows the sag departure of the second test case that is an 80 mm diameter aperture F/1 parabola with the Gaussian bump near the edge of the aperture.

Similar to the Eq. (4.2) in [24], the curvature of the best-fit sphere (bfs) is computed as

$$c_{bfs} = \frac{2\langle f(\rho_{\max}, \theta) \rangle}{\rho_{\max}^2 + \langle f(\rho_{\max}, \theta) \rangle^2}, \quad (4.3)$$

where the angle brackets denote the average of the sag at the edge of the aperture over θ . Because for freeform surfaces the sag also depends upon θ , correctly computing the curvature of the best-fit sphere has a profound effect on the computations associated with fitting of surfaces with gradient-orthogonal Q-polynomials, especially when the surface to be fitted has asymmetric components at the edge of the aperture.

Numerical Simulations for the Efficacy of Zernike versus Q-polynomials

We investigated the fidelity of creating freeform optical surface descriptions based on the gradient-orthogonal Q-polynomials and the Zernike polynomials with data points sampled on the hexagonal uniform and the edge clustered grids. We have carried out the least squares fits with the increasing numbers of basis elements (coefficients). The relation between the number of samples and the number of basis elements was established empirically as $9*k^2$, where k is the highest order of the polynomial in the polynomial fit. Truncation of the sums is carried out based upon the condition $k < T$, for some given integer T , and k equals $m+2n$ for the gradient-orthogonal Q-polynomials.

In Figure 26, we illustrated the effect of sampling on the fidelity of the surface representation with both sets of ϕ -polynomials for the f_I test case. We found that both sets of ϕ -polynomials performed about identically for this test case. We have made use of

approximately 54845 samples and 3320 elements from either set of ϕ -polynomials. We have seen that for the f_1 test case, which is the 80 mm diameter F/1 parabola with a Gaussian bump placed away from the edge of the aperture, hexagonal uniform ray grids yield for both ϕ -polynomials a Peak to Valley (PV) fit residual around ~ 10 nm (see Figure 26 (a)). Edge clustered ray grids result in a remarkable improvement on the overall fit residual profile, as shown in Figure 26 (b). Both sets of the ϕ -polynomials produced PV fit residuals on the order of sub-nanometers with edge clustered ray grids.

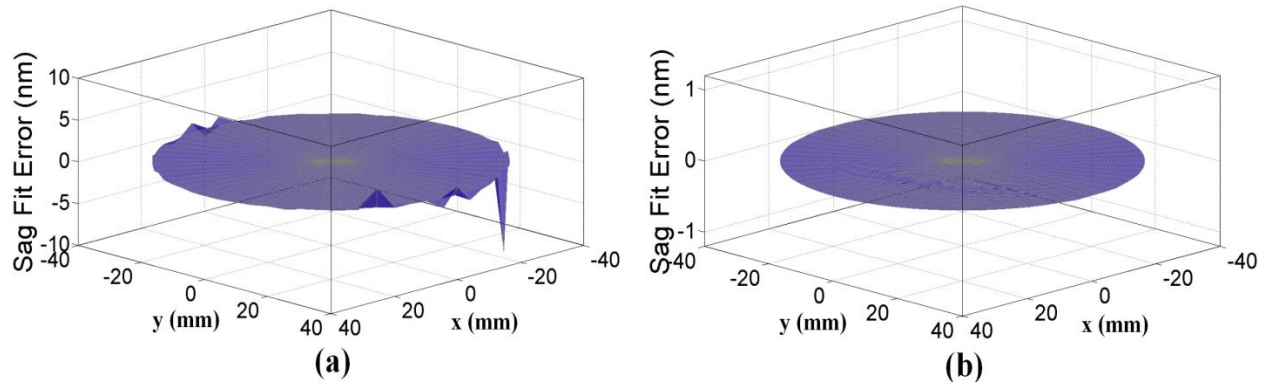


Figure 26 Sag fit residual profiles for f_1 ; the F/1 parabola with a Gaussian bump away from the edge of the aperture with $T=80$; (a) fit residual with hexagonal uniform sampling, (b) fit residual with edge clustered sampling. The gradient-orthogonal Q-polynomial and the Zernike polynomial representations give indistinguishable results, so only one is shown.

In Figure 27, we have displayed the effect on the fit residual for test case f_2 , when the Gaussian bump is placed near the edge of the aperture. Also in this case, the fit residuals for gradient-orthogonal Q-polynomials and Zernike polynomials are indistinguishable. In Figure 27 (a), the hexagonal uniform ray grid is used to create samples for the least squares fitting, and we observe that the PV fit residuals are around ~ 4 nm with the gradient-orthogonal Q-polynomials and Zernike polynomials. The outcome is more compelling with the edge

clustered ray grid, which increases the density of samples towards the edge of the aperture. As seen in our earlier work [18], this ray grid strategy significantly reduces the PV fit residuals, which is observed in Figure 27 (b) as both ϕ -polynomial sets produced a sub-nanometer fit residual with the edge clustered sampling grid.

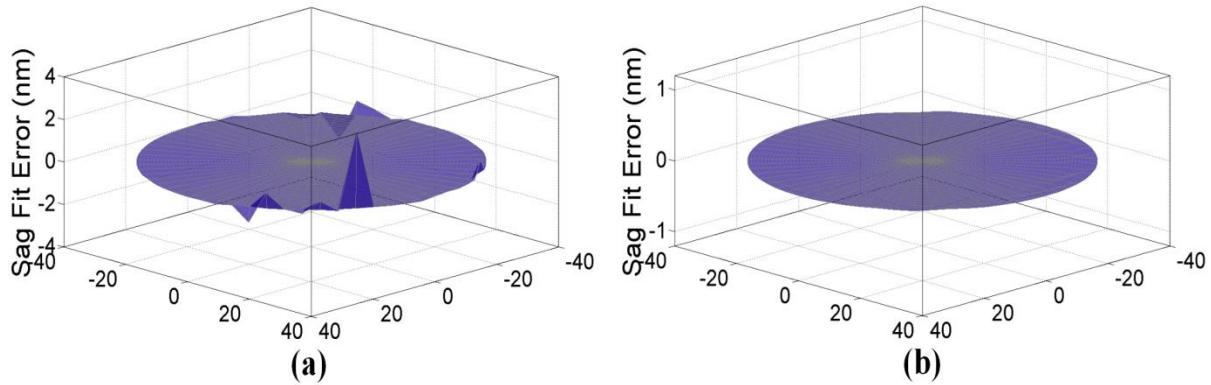


Figure 27 Sag fit residual profiles for f_2 ; the F/1 parabola with Gaussian bump near the aperture edge with $T=80$; (a) fit residual with hexagonal uniform sampling, (b) fit residual with edge clustered sampling. Zernike and gradient-orthogonal Q-polynomials perform similarly, so only one is shown.

In Figure 28, we have compared RMS fit residuals resulting in fitting test cases f_1 and f_2 with hexagonal uniform and edge clustered ray grids with the two sets ϕ -polynomials. We have gradually increased the degree of the Zernike polynomials and the gradient-orthogonal Q-polynomials as the truncation parameter in the sum is moved from $T=5$ to $T=80$ in steps of 10. As the number of basis elements goes up from 19 to 3319, the number of data samples in the fit increases from 226 to 54845. As stated previously [18], we observe that the edge clustered sampling consistently produces better fits when compared to the hexagonal uniform sampling as demonstrated by the solid black lines that are compared to the solid or dashed blue lines in Figure 28 (a) and Figure 28 (b). We have also shown that gradient-orthogonal Q-polynomials

and Zernike polynomials produced effectively exact representations with edge clustered sampling for both the less stressing f_1 case, with the bump away from the edge, and the more stressing f_2 case, with the bump near the edge of the aperture as marked with solid black lines in Figure 28 (a) and Figure 28 (b).

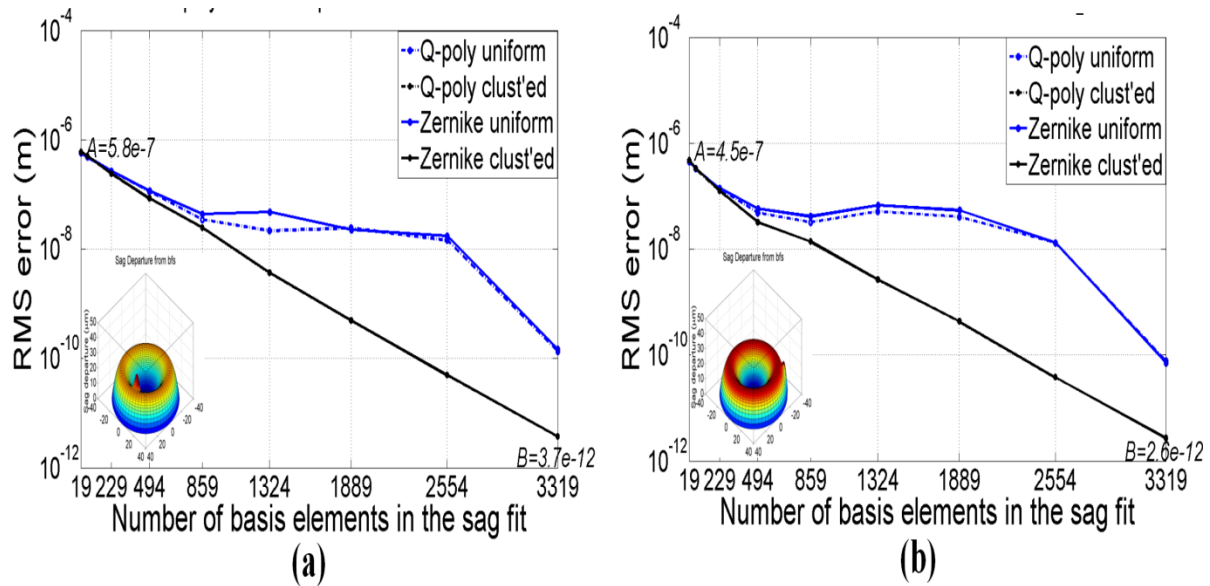


Figure 28 Comparing Zernike and gradient-orthogonal Q-polynomials as freeform surface representations. The fidelity is investigated with both edge clustered and hexagonal uniform sampling in the case of fitting analytical functions with these surface descriptions; the evolution of the RMS fit residual vs. the number of coefficients for the test case (a) f_1 , F/1 parabola with the bump away from the edge, (b) f_2 , F/1 parabola with the bump at the edge.

In the case of fitting an analytical surface using the hexagonal uniform sampling, which is known to be less efficient, the gradient-orthogonal Q-polynomials fit residuals are slightly more accurate than that of the Zernike polynomials for both the f_1 and the f_2 test cases, as shown by the dash-dot blue curves in Figure 28 (a) and Figure 28 (b). Zernike polynomials and gradient-orthogonal Q-polynomials combined with edge clustered sampling consistently produced significantly better fits as the maximum degree of the polynomial is increased from

T=5 to T=80 (see the black solid curves in Figure 28 (a) and Figure 28 (b)). For both test cases f_1 and f_2 , the ϕ -polynomials fits reached the required subnanometer levels (see point B in Figure 28 (a) and Figure 28 (b)). The gradient-orthogonal Q-polynomials performed as well as Zernike polynomials in achieving the accuracy levels for the optical surface descriptions, as illustrated here test cases f_1 and f_2 .

