Research Article | Vol. 27, No. 5 | 4 Mar 2019 | OPTICS EXPRESS 6815
Optics EXPRESS

Freeform surface selection based on parametric fitness function using modal wavefront fitting

ISAAC TRUMPER,[1] MAHAM AFTAB,[1,2] AND DAE WOOK KIM[1,3,*]

[1] College of Optical Sciences, University of Arizona, 1630 E. University Blvd., Tucson, AZ 85721, USA
[2] Optical Sensors and Metrology, Design & Competence Group, ASML US Inc., USA
[3] Steward Observatory, University of Arizona, 933 N. Cherry Ave., Tucson, AZ 85719, USA
[*] letter2dwk@hotmail.com

Abstract: We present an analytic methodology to guide the selection of a surface within an optical design to apply freeform optimization. The methodology is discussed in the context of other means currently available, such as human intuition, aberration theory, and other direct surface construction methods. We describe the selection criteria for our proposed method and provide the form of the parametric fitness function used to combine the criterion. Finally, a case study comparing a design optimization procedure guided by the proposed methodology to human intuition is presented based on a real instrument designed for a millimeter-wave astronomy application. The methodology is shown to be effective even in the case of an optical system with a large number of freeform/optical surfaces. The proposed approach provides an objective and scalable solution to guide freeform optical system design by aiding a human's design intuition.

1. Introduction

Freeform optics are revolutionizing the capabilities of optical systems with applications across the entire optical design spectrum [1]. Publications on the performance of freeform optical systems highlight the improved field of view (FOV), system compactness, and imaging [2–4]. Researchers are developing and improving the theory to describe the aberration behavior of systems that lack rotational symmetry [5–7]. The prescription of the freeform surface is also an active area of research, where many different surface descriptions exist, some of which are: Zernike polynomials, *XY* polynomials, NURBS, and Q polynomials.

Creating a first-order design is critical to achieving the optical requirements in an optimal manner because the first-order principles constrain the design form of the optical system. In the space of freeform optical design, methods to directly generate a starting configuration from a set of planes have been developed [8–10]. Some methods solve analytic equations to set the second-order imaging properties of the system [11, 12]. Further methods develop the surface in a step-by-step method, growing the FOV and therefore the surface [13]. Another method defines a construction technique to generate the geometry and power distribution in a design optimally suited for freeform optimization [14]. Aberration theory may be used to determine which surface limits the optical performance and which freeform terms to apply, or even to determine the initial system design [15]. However, once the system has many optical surfaces and the design matures, numerical optimization is still necessary. Unlike other analytical approaches, we present a parametric numerical alternative methodology to guide the process of choosing an optimal surface in an existing design to optimize into a freeform optic.

Optical design is an art form that requires intuition and understanding from the optical designer [16, 17]; for now, there is no magic button to push in a program that returns the ideal optical system. Instead, the software must be watched and guided with care to produce the optimal results that many designers achieve. Choosing a surface in a design to make freeform and optimize is no different. The designer must use their knowledge combined with the available data obtained from the program (e.g. aberrations, pupil maps) to decide where to apply the freeform

#356046 https://doi.org/10.1364/OE.27.006815
Journal © 2019 Received 26 Dec 2018; revised 5 Feb 2019; accepted 6 Feb 2019; published 22 Feb 2019

Optics EXPRESS

terms or rely on numerical optimization methods to guide the selection process [18–20]. The choice of surface impacts the outcome of the design by constraining the design space, which is potentially detrimental. Some commercial software provide similar or limited features, but the methodology and specifics are trade-secrets.

2. Background

Modern optical design is enabled by computer modeling and optimization of the optical system, where rays are traced and used to compute various parameters and metric concerning the design. In the presented work, we leverage the computational power and ray tracing capabilities found in the software package OpticStudio. At its essence, ray tracing involves computing the interaction of a vector (ray) and the optical elements that comprise the optical system. The ray is propagated from one surface to another, where its direction and location are recorded. This ray data can be used to compute most properties of the optical system that concern the optical designer. The raw ray data may also be accessed and used to compute a custom metric for the optical system. We employ this style of computation, where our fundamental data is the ray information at a given surface in the optical system. The ray data is composed of a location and direction in three dimensions, where typically the location is expressed in Cartesian coordinates (x, y, z) and the direction is given by the direction cosines (L, M, N). Rays from a point in the FOV (Field of View) are normal to a wavefront, which is a surface of constant phase in the oscillations of the electromagnetic field. A plane that is normal to the ray defines the local slope of the wavefront. In this way, a ray direction is equivalent to the wavefront normal, which can be used to calculate the local wavefront slope. Shown in Fig. 1 is a schematic of how the rays in an optical system relate to the wavefront and its slope. To calculate the wavefront slope S_x and S_y in the x and y directions, respectively, we use the ratio of the direction cosine in the direction of interest to the direction cosine of the z-axis. The ray direction vector is a unit vector given by the direction cosines (L, M, N), which leads to the vector $\vec{H} = L\hat{x} + M\hat{y} + N\hat{z}$ whose magnitude $H = 1$. Eq. (1) gives the relationship between wavefront slope in the y-direction and the ray direction cosines,

$$S_y = \tan\left(\theta_y\right) = \frac{Y}{Z} = \frac{Y/H}{Z/H} = \frac{\cos\left(\phi_y\right)}{\cos\left(\phi_z\right)} = \frac{M}{N}, \tag{1}$$

where Y and Z are the projection of the ray onto the y and z axes, respectively, H is the length of the vector, ϕ_y and ϕ_z are the angles between the vector and the y and z axes, respectively, and θ_y is the wavefront slope angle in the y-direction. An equivalent formula is used for the x-direction, where the ratio is between the x-direction cosine (L) and the z-direction cosine (N).

2.1. Chebyshev polynomials and their gradients

Aftab et al. derive a vector polynomial basis set, called $\vec{G}$ polynomials [21]. This set is obtained from gradients of the two-dimensional Chebyshev polynomials of the first kind, named F polynomials. Both sets are complete and orthogonal across rectangular apertures. Chebyshev polynomials are well known for their data fitting properties. They generally show a fast rate of convergence [22] and hold discrete orthogonality in several cases [23]. F and $\vec{G}$ polynomials form a modal data fitting model that is valuable in many situations. They are well-suited for applications where a very large number of polynomials are needed for fitting data such as freeform optics or the reconstruction of surfaces from slope data, where mid-to-high spatial frequencies must be correctly represented. Some features of these polynomial sets that make this possible and efficient [21] are: (a) development of recursive relations for both F and $\vec{G}$ polynomial sets (b) a one-to-one correspondence between the coefficients of the two polynomial sets (c) ease and accuracy of generating the gradient $(\vec{G})$ polynomials which, unlike most other gradient polynomial sets, do not need an orthogonalization process. Shown in Eq. (2) is a mathematical

