Fast wave-front reconstruction in large adaptive optics systems with use of the Fourier transform

Lisa A. Poyneer, Donald T. Gavel, and James M. Brase

Lawrence Livermore National Laboratory, Livermore, California 94550

Received March 22, 2002; revised manuscript received May 14, 2002; accepted May 20, 2002

Wave-front reconstruction with the use of the fast Fourier transform (FFT) and spatial filtering is shown to be computationally tractable and sufficiently accurate for use in large Shack–Hartmann-based adaptive optics systems (up to at least 10,000 actuators). This method is significantly faster than, and can have noise propagation comparable with that of, traditional vector–matrix-multiply reconstructors. The boundary problem that prevented the accurate reconstruction of phase in circular apertures by means of square-grid Fourier transforms (FTs) is identified and solved. The methods are adapted for use on the Fried geometry. Detailed performance analysis of mean squared error and noise propagation for FT methods is presented with the use of both theory and simulation. © 2002 Optical Society of America

OCIS codes: 010.0010, 010.1080.

1. INTRODUCTION

Current adaptive optics (AO) systems use vector–matrix-multiply (VMM) reconstructors to convert gradient measurements to wave-front phase estimates. As the number of actuators n increases, the time to compute the reconstruction by means of the VMM method scales as $O(n^2)$. The number of actuators involved in AO systems is expected to increase dramatically in the future. In astronomical applications, this is due to both increasing telescope diameters and new higher-resolution applications on existing systems. There are many other applications, including high-resolution laser beam control and communications systems. This increase in size, from hundreds up to tens of thousands of actuators, requires a faster method for wave-front reconstruction.

A wave-front reconstruction method with use of the discrete Fourier transform (DFT) was suggested by Freischlad and Koliopoulos.[1] This method is for square apertures on the Hudgin geometry. In a further paper,[2] the same authors derived methods for additional geometries, including the Fried geometry, which uses one Shack–Hartmann (SH) sensor in each subaperture to produce gradient measurements. Freischlad also considered the case of small circular apertures.[3] This paper builds on that work in four important ways. First, the circular-aperture case is thoroughly examined. The boundary problem is identified, showing that use of only in-aperture data for circular apertures leads to large errors. Two methods for solving this boundary problem are presented. They both provide perfect reconstruction of sensed modes when no noise is present. Second, the Fourier transform (FT) method and boundary techniques are adapted for use on the Fried geometry. Third, the performance of these methods, in terms of both speed and reconstruction error, is analyzed. Reconstruction is treated as an estimation problem, which leads to a linear model of system error in response to noise. Theoretical results for small apertures are confirmed by simulation. Finally, the performance of large systems (up to 50,000 actuators) is examined through simulation. With the use of an FT method presented in this paper, the implementation of a 10,000-actuator system with satisfactory speed and reasonable error performance is feasible given current technology.

This paper is focused on the performance of FT reconstructors with the use of specific discrete models. Therefore it will not directly address phenomena associated with the correction of continuous wave fronts such as branch points or partially obscured subapertures. In particular, deformable mirror (DM) influence functions are not considered. Unlike VMM reconstructors, which are in practice obtainable directly from the AO system DM and sensors, FT methods are filters derived to fit certain specific sensor models. Current research by the authors addresses these more complex issues and will be presented in a subsequent paper.

2. INVERSE SPATIAL FILTER

The basic inverse spatial filter, first derived by Freischlad,[1] is presented here again for reference, with an emphasis on its derivation in terms of an inverse filtering problem, as opposed to modal expansion over Zernike polynomials.[2] Using the Hudgin geometry,[4] we model the sensor measurements as the first differences of the wave-front phase. This corresponds to wave-front sensors centered between each pair of points (see Fig. 1). The piston-removed (across the aperture) phase $\phi[m, n]$ is an $N \times N$ discrete signal. The gradients $s_x[m, n]$ and $s_y[m, n]$ are simply the first differences between adjacent phase points:

$$s_x[m, n] = \phi[m + 1, n] - \phi[m, n], \tag{1}$$

$$s_y[m, n] = \phi[m, n + 1] - \phi[m, n]. \tag{2}$$

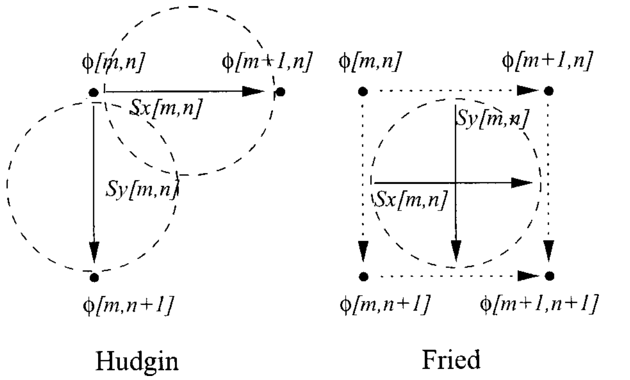


Fig. 1. Hudgin and Fried sensor geometries. The circles represent the Shack–Hartmann wave-front sensor locations. In the Hudgin geometry, the gradients are first differences. In the Fried geometry, the gradients are the average of the two nearest first differences to the subaperture.