The Effect of Irregular Surface Features Height on ϕ -polynomials Surface Description Efficacy

We expanded the test case study given in Figure 28 to quantify the fit residuals for Zernike polynomials and gradient-orthogonal Q-polynomials by systematically doubling the height of the Gaussian bump. We have quantified the minimal RMS fit residuals in the fits for when the truncation point in the expansion is determined by T=80, with hexagonal uniform and edge clustered sampling with both ϕ -polynomials sets for the height of the bump set at 12.5 μm , 25 μm , 50 μm , and 100 μm as shown in Figure 29.

In Figure 29, dash-dot lines show the RMS fit residuals in the least-squares approximations with gradient orthogonal Q-polynomials whereas solid lines illustrate the performance of the Zernike polynomials. Results show that there is a linear relationship between the minimum RMS fit residual and the height of the bump. Specifically, in Figure 29 (a) that addressed a bump away from the edge of the aperture (i.e. case f_1) Point A shows the RMS fit residual when the height of the bump is 12.5 μm using Zernike polynomials with edge clustered sampling. Point B shows the RMS fit residual when the height of the bump is 100 μm . The RMS fit residual increased from 4.5×10^{-12} m to 3.6×10^{-11} m that is 8 times. An equivalent relation is also found for the Points C and D. Moreover, we observe that with edge clustered sampling both

ϕ -polynomials produced two orders of magnitude better RMS fit residuals when compared with either ϕ -polynomials performance with hexagonal uniform sampling (see blue and black curves in Figure 29). The Point A records a RMS fit residual 4.5×10^{-12} m; Point C shows 1.8×10^{-10} m RMS fit residual.

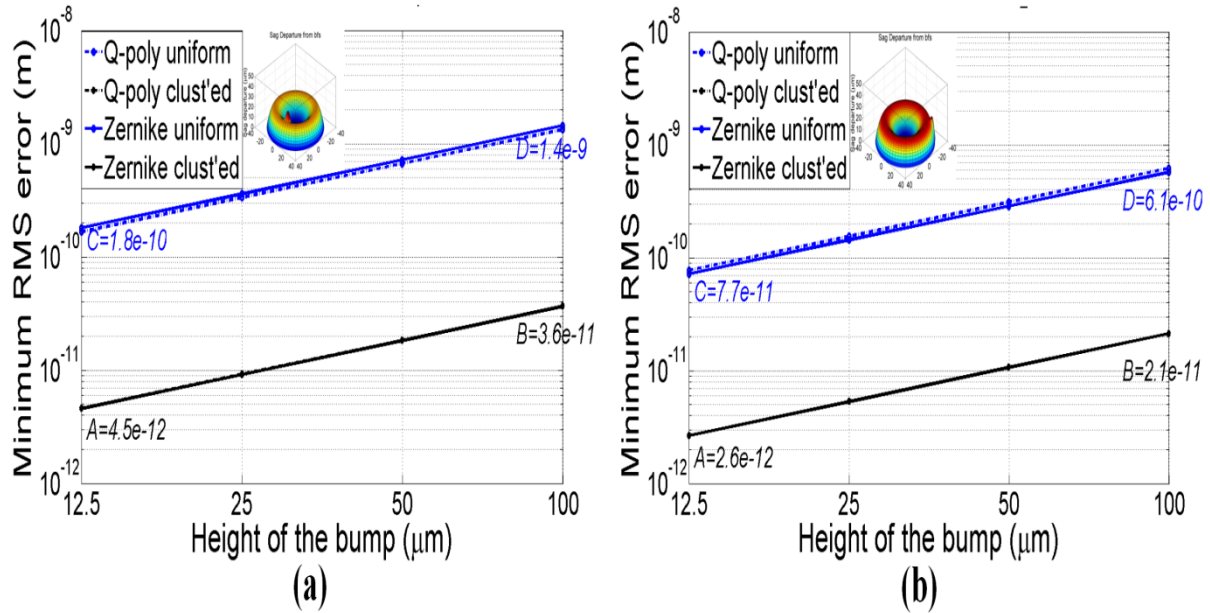


Figure 29 Zernike (solid lines) and gradient-orthogonal Q-polynomials (dash lines) surface approximation performance over a range of heights of the rotationally nonsymmetric bump with hexagonal uniform and edge clustered sampling for the test cases (a) f_1 , (b) f_2 .

The blue curves show RMS fit residuals in the approximants when hexagonal uniform sampling is used for both sets of ϕ -polynomials. In Figure 29 (a) the blue dash-dot curve is slightly lower than the solid blue line indicating the gradient-orthogonal Q-polynomials performed slightly better, while not significantly, with hexagonal uniform sampling for the test case f_1 . Similarly for Figure 29 (b), the Zernike polynomials performed slightly better, while not

significantly, with hexagonal uniform sampling for the test case f_2 (see blue lines in Figure 29 (b)). Similarly the black curves demonstrate the improved performance with edge clustered sampling. We can see for both Figure 29 (a) and Figure 29 (b) that the black dash-dot line and the black solid line coincide, which suggest that with edge clustered sampling, Zernike polynomial and gradient-orthogonal Q-polynomials provide fits with identical fidelity for the test cases f_1 and f_2 .

Conclusion

In this chapter we have seen that in order to achieve an acceptable ϕ -polynomial fit to an asymmetric localized feature any single additive polynomial requires many terms, on the order of thousands, if subnanometer accuracy is required, as is the case in precision optics. We have also observed that Zernike polynomials placed additively on a base conic section and gradient-orthogonal Q-polynomials with a best fit sphere base are able to equally represent the nonsymmetric features of the surface no matter where these features might be positioned over a significant range of feature height and slope. One crucial step working with Q-polynomials is to accurately calculate the curvature of the best fit sphere, which later on affects the sag computation significantly (see Eqs. (4.3) and (2.32)). Also, in both cases, the use of recurrence formula is a key enabler to nanometer accuracy when representing high frequency features in an aperture.

In all the analyses carried out, we have used least-squares methods in arriving at the coefficients of fit. In a real optical design environment, these approximations are the results of optimization procedures involving not only the polynomials, but also their first and second

derivatives. Hence a next level of comparison takes into account the first and second derivatives of the polynomials under evaluation. Also, the offset Gaussian bump may be considered as a possible extreme feature to fall beyond a departure that would be seen in a freeform optical design for an imaging application. In addition, while representing a surface with thousands of coefficients such as given in this paper currently exceeds the capabilities of commercial optical design optimization, it does not exceed their analysis. An alternative to thousands of terms for representing a generic asymmetric feature, while perhaps not as narrow as in this chapter, is to consider using a number (tens) of multi-centric additive bases. An initial evaluation of this approach is found in [10].

The capability to fabricate rotationally nonsymmetric surfaces for imaging applications is a new capability for the industry and as a result there are currently few examples. However, the generation of aspheric surfaces with small tool grinding and polishing provides an early set of surface examples that often suffer from significant mid spatial frequencies. Also bump generation with small tools polishing may occur during the fabrication process. For this study, the stressing asymmetric surface was used to establish that there are no limits to the application of the results in the context of current or future rotationally nonsymmetric surfaces in image forming optical systems. The offset Gaussian bump may be considered as a possible feature during fabrication if considered as an isolated bump. However, there is some anticipation that the Gaussian bump used in this simulation could represent a limiting spatial frequency in the aperture, but, as part of an imaging surface departure, it would be expected that there may be tens of, or perhaps even hundreds of features with this limiting geometry on a future surface. Future work will investigate the application of the tools developed under this work to fitting mid spatial

frequencies on measured surface data with the goal to set tolerances for fabrication. The application of freeform surfaces in advanced optical system design also requires establishing quantitatively the equivalence between various freeform surface descriptions.

Finally we can clearly observe with all the experiments carried out in this chapter that it is not so much the type of the ϕ -polynomials but the type of the sampling grid that dominates the magnitude of the errors in RMS fit residuals, thus the level of accuracy obtained through the fitting process.

CHAPTER FIVE: HYBRID RADIAL BASIS FUNCTIONS AND LOCAL ϕ -POLYNOMIALS METHOD

In this chapter, we develop an efficient and accurate localized hybrid optical surface characterization method combining in one implementation assets of both Radial Basis Functions (RBFs) and ϕ -polynomials, in another implementation RBFs are locally deployed instead of ϕ -polynomials with a locally varying spatial shape optimization. This local description method is based upon the partitioning of an aperture into smaller subapertures as opposed to a global description. The new method not only has a striking significance in the reduction of the order of ϕ -polynomials terms used for the description of optical surfaces but also it is applicable to any overall shaped aperture. Initial results show that the proposed method yields sub-nanometer accuracy with as few as 25 terms of local ϕ -polynomials used in each subaperture. Sub-nanometer accuracy is required for the stringent conditions of lithography and related precision optics applications. Less stringent conditions are also shown to be achieved with as few as 16 terms ϕ -polynomials deployed in each subaperture. The content for this chapter was recently published in [19].

The method is based upon the partition of unity principle employing RBFs as weights for local partitions, and ϕ -polynomials or RBFs as local surface descriptors for freeform optical surfaces. The concept of local shape descriptors is not new as it is central to differential geometry [42]. In optics, it has been adopted in surface metrology using curvature sensing [43, 44] and in some cases combined with wavefront reconstruction [45, 46], and stitching interferometry [47].

Because the state of the art optical surface description methodologies are largely covered in chapter 2, we skipped the review of RBFs and orthogonal φ -polynomials in this chapter. This chapter starts with the description of the localized hybrid RBF and φ -polynomials method. In the following section, we describe another implementation of this method with local Gaussian RBFs with local shape optimization. In the section preceding the conclusion, the numerical experiments showing the details of deploying only low-order φ -polynomials to describe a fairly complicated freeform surface are described. The last section concludes this chapter.

Hybrid RBF and Local φ -polynomials Method

Inspired by the intuitive notion of local shape descriptions for freeform surface together with some of the fundamental ideas associated with RBF-QR presented by Fornberg et al. [36] and Fasshauer and McCourt [37], we have developed a hybrid method employing local φ -polynomials as orthogonal polynomials for local surface description and combining the local descriptions based upon these polynomials over circular subapertures with Wendland's compactly supported RBFs (CSRBFs) as a global description. In this form, this method may be applied to any overall shaped aperture. Conceptually, the method can be thought of as follows. Instead of translating RBFs with their associated centers over the aperture of the optical elements, we translate the coordinate origin of the φ -polynomials to the centers of the local circular subapertures. Then we carry out a polynomial regression fit over the local subapertures. The contributions of each subaperture are combined with Wendland's CSRBFs that serve as the weights in order to render the overall surface description. Wendland's CSRBFs have been used as weights for the local partitions in the context of RBF interpolation techniques [39]. We

summarize here a partition unity method [39] with a local ϕ -polynomials approximation for freeform optics surfaces. It is important to note that accuracy obtained over the local subapertures is carried over to the global description as for any other partition of unity method. The algorithm associated with this hybrid method for the description of the freeform surfaces can be summarized in four steps:

– *Step 1:* Decompose the domain into smaller circular subapertures and record the centers and radii of the subapertures. Depending upon the required accuracy over the overall fit, the radii of the subapertures can be adjusted. Sample domain decomposition is shown in Figure 30.

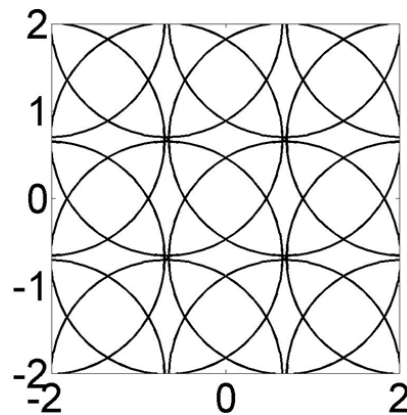


Figure 30 Domain decomposition with circular subapertures of radius 1.33 mm over a 4 mm x 4 mm square aperture.