Optics EXPRESS

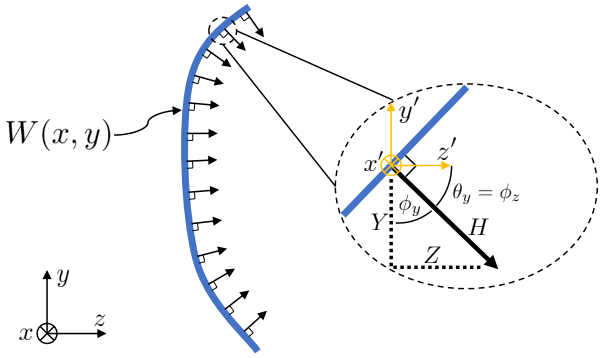


Fig. 1. Relationship between the ray model (black arrows) and wavefront ($W(x, y)$), from which the wavefront slope angle ($\theta_{x,y}$) in each direction is obtained. Rays are normal to the wavefront, and therefore their direction is related to the wavefront slope in the x and y directions. The local ray coordinates in the intersection point are labeled with primes, while the coordinates of the surface-vertex are un-primed. The x, y, z axes are oriented in the same direction, but the origin is shifted.

description of the $\vec{G}$ polynomials in terms of the scalar F basis set, while Eq. (3) gives this description in terms of the standard, recursive 1-D Chebyshev polynomials of the first kind, called $T(x)$ and $T(y)$.

$$\vec{G}_n^m(x, y) = \nabla F_n^m(x, y) = \frac{\partial}{\partial x} F_n^m(x, y)\,\hat{i} + \frac{\partial}{\partial y} F_n^m(x, y)\,\hat{j} \tag{2}$$

$$\vec{G}_n^m(x, y) = \frac{m}{2y\left(1 - x^2\right)}\left(T_{n+1}(y) - T_{n-1}(y)\right)\left(T_{m-1}(x) - xT_m(x)\right)\hat{i} + \frac{n}{2x\left(1 - y^2\right)}\left(T_{m+1}(x) - T_{m-1}(x)\right)\left(T_{n-1}(y) - yT_n(y)\right)\hat{j}, \tag{3}$$

where n and m are double index variables describing the polynomials' order [21] while $\hat{i}$ and $\hat{j}$ are unit vectors indicating a Cartesian coordinate system's axes. Quiver plots of the first 5 non-trivial $\vec{G}$ polynomials are given in Fig. 2.

F and $\vec{G}$ polynomials based data processing has also shown to be adaptable and efficient in various practical situations [21, 24] such as the presence of NaNs (not a number) in data (representing missing data, apertures, blockers etc.), uneven sampling or a different sampling in two directions (e.g. x and y in Cartesian coordinates).

3. Optimal surface selection for freeform optimization

In this work, we present a methodology to select the optimal surface within an existing optical design to optimize into a freeform shape. We use real ray data from design software (e.g. Code V, OpticStudio) to inform and guide this methodology.

3.1. Underlying principles

The fundamental principle that motivates this work is the time-reversible nature of electromagnetic (EM) radiation, or that the forward propagation of an EM wave is equivalent to its reverse propagation. A similar approach is often used to design Computer Generated Holograms (CGH) in order to create an aberration-free null condition. This can be expressed in terms of ray-tracing

Optics EXPRESS

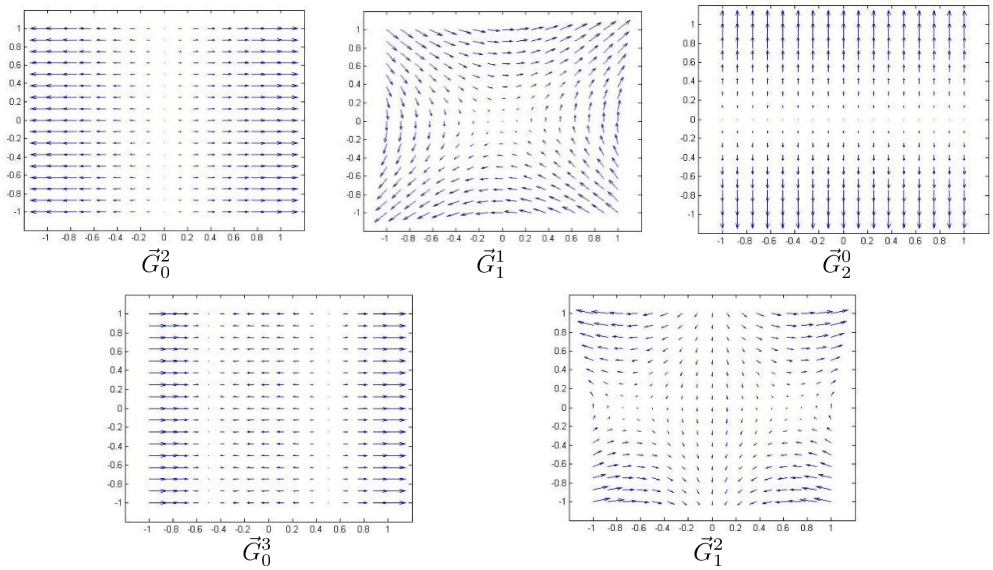


Fig. 2. First 5 non-trivial $\vec{G}$ polynomials plotted in quiver format with their corresponding labels. Note that the x, y coordinates are all normalized from -1 to 1.

by stating that a model tracing rays from the source to the image is equivalent to tracing rays from the image to the source. Forward traced rays from a perfect optical system (i.e. aberration free) will start from a source plane and end at a point on the image plane as a point-to-point mapping. If that same system is ray-traced from the same image point (i.e. no distortion) to the source plane (reverse ray-trace) the rays will take the same path as those from the forward case. At every surface within the optical system the rays will have the same position and direction. In other words, the wavefronts in both the forward and reverse cases are identical at each surface in the system. This principle is shown pictorially in Fig. 3, where an off-axis (in aperture) Cassegrain telescope is simulated in OpticStudio and the wavefront at an intermediate surface is sampled in both the forward and reverse models using the real ray data. In the zoom-in graphic of Fig. 3 we see zero difference between the on-axis wavefronts (ray position and direction) of the forward and reverse models (blue and yellow, respectively). However, as the conic are used away from their geometric conditions, we see a disagreement between the forward and reverse models. This demonstrates that the proposed method is able to include all the relevant ray information across field points, but it is equally valid for any set of wavelengths, configurations, etc. that are ray traced.

We utilize this principle by recognizing that a difference in wavefronts between the forward and reverse case signifies a departure from the aberration and distortion free conditions on the optical system. The difference between these cases contains data that can be used to generate an initial freeform surface, as shown in the work by Yang [8–10], but in this work we demonstrate another use case that compliments prior work by answering a more fundamental question of where to optimally place the freeform surface within the design. Once a surface has been selected, any number of methods [25] to generate that freeform surface may be used.