The DFT is applied to a finite-duration signal, under the assumption that the signal is periodic. The forward transform of a spatial signal of size $N \times N$ is

$$X[k,\, l] = \mathcal{F}\{x[m,\, n]\} = \frac{1}{N^2}\sum_{p=0}^{N-1}\sum_{q=0}^{N-1} x[p,\, q]\exp\left[-\frac{j2\pi}{N}(kp + lq)\right]. \tag{3}$$

With use of the shift property of the DFT, Eqs. (1) and (2) become

$$S_x[k,\, l] = \Phi[k,\, l]\left[\exp\left(\frac{j2\pi k}{N}\right) - 1\right], \tag{4}$$

$$S_y[k,\, l] = \Phi[k,\, l]\left[\exp\left(\frac{j2\pi l}{N}\right) - 1\right]. \tag{5}$$

To get the inverse filter, multiply each of the above equations by the complex conjugate of its exponential term and combine them, solving for $\Phi[k,\, l]$:

$$\hat{\Phi}[k,\, l] = \begin{cases} 0, & k,\, l = 0 \\ \left\{\left[\exp\left(-\frac{j2\pi k}{N}\right) - 1\right]S_x[k,\, l] + \left[\exp\left(-\frac{j2\pi l}{N}\right) - 1\right]S_y[k,l]\right\} \times \left[4\left(\sin^2\frac{\pi k}{N} + \sin^2\frac{\pi l}{N}\right)\right]^{-1}, & \text{else} \end{cases}. \tag{6}$$

The pole of the filter at $k,l = 0$ is fixed by making that value zero, which sets the dc gain (or piston) of the wave-front phase across the whole square grid to zero. As this mode is disregarded in reconstruction, it does not add any error. Taking the inverse transform produces the estimate $\hat{\phi}[m,\, n]$.

The performance of this reconstruction method has already been analyzed for square apertures[1,2] and on small circular apertures.[3] However, a systematic study of the applicability of this filter on large circular apertures has not been done. Section 3 presents the results of such a study.

3. RECONSTRUCTION ON A CIRCULAR APERTURE

The inverse filter was derived for a regular grid of gradient measurements. When one is dealing with a real AO system (astronomical telescopes in particular), the gradients are typically available only on a circular aperture. The measurement data cannot be simply zero padded and filtered. Doing so produces huge errors. See Fig. 2 for an illustration of these errors.

First, this boundary problem is identified and explained. Then two methods for altering the gradient data of a circular aperture are presented. These methods produce perfect reconstruction of sensed modes in the absence of noise.

There are two key assumptions in the inverse filter derivation that must be satisfied for it to work. The first assumption is that ϕ is spatially periodic. This assumption is necessary for use of the DFT method, and it must be maintained for a set of gradient measurements. A check on this condition is that the sum of every row (for x gradients) or column (for y gradients) in the $N \times N$ gradient signal equals zero. The second assumption is based on the modeling of the gradients as first differences. Any closed path of gradients must sum to zero. Both of these conditions have been identified in earlier work.[1,3]

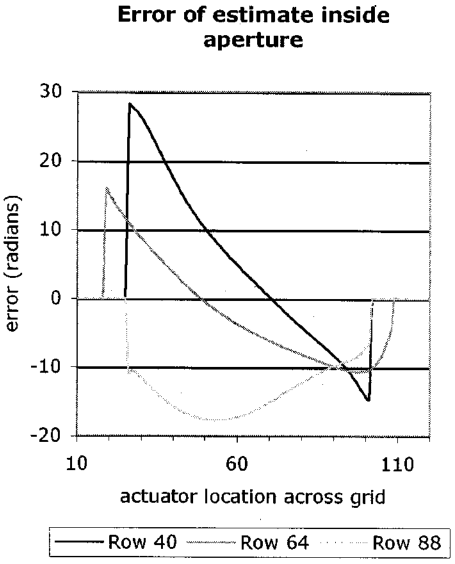


Fig. 2. Estimate error from reconstruction of just the gradients inside the aperture, with the rest set to zero. No noise was added so as to clearly isolate the effects of the boundary. The gradients were calculated directly from phase points by using Eqs. (1) and (2). Each curve in the plot is a slice across the aperture along a row of actuators. This simulation was done on a 6376-actuator system, which was 90 actuators in diameter on a 128×128 grid. The input phase aberration had an rms error of 1690 nm. The reconstruction had an rms error of 1002 nm for this trial. The error spans the aperture and is not easily removed. (Note how the error changes shape and sign from row 40 to row 88.) With the use of either of the boundary methods, the reconstruction error was essentially 0 nm.