– *Step 2:* For each point where the global fit is evaluated, find the subapertures that it belongs to and build a weight matrix. The weight matrix is built to identify the contributions of local fits to the overall global fit. The point may be located at the intersection of the more than one overlapping subapertures. In this case, each intersecting subaperture will contribute to the overall surface at the points in the intersection according to the weights defined in this step. The radii of the subapertures match the compact support of Wendland’s CSRBFs. Wendland’s CSRBFs are

utilized for the weight assignment as they provide sparse band-diagonal approximation matrices through omitting the points falling beyond their compact support. This approach makes the method even more local as compared to that of Gaussians, because Gaussians include a tail section spanning the whole aperture. This is especially useful when large sets of sampling and evaluation points are used. A Wendland's C^2 CSRBF, as a weight function, is given as [39]

$$w(\mathbf{x}_i, \mathbf{x}) = \left(1 - \varepsilon \|\mathbf{x} - \mathbf{x}_i\|\right)_+^4 \left(4\varepsilon \|\mathbf{x} - \mathbf{x}_i\| + 1\right), \quad (5.1)$$

where \mathbf{x}_i denotes the center of the subaperture that the point \mathbf{x} falls within. The ε is the shape parameter as before. The subscript after the first term shows that there is a cut-off after the compact support in the function declaration. In Figure 31, several Wendland's CSRBFs that might be used as weight functions are shown.

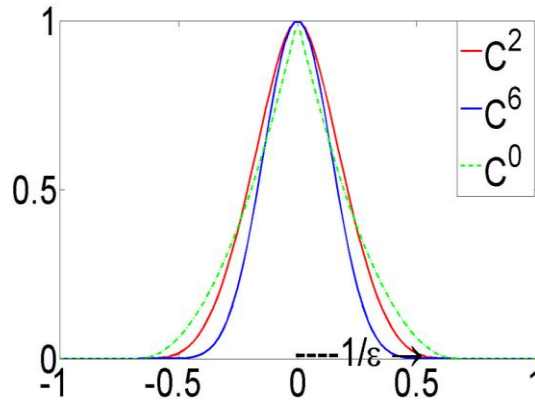


Figure 31 Wendland's CSRBFs for weight assignment, adapted [39].

Weights are assigned after they are normalized according to the *Shepard method*, a moving least-squares method, such that contributions from multiple subapertures add up to unity.

The Shepard method is formulated in [39] as

$$s(\mathbf{x}) = \sum_{i=1}^N f(\mathbf{x}_i) \frac{w(\mathbf{x}_i, \mathbf{x})}{\sum_{j=1}^N w(\mathbf{x}_j, \mathbf{x})}, \quad (5.2)$$

where \mathbf{x}_i denotes the center of the subaperture that the point \mathbf{x} falls within, $f(\mathbf{x}_i)$ represents the contribution of i^{th} intersecting subaperture (local approximation in the i^{th} subaperture), and $s(\mathbf{x})$ is the final outcome at point \mathbf{x} , i.e. global approximation.

– *Step 3*: For each subaperture, carry out a local least squares approximation with local ϕ -polynomials shifted to the centers of subapertures. A least-squares matrix can be formed with local samples within each subaperture and as many ϕ -polynomials as desired with the recurrence relations. We established, for example, that a small subset of FRINGE Zernike polynomials provides subnanometer accuracies within each subaperture. To speed up the determination of sample and evaluation points within each subaperture, a kd-tree data structure [48] is utilized for all sample and evaluation points separately after the domain decomposition step. Local samples and local evaluation points are found by querying the kd-trees for each subaperture. In this way, we can locate the points inside the subaperture in a fast and efficient manner. The algorithm complexity reduces from an $O(N)$ procedure to an $O[\log(N)]$ process.

– *Step 4*: By combining the local surface descriptions computed in step 3 with the normalized weights computed in step 2, stitch the overall surface description. With the weights and local surface descriptions computed in previous steps, this step reduces to accumulating weighted local results.

Hybrid RBFs and Local RBFs with Locally Shape Optimization

A variation of the hybrid RBF and local ϕ -polynomials method is obtained by using Gaussian RBFs instead of local ϕ -polynomials for the local surface descriptions with locally optimized shape parameters in subapertures. Step 3 of the algorithm presented in previous section is modified to approximate the surface with local Gaussian RBFs in place of local ϕ -polynomials. The contribution of this method is its capability to allow utilizing a locally varying spatial shape parameter assigned to the each subaperture as opposed to having a unique shape parameter used in global RBF approximations. We have optimized this contiguously varying spatial shape parameter for each subaperture in order to yield minimum least square approximation errors. After the local samples and local evaluation points for each subaperture are determined through querying the kd-trees, a local least squares approximation matrix is formed with these local samples and as many as local Gaussian RBFs as required. In our implementations, local Gaussian RBF centers are created uniformly across each subaperture in addition to sample and evaluation points. When a subaperture is at the intersection with the aperture boundary, then Gaussians RBF centers are shifted into the part of the subaperture located inside the aperture.

The local approximations with the Gaussian RBFs with uniformly located centers is optimized over the shape parameter w_i to achieve best accuracy over each subaperture. An approximation with locally optimized Gaussian RBFs can be formalized as

$$f(\mathbf{x}) = \sum_{n=1}^N \phi\left(\omega_i^2 \|\mathbf{x} - \mathbf{x}_n\|_2\right) a_n, \mathbf{x} \in R^s, \quad (5.3)$$

where a_n represents the weights in the approximation, \mathbf{x}_n represents the centers and \mathbf{x} a point in the subaperture, w_i is the shape factor constant across the subaperture, and ϕ are the local RBFs within the subaperture. A brute force approach is adapted for shape parameter optimization within the subapertures. For each subaperture, the shape parameter interval is divided into equally spaced points and for each shape parameter an approximation is carried out within the subaperture. The approximation which minimizes the least square error is selected as the best shape parameter value for that subaperture. This way locally varying spatial shape parameters are optimized for each and every subaperture. An illustration of this method with 16 subapertures and 16 different optimized shape parameters is shown in Figure 32.

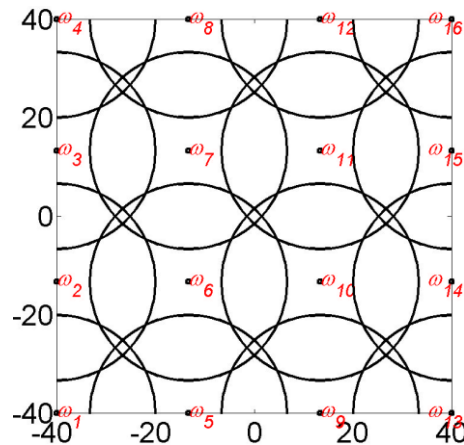


Figure 32 Locally optimized shape parameters for hybrid RBFs.

After optimizing shape parameters for the best RBF approximation in each subaperture, the local approximations are combined with the weights calculated according to the CSRBFs and Shepard's method. The global surface description is the accumulation of the weighted local approximations as before.

Numerical Experiments for Hybrid RBF and Local ϕ -polynomials Method

In this section, we describe an application of the hybrid RBF and local ϕ -polynomials method, specifically Zernike polynomials, for the description of an extremely asymmetric surface. The surface is chosen to be a stressing example of departure from rotational symmetry. It does, however, represent a descriptive case for spatial frequency. The surface is an F/1 parabola over an 80 mm \times 80 mm rectangular domain with several 12.5 μm – 100 μm isotropic and anisotropic bumps distributed over the aperture. An analytical description of the surface is given as follows

$$\begin{aligned}
f_1(x, y) = & \frac{x^2 + y^2}{320} + 0.0125e^{-0.09(x-20)^2 - 0.25(y-5.5)^2} \\
& + 0.025e^{-0.49(x+26)^2 - 0.04(y-26)^2} + 0.05e^{-0.01(x)^2 - 0.01(y)^2} \\
& + 0.05e^{-0.09(x+33)^2 - 0.49(y+12)^2} + 0.1e^{-0.04(x-18)^2 - 0.09(y+39)^2} \\
& - 0.1e^{-0.09(x-28)^2 - 0.01(y-5)^2} - 0.05e^{-0.04(x+15)^2 - 0.09(y)^2} \\
& - 0.1e^{-0.0144(x)^2 - 0.0064(y-22)^2} .
\end{aligned} \tag{5.4}$$

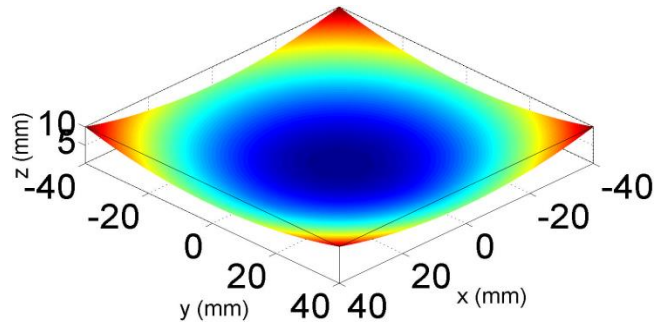


Figure 33 The F/1 Parabola where $12.5\ \mu\text{m} - 100\ \mu\text{m}$ bumps may be visualized in Figure 34.

In Figure 33 we show the test function, the F/1 parabola with several bumps over the rectangular aperture, to mainly show the overall sag of the surface. In Figure 34, several isolated bumps on the surface are shown. There are several radially symmetric and anisotropic bumps of different heights in the range between $12.5\ \mu\text{m}$ and $100\ \mu\text{m}$. We have sampled the representative freeform surface with 600×600 uniform samples, and evaluated the overall fit with 120×120 uniform points.

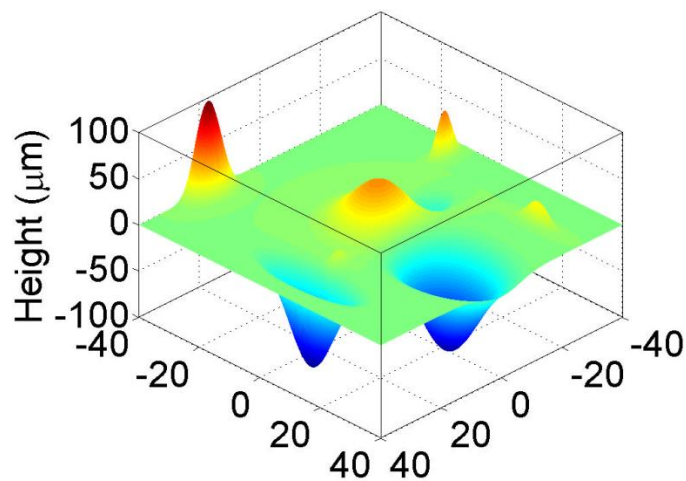


Figure 34 $12.5\ \mu\text{m}$ to $100\ \mu\text{m}$ isotropic and anisotropic bumps on F/1 parabola over an $80\ \text{mm} \times 80\ \text{mm}$ square aperture.

The shape parameter for the weight function, which is a Wendland's C^2 CSRBF, is 1.25 mm^{-1} . As for the local approximations, we have used 36 FRINGE Zernike polynomials. We have decomposed the domain with 100×100 overlapping circular subapertures. The radius of each subaperture is $800 \text{ }\mu\text{m}$. In Figure 35, we show a subregion of the aperture that is located within -2 mm to 2 mm . The decomposition of the subregion with circular subapertures of radii $800 \text{ }\mu\text{m}$ is also exhibited in Figure 35. We can observe the uniform distribution of samples (blue points) along with a grid of uniform evaluation points (red points). As shown in chapter 4, there is actually no significant difference in surface approximation performance of ϕ -polynomials with uniform or clustered samples when the number of ϕ -polynomials is low and the number of samples are abundant (see Figure 28), such as in this case 36, we make use of a uniform grid of samples across the overall aperture.