Our proposed methodology is able to synthesize all ray data at a selected surface across field points, wavelengths, configurations, etc. that may be desired. As long as the ray positions and directions are recorded at the surface, the information can be utilized cohesively when deciding which surface to select. We are able to combine such disparate ray data by leveraging the flexibility of modal fitting using the $\vec{G}$ polynomials described in Section 2.1. When we collect

Optics EXPRESS

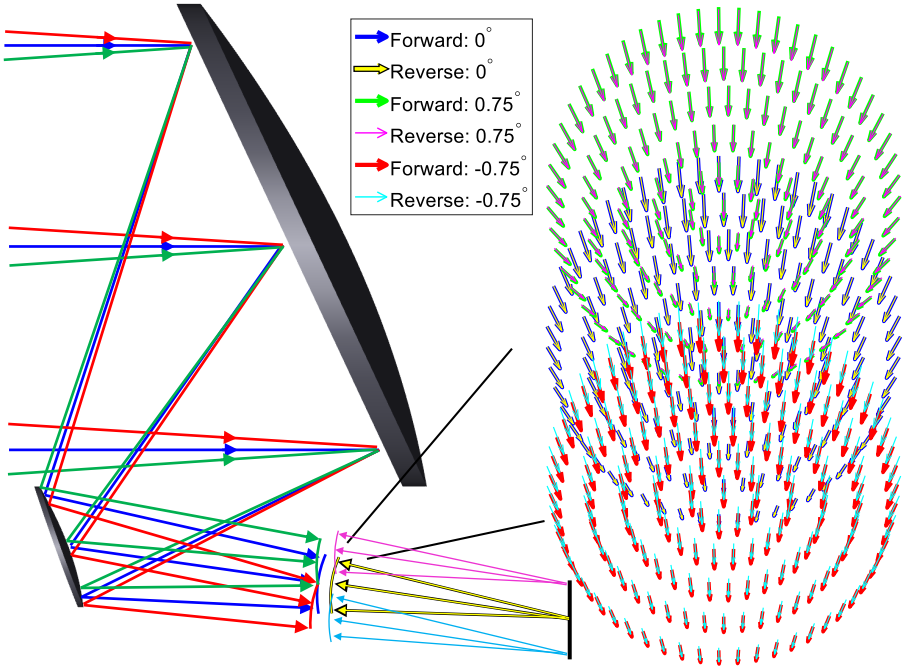


Fig. 3. Graphical representation of the time-reversibility of ray-tracing, where the red, green, and blue rays are traced in a time-forward sense, from the source to the image, while the cyan, magenta, and yellow rays are traced in a time-reversed sense, from the image point to the source. The ray locations and directions at the surface of interest are shown as an insert quiver plot. The rays from the reverse direction are plotted in the coordinate frame of the forward direction rays, which allows a direct comparison of the two ray sets.

the ray data from all relevant sources, we simply use the entire data set at once, as described in the following Section 3.2. If desired, the ray data can be weighted during this process to achieve finer control over the surface selection criteria.

3.2. *Parametric freeform surface selection criteria*

At a surface within the optical system, the real ray data from forward and reverse ray-traces are used to compute multiple parameters that inform the fitness of, or how well suited, the surface is to the application of freeform terms. To compute these parameters, a set of steps shown below are followed:

I Ray data is collected at every surface of interest, in the forward and reverse models, which results in a discrete collection of x, y, z intersection points and L, M, N direction cosines after interacting with the surface. A subscript f or r designates the forward and reverse data, respectively. This ray data may be obtained by tracing multiple field points, wavelengths, system configurations, etc.

II The ray data is used to define a discrete 3D vector field on each surface

$$\vec{A}_f\left(x_f^i, y_f^i, z_f^i\right) = L_f^i\hat{x} + M_f^i\hat{y} + N_f^i\hat{z} = \vec{A}_f(i) \tag{4}$$

$$\vec{A}_r\left(x_r^j, y_r^j, z_r^j\right) = L_r^j\hat{x} + M_r^j\hat{y} + N_r^j\hat{z} = \vec{A}_r(j), \tag{5}$$

Optics EXPRESS

where $i = 1 \ldots I$ and $j = 1 \ldots J$, which represent the discrete sample index in the forward and reverse directions, respectively.

III The discrete forward and reverse vector fields are then fit separately using a least squares technique with a continuous vector field defined by the $\vec{G}$ polynomials (see Section 2.1 for a description) whose polynomial coefficients are given by G_t for $t = 1 \ldots T$.

$$\begin{bmatrix} \vec{A}_f(1) \\ \vdots \\ \vec{A}_f(I) \end{bmatrix} = \begin{bmatrix} \vec{G}_1(1) & \ldots & \vec{G}_T(1) \\ \vdots & \ddots & \vdots \\ \vec{G}_1(I) & \ldots & \vec{G}_T(I) \end{bmatrix} \begin{bmatrix} G_{1,f} \\ \vdots \\ G_{T,f} \end{bmatrix} \tag{6}$$

$$\begin{bmatrix} \vec{A}_r(1) \\ \vdots \\ \vec{A}_r(J) \end{bmatrix} = \begin{bmatrix} \vec{G}_1(1) & \ldots & \vec{G}_k(1) \\ \vdots & \ddots & \vdots \\ \vec{G}_1(J) & \ldots & \vec{G}_k(J) \end{bmatrix} \begin{bmatrix} G_{1,r} \\ \vdots \\ G_{T,r} \end{bmatrix} \tag{7}$$

IV The coefficients G_t in the forward and reverse direction are used to define new vector fields $\vec{G}_f$ and $\vec{G}_r$, respectively. These vector fields are sampled at K discrete points, which at least include the same points as the measured vector fields, $\vec{A}_f$ and $\vec{A}_r$.

V The $\vec{G}$ polynomials have a direct correspondence to the F polynomials, which allows us to integrate the vector field and obtain the wavefront described by the F polynomials at the surface of interest.

$$\begin{bmatrix} \vec{G}_f(1) \\ \vdots \\ \vec{G}_f(K) \end{bmatrix} \rightarrow \begin{bmatrix} F_f(1) \\ \vdots \\ F_f(K) \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} \vec{G}_r(1) \\ \vdots \\ \vec{G}_r(K) \end{bmatrix} \rightarrow \begin{bmatrix} F_r(1) \\ \vdots \\ F_r(K) \end{bmatrix} \tag{9}$$

VI By computing the difference in the reconstructed forward and reverse wavefronts at the surface of interest, we can obtain a new wavefront (F_Δ) that measures the degree of mismatch between the forward and reverse wavefronts in the optical system, which we want to fix through freeform optimization.

$$\begin{bmatrix} F_f(1) \\ \vdots \\ F_f(K) \end{bmatrix} - \begin{bmatrix} F_r(1) \\ \vdots \\ F_r(K) \end{bmatrix} = \begin{bmatrix} F_\Delta(1) \\ \vdots \\ F_\Delta(K) \end{bmatrix} \tag{10}$$

The quantities listed above form the basic building blocks upon which all the data processing is founded. To fully describe the optical system, the FOV and pupil (for each field) is sampled in both forward and reverse directions (Step I) and used for form the discrete vector fields given by Eqs. (4) and (5) in Step II. Step III is applied to all forward or reverse direction data (both field and pupil) at given surface to generate a continuous vector field that represents the normal (slope)

Optics EXPRESS

of the wavefront. A continuous vector field is critical for this application since we are modeling the potential effects of a freeform surface in the optical system, which requires a continuous surface and a continuous first derivative (slope) for physical realization. We are not proposing this solution for a Diffractive Optical Element or CGH or hybrid-type optics application in this respect. Ensuring that the wavefront we generate through fitting is continuous (Step V), we guarantee that there is a physical surface that can achieve or approximate the necessary wavefront manipulation, disallowing any unphysical cases. The final process is taking the difference between the forward and reverse wavefronts (Step VI).