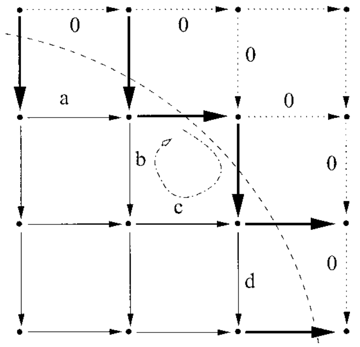


Fig. 3. The three types of gradients are shown in the Hudgin geometry at the edge of an aperture. Bold lines are the boundary gradients b_x and b_y. These connect phase points across the aperture edge. Thin solid lines are the inside gradients i_x and i_y, which can be obtained from measurement. The dotted lines are the outside gradients. Note that a closed loop across the aperture edge does not sum to zero if the boundary gradients are set equal to zero.

It is important to note that this spatial periodicity assumption does not require that the sensed wave front be inherently periodic or that reconstruction be done on a very large grid to obtain correct low-frequency components. Just as VMM reconstructors work while estimating only the phase values inside the aperture, FT methods accurately reconstruct on grids of aperture size. Consider an aperture with N actuators (phase points) across the diameter. In discrete space, the lowest frequency that can be represented is $2\pi/N$, which has period N. A phase aberration of a pure sinusoid $\sin(\pi x/N)$, where x represents the actuator index across the aperture, has period $2N$, which is twice the width of the aperture. It can still be sensed and reconstructed correctly, however. That is because only one half of a period of the sinusoid is sensed in the aperture, and it is this segment of the sinusoid that is repeated, on account of spatial periodicity. If gradients are taken exactly with Eqs. (1) and (2), any discrete signal will be exactly reconstructed by the filter in Eq. (6), minus the piston.

Because of the spatial periodicity and closed-path-loop conditions, zero-padding the gradient measurements is incorrect. Doing so violates both conditions in general. These inconsistencies manifest themselves in errors that span the aperture. The errors do not become less significant as the aperture size increases. Unlike the square-aperture case,[2] the amplitude of the error remains large and spans the circular aperture.

Proof of this comes from an examination of the Hudgin geometry and reconstruction process. Consider the gradients taken from the wave-front phase $\phi[m, n]$ across a circular aperture on a square grid. There are three types of gradients, which are illustrated in Fig. 3. The gradients from sensors inside the aperture are the inside gradients $i_x[m, n]$ and $i_y[m, n]$. The gradients that cross the aperture edge are the boundary gradients $b_x[m, n]$ and $b_y[m, n]$. When the data are zero padded outside the aperture, these boundary gradients are incorrectly set to zero. Last, the gradients wholly outside the aperture are the outside gradients. These can safely be considered to be zero everywhere. By linearity, the correct gradient sets can be written as a sum of the inside and boundary gradients:

$$s_x[m, n] = i_x[m, n] + b_x[m, n], \tag{7}$$

$$s_y[m, n] = i_y[m, n] + b_y[m, n]. \tag{8}$$

Because the filtering process is linear, we can consider the results of filtering each part separately. Filtering the whole $s_x[m, n]$ and $s_y[m, n]$ will produce an exact reconstruction. Taking only the values inside the aperture and zero padding is equivalent to filtering just $i_x[m, n]$ and $i_y[m, n]$. This means that the error of this estimate is exactly the result of filtering $b_x[m, n]$ and $b_y[m, n]$. It is this term that generates the large errors when zero padding.

The previous development suggests that a method for obtaining a correct set of gradients involves estimation of the boundary and/or outside gradients. Two different methods are now presented that generate consistent sets of gradients. With no noise, both methods produce perfect reconstruction of all the sensed modes. Furthermore, it will be shown below that the methods require only $O(n)$ operations, preserving the speed advantage of the FT methods.

A. Boundary Method

The first method will be called the boundary method because it estimates the gradients that cross the boundary of the aperture. It follows directly from the development above of inside, boundary, and outside gradients. This process is shown in Fig. 4.