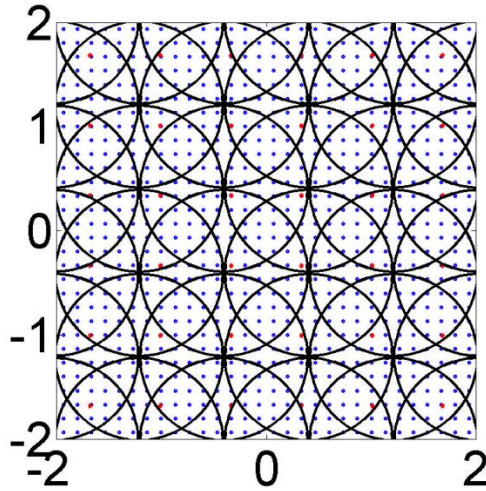


Figure 35 Decomposition of the aperture of an F/1 parabola into circular subapertures of radii $800 \text{ }\mu\text{m}$ along with uniformly distributed sample points, shown only for a -2 mm to 2 mm subregion.

Although the number of the FRINGE Zernike polynomials used in local subapertures is only 36, we have obtained excellent approximation errors on the orders of a subnanometer. In

Figure 36, we show the Peak to Valley (PV) approximation error profile for the F/1 parabola with several bumps. Results show that the approximation PV errors are less than or equal to around 0.3 nm. There are no edge-related oscillation errors even with uniform sampling. Nonetheless, errors concentrate around the 25 μm and 50 μm anisotropic bumps whose slopes are the largest [see lines 2 and 3 in Eq. (5.4)]. The overall RMS error for this description is 0.01 nm. Given the oscillatory nature of the errors, even if at subnanometer scale, smoothing of the computed surface would further decrease the magnitude of these errors.

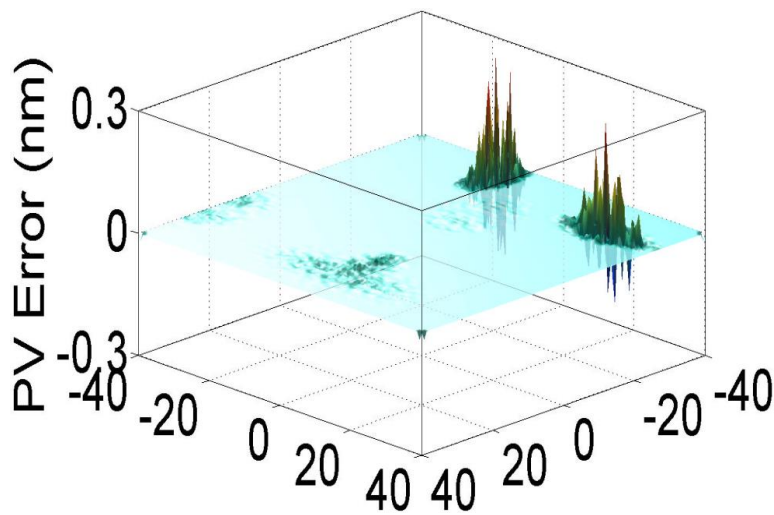


Figure 36 Approximation error profile for an F/1 parabola with several bumps showing the maximum PV errors on the orders of the subnanometer with only 36 local FRINGE Zernike polynomials across an 80 mm x 80 mm aperture.

We have carried out the F/1 parabola with bumps test with 25, 36, and 64 Zernike polynomials within the local subapertures in various surface approximations. We have recorded the radius of the subapertures along with the number of subapertures that is sufficient for reaching the subnanometer PV errors. Table 4 summarizes the overall results. The shape

parameter for Wendland's CSRBFs for each experiment is the inverse of the radius of the subapertures shown in Table 4.

The more Zernike polynomials that are added into the approximation set, the more capable the method becomes in terms of locally describing a freeform surface, and thus we can increase the radius of the subapertures. This trade-off is well captured in Table 4 and Figure 37. In Figure 37 (a), as the number of Zernike polynomials in the local approximation increases from 25 to 64, the radii of the subapertures increase from 610 μm to 1.33 mm. Meanwhile, the number of subapertures decreases from 16900 to 3600, see Figure 37 (b).

Table 4 Subnanometer PV errors with a small set of Zernike polynomials (4th column) or Gaussian RBFs (5th column) in each subaperture.

Cell count	Cell radius (mm)	Number of local basis elements	PV error Zernikes (nm)	PV Error Gaussians (nm)
60x60	1.33	64	0.2	1.01
100x100	0.8	36	0.31	2.29
130x130	0.61	25	0.78	3.59

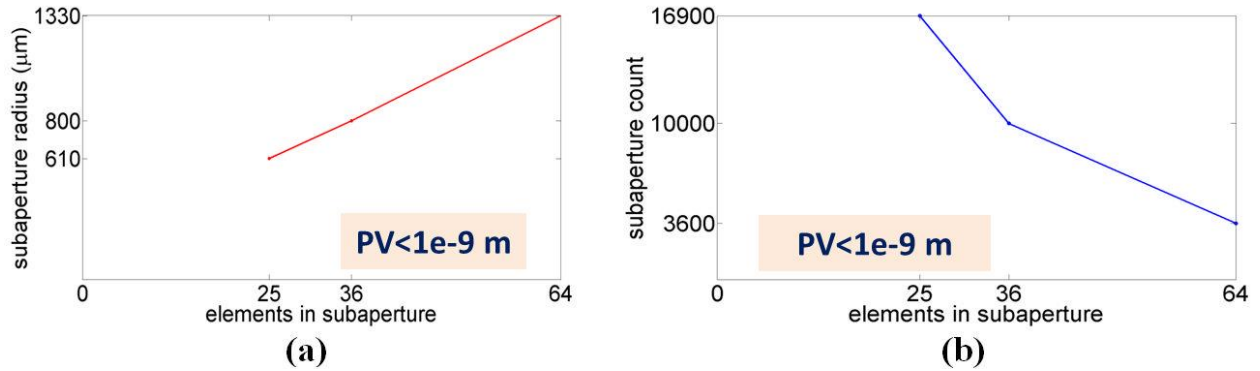


Figure 37 The trade-off in the subaperture radius (a) and subaperture count (b) vs. the number of basis elements within subapertures.

In Table 4, we also compare a simple version of the Gaussian RBFs as local surface approximants to that using a small number of Zernike polynomials in each subaperture needed to represent the surface using the hybrid method in both cases. In this investigation, we included a shape optimization, where w_i was varied from 0.01 mm^{-1} to 10 mm^{-1} for Gaussians (with uniformly distributed centers) over each subaperture. Every other parameter, i.e., the number of samples and their uniform distribution, was kept the same. As we varied the number of Gaussians in each subaperture, the results show that although local Gaussians within subapertures provide accuracies around nanometers, Zernike polynomials provide accuracies on the orders of subnanometers. However, Gaussian RBF implementation may still be improved with different sampling, different center distributions and a modified shape optimization. For example in [29], different grid types were applied for approximation with Gaussian RBFs and results concluded that residual errors are comparable to polynomial approximation counterparts. Furthermore in [49], the authors proposed and compared different edge remedies for improving the errors of RBF approximations, including clustering the sample points towards the boundary and two possible Not-a-Knot implementations. Without incurring any additional computational

cost, methods mentioned in [49] for improving Gaussian RBF approximation errors may be applied within each subdomain to further increase the level of accuracy obtained through Gaussian approximations.

In order to achieve more accurate approximants with Gaussian RBFs whose PV errors are shown in table 4, an adaptive approach may be used for the domain decomposition and sampling steps instead of a fixed uniform grid of samples and a uniform decomposition of the overall aperture into subapertures. For example, Driscoll and Heryudono present an adaptive refinement method based upon residual subsampling [30]. This adaptive method clusters the samples and the subapertures around the steep regions, whereas it coarsens the samples and subapertures around the large smooth areas. By clustering the samples towards the steep gradients and locally increasing the density of subapertures around the steep regions, more accuracy can be achieved and significant benefits in terms of cost, especially for local RBF approximations, can be gained, where a local shape optimization can be carried out based on the density of the RBF centers.

Another possible improvement for the Gaussian RBF approximations is to use an approach that is presented in [31] that allows selecting sample and center locations along with an optimum shape parameter for each and every RBF basis. This method has shown significant advantages compared to orthogonal polynomials in 1D [31]. Fornberg and Zuev show that a small set of Gaussian RBFs with properly optimized basis center locations and shape parameters are able to achieve the level of accuracies that is only to be matched by utilizing a high number of Chebyshev polynomials [31].

In a large range of optics applications termed as illumination optics, a surface description is considered acceptable if the surface is approximated with 10 nm accuracy. Hence, as an additional experiment, we have carried out an F/1 parabola with bumps approximation test with 16, 25, 36, and 64 Zernike polynomial terms in local subapertures while recording the radius of the subapertures and their total number in order to reach a PV accuracy of 10 nm. Results reported in Table 5 show that as few as 16 ϕ -polynomials terms in each subaperture can describe this surface with 10 nm PV errors. A similar trade-off to the one shown in Figure 37 also exists between the radius of subapertures and the subaperture count versus the number of local basis elements in each subaperture. As the number of ϕ -polynomials in each subaperture decreases from 64 to 16, the radius of the subapertures decreases from 2.29 mm to 670 μm , whereas the number of subapertures increases from 35×35 to 120×120 to reach 10 nm PV errors.

Table 5 Showing 10 nm PV errors with a small set of Zernike polynomials.

Cell count	Cell radius (mm)	Number of local basis elements	PV Error (nm)
35x35	2.29	64	10.06
57x57	1.40	36	6.07
75x75	1.07	25	9.35
120x120	0.67	16	8.41

In summary, the numerical experiments quantify that a small set of FRINGE Zernike polynomials, 25 in each subaperture, are able to describe the overall surface within a non-circular aperture with the hybrid RBF local φ -polynomials method within subnanometer accuracies. The analysis further shows that fewer polynomials are needed if the requirement on accuracy is loosened such as to satisfy only illumination optics application requirements.

Conclusion

As the optics manufacturing industry is forging ahead in the advancement of their methods, freeform optical elements are going to be key components of optical systems in the near future. In this paper, we describe a fast, efficient hybrid method combining local approximants (i.e., RBFs or φ -polynomials) and RBF global approximants. With this method, we are able to describe a freeform surface by using only 25 FRINGE Zernike polynomials in each subaperture within subnanometer accuracy. With a simple local RBF approximant or with a low order 25 to 64 Zernike basis functions in subapertures, nanometer-level accuracy was achieved. Because of its local nature and the ability to carry the local accuracies over to the overall surface description, this hybrid method reduces the order of the φ -polynomials required to describe a freeform surface. The method is highly efficient mainly because φ -polynomials of three inherent properties. First, the method makes use of Wendland's CSRBFs that are known to handle best the large datasets; also they result in band diagonal approximation matrices that are simple to manipulate in algebraic systems since they do not have a tail section spanning the whole aperture like Gaussians. Second, to find the local samples and evaluation points, we make use of kd-trees location queries, which reduces computational complexity to an $O[\log(N)]$ process. A third

reason for efficiency is the fact that the number of local basis functions is kept to a minimum, which results in small approximation matrices. In the case of φ -polynomials, using a small number of them allowed us using uniform sampling without an approximation performance penalty. Reducing the number of φ -polynomials to a minimum in subapertures and using only lower order φ -polynomials is also important because it facilitates understanding of the local optical properties of the surface for optical designers while providing computational advantages. We also note that there is no inherent limit in terms of the number of local φ -polynomials that may be used in the method; in order to achieve better accuracies than subnanometers, that is, machine precision, high-order polynomials may also be used within the local subapertures. Finally, the Q-polynomials or other forms of φ -polynomials may substitute within local subapertures the Zernike polynomials. However, a crucial step working with Q-polynomials is to accurately compute the curvature of the best-fit sphere, which requires mean sag over the perimeter of the local subaperture as a targeted step into the hybrid algorithm.