With the data processed in the manner described above, we compute:

1. $\Delta S_{f,r}^{\mathrm{RMS}}$: the root-mean-square (RMS) error between the data and the fit in the forward and reverse directions, respectively. This quantity is a metric of how well a single surface represents the ensemble of all simulated wavefront slope data. This metric is only evaluated at the spatial points where both $\vec{A}_{f,r}$ and $\vec{G}_{f,r}$ are defined, which by definition is the set of $\vec{A}_{f,r}$, but could be a subset of $\vec{G}_{f,r}$.

$$\Delta S_f^{\mathrm{RMS}} = \mathrm{RMS}\left(\begin{bmatrix}\vec{A}_f(1)\\ \vdots \\ \vec{A}_f(I)\end{bmatrix} - \begin{bmatrix}\vec{G}_f(1)\\ \vdots \\ \vec{G}_f(I)\end{bmatrix}\right) \tag{11}$$

$$\Delta S_r^{\mathrm{RMS}} = \mathrm{RMS}\left(\begin{bmatrix}\vec{A}_r(1)\\ \vdots \\ \vec{A}_r(J)\end{bmatrix} - \begin{bmatrix}\vec{G}_r(1)\\ \vdots \\ \vec{G}_r(J)\end{bmatrix}\right) \tag{12}$$

2. $\Delta S_{x,y}^{\mathrm{RMS}}$: the RMS of the slope difference between the forward and reverse vector fields describes a metric of the amount of potential aberration correction over the entire FOV at the surface. The slope of the fitted vector field is computed by using the ratio of the vector component along the desired slope axis to the z vector component.

$$\begin{bmatrix}\Delta G_x\\ \vdots \\ \Delta G_x\end{bmatrix} = \begin{bmatrix}G_f^x(1)/G_f^z(1)\\ \vdots \\ G_f^x(K)/G_f^z(K)\end{bmatrix} - \begin{bmatrix}G_r^x(1)/G_r^z(1)\\ \vdots \\ G_r^x(K)/G_r^z(K)\end{bmatrix} \tag{13}$$

$$\begin{bmatrix}\Delta G_y\\ \vdots \\ \Delta G_y\end{bmatrix} = \begin{bmatrix}G_f^y(1)/G_f^z(1)\\ \vdots \\ Gf^y(K)/G_f^z(K)\end{bmatrix} - \begin{bmatrix}G_r^y(1)/G_r^z(1)\\ \vdots \\ G_r^y(K)/G_r^z(K)\end{bmatrix} \tag{14}$$

$$\Delta S_x^{\mathrm{RMS}} = \mathrm{RMS}\left(\begin{bmatrix}\Delta G_x\\ \vdots \\ \Delta G_x\end{bmatrix}\right) \tag{15}$$

$$\Delta S_y^{\mathrm{RMS}} = \mathrm{RMS}\left(\begin{bmatrix}\Delta G_y\\ \vdots \\ \Delta G_y\end{bmatrix}\right) \tag{16}$$

Optics EXPRESS

3. $F_{\Delta}^{\mathrm{PSD}}$: the Power Spectral Density (PSD) of the wavefront difference evaluated at a spatial frequency f_1. The PSD of the surface is computed by taking the difference in forward and reverse wavefronts (F_{Δ}). This calculation provides a tool to determine if the surface will be manufacturable. A set of these parameters evaluated at multiple spatial frequencies may be linearly combined if more than one spatial frequency are of interest.

$$F_{\Delta}^{\mathrm{PSD}} = \mathrm{PSD}\left(\begin{bmatrix} F_{\Delta}(1) \\ \vdots \\ F_{\Delta}(K) \end{bmatrix}\right)\Bigg|_{f1} \tag{17}$$

The three criteria described above combine to give a metric for selecting the optimal surface at which to apply the freeform terms.

3.3. Parametric fitness function definition

The four criteria defined in Section 3.2 are used in a fitness function to determine the optimal surface for freeform optimization. As such, specific weight values and a method of combining the parameters was chosen for this work. Furthermore, additional scaling factors based on the quality of data (see Sec 3.4) obtained from ray-tracing were included in the computation of the fitness function. The form of fitness function is given in Eq. (18),

$$f = w_1 U_{fr} - w_2\sqrt{\left(\Delta S_f^{\mathrm{RMS}}\right)^2 + \left(\Delta S_r^{\mathrm{RMS}}\right)^2} + w_3\Delta S_x^{\mathrm{RMS}} + w_4\Delta S_y^{\mathrm{RMS}} - w_5 F_{\Delta}^{\mathrm{PSD}}, \tag{18}$$

where U_{fr} is the minimum fractional overlap between the forward (f) and reverse (r) data, described in more detail in Section 3.4. $\Delta S_f^{\mathrm{RMS}}$ and $\Delta S_r^{\mathrm{RMS}}$ are the parameters calculated by Eqs. (11) and (12), $\Delta S_x^{\mathrm{RMS}}$ and $\Delta S_y^{\mathrm{RMS}}$ are the parameters computed in Eqs. (15) and (16), and $w_{1\ldots5}$ are the weight values applied to the parameters. Based on these parameters, a larger fitness function value represents a more optimal candidate to be a freeform surface.

The fitness function incorporates the fractional overlap between the forward and reverse data (U_{fr}) because a surface with more overlap (closer to unity) has the potential to correct a larger portion of the aberrations in the system compared to a surface with less overlap (U_{fr} closer to zero). This is a representation of the spatial aspects of aberrations, where both the position and direction of a ray must be considered. Note that as the design matures, and the aberrations in the system are reduced, the overlap between the forward and reverse data will increase at all surfaces as shown in Fig. 4. The terms $\Delta S_x^{\mathrm{RMS}}$ and $\Delta S_y^{\mathrm{RMS}}$ describe the amount of potential aberration correction at the surface. A larger RMS difference states that the surface can affect more meaningful change and correspondingly the fitness function value increases. Note that this factor deals with the direction of the ray, in contrast to its position, as addressed above. By combining these three parameters, when there is little aberration correction to be had at the surface the fitness of the surface is reduced. This places an emphasis on getting the most aberration correction out of a single freeform surface as possible. We then include a metric concerning the manufacturability of the surface by computing the PSD ($F_{\Delta}^{\mathrm{PSD}}$) at specific spatial frequencies and subtracting it from the fitness function. A larger PSD value at a frequency is harder to produce due to larger slopes and therefore is counted negatively in the fitness function. The PSD directly describes the magnitude of variation of the surface across the spatial frequency domain. As the value of the PSD at a specific frequency increases, the slope of the surface described by that spatial frequency also increases. Surface slope plays a critical role when determining the manufacturability of a surface because of difficulty fabricating and testing high slope surfaces. This is because creating a high slope requires challenging tool paths and removal

Optics EXPRESS

while simultaneously pushing the needed dynamic range of the metrology tool. Therefore, most freeform surfaces have a maximum allowable surface slope. The root-sum-square (RSS) combination of $\Delta S_{f,r}^{\mathrm{RMS}}$ (RMS difference between the fit and measured slopes in the forward and reverse directions) is subtracted from the fitness function because it is a metric of the quality of the processed data. To produce reasonable and trustworthy results, the fit of the measured data needs to be as good as possible. As the fit quality is reduced, the RMS error will increase and therefore lower the fitness of the surface.