Only the inside gradients are known from measurement. The outside gradients can all be set to zero. This leaves the boundary gradients undetermined. A loop continuity equation can be written for each of the two smallest loops that involve a boundary gradient. Setting each of these equations to zero describes a solution that satisfies loop continuity across the whole grid. With use of the configuration shown in Fig. 4, a partial list of the loop equations is

$$\begin{aligned} &-u1 + u2 = a, && u2 + u3 = 0, && \\ &u3 + u4 = c + b && u4 + u5 = 0 && u5 - u6 = d. \end{aligned} \tag{9}$$

All of these loop continuity equations involving the boundary gradients combine to form a linear system. Where $\mathbf{u}$ is the vector of all boundary gradients and $\mathbf{c}$ is a vector containing sums of measured gradients, the system can be expressed as

$$\mathbf{Mu} = \mathbf{c}. \tag{10}$$

The matrix $\mathbf{M}$ is fixed for a given geometry. The vector $\mathbf{c}$ has a fixed combination of gradients, but the value of these gradients depends on the actual measurement. If there is no noise, the system has an infinite number of valid solutions. Each of these solutions represents, in essence, a different piston offset of the aperture from zero phase.

When there is noise on the measurements, this system has no exact solution in general. The boundary gradients can instead be estimated by using a least-squares fit. This estimation can be solved by such methods as using

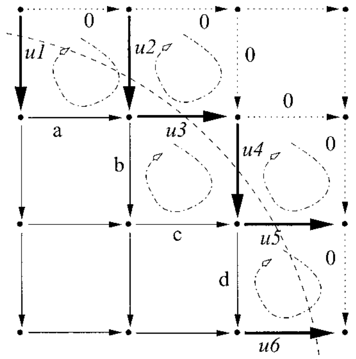


Fig. 4. Boundary method. Setting each closed loop across the aperture edge to zero results in an equation relating the unknown boundary gradients to the measured inside gradients and the zeroed outside gradients. In this example, the equations for the boundary gradients $u1$, $u2$,... are as follows, starting from the upper left corner: $-u1 + u2 = a$, $u2 + u3 = 0$, $u3 + u4 = c + b$, $u4 + u5 = 0$, and $u5 - u6 = d$. The complete set of equations for the whole aperture forms a linear system, which is then solved for the estimate of the boundary gradients.

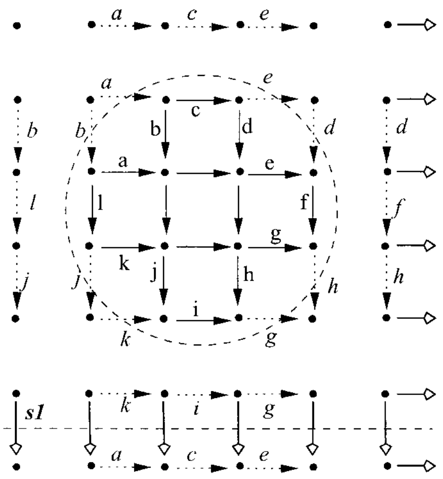


Fig. 5. Extension method, shown for $N = 6$. The values of the gradients closest to the aperture edge are repeated outside the aperture. For example, gradients a, c, and e are each topmost in their columns and are extended upward out of the aperture. The unmodified gradients, which are left as zeros, are not shown in this figure for clarity. The seam gradients are along the right and bottom edges. These gradients "connect" the spatially periodic copies of the wave-front phase, and they must be set so that every row of the x gradients and every column of the y gradients sum to zero. For example, the leftmost column in this case must satisfy the equation $s1 = -b - l - j$. Examination of this figure shows how loop continuity is satisfied exactly by the extension method.

the pseudoinverse of $\mathbf{M}$. In this case, the boundary gradients are correlated with the noisy measurements but generally have much higher variance than the noise.

B. Extension Method

The second method of obtaining a consistent set of gradients will be called the extension method because it extends out the gradients from inside the aperture. It is illustrated in Fig. 5.

The extension method extends the wave-front shape to outside the aperture. It does this by a simple method that is based on preserving loop continuity. The x gradients are extended up and down out of the aperture, while the y gradients are extended to the left and the right. For example, the uppermost x gradient in a given column in the aperture has its value repeated in all the outside gradients above it. Loop continuity is preserved, even where the extended gradients cross each other, such as at the "corners" of the aperture. All the smallest loops involving these new extended gradients will automatically sum to zero. Each extended gradient is canceled out in the smallest-loop equation by one with the identical value below or above it.