A consideration may arise with the total number of the subaperture count. As this method utilizes only the lower order φ -polynomials and it deploys 16 to 64 φ -polynomials in each subaperture, it accomplishes a level of subnanometer surface approximation accuracy with a finer set of subapertures. The total number of basis elements used in the overall surface approximation maybe calculated as the total subaperture count times the local number of basis elements. A global surface fit such as the one described in [41] over a circular aperture may result in a smaller number of terms globally then the subaperture count times the local basis elements in the method shown here. For example, tens of thousands of global φ -polynomials consisting of a major count of high order φ -polynomials may be less than the total subaperture

count times the number of local basis elements. However, in this text we are working with least squares approximations for optical surface descriptions. A least square matrix of a global approximation consisting of tens of thousands of higher order φ -polynomials with a tens of thousands or even hundreds of thousands of samples points means working with a huge least squares matrix in surface descriptions. Even the formation of this matrix is computationally intensive because of the computation of the higher order polynomials even with the recurrence relations leaving aside the computational cost to solve this huge global approximation matrix, which is $O(N^3)$ N being the size of the matrix, i.e. tens of thousands. It might be the case that author in [41] suggests the usage of Fourier methods to compute the fit coefficients due this significant computational cost. As opposed to these global methods, local method such as the one described in this chapter has the advantage of the divide and conquer approach. Decomposition step does not incur any significant cost due to the kd-trees to locate the local samples within subapertures. Also forming and solving the least square matrix of local subapertures consisting of only 25 polynomials and around 120 samples is no cost at all compared to the solution of the global approximation matrix consisting of tens of thousands rows or columns. Also we may keep in mind that computing a lower order φ -polynomial utilized in the local hybrid method shown in this chapter is several times easier than computing a higher order φ -polynomial needed in a global approximation method in terms of least squares. In terms of conditioning of linear systems, lower order polynomials do not suffer from ill-conditioning even when they are computed with explicit expressions. Also round off errors are minimized when working with computational methods for smaller size matrixes.

CHAPTER SIX: ACCELERATION OF COMPUTATION OF φ -POLYNOMIALS

In this chapter, we investigate the benefits of making an effective use of impressive computational power offered by multi-core platforms for the computation of φ -polynomials used in the description of freeform surfaces. Specifically, Zernike polynomials and gradient orthogonal Q-polynomials are implemented through a set of parallel algorithms on Graphical Processing Units (GPUs), with their respective recurrence relations. The results show that more than an order of magnitude improvement is achieved in computational time over a sequential implementation if recurrence-based parallel algorithms are adopted in the computation of the φ -polynomials. The results reported in this chapter are under review in the literature [50].

As the demands of the optical surface descriptions increase, more terms of φ -polynomials may be required in order to accurately express the surface departure from a base surface that may typically be a sphere, a conic, or a best fit sphere. Both the total number of terms and the higher order φ -polynomials themselves become computationally intensive in their inclusion for describing the surface. Furthermore, multi-dimensional optical surface optimization with the full aperture φ -polynomials is a highly challenging and computationally intensive task. Optimization cycles may become a major bottleneck for the optical design process.

In order to reduce the computational time for the computationally intensive scientific problems, such as the computation of φ -polynomials and their utilization in global multi-dimensional surface optimization, GPUs may be utilized with several parallel algorithms. In many different fields of science ranging from computational dynamics [51] to optical imaging

applications [52] GPUs are reported to accelerate applications more than one order of magnitude or achieve the fastest data processing and visualization rates [61]. Through parallel algorithms designed to work on Single Instruction Multiple Thread (SIMT) GPU architecture, the computation of the full aperture ϕ -polynomials may achieve a significant pace by leveraging the commodity graphics hardware. The main contribution presented in this chapter is to devise and implement several recurrence-based data-parallel algorithms for the computation of Zernike and gradient orthogonal Q-polynomials and show that an order of magnitude speedup is possible in the computation of these ϕ -polynomials.

This chapter is organized as follows: In the next section we briefly review the general purpose computation on GPUs. The following section summarizes the details of the parallel algorithms to implement the recurrence relations of ϕ -polynomials on a SIMT architecture. Prior to the conclusion, we show the computational results of executing the specifically designed parallel algorithms for Zernike polynomials and gradient orthogonal Q-polynomials on GPUs and report the speedups as compared to that of a sequential implementation of the recurrence relations on multi-core Central Processing Units (CPU). The last section concludes this chapter.

General Purpose Computation on GPUs

GPUs are invented to tailor the rendering of computer graphics. The rendering of visual scenes, effects and artificial environments on the computer is a computationally intensive task which is inherently parallel. GPUs are specially built to take advantage of parallelism to render these graphics. In the last decade, graphics programmers tricked the GPUs into handling more general purpose computation instead of just executing specific tasks in graphics pipeline.

Executing general purpose programs other than graphics on GPUs is called general purpose computation on GPUs. A detailed survey of general purpose computation on GPUs can be found in [53]. The reason that we are today able to do general purpose computation on GPUs is that the graphics pipeline has been made gradually programmable over the last 15 years. One of the recent realizations of the graphics pipeline on GPUs is shown in Figure 38.

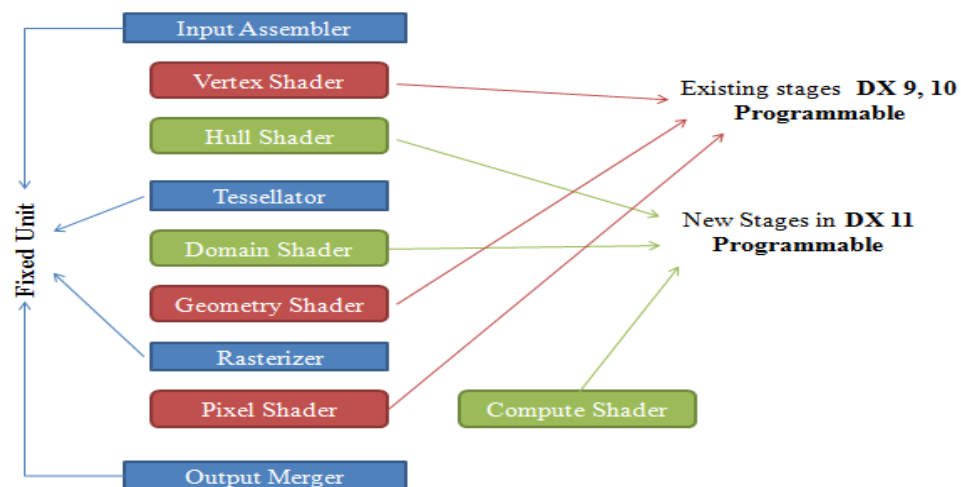


Figure 38 Microsoft's DirectX 11 standard graphics pipeline adapted [54].

Vertices of 3D scenes are fed into this pipeline and pixels of images are obtained as the outcome in the end of the pipeline shown in Figure 38. Each stage of the graphics pipeline has a specialized task to perform, i.e. vertices transformation in the vertex shader, illumination and lighting models in the pixel shader. In early implementations, each stage in the pipeline had its own allocated resources. However recently, all the resources on the GPUs are unified to be allocated on demand. The new type of architectural models is called as Compute Unified Device Architecture (CUDA) shown in Figure 39.

We can see in Figure 39 that a GPU is organized into many streaming multiprocessors. Each multiprocessor consists of many thread processors with a double precision arithmetic logic unit for executing instructions, and a shared memory to act as a mediator for communication among the threads of a thread block.

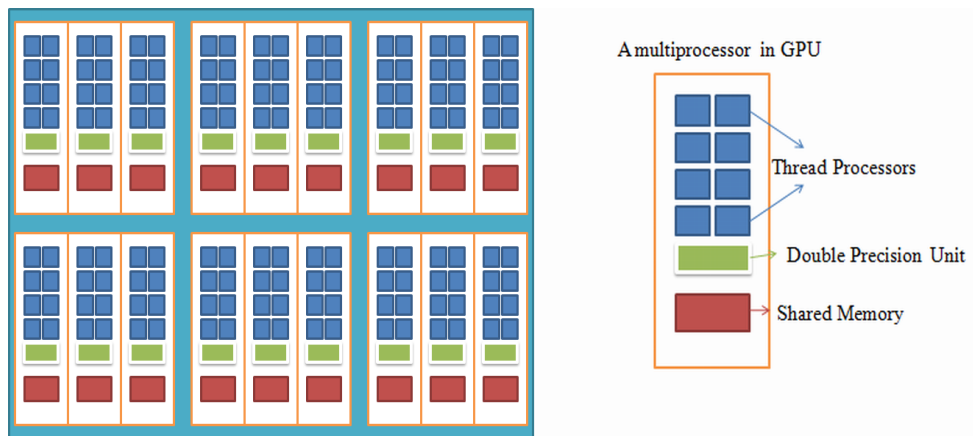


Figure 39 an example GPU architecture adapted [55].

The data-parallel programming model built on top of CUDA to take advantage of parallelism is referred as Single Instruction Multiple Thread (SIMT) model. In order to satisfy thread level synchronization and a fine level of granularity, the threads are grouped into the thread blocks, and all the threads within a block executed on the same multiprocessor. In this way, when a thread branches on the pipeline, only the threads within the block are affected. Threads within the same block use the shared memory for common read and write purposes in synchronization.

With the advent of CUDA also a CUDA C programming language became available to easily harness the power of GPUs in general purpose computation [56]. With this programming language, a programmer without a graphics programming background is able to start taking

advantage of parallelism and computational power of GPUs. As a consequence, the applications taking advantage of general purpose computation on GPUs started to emerge in all fields of science and engineering. For instance, a GPU cluster is used to implement a FEM based seismic wave propagation and 12 to 20 times speedup is reported in [57]. Similarly, Walsh et al. report 10 to 30 times performance improvement in spectral finite element method and least square minimization applications in fluid dynamics and geosciences with general purpose computation on GPUs [58]. Another similar speedup figure (about 10 times) is reported in a medical image segmentation application accelerated through parallelism on GPUs by Kaufmann et al [59]. There are many other applications which we have not listed in this thesis report similar or more speedup figures in different areas of computational science and engineering through parallelizing the applications on GPUs. For further list of examples, please visit the showcase at [60].

Parallelization of Recurrence Relations of ϕ -polynomials

The two types of ϕ -polynomials specifically Zernike and gradient orthogonal Q-polynomials are considered in order for parallelization on GPUs in this chapter. Briefly, Zernike polynomials consist of orthogonal polynomials in the radial direction and Fourier series in the angular direction. The orthogonal polynomials in the radial direction are strongly related to Jacobi polynomials and sometimes they are called one-sided Jacobi polynomials. Zernike polynomials definition in explicit form Eq. (2.22) and Eq. (2.23) and their relation to Jacobi polynomials Eq. (2.24) along with their recurrence relations Eq. (2.26) to Eq. (2.29) are given in chapter 2.