3.4. Data quality check factor

We implement data quality checks per evaluated surface after obtaining the forward and reverse ray-trace data. We impose a minimum fractional overlap criterion between each sample in the field of view to ensure that the data accurately represents and sufficiently samples the wavefronts. If there were no overlap between field points, a freeform surface can perform highly localized correction to minimize the aberrations across the discretely sampled field, but in reality the continuous field will still contain significant aberrations. If the design was utilizing multiple configurations or wavelengths that also sampled the surface at different locations, a fractional overlap criterion should be included. The fractional overlap between the data sets is defined by drawing a polygon encompassing the individual data and computing the overlap between these regions. In this way, any grouping of rays may be formed, and the overlap between other groups can be computed and used in the fractional overlap parameter. By specifying a minimum amount of overlap between field sampling at every surface of interest we help ensure realistic design guidance. Furthermore, we implement a method to control for the uncertainty caused by a mismatch in data location (sampling of the wavefront) between the forward and reverse cases. For any system with aberrations, the forward and reverse wavefronts will have regions where they do not overlap. This creates a problem when evaluating the surface selection criteria because the criteria depend on the difference between the forward and reverse cases, but the non-overlapping sampled wavefronts have no physically meaningful data with which to take the difference. To remedy such a case, we simply down-select the forward and reverse data until we are working with the union of the data (i.e. the overlap region). As the design matures, and the aberrations are reduced, the overlap region will grow, as shown pictorially in Fig. 4.

The fitness function definition that we have selected is based on our reasoning of how the surface criterion should be combined. We recognize that a different method of combining the terms and their specific weight values will change the outcome of this analysis. Furthermore, a different set of criterion may be defined for a different application based on its needs. All these variations should be investigated and optimized such that the surface recommendation is robust and reliable.

4. Methodology case study: optimization path comparison

To validate the proposed methodology of selecting an optimal surface to apply freeform optimization at, we present a case study of a millimeter-wave instrument for the Tomographic Ionized-carbon Mapping Experiment (TIME). This design was generated by the authors as part of a separate body of work. The instrument design required the use of freeform optics to fit the system inside a spatially-constrained volume (telescope cabin) without sacrificing optical performance. A design was optimized using human guided intuition that achieved this goal, as shown in Fig. 5, and whose basic optical prescription is given in Table 1. The instrument has a linear field-of-view of $\pm 0.023^{\circ}$, which is rotated to track celestial bodies [26,27] using a K-mirror (K1–3 in Fig. 5). The orientation of the K-mirror is sampled at 5 different angles over its range of -45 to 45 degrees as separate configurations in the optical design. The design has an entrance pupil diameter of 12 m (telescope aperture), and a final working F/# of F/3 to couple into the detector modules. The performance metrics of this millimeter-wave instrument are primarily

Optics EXPRESS

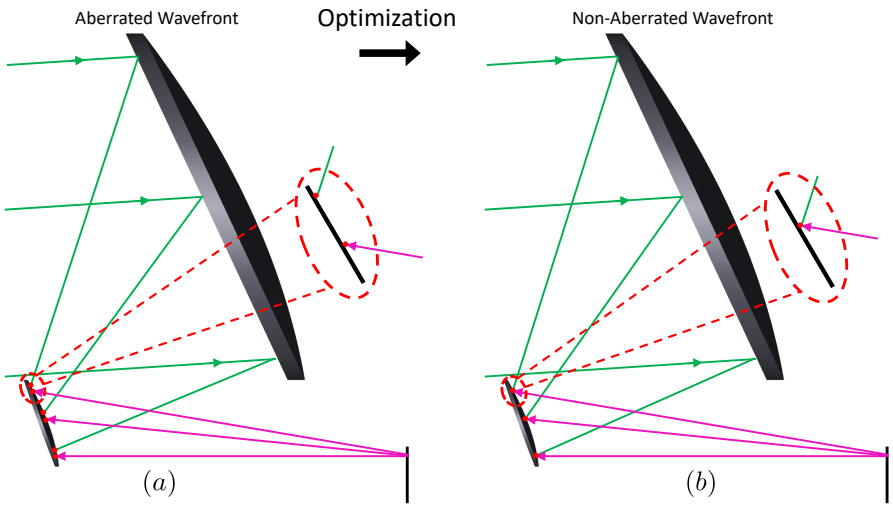


Fig. 4. Diagram showing the development of the spatial overlap region between the forward and reverse ray traces at a surface. The system in (a) is aberrated, and at the surface of interest, the sampled rays in the aperture do not overlap. However, after optimization (b), the ray footprint on the surface matches.

dictated by eliminating unwanted signals coupling to the detectors, and distortion. The unwanted signals can come from aberrations, both at the image and pupil. Image space telecentricity is also important to ensure the desired signal is coupled to the detectors.

The instrument designed for TIME has a total of 6 mirror surfaces: three powered, three flat, and a single cold lens made from high-density polyethylene (HDPE), arranged in a folded-path geometry about the full volume of the cabin. Each powered surface employs a freeform, or non-rotationally symmetric, surface to control the aberrations and obtain the required performance in such a folded geometry. The second surface within the K-mirror uses a powered freeform surface as part of the image and pupil relay comprised of K2, P1, and P2. The flat mirror (F1) folds the optical path so that the relay can fit within the cabin volume. A Lyot stop is formed within a cryostat after the second powered mirror (P2) to control stray light and eliminate unwanted loading on the detectors. The cold lens serves to form the final image and create a telecentric object space in order to couple the radio waves to the detector modules within the cryostat.

When a clear global optimum to an optical design problem is not readily available, either through human intuition or numerical optimization, a methodology to guide the optimization process is desirable. Finding the global optimum of a complex optical system is typically limited by numerical power or time, and is therefore often confined within a local optimum. This study attempts to provide guidance in the optimization process in a way that can lead to a high-potential local optimum space with more objective and parametric methods. We will now present three cases in which the same starting point design was optimized using the same merit function, but the optimization procedure was guided by different sources. Please, note that this merit function is not the fitness function defined in Eq. (18). All designs start from the same place, but the order of freeform surface optimization changes. The first case follows the same procedure as the original instrument design, which was guided by human intuition and met the requirements and is being manufactured now. The second case employs the methodology introduced in this paper to optimize for the lowest merit function (i.e. best optical performance), while the third case also employs the proposed methodology but with the goal of finding the highest merit function