The final step is to fix the spatial periodicity. The seam gradients are those that connect one copy of the phase signal to the next. Observe in Fig. 5 at the bottom how the extended values from the top of the aperture meet those from the bottom. The seam gradients connect them. These seam gradients are set from the spatial periodicity condition by setting the sum of their row or column to zero. This satisfies the smallest-loop conditions for the seam gradients as well.

The extension method produces a completely consistent set of gradients. These provide perfect reconstruction of the phase when there is no noise, except for the piston. If there is noise, the same procedure is done, though loop continuity will not hold on loops involving the seam gradients, just as the boundary gradients in the boundary method were the best, but not an exact, solution when there was noise.

The principles and the methods used above can be applied to fill in missing information in general, such as with different aperture shapes or when there is central obscuration due to the secondary mirror in a large telescope.

4. ADAPTING TO THE FRIED SENSOR GEOMETRY

The inverse filter as described above is specific to the Hudgin geometry. The Fried geometry[5] is frequently used, however, in modeling the behavior of SH sensors. It models the gradients that are generated by SH sensors, which are centered between phase points. This allows one sensor to provide both x- and y-gradient measurements. See Fig. 1 for an illustration and a comparison with the Hudgin geometry. This sensor configuration is common in astronomical AO systems. Its features and implications have to be considered if an FT method is used for reconstruction. In this section, the inverse filter is derived for the Fried geometry and the boundary and extension methods are adapted to it. This treatment has some similarities to, but also significant differences from, the Fried-geometry consideration of Freischlad.[2]

A. Filter Derivation

In the Fried geometry, the gradient is modeled as the average of the two nearest first differences:

$$s_x[m, n] = \tfrac{1}{2}(\phi[m + 1, n] - \phi[m, n] + \phi[m + 1, n + 1] - \phi[m, n + 1]), \tag{11}$$

$$s_y[m, n] = \tfrac{1}{2}(\phi[m, n+1] - \phi[m, n] + \phi[m+1, n+1] - \phi[m+1, n]). \tag{12}$$

Using the same method as that above, we can derive the inverse spatial filter to reconstruct the phase:

$$\hat{\Phi}[k, l] = \begin{cases} 0, & k, l = 0, k, l = N/2 \\ \left\{\left[\exp\left(-\frac{j2\pi k}{N}\right) - 1\right]\left[\exp\left(-\frac{j2\pi l}{N}\right) + 1\right]S_x[k, l] \right. \\ \left. +\left[\exp\left(-\frac{j2\pi l}{N}\right) - 1\right]\left[\exp\left(-\frac{j2\pi k}{N}\right) + 1\right]S_y[k, l]\right\} \\ \times \left[8\left(\sin^2\frac{\pi k}{N}\cos^2\frac{\pi l}{N} + \sin^2\frac{\pi l}{N}\cos^2\frac{\pi k}{N}\right)\right]^{-1}, & \text{else} \end{cases}. \tag{13}$$

This filter has not been previously derived. This filter also has the pole at the piston mode and an additional pole at the highest-frequency, or waffle, mode. Both poles are zeroed out. This latter mode is not sensed in the Fried geometry. Nor is it normally controlled for in AO systems. A waffle error will therefore be present, the magnitude of which depends on the amount of waffle in the input phase. In practice, under atmospheric turbulence this waffle component is actually quite small.

We can calculate the variance of the Kolmogorov waffle component as follows. The Kolmogorov spectrum is given by[6]

$$S_\phi(k) = 0.023k^{-11/3}r_0^{-5/3}, \tag{14}$$

where k is the spatial frequency. The variance is the integral of this spectrum over a region of k-space centered at the waffle frequency $1/2d$ and having an extent Δk appropriate to the aperture size, $\Delta k \approx 1/D$; that is,

$$\sigma_{\phi_w}^2 \approx 4S_\phi(1/2d)\Delta k^2. \tag{15}$$

The piston-removed wave-front variance as derived by Noll[6] is given by

$$\sigma_\phi^2 = 1.03(D/r_0)^{5/3} \tag{16}$$

and dividing produces

$$\sigma_{\phi_w}^2 \approx 1.13(d/D)^{11/3}\sigma_\phi^2. \tag{17}$$

This means that for 13 or more subapertures across the diameter of the telescope, the waffle component is at most 0.01% of the total piston-removed wave-front variance. Therefore the error due to this missed waffle component is very small.