Similarly to the Zernike polynomials, gradient orthogonal Q-polynomials satisfy a recurrence relation (see Eq. (2.43)). Their recurrence relation however is an unconventional recurrence relation that works with an auxiliary polynomial in tandem. The auxiliary polynomial itself satisfies a 3-term recurrence relation given in Eq. (2.37) which is very similar to 3-term Zernike polynomials recurrence. The coefficients for the auxiliary polynomial recurrence relation can be found in chapter 2 (see Eq. (2.38) to Eq. (2.41)). The unconventional recurrence relation for the gradient orthogonal Q-polynomials along with the coefficients for the unconventional recurrence relation can be found from Eq. (2.43) to Eq. (2.45) in chapter 2. In this chapter, we only show the parallel algorithms to implement recurrence relations.

In order to investigate the parallelization and possible speedups in the computation of the ϕ -polynomials, recurrence relations shown in Eq. (2.26) and Eq. (2.43) are parallelized on a SIMT architecture. There are two promising ways to accelerate the computation of the recurrence shown in Eq. (2.26). The first one is that instead of computing the coefficients shown in Eq. (2.27) to Eq. (2.29) sequentially as the recurrence is run, all the coefficients up until the $(n-m)/2^{\text{th}}$ execution of the recurrence relation are computed together at once:

for each thread

get the local id corresponding to the n_f^{th} recurrence run,

compute the rv_{1n_f} , rv_{2n_f} , rv_{3n_f} locally (see Eq.(2.27) to Eq. (2.29))

store them

end

In the above algorithm, each thread operates for a specific run of the recurrence relation, computes the coefficients required only for that specific run. When all the threads return, all the coefficients for the recurrence are ready to use. This is a data-parallel SIMT algorithm, since a single compute instruction is executed on each and every thread with different data corresponding to the specific recurrence runs, nf .

The second way to accelerate the recurrence relation shown in Eq. (2.26), thus the computation of Zernike polynomials shown in Eq. (2.24) and Eq. (2.22), is to compute the recurrence relation on each thread for each sample ray position in the ray grid. In other words, each thread computes the recurrence relation, thus the Zernike polynomial, for each sample location of the rays on the ray grid over the aperture. Each thread not only computes the recurrence relation shown in Eq. (2.26) but also the power term in Eq. (2.24) and sine or cosine terms in Eq. (2.22) on the sample ray point (r, θ) . Hence once all the threads return, the computation of the Zernike polynomial, $Z_n^m(r, \theta)$ is completed across the aperture of the optical element. This data parallel SIMT algorithm is shown below:

for each thread

get the local sample ray point to operate on (r, θ) .

create a local data cache[3].

store cache[0] the first Jacobi, J_0 , cache[1] the second Jacobi, J_1 at (r, θ) .

```

while (recurrence exec num < (n-m)/2)

    run the recurrence, store it cache[2].

    swap cache[0], cache[1], swap cache[1], cache[2].

    recurrence exec number ++.

end

compute power, r^m

compute sine/cosine (mθ).

store the result

end

```

Similar algorithms are written for the gradient orthogonal Q-polynomial recurrence relation shown in Eq. (2.43) and also for the auxiliary polynomial working in tandem. Two recurrence relations are implemented together in parallel for the Q-polynomials. Specific details might be a little different depending on the definition of special cases for the recurrence relation in [12]. Similarly to the Zernike polynomials, the coefficients of the recurrence relation are computed in parallel much the same way as the first algorithm given in the previous page. However not all the coefficients can be implemented in parallel for the gradient orthogonal Q-polynomials because there are interdependencies between the coefficients. In that case a single thread is chosen to compute those coefficients, for instance f and g computation shown in Appendix A of [12].

Numerical Results of SIMT Parallelization of ϕ -polynomials

In this section, we present the results of the implementation of the algorithms shown in the previous section on commodity graphics hardware. We have used MATLAB® and CUDA™ C programming languages [56] to implement and run the parallel and sequential algorithms for the ϕ -polynomials. All the implementations are run on a middle ranking laptop computer with a GeForce™ GT 650M GPU and a CPU of Intel® Core™ i7-3610QM.

The first step is to validate that the result that is obtained out of parallelization is the same as the result that comes out of the sequential algorithm. For this purpose, a low order Zernike and Q-polynomial is computed and the results are compared to sequential counterparts and displayed in Figure 40. Since the sequential and parallel computed ϕ -polynomials coincide visually, only the GPU versions are shown. However, to quantify the differences, Figure 40 also shows the difference between the sequential and the parallel versions and results show that they are in correspondence within 14 significant digits. This is because of the IEEE compliant double precision support inherent on both chips.

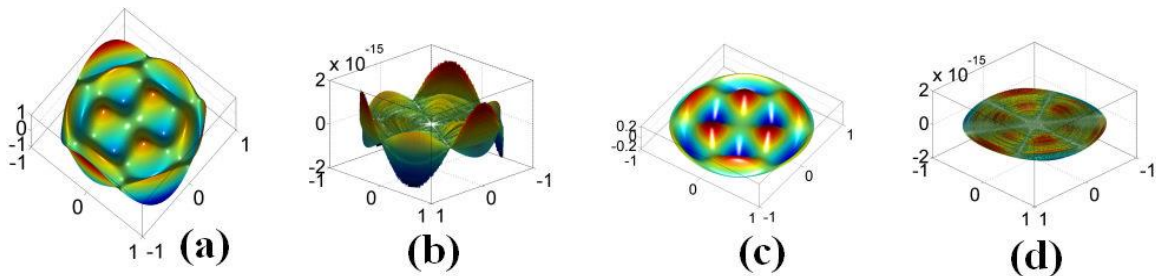


Figure 40 GPU computed low order ϕ -polynomials (a) Zernike, Z_9^3 , (c) gradient orthogonal Q-poly, Q_3^3 ; the difference between the parallel and sequential implementations within 14 significant digits (b) Zernike (d) Q-poly.

The second investigation is to analyze the effect of the total number of ray points across the aperture over the ϕ -polynomials computation time. Naturally, as the number of ray points increases, the total time to compute a specific ϕ -polynomial increases. In order to quantify however the effect of the ray grid size on the computation time, a high order Zernike, $Z_{110}^{10}(r, \theta)$ and a high order gradient orthogonal Q-polynomial, $Q_{50}^{10}(r, \theta)$ were computed both sequentially and in parallel. The results are shown in Table 6 and Figure 41.

Table 6 Effect of the size of the ray grids on the speedup of the computation of ϕ -polynomials.

Grid-size	Zernike polynomial			Gradient orthogonal Q-poly		
	<i>CPU</i> (ms)	<i>GPU</i> (ms)	<i>Speedup</i>	<i>CPU</i> (ms)	<i>GPU</i> (ms)	<i>Speedup</i>
256x256	43.0	5.0	8.6	53.0	5.9	8.9
512x512	107.8	8.3	13.0	162.9	13.2	12.4
1024x1024	316.3	21.1	15.0	512.7	25.2	20.4
2048x2048	1219.7	71.2	17.1	1967.3	78.9	24.9

Table 6 shows that the computation time for both the parallel and the sequential algorithms increases as the number of rays quadruples at each row for both of the ϕ -polynomials. However the time for the sequential algorithm increases more in proportion to the parallel algorithm time. The ratio of the total time for the sequential algorithm execution over the total time that it takes to execute the parallel algorithm is defined as the *speedup* and this parameter increases as the ray grid size increases. Figure 41 shows the total execution times of sequential and parallel algorithms on CPU and GPU and corresponding speedups of the ϕ -polynomials with respect to the ray grid size.

In Figure 41, we can clearly see that the computational time for the gradient orthogonal Q-polynomials is higher than for the Zernike polynomials (see dash-dot blue line in Figure 41 (a)), although the recurrence relations are run exactly 50 times for both of the ϕ -polynomials. The reason for the compute intensive nature of the gradient orthogonal Q-polynomial is because of the unconventional recurrence relation and the necessity of an auxiliary polynomial computation through another recurrence. This computationally expensive operation causes significant overhead for the sequential algorithm on CPU; however it is not a significant burden for the parallel algorithm running on GPU. This can be observed with the almost coincident red dash-dot and solid lines on Figure 41 (a) showing the parallel execution times of Q-polynomial and Zernike polynomial, respectively. Figure 41 (b) quantifies the speedup for both the gradient orthogonal Q-polynomial and the Zernike polynomial. Results show that the speedup increases with the total number of ray samples and grows significantly in average as the number of rays quadruples across the aperture.

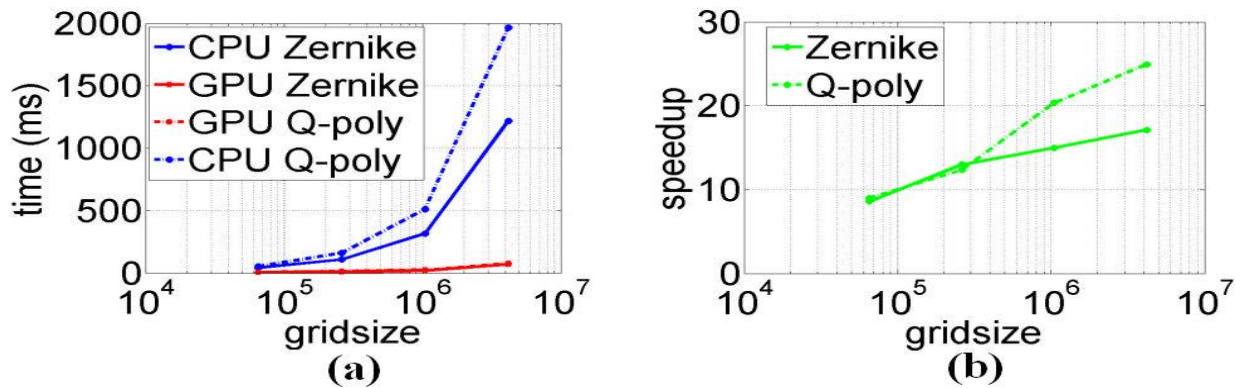


Figure 41 (a) Total execution time of the sequential and parallel algorithms of ϕ -polynomials on both CPU and GPU as a function of the grid size (b) speedups of ϕ -polynomials with grid size.

Another aspect of inquiry for the ϕ -polynomials computation is the order of the ϕ -polynomials. As higher order polynomials may occur in optical surface description, it is desirable to determine if parallelization and speedup in computation time is affected by the order of the ϕ -polynomials. In table 7 and Figure 42 we show the computation times of sequential and parallel algorithms as the order of the ϕ -polynomials is increased. For this experiment, the total number of ray points is kept fixed at 1024×1024 over the circular aperture, and azimuthal order is fixed at $m=2$.

Table 7 Effect of the order of the ϕ -polynomial over the computation time and speedup.

Polynomial order	Zernike polynomial			Gradient orthogonal Q-poly		
	<i>CPU (ms)</i>	<i>GPU (ms)</i>	<i>Speedup</i>	<i>CPU (ms)</i>	<i>GPU (ms)</i>	<i>Speedup</i>
50	177.7	21.7	8.2	262.3	25.2	10.4
100	289.4	21.8	13.3	484.1	25.5	19.0
150	406.5	22.1	18.4	751.2	25.1	29.9
200	514.7	22.6	22.8	953.7	24.7	38.6

Table 7 shows that the speedup for the ϕ -polynomials increases as the order of the polynomials increase. It takes gradually more time to compute the ϕ -polynomial sequentially if the order of the polynomial is increased (see Figure 42 (a), blue lines). However the ϕ -polynomial computational time does not grow at all if parallel algorithms are utilized (see Figure 42 (a), red lines). Consequently, this finding leads to speedups with parallelization of an order of magnitude, i.e. 10 to 40 times, in computation of Zernike or gradient orthogonal Q-polynomials over the polynomial order (see Figure 42 (b)).