Optics EXPRESS

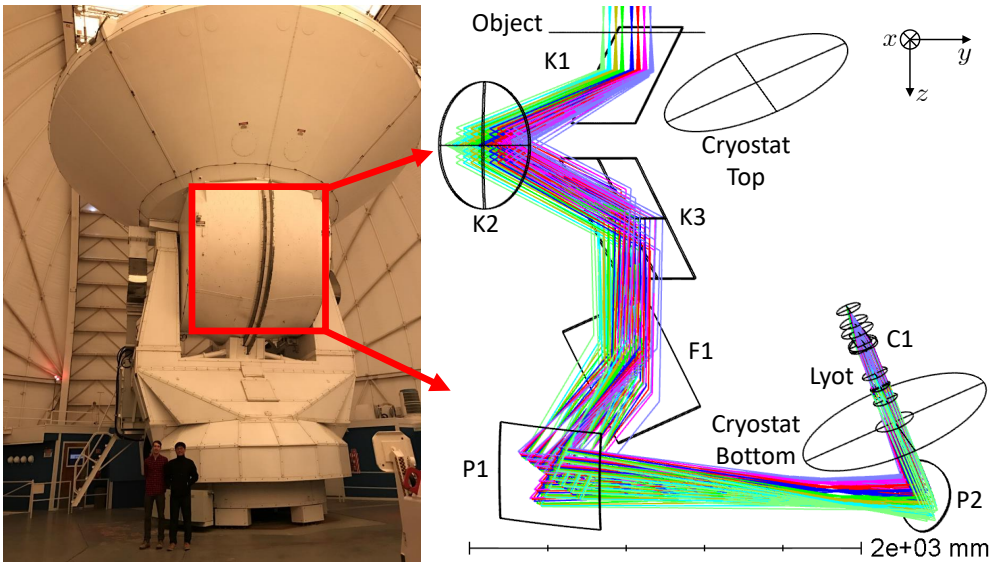


Fig. 5. 3D wire frame view of the millimeter-wave instrument design created for the Tomographic Ionized-carbon Mapping Experiment (TIME). The folded path was required to fit within a cabin of an existing telescope, and as such needed freeform optics to maintain optical performance. This design was optimized using human intuition, where the surfaces K2, P1, P2, and C1 are freeform. The remaining surfaces (K1, K3, and F1) are flat. Note that surface C1 is a refractive optic where one side is flat and the other is freeform.

Table 1. Basic optical prescription of the TIME instrument which is used in the case study on optimal freeform surface selection. Up to 6th order XY-polynomials and the first 36 Zernike terms were used during optimization.

Surface	Type	Spacing (mm)	Tilt X (deg)	Tilt Y (deg)	Tilt Z (deg)
Object	–	-150	0	0	0
K1	Flat	1100	55	0	[-45, 45]
K2	XY-polynomial	1100	20	0	[-45, 45]
K3	Flat	750	55	0	[-45, 45]
F1	Flat	1650	55	0	120
P1	Standard Zernike	2600	-10	-17	0
P2	XY-polynomial	873.65	-28.75	-1.6	60
C1 (front)	XY-polynomial	40	0	0	0
C1 (back)	Flat	201.32	0	0	0
Image	Flat	0	0	0	0

Optics EXPRESS

(i.e. worst case). The third case was selected to demonstrate that the TIME design is not in a gentle and wide local solution space, which would lead to a final design that is independent of the path guidance. These three cases were chosen to compare the human guided (case 1) to the fitness function guided (case 2) and show that by implementing an analytic method of choosing the surface to apply freeform optimization, we can obtain a better performing optical system. Furthermore, the third case was selected to demonstrate that the proposed methodology was not obtaining a result by coincidence, that in fact our fitness function definition given by Eq. (18) was seeking the specified goal. A summary of the results from this case study is shown in Table 2, where the order of surface selection and final merit function are listed.

Table 2. Overview of the case study results showing the final merit function value obtained by following the specified surface order and methodology. The same merit function definition is used in all cases and is a combination of multiple performance metrics: spot size, Lyot stop quality, telecentricity, and distortion, which are set by the TIME science goals [28, 29].

Surface Order	Methodology	Merit Function Value
K2 P1 P2 C1	Human Intuition	0.0151
P1 K2 P2 C1	Fitness (Optimal)	0.0124
P2 C1 K2 P1	Fitness (Worst-case)	0.0153

The fitness function given by Eq. (18) combines parameters with different units and individual weights per surface. This allows for maximum flexibility when defining what metrics are important to the designer, but also leads to larger uncertainty about how to choose the weighting. In the presented cases, we decided that the weighting was best expressed as the value with which the parameter was normalized. For instance, the RMS of the slope difference was weighted by the factor that made the parameter expressed in terms of a normalized RMS. This weighting allowed us to concisely compare the parameters, but also did not select against large surface slopes, which is not typically allowable. We were able to not consider the impact on surface deviation and slope because the fabrication methods used can handle practically any surface shape required. Furthermore, because the manufacturability of the surfaces was not a driving metric, the PSD metric was weighted by zero, and therefore not included in the fitness function.

The general procedure for optimizing each cases design is as follows. The starting point design is loaded into OpticStudio and the chosen methodology is used to select a surface at which to apply freeform optimization. The selected surface definition is changed to handle freeform terms. For the standard Zernike surface (P1), we allowed terms Z5–36 and up to 6th order aspheric terms to vary, while for an XY-polynomial surface (K2, P2, and C1) we varied all terms up to 6th order excluding terms with a linear component (i.e., xy^n or $x^n y$). The design is then optimized given a fixed merit function definition until there is no significant change in the merit function value over multiple cycles of the optimizer. This process repeats until all four potential surface candidates are optimized, where each previously optimized surface is allowed to continue to vary its prescription as further surfaces are optimized. We allow all non-flat surfaces to vary their respective parameters (freeform and non-freeform) during the optimization process so that the design can mature as more surfaces are made freeform. As each surface is selected for freeform optimization, the same parameters are chosen to describe the surface, maintaining consistency across each case. We evaluate the design's final merit function value, which determines how well the methodology performed given the desired outcome. Since the merit function definition is the same across all cases, this single number represents the optimized performance of the design.

Optics EXPRESS

4.1. Case-1: human intuition guided

The human guided case follows the same surface selection ordering as used in the design of the TIME instrument, but uses a slightly modified merit function definition and employs a different style of optimization. These changes from the real instrument design were made in order to eliminate systematic sources of error and create an apples-to-apples comparison in this work. Furthermore, we acknowledge that this case result may vary depending on the designer, since this approach is by default, subjective. Since the final performance already met the required science performance, we stopped optimization and it is now being manufactured.

4.2. Case-2: fitness function guided for best performance

To guide the design towards the lowest merit function, we use the results from the fitness function methodology present in Sec 3.2 to select the surface for freeform optimization that obtains the largest fitness function value. The surface selection ordering and the value for each evaluation of the fitness function on the freeform candidate surfaces is summarized in Table 3, where after a surface has been selected, it is removed from further evaluations, as represented by the dashes.

Table 3. Fitness function evolution throughout the objectively guided case for best performance. The largest value of fitness function determined which surface would be chosen next for freeform optimization.

		Fitness Value			
Step	Selected Surface	K2	P1	P2	C1
1	P1	0.688	0.772	0.415	0.662
2	K2	0.791	–	0.539	0.647
3	P2	–	–	0.565	0.393
4	C1	–	–	–	–

4.3. Case-3: fitness function guided for worst performance

After obtaining the result from the fitness function guided method for best performance and finding that it led to a more optimal design, we wanted to check if the methodology was actually guiding the optimization in the direction we expected, even for the worst case. To accomplish this, we selected the least optimal surface for freeform optimization at each step in the process. A summary of these results is shown in Table 4, where the selected surface and fitness function values for each surface are listed.