B. Dealing with a Circular Aperture

We now know that the values of the gradients outside the aperture must be estimated with some method for proper reconstruction. But in the Fried geometry the gradients do not connect points, so the boundary and extension methods cannot be applied directly. Using the same method as Freischlad,[2] a simple invertible linear transform can convert the gradients into two sets that do connect points on the grid. The gradients are now oriented along a different orthogonal basis, in the directions referred to as a and b. Figure 6 illustrates the two new unconnected grids:

$$s_a[m, n] = s_x[m, n] + s_y[m, n], \tag{18}$$

$$s_b[m, n+1] = s_x[m, n] - s_y[m, n]. \tag{19}$$

Here this method diverges from Freischlad's. Instead of using a filter for this new geometry,[2] we treat the two grids separately as cases of the Hudgin geometry. With a few minor modifications, such as making the new grids square shaped, the boundary and extension methods can be applied directly to these two sets in the a and b directions.

The primary difficulty in using the Fried geometry arises from the recombination of two disjoint grids. These grids are independent of each other, which is a direct result of the waffle mode being unsensed and uncontrollable. Processing each grid independently and then recombining them can result in large waffle errors. Fortunately, when no noise is present, this waffle error is simple to remove, as it has constant magnitude in the aperture.

When there is noise, matters become more difficult. In practice, the boundary and extension methods have dissimilar noise properties after filtering. These error terms are shown in Fig. 7. The boundary method was observed to cause wafflelike errors that varied widely across the aperture. The magnitude and the sign of the waffle typically changed across the aperture in ways dependent on the noise and the specific realization of phase aberra-

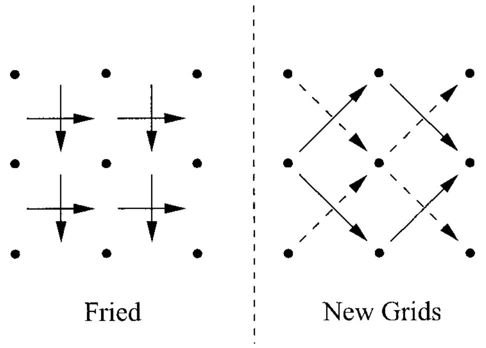


Fig. 6. The coordinate transform from the Fried geometry produces two disjoint grids. One grid is connected by the dashed gradients, the other by the solid gradients. Combining these uncoupled grids introduces an unknown waffle error.

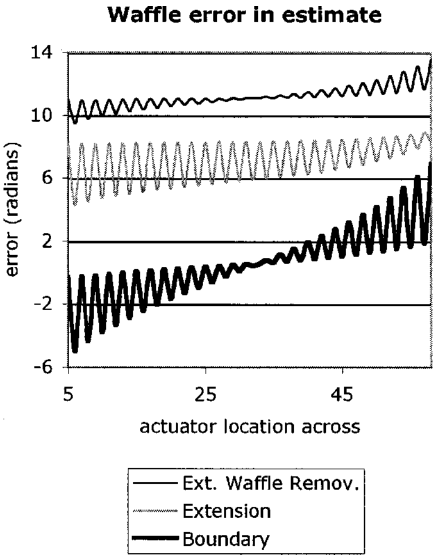


Fig. 7. Waffle is a significant concern in Fried-geometry reconstructions in noisy conditions. For a representative sample case with noise, the errors of the boundary and extension reconstructions, as compared with a perfect reconstructor, are shown. At the top is the reduced error in the extension case after waffle removal. Global waffle is completely removed, while local waffle, which is due to noise, remains.

tion. The noise could not be cleanly removed. The extension method produced more correlated noise on the estimate. It did not normally vary in sign, and the magnitude variations were much less severe. The extension method's noise could be removed for the most part in a simple waffle-removal step. The boundary method's error was severe enough to limit the usefulness of the reconstruction. This modal removal step is discussed in Section 5.

5. MODAL REMOVAL

As Section 4 has demonstrated, it is desirable to discard certain modes from the reconstruction. Piston is a mode that is discarded in the general case. A second mode that is normally removed is waffle. This is because Fried-geometry reconstructions, including FT methods, can introduce large waffle errors into the estimate. In a VMM reconstructor, this modal removal can be built directly into the matrix. Modal removal in an FT-based method is a separate step but is quick and easy to implement in $O(n)$ steps.

The fastest way to remove a mode from the spatial filtering estimate would be to identify the frequency coefficients for that mode and zero them. The spatial filter already does this for piston (and waffle in the Fried-geometry case) across the whole square grid. However, simply zeroing the frequency coefficients is not a good method in practice. This is because modes of interest inside the aperture are not compactly represented in frequency space, so determining the correct amount to remove is nontrivial. Second, though the power at high frequencies is low, it is essential for sharp features. Incorrectly removing high-frequency components can have deleterious effects on the accuracy of the estimate, especially at the edges.