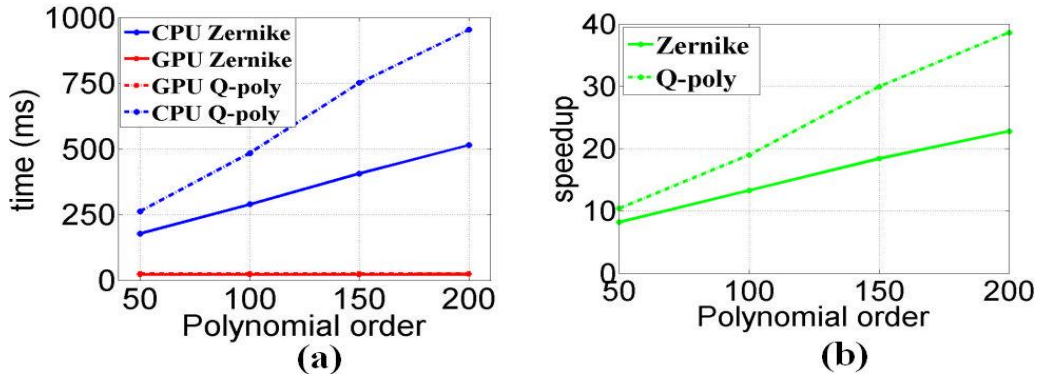


Figure 42. Effect of the polynomial order on the computation of ϕ -polynomials (a) computation times on CPU and GPU (b) speedups through parallelization.

The speedups reported in table 6 and table 7 may be associated with the specific features of the GPUs in executing the parallelized recurrence algorithms. A modern GPU is able to run many more concurrent active threads compared to the number of concurrent threads a CPU can run [62]. The GeForce 650M GPU has 2 multiprocessors each having the ability to support 1536 active concurrent threads [62] whereas Intel® Core™ i7 3610QM consists of 4 cores running 8 concurrent threads in total with hyper-threading [63]. Furthermore, GPU threads are lightweight, and context switches are faster. Although the computational load is increased gradually in table 7 by incrementing the polynomial order, the computational time to compute the ϕ -polynomials on GPU does not change due to the GPU ability to execute more instructions in one clock cycle. The Clenshaw process [64] to compute the linear combination of ϕ -polynomials based upon the recurrence relations implemented in this chapter may also be carried out on GPUs. It seems natural that the extension of the Clenshaw algorithm in parallel would yield similar speedups as reported in this chapter.

Conclusion

In this work, we have investigated the effects of parallelizing the algorithms of φ -polynomial computation with the recurrence relations as they provide more robust and efficient results. Also the effects of ray grid sizes and the orders of the φ -polynomials on the computational time are examined. We have quantified the increased benefits through parallelization as the intensity of the computation grows, such as higher orders and finer ray-grid resolutions. Furthermore, the parallel algorithms proposed in this research were validated to be in excellent correspondence with the sequential implementations. We have utilized the many-core highly threaded GPU for parallel executions, and used a multi-core CPU for the sequential algorithms. This by no means states that CPUs should not be utilized for parallelization with appropriate hyper-threading libraries. Just the contrary, the future computation of the φ -polynomials should take advantage of parallel algorithms running on both highly threaded many-core GPUs and CPUs.

CHAPTER SEVEN: CONCLUSION

Freeform optical elements, which are intrinsically not rotationally symmetric, will not only be the most prevalent components but also they will play a key role in the future of optical systems. They provide a reduction in the physical size of the optical elements through compact and lightweight designs. In this dissertation, we have investigated mathematical and computational propositions for freeform optical surface description with a motivation for most economical, efficient and beneficial methods in terms of precision, accuracy, computation and general applicability.

In the second chapter, we presented a review of the state of the art surface description methodologies for optical surfaces. We have elaborated on the major short comings of power series description including ill-conditioning and numerical artifacts associated with this method. We surveyed ϕ -polynomials for both rotationally symmetric and freeform surfaces. The complete and orthogonal Zernike polynomials are described in this chapter. We detailed the major drawbacks of Zernike polynomials such as the round off errors produced by numerical cancelation, which are prevented if the recurrence relations are deployed in the computation of these polynomials. Also recently introduced slope and gradient orthogonal Q-polynomials are summarized and reviewed along with the recurrence relations developed for them. Finally the Radial Basis Functions (RBFs) and QR based algorithms in order to remove the ill-conditioning associated with the near flat basis functions are described for the freeform shape description for generally shaped apertures.

One of the major contributions of this dissertation is described in chapter three. The effect of the type of ray grids in sampling an optical surface to perform a φ -polynomial fit is investigated in this chapter. It is shown in this chapter that an edge clustered fitting grid effectively removes the edge ringing that arises as the polynomial adapts to the fully nonsymmetric features of the freeform surface. The impact of this fitting grid on the reduction of edge ringing and improvement of the accuracy several orders of magnitude is compared to different types of sampling grids including but not limited to hexagonal uniform subgrids centered on the rectangular uniform grids, Chebyshev based radial sampling, and random grids. Also shown in this part of the dissertation is that for rotationally symmetric surfaces, the type of the fitting grid does not have a particular influence on the quality of the fit as all types of ray grids produced very good approximants with a small number of φ -polynomial terms and samples. Another outcome of this study is that a large number of φ -polynomial terms, i.e. thousands might become necessary in order to represent a freeform surface. Nonetheless, the significant observation established in this chapter is that edge-clustered fitting grids produced very good approximants, and improved the approximation performance by several orders of magnitude compared to that of fitting grids without edge clustering in addition to providing stable and exponentially decreasing errors.

As part of the investigation of efficient ray grids for freeform surface description, we have assessed the merits and drawbacks of two different sets of φ -polynomials in terms of least squares. Results obtained in chapter four show that Zernike polynomials and gradient orthogonal Q-polynomials added on top of a best fit sphere are able to represent freeform surfaces with similar if not identical accuracies with edge clustered fitting grids over a significant range of

heights and locations for the nonsymmetric features of the optical surfaces. Finally we clearly observed with all the experiments carried out in chapter four that it is not so much the type of the ϕ -polynomials but the type of the sampling grid that dominates the magnitude of the errors in Root Mean Square (RMS) fit residuals, thus the level of accuracy obtained through the fitting process. It is also found out in this chapter that accurately computing the best fit sphere radius has a profound effect on the accuracy of the approximant obtained with gradient orthogonal Q-polynomials.

Another contribution presented in chapter five is a hybrid, local, and efficient method combining assets of both RBFs and ϕ -polynomials for the description of freeform surfaces over more general aperture shapes. The method proposed is based upon the partition of unity approach acting on decomposition of aperture into smaller overlapping subapertures. This method is applicable to more general aperture shapes given its decomposition and stitching properties as opposed to the global application of ϕ -polynomials for predefined geometries. This method is not only applicable to more general shaped domains, but also it reduces the order of the ϕ -polynomials deployed. In other words, it does not require higher order terms in the description as opposed to globally employed ϕ -polynomials. In fact, initial results show that the proposed method yields sub-nanometer accuracy with as few as 25 terms of local ϕ -polynomials utilized in each subaperture. Sub-nanometer accuracy is required for the stringent conditions of lithography and related precision optics applications. Less stringent conditions are also shown to be achieved with as few as 16 terms of ϕ -polynomials deployed in each subaperture. Conceptually the method can be thought of as deployments of groups of ϕ -polynomials whose origin is translated across the aperture into the centers of the subapertures. Instead of translating

the centers of RBFs across the aperture, in this method the origin of the ϕ -polynomials are shifted into the centers of subapertures.

A variation of this method can be obtained when RBFs are deployed locally inside the subapertures instead of local ϕ -polynomials. This method offers the opportunity of shape optimization for RBFs over each local subaperture. In other words, locally varying shape parameters are optimized and assigned to all RBFs inside each subaperture. This is a spatially varying shape optimization technique but yet different from the method presented in [31], since all RBFs within a subaperture still assigned the same locally optimized shape parameter. Initial numerical results are presented in chapter five of this dissertation. As a future work of this local hybrid method, an adaptive approach may be used for the domain decomposition and sampling steps instead of a fixed uniform grid of samples and a uniform decomposition of the overall aperture into subapertures. For example, an adaptive refinement method based on the residual subsampling presented in [30] may provide significant advantages especially when finer sampling and subaperture decomposition is achieved in steep regions where errors are augmented. Another possible improvement maybe implemented through a different shape optimization (such as shown in [31]) over each subaperture.

A discussion point may be raised on the total number of subaperture count for this local method. The method significantly reduces the order of the local ϕ -polynomials at the expense of the subaperture count. It works by decomposing the domain into smaller subapertures. As the level of precision and accuracy is increased, it is possible to obtain that level of accuracy through another level of decomposition of subapertures. This means more subapertures and thus an

increase in the total number of φ -polynomials over the global aperture. However, even then since the local least square matrix size is now small, i.e. on the order of tens or hundreds, as opposed to the global fit using global φ -polynomials requiring extreme high orders and all the samples over the aperture, i.e. the size of tens of thousands if not hundreds of thousands. Computationally the hybrid method is much more efficient compared to the global approach because the solution of a fully populated least square matrix depends on the size of the matrix, N , as $O(N^3)$. We propose as a future work another implementation that takes into account the total number of subapertures, and whenever a finer set of subdivision is required, it only produces another level of division over the subapertures and regions where the accuracy is lower than that required, similar to the adaptive gridding used in [30]. This way total number of subapertures may be kept within a smaller bound if desired.

A significant outcome of this research conducted in this dissertation is the devised and implemented recurrence based parallel algorithms for the computation of φ -polynomials to take advantage of highly threaded many-core computational platforms i.e. Graphical Processing Units, (GPUs). Specifically, Zernike polynomials and gradient orthogonal Q-polynomials are implemented through a set of parallel algorithms on GPUs, with their respective recurrence relations by using CUDA C programming. The results presented in chapter six show that more than an order of magnitude improvement is achieved in computational time over a sequential implementation if recurrence-based parallel algorithms are adopted in the computation of the φ -polynomials.

In our implementations, we verified that the accuracy obtained through a sequential implementation was guaranteed by the parallel algorithm. Our results showed that there was an excellent correspondence within 14 significant digits obtained over the sequential and parallel implementations.

We investigated the effect of the size of the ray grid on the speedups achieved with parallelism in the computational time for the recurrence based ϕ -polynomials implementations. The results established that the finer set of ray grids increased the speedup in computational time, and speedup grew significantly in average as the size of the ray grid quadruples in each numerical experiment.

Another analysis is carried out for the effect of order of the ϕ -polynomial on the speedup obtained on GPUs in computational time. It is interesting to note that as the order of a ϕ -polynomial is increased the computational time of the sequential algorithm grows in proportion to the order of the ϕ -polynomial whereas the parallel implementation computational time almost does not change. This observation on the speedups with parallel implementation of recurrence based ϕ -polynomial algorithms is shown in chapter six.

Conclusively, we have quantified the increased benefits through parallelization as the intensity of the computation grows, such as higher orders and finer ray-grid resolutions. As a result, the future computation of the ϕ -polynomials should take advantage of parallel algorithms, such as the ones devised for this dissertation, running on both highly threaded many-core GPUs and CPUs with appropriate hyper threading support.