5. TIME optical performance comparison

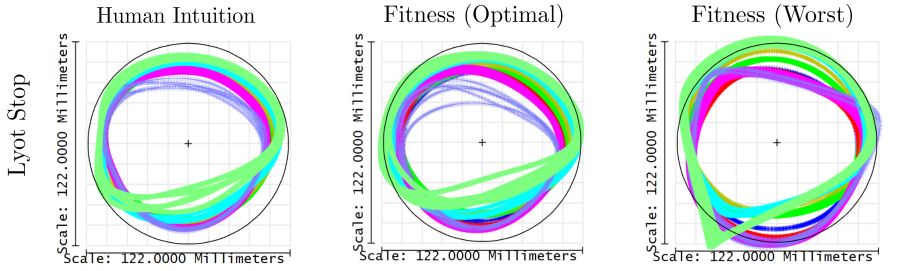
The three cases presented above show that the fitness function method provides a means of analyzing a design in a meaningful way. The fitness function definition and analysis methodology work together to give the designer a non-biased and objective way of selecting a surface to optimize. But what is the actual performance difference between using the analytic method and a human's intuition? The answer depends strongly on the merit function definition, as well as the specific design, but we now present a comparison between the three cases to give the proposed methodology more grounding in optical performance. The merit function we used during optimization was based on a combination of minimizing the geometric spot size, uniformity of the Lyot stop, image-space telecentricity, and distortion as the TIME science

Table 4. Fitness function evolution throughout the objectively guided case for worst performance. The lowest value of fitness function determined which surface would be chosen next for freeform optimization.

		Fitness Value			
Step	Selected Surface	K2	P1	P2	C1
1	P2	0.688	0.772	0.415	0.662
2	C1	0.732	0.561	–	0.516
3	K2	0.270	0.756	–	–
4	P1	–	–	–	–

requirements dictate. These four metrics are shown for each of the three cases used to validate the proposed methodology in Fig. 6 and 7.

We have separated the metrics into two categories, where the imaging and distortion metrics were straight-forward and intuitive to optimize for, while the Lyot performance and the telecentricity were complex and less-stable merit function definitions. The imaging performance is



Telecentricity

Configuration	Maximum Chief Ray Angle (deg)	Configuration	Maximum Chief Ray Angle (deg)	Configuration	Maximum Chief Ray Angle (deg)
1	0.053	1	0.038	1	0.031
2	0.053	2	0.038	2	0.033
3	0.057	3	0.035	3	0.035
4	0.061	4	0.034	4	0.043
5	0.058	5	0.031	5	0.050

Fig. 6. Complex optical performance specification comparison between the three cases of guiding methodology. Each color represents a different field point in the design, while the five configurations of the K-mirror are plotted on top of one another in the Lyot metric. For each K-mirror configuration a maximum chief ray angle on the image plane is reported.

measured by the fraction of enclosed energy within a given radius, as computed using diffraction and aberrations. Distortion is measured by the deviation of the centroid location of the spot to the ideal image location, which for our case should be a straight vertical line. The Lyot uniformity is measured by looking at the circularity of the rays at the edge of the pupil and their deviation from one another. Telecentricity is measured by calculating the maximum chief ray angle on the image plane surface. We see that the human intuition case has the best Lyot performance, its internal pupil is most circular and uniform across the field and configurations, but suffers in all other metrics. The fitness function guided for worst performance has the poorest imaging performance and distortion, while its telecentricity is the 2nd best. The optimally guided design has the best telecentricity, imaging, and distortion, while its Lyot uniformity is worst. We note that depending

Optics EXPRESS

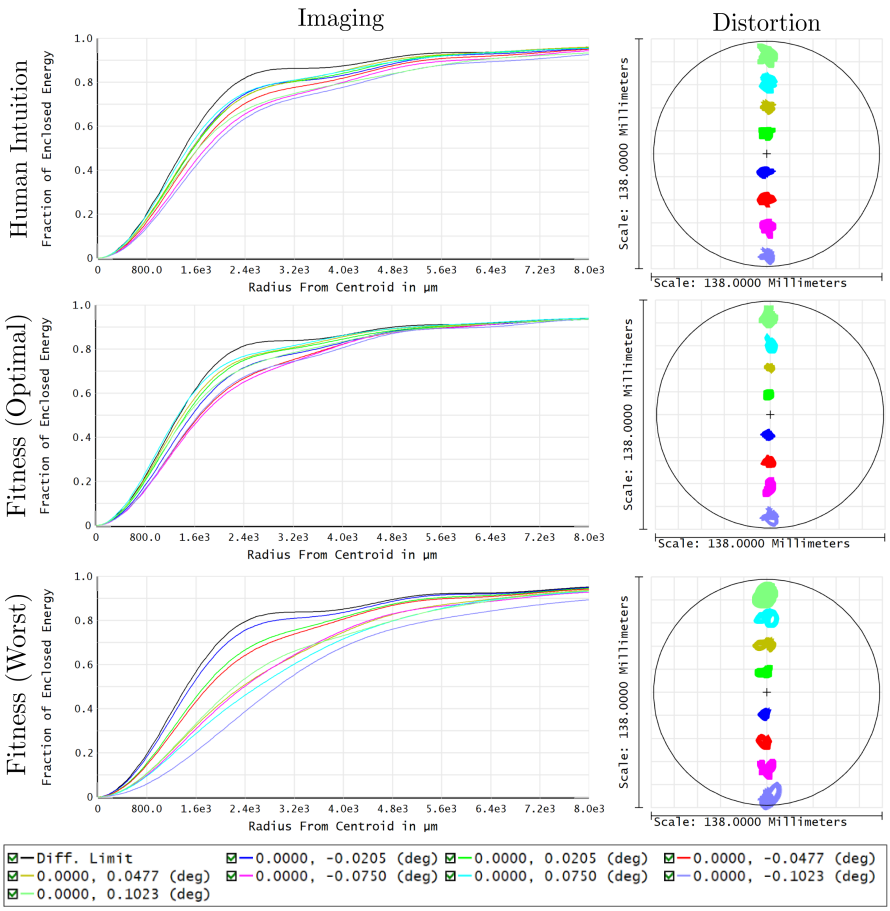


Fig. 7. Typically defined optical performance comparison between the three cases of guiding methodology. Each color represents a different field point in the design, while the imaging performance shows the diffraction limit as a solid black line and the distortion metric shows the overlay of rays at the image plane with a circle of diameter 135 mm in black.

on how the merit function is constructed, a different result is obtained. Therefore, while the fitness function guided for optimal performance has lowest Lyot performance, the structure of the merit function weights the other metrics as higher priority. Furthermore, the Lyot uniformity is a complex merit function to optimize for, and is more unstable than the other metrics.