Instead, the modal removal process is applied in the spatial domain. The coefficient for the mode is determined by projection. Where $v[m, n]$ is the mode of concern,

$$c_v = \frac{\sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \hat{\phi}[m, n] v[m, n]}{\sum_{m=0}^{N-1} \sum_{n=0}^{N-1} v[m, n] v[m, n]}. \tag{20}$$

Then the estimate with that mode removed is

$$\hat{\phi}_{-v}[m, n] = \hat{\phi}[m, n] - c_v v[m, n]. \tag{21}$$

This method completely removes all global waffle across the entire aperture. Local waffle will still remain. An orthogonal basis set must be used for modal removal. For example, a waffle-mode vector of ±1 over an odd number of actuators has a piston component that must be removed first.

6. COMPLETE METHODS

The previous sections have identified the boundary problem with circular apertures and presented methods to solve it. Both the Hudgin and Fried geometries have been considered. For the performance analysis, the best overall method for each geometry, as determined by analysis and simulation, is now presented. The complete methods are illustrated as a flow chart in Fig. 8.

The Hudgin geometry method will be called the Hudgin-FT method. Assuming this geometry, the FT reconstruction consists of the following steps. The extension method is applied to the measured gradients. Then the gradients are Fourier transformed, and the inverse filter [Eq. (6)] is applied. The result is inverse transformed, and piston is removed to obtain the final estimate.

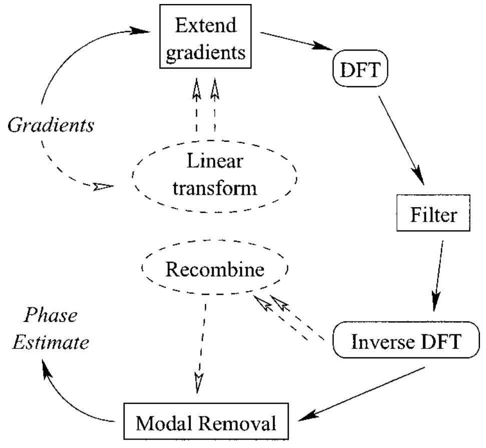


Fig. 8. Complete process of FT reconstruction. For the Hudgin-FT method, which is based on the Hudgin geometry, the gradient measurements are first extended. They are Fourier transformed, filtered, and inverse transformed before piston is removed. For the Fried-FT method, which is based on the Fried geometry, a few more steps are required. These are illustrated by the dashed arrows. The gradients are first converted to the two grids before they are extended and filtered. The two results are recombined, and then waffle and piston are removed.

The Fried-geometry method will be called the Fried-FT method. The gradients are first converted from the *x, y* directions to the *a, b* directions [recall Eq. (18)]. The two grids are made consistent by using the extension method. These two grids are Fourier transformed and then filtered individually with the use of Eq. (6). The results are inverse transformed and recombined. Waffle and piston are removed last.

These two methods reconstruct all sensed modes perfectly when there is no noise. Next, their speed and performance in the presence of noise will be analyzed.

7. TIME ANALYSIS

The whole purpose of exploring an FT approach to wave-front reconstruction is to exploit the speed of the fast Fourier transform (FFT). For sizes that are powers of 2, this implementation of the DFT is $O(n \log n)$ to compute, where n is the total number of elements. A VMM reconstruction on the same number of elements is $O(n^2)$. Finding a faster way to reconstruct than the VMM method is essential for large AO systems.

We must also consider the time cost of the extra processing steps in the FT reconstruction to ensure that it remains fast. These extra operations must be at most $O(n \log n)$ so as to not adversely affect performance. For the analysis of the extra processing, the following parameters are used. The width (in phase points) of the square grid is N. When the FFT is used, N is a power of 2. The radius of the aperture is R. The number of total elements is $n = N^2$. The number of actuators is approximately $a = \pi R^2$. When the FFT is used, the radius R varies from $N/4$ to $N/2$. If the radius were smaller than $N/4$, the aperture could be fitted on a smaller power-of-2-sized grid. A radius bigger than $N/2$ means the aperture is too large for the grid and must be reconstructed on the next power-of-2-sized grid. For the DFT, the radius R is slightly smaller than $N/2$.

For the Hudgin-FT method, the time analysis is simple. The first step is the extension process. In each dimension, $4R$ gradients are extended to fill out $2RN - a$ other gradients. Then $2N$ seam gradients are set, each by addition of $N - 1$ other gradients. If all of these operations are done sequentially, this amounts to $2(2RN - a) + 2N(N - 1 + 1)$ fundamental operations. For small apertures on large grids, R is at least $N/4$, making the number of the above operations $(3 - \pi/8)N^2$. This is $O(n)$. For very large apertures, where R is nearly $N/2$, there are $(4 - \pi/2)N^2$ operations. This is also $O(n)$. For example, as shown in Fig. 5, $N = 6$ and $R = 2$, leading to ten gradients set by extension, and 12 assignments to sums of five elements each, which is $10 + 12 \times (5 + 1) = 82$ operations.

After the signals are Fourier transformed, the filter is applied by multiplication and addition over the $N \times N$ grid. This is $O(1)$ operations over n elements, or $O(n)$. After the inverse transform, the piston is removed. This removal is an addition of all of the elements and a subtraction of the result from each one, which is again $O(n)$. Therefore the extra processing in the Hudgin-FT method is $O(n)$.

The Fried-FT method is not so simple. The first step is the linear transform of the gradients. This requires $2n$ additions. Then the extension process is applied to the two separate grids. In this case, R varies from $N/(4\sqrt{2})$ to $N/(2\sqrt{2})$. This results (as above) in $O(n)$ operations. Setting the two seam gradients for the two grids is $O(n)$. The same filtering as that in the Hudgin-FT method is done, but twice. The final waffle removal is the same amount of computation as that for the piston removal, namely, $O(n)$. So the extra processing in the Fried-FT method is also $O(n)$, though for a given geometry the Fried-FT method requires more than twice as many operations.

The processing for both the Hudgin-FT and Fried-FT methods is $O(n)$. The overall processing time for FT reconstruction is therefore dominated by the actual DFT implementation. For power-of-2-sized grids, the FFT is $O(n \log n)$. Using other size grids requires having a fast implementation, such as one based on prime factors.

Not only are the FT reconstructors faster in theory, but the implementation of a large system is within computational reach today. An estimate of performance requirements illustrates this. For an AO system running at 200 frames per second, the entire reconstruction must be done in 5 ms. This means that for the Hudgin-FT method the data extension, two forward FFTs, the filter application, and one inverse FFT must be completed in at most 5 ms. A good estimate of FFT time is 1 ms. For the Fried-FT method, there are twice as many operations, so a single FFT needs to be done in 0.5 ms. A 10,000-actuator system requires a 128 × 128 element FFT. This can be done in 1DL on a 1.7-GHz Pentium 4 in 1.8 ms. This result could be reduced by faster hardware or a different implementation.

8. PERFORMANCE ANALYSIS

It is clear that the FT methods are fast. It must now be shown that they are reasonably accurate as well. This is particularly important because these methods were not derived or proven to be optimal; they were developed to be fast. By modeling the reconstruction process as an estimation problem, a powerful framework for performance analysis can be established. In the case of white noise on the gradient measurements, the mean squared error of the reconstruction can be expressed as a linear function of the noise variance. This allows easy comparison of different methods. This section will develop the estimation-problem analysis and discuss the significance of the various error metrics.

A. Modeling the Measurement and Reconstruction Processes

The wave-front sensing and reconstruction process can be modeled by using vectors and matrices. This model is a simple one—it treats the aberrated wave-front phase as a set of discrete points. It allows for different geometries for gradient measurement (e.g., Hudgin and Fried) to be used, as well as arbitrary reconstruction methods.

The first step in the model is the conversion of the wave-front phase to gradient measurements. The gradients **g** are generated from the piston-removed wave front

ϕ through matrix $\mathbf{H}$ under additive noise $\mathbf{n}$. This noise is assumed to be zero mean and uncorrelated with the measurements and the wave front, with covariance matrix $\mathbf{\Lambda_n}$:

$$\mathbf{g} = \mathbf{H}\phi + \mathbf{n}. \tag{22}$$

The estimator function (e.g., reconstructor) can also be expressed as a matrix $\mathbf{M}$. If $\mathbf{M}$ is complex, as it is in the case of the FT method, it must be conjugated when it is transposed. However, the derivation below will omit notation of conjugation for clarity. The actuator commands $\hat{\phi}$ are estimated from measurements $\mathbf{g}$ with the matrix:

$$\hat{\phi} = \mathbf{Mg}. \tag{23}$$

Note that this matrix $\mathbf{M}$ can express nearly any linear reconstruction method, whether a standard minimum-least-squares VMM reconstruction, or the complete Hudgin-FT or Fried-FT method. It is required that the reconstructor $\mathbf{M}$ produce an estimate with