REFERENCES

- [1] O. Cakmakci, and J.P. Rolland, “Head worn displays: a review,” *J. Disp. Technol.* **2**(3), 199-216 (2006).
- [2] J. C. Miñano, P. Benitez, and A. Santamaria, “Freeform optics for illumination,” *Opt. Rev.* **16**(2), 99–102 (2009).
- [3] K. Fuerschbach, J. P. Rolland, and K. P. Thompson, “A new family of optical systems employing ϕ -polynomial surfaces,” *Opt. Express* **19**(22), 21919–21928 (2011).
- [4] J. P. Rolland and K. P. Thompson, “Freeform optics: Evolution? No, revolution,” *Proc. of the SPIE*, DOI: 10.1117/2.1201207.004309, (2012).
- [5] E. Abbe, Lens system, U.S. Patent No. 697,959, (April, 1902).
- [6] G. W. Forbes, “Shape specification for axially symmetric optical surfaces,” *Opt. Express* **15**(8), 5218-5226, (2007).
- [7] B. Ma, L. Li, K.P. Thompson, and J.P. Rolland, “Applying Slope Constrained Q-type Aspheres to Develop Higher Performance Lenses,” *Opt. Express* **19**(22), 21174-21179 (2011).
- [8] B. Ma, K. Sharma, K. P. Thompson, and J. P. Rolland, “Mobile device camera design with Q-type polynomials to achieve higher production yield,” *Opt. Express* **21**(15), 17454-17463 (2013).
- [9] Y. Tohme and R. Murray, “Principles and Applications of the Slow Slide Servo,” *Moore Nanotechnology Systems White Paper* (2005).

- [10] O. Cakmakci, B. Moore, H. Foroosh, and J. P. Rolland, “Optimal local shape description for rotationally nonsymmetric optical surface design and analysis,” *Opt. Express* **16**(3), 1583–1589 (2008).
- [11] O. Cakmakci, S. Vo, H. Foroosh, and J. P. Rolland, “Application of radial basis functions to shape description in a dual-element off-axis magnifier,” *Opt. Letters* **33**(11), 1237-1239 (2008).
- [12] G. W. Forbes, “Characterizing the shape of freeform optics,” *Opt. Express* **20**(3), 2483–2499 (2012).
- [13] F. Zernike, “Beugungstheorie des schneidenverfahrens und seiner verbesserten form, der phasenkontrastmethode,” *Physica* **1**(7-12), 689-704 (1934).
- [14] L. Seidel, “Zur Dioptrik. Ueber die entwicklung der glieder 3ter ordnung, welche den weg eines ausserhalb der ebene der axe gelegenen lichtstrahles durch ein system brechender medien bestimmen,” *Astr. Nachr.*, **43**(21), 321-332 (1856).
- [15] H. H. Hopkins, *Wave Theory of Aberrations*, (Oxford, Clarendon Press, 1950).
- [16] R. W. Gray, C. Dunn, K. P. Thompson and J. P. Rolland, “An analytic expression for the field dependence of Zernike Polynomials in rotationally symmetric optical systems,” *Opt. Express* **20**(15), 16436 – 16449 (2012).
- [17] G. W. Forbes, “Robust and fast computation for the polynomials of optics,” *Opt. Express* **18**(13), 13851–13862 (2010).

- [18] I. Kaya, K. P. Thompson, and J. P. Rolland, "Edge clustered fitting grids for ϕ -polynomial characterization of freeform optical surfaces," *Opt. Express* **19**(27), 26962–26974 (2011).
- [19] I. Kaya and J. P. Rolland, "Hybrid RBF and local ϕ -polynomial freeform surfaces," *Adv. Opt. Techn.* **2**, 81-88 (2013).
- [20] V.N. Mahajan and G. Dai, "Ortho-normal polynomials in wavefront analysis: analytical solution," *J. Opt. Soc. Am.* **A24**, 2994-31016 (2007)
- [21] I. Kaya, K. P. Thompson and J. P. Rolland, "Comparative assessment of freeform polynomials as optical surface descriptions," *Opt. Express* **20**, 22684 – 22691 (2012).
- [22] L.N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, (SIAM, Philadelphia, 1997).
- [23] G.W. Forbes and C.P. Brophy, "Asphere, o asphere, how shall we describe thee," *Proc. of the SPIE* **7100**, (2008).
- [24] G. W. Forbes, "Robust, efficient computational methods for axially symmetric optical aspheres," *Opt. Express* **18**(19), 19700–19712 (2010).
- [25] J.P. Rolland, C. Dunn, and K.P. Thompson, "An Analytic Expression for the Field Dependence of FRINGE Zernike Polynomial Coefficients in Rotationally Symmetric Optical Systems," *Proc. of the SPIE* **7790**, 7790-22 (2010).
- [26] J. Wyant and K. Creath, "Basic wavefront aberration theory for optical metrology," in *Applied Optics and Optical Engineering XI*, R. R. Shanon and J. Wyant, eds. (Academic Press, New York, 1992), pp. 1-53.

- [27] M. Born and E. Wolf, *Principles of Optics*, (Cambridge University Press, Cambridge, 1999).
- [28] A. B. Bhatia and E. Wolf, “On the circle polynomials of Zernike and related orthogonal sets,” *Proc. Camb. Philos. Soc.* **50**(1), 40-48, (1954).
- [29] J. P. Boyd and F. Yu, “Comparing seven spectral methods for interpolation and for solving the Poisson equation in a disk: Zernike polynomials, Logan–Shepp ridge polynomials, Chebyshev–Fourier Series, cylindrical Robert functions, Bessel–Fourier expansions, square-to-disk conformal mapping and radial basis functions,” *J. Comput. Phys.* **230**(4), 1408 – 1438 (2011).
- [30] T. A. Driscoll and A. R. H. Heryudono, “Adaptive residual subsampling methods for radial basis function interpolation and collocation problems,” *Comp. Math. Appl.* **53**, 927 – 939 (2007).
- [31] B. Fornberg and J. Zuev, “Runge phenomenon and spatially variable shape parameters in RBF interpolation,” *Comp. Math. Appl.* **54**, 379 – 398 (2007).
- [32] O. Cakmakci, K. Thompson, P. Vallee, J. Cote, and J. P. Rolland, “Design of a freeform single-element head-worn display,” *Proc. of SPIE* **7618**, 761803 (2010).
- [33] O. Cakmakci, G. E. Fasshauer, H. Foroosh, K. P. Thompson, and J. P. Rolland, “Meshfree approximation methods for freeform surface representation in optical design with applications to head worn displays,” *Proc. of SPIE* **7061**, 70610D-1-15 (2008).

- [34] J.P. Rolland and O. Cakmakci, “Head-worn displays: the future through new eyes,” *Optics and Photonics News*, **20** (4), 20-27 (2009).
- [35] T. A. Driscoll and B. Fornberg, “Interpolation in the limit of increasingly flat radial basis functions,” *Comput. Math. Appl.* **43**, 413 – 422 (2002).
- [36] B. Fornberg, E. Larsson, and N. Flyer, “Stable computations with Gaussian radial basis functions,” *SIAM J. Sci. Comput.* **33**, 869-892, (2011).
- [37] G. Fasshauer and M. McCourt, “Stable evaluation of Gaussian RBF interpolants,” *SIAM J. Sci. Comput.* **34**, A737 – A762 (2012).
- [38] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, chap. 22, (Dover, 1972).
- [39] G. E. Fasshauer, *Meshfree Approximation Methods with MATLAB*, (World Scientific Publishing, Singapore, 2007).
- [40] R. Platte, *Accuracy and Stability of Global Radial Basis Function Methods for the Numerical Solution of Partial Differential Equations*, Ph.D. Thesis, (University of Delaware, 2005).
- [41] G.W. Forbes, “Fitting freeform shapes with orthogonal bases,” *Opt. Express* **21**(16), 19061-19081 (2013).
- [42] J. J. Koenderink, *Solid Shape*, chap. 6, (The MIT Press, Cambridge, 1990).

- [43] F. Roddier, “Curvature sensing and compensation: a new concept in adaptive optics,” *Appl. Optics* **27**, 1223 – 1225 (1988).
- [44] P. E. Glenn, “Robust, sub-angstrom level mid-spatial frequency profilometry,” *Proc. SPIE* **1333**, 230–238 (1990).
- [45] W. Zou and J. Rolland, Differential Shack–Hartmann curvature sensor, U.S. patent 7,390,999 (June 2008).
- [46] W. Zou, K.P. Thompson, and J.P. Rolland, “Differential Shack-Hartmann curvature sensor: local principal curvature measurements,” *JOSA A* **25**, 2331-2337 (2008).
- [47] Y.-M. Liu, G. Lawrence and C. Koliopoulos, “Subaperture testing of aspheres with annular zones,” *Appl. Optics* **27**, 4504 – 4513 (1988).
- [48] G. Shechter, “k-D tree”, (Mathworks, 2004),
<http://www.mathworks.com/matlabcentral/fileexchange/4586-k-d-tree>.
- [49] B. Fornberg, T. A. Driscoll, G. Wright and R. Charles, “Observations on the behavior of radial basis function approximations near boundaries,” *Comp. Math. Appl.* **43**, 473 – 490 (2002).
- [50] I .Kaya and J.P. Rolland, “Acceleration of computation of φ -polynomials,” *Opt. Express* (under review).
- [51] C. L. Phillips, J.A. Anderson and S. C. Glotzer, “Pseudo-random number generation for Brownian dynamics and dissipative particle dynamics simulations on GPU devices”, *J. Comput. Phys.* **230**(19), 7191-7201 (2011).

- [52] S. Liu, P. Li and Q. Luo, “Fast blood flow visualization of high resolution laser speckle imaging data using graphics processing units”, *Opt. Express* **16** (19), 14321-14329 (2008).
- [53] J. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Kruger, A. Lefohn, and T. Purcell, “A survey of general-purpose computation on graphics hardware,” *Comput. Graph. Forum* **26** (1), 80-113 (2007).
- [54] A. Klein, “Introduction to Direct3D 11 graphics pipeline,” (Microsoft, 2008), <http://www.microsoft.com/en-us/download/details.aspx?id=15051>.
- [55] NVIDIA, “CUDA programming model overview,” (NVIDIA, 2008), <http://www.sdsc.edu/us/training/assets/docs/NVIDIA-02-BasicsOfCUDA.pdf>.
- [56] NVIDIA, *CUDA C programming guide*, (NVIDIA Corp., 2012).
- [57] D. Komatitsch, G. Erlebacher, D. Goddeke, and D. Michea, “High-order finite element seismic wave propagation modeling with MPI on a large GPU cluster,” *J. Comput. Phys.* **229**(20) 7692-7714 (2010).
- [58] S. Walsch, M. Saar, P. Bailey, and D. Lilja, “Accelerating geoscience and engineering system simulations on graphics hardware,” *Computers & Geosciences* **35**(12), 2353-2364 (2009).
- [59] C. Kauffmann and N. Piche, “Seeded ND medical image segmentation by cellular automaton on GPU,” *Int. J. CARS* **5**(3), 251-262 (2010).

[60] NVIDIA, “CUDA community showcase,” (NVIDIA, 2013)

<http://www.nvidia.com/object/cuda-apps-flash-new.html#>.

[61] Y. Jian, K. Wong, and M. V. Sarunic, “Graphics processing unit accelerated optical coherence tomography processing at megahertz axial scan rate and high resolution video rate volumetric rendering,” *J. Biomed. Opt.* 18(2), 026002-1-4 (2013).

[62] NVIDIA, *CUDA C best practices guide*, (NVIDIA, 2012).

[63] Intel, “Intel core 7 processor specifications,” (Intel, 2013),

<http://www.intel.com/content/www/us/en/processors/core/core-i7-processor/Corei7Specifications.html>.

[64] C.W. Clenshaw, “A note on the summation of Chebyshev series,” *Math. Tables Other Aids Comput.* 9, 118-120 (1995).