6. Conclusion

In this work we present a methodology to objectively guide the selection of a surface in an optical design for freeform optimization. The methodology is based in the real ray data obtained from ray tracing an optical system in a time-forward and time-reversed sense. We show that for our proposed selection criteria, the methodology correctly informs the designer on what ordering of surface optimization will yield the desired performance outcome, whether that be optimal or worst. This work attempts to navigate the space between local and global optimization, providing overall design insights. We hope that this work may be used by the optical design community to inform the continuing development and blossoming of the freeform design space. The goal of the demonstrated parametric fitness approach is not to replace analytical or intuition based

approaches. Instead, this objective methodology will help a designer to check other possible optimization paths, which may lead to a hidden local minimum. This is especially true when the design has many optical surfaces and the designer has too many options. In this case, our proposed numerical approach can provide top candidates to the designer so that they can apply theoretical or analytical approaches to the filtered options.

Acknowledgments

The authors would like to acknowledge the II-VI Foundation Block-Gift Program for helping support general deflectometry research in the LOFT group, making this research possible. We would also like to thank Zemax for their student license of OpticStudio. Furthermore, we would like to acknowledge the support of the TIME collaboration.

References

1. K. P. Thompson and J. P. Rolland, "Freeform optical surfaces: a revolution in imaging optical design," Opt. Photonics News **23**, 30–35 (2012).
2. K. Fuerschbach, J. P. Rolland, and K. P. Thompson, "Theory of aberration fields for general optical systems with freeform surfaces," Opt. express **22**, 26585–26606 (2014).
3. J. Zhu, W. Hou, X. Zhang, and G. Jin, "Design of a low F-number freeform off-axis three-mirror system with rectangular field-of-view," J. Opt. **17**, 015605 (2015).
4. M. Beier, J. Hartung, T. Peschel, C. Damm, A. Gebhardt, S. Scheiding, D. Stumpf, U. D. Zeitner, S. Risse, R. Eberhardt, and A. Tünnermann, "Development, fabrication, and testing of an anamorphic imaging snap-together freeform telescope," Appl. Opt. **54**, 3530–3542 (2015).
5. K. Thompson, "Description of the third-order optical aberrations of near-circular pupil optical systems without symmetry," JOSA A **22**, 1389–1401 (2005).
6. K. Fuerschbach, G. E. Davis, K. P. Thompson, and J. P. Rolland, "Assembly of a freeform off-axis optical system employing three φ-polynomial Zernike mirrors," Opt. Lett. **39**, 2896–2899 (2014).
7. B. D. S. amd Joseph M. Howard, "Low-order aberration coefficients applied to design of telescopes with freeform surfaces," in *International Conference on Space Optics,* vol. 10562 (2017), pp. 10562 – 10562 – 9.
8. T. Yang, J. Zhu, X. Wu, and G. Jin, "Direct design of freeform surfaces and freeform imaging systems with a point-by-point three-dimensional construction-iteration method," Opt. Express **23**, 10233–10246 (2015).
9. T. Yang, J. Zhu, and G. Jin, "Starting configuration design method of freeform imaging and afocal systems with a real exit pupil," Appl. Opt. **55**, 345–353 (2016).
10. T. Yang, G.-F. Jin, and J. Zhu, "Automated design of freeform imaging systems," Light. Sci. & Appl. **6** (2017).
11. B. D. Stone and G. W. Forbes, "Foundations of first-order layout for asymmetric systems: an application of Hamilton's methods," J. Opt. Soc. Am. A **9**, 96–109 (1992).
12. B. D. Stone and G. W. Forbes, "Foundations of second-order layout for asymmetric systems," J. Opt. Soc. Am. A **9**, 2067–2082 (1992).
13. R. Tang, B. Zhang, G. Jin, and J. Zhu, "Multiple surface expansion method for design of freeform imaging systems," Opt. Express **26**, 2983–2994 (2018).
14. A. Bauer, E. M. Schiesser, and J. P. Rolland, "Starting geometry creation and design method for freeform optics," Nat. Commun. **9** (2018).
15. Y. Zhong and H. Gross, "Initial system design method for non-rotationally symmetric systems based on Gaussian brackets and nodal aberration theory," Opt. Express **25**, 10016–10030 (2017).
16. J. L. Bentley and C. Olson, *Field guide to lens design* (SPIE, 2012).
17. R. R. Shannon, *The art and science of optical design* (Cambridge University, 1997).
18. A. Yabe, "Method to allocate freeform surfaces in axially asymmetric optical systems," Proc. SPIE **8167**, 816703–1–10 (2011).
19. A. Yabe, "Representation of freeform surfaces suitable for optimization," Appl. Opt. **51**, 3054–3058 (2012).
20. C. Liu and H. Gross, "Numerical optimization strategy for multi-lens imaging systems containing freeform surfaces," Appl. Opt. **57**, 5758–5768 (2018).
21. M. Aftab, J. H. Burge, G. A. Smith, L. Graves, C. J. Oh, and D. W. Kim, "Modal data processing for high resolution deflectometry," Int. Jnl Precis. Eng. Manuf. Technol. (2018).
22. L. Fox and I. Parker, *Chebyshev polynomials in numerical analysis*, Oxford mathematical handbooks (Oxford University, 1968).
23. J. C. Mason and D. C. Handscomb, *Chebyshev polynomials* (Chapman and Hall/CRC, 2002).
24. M. Aftab, J. H. Burge, G. A. Smith, L. Graves, C. jin Oh, and D. W. Kim, "Chebyshev gradient polynomials for high resolution surface and wavefront reconstruction," in *Optical Manufacturing and Testing XII,* vol. 10742 (2018), pp. 10742 – 10742 – 7.
25. I. Kaya, K. P. Thompson, and J. P. Rolland, "Comparative assessment of freeform polynomials as optical surface descriptions," Opt. Express **20**, 22683–22691 (2012).

Optics EXPRESS

26. D. Swift, "Image rotation devices–a comparative survey," Opt. & Laser Technol. **4**, 175 – 188 (1972).
27. P. T. Wallace, "A rigorous algorithm for telescope pointing," in *Astronomical Telescopes and Instrumentation,* vol. 4848 (2002), pp. 4848 – 4848 – 12.
28. J. Hunacek, J. Bock, C. M. Bradford, B. Bumble, T.-C. Chang, Y.-T. Cheng, A. C. Asantha Cooray, S. Hailey-Dunsheath, Y. Gong, C.-T. Li, R. O'Brient, E. Shirokoff, C. Shiu, J. Sun, Z. Staniszewski, B. Uzgil, and M. Zemcov, "Detector modules and spectrometers for the TIME-Pilot [CII] intensity mapping experiment," in *Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy VIII,* vol. 9914 (2016), pp. 9914 – 9914 – 10.
29. J. Hunacek, J. Bock, C. M. Bradford, V. Butler, T.-C. Chang, Y.-T. Cheng, A. Cooray, A. Crites, C. Frez, S. Hailey-Dunsheath, B. Hoscheit, D. W. Kim, C.-T. Li, D. Marrone, L. Moncelsi, E. Shirokoff, B. Steinbach, G. Sun, I. Trumper, A. Turner, B. Uzgil, A. Weber, and M. Zemcov, "Hafnium films and magnetic shielding for TIME, a mm-wavelength spectrometer array," J. Low Temp. Phys. (2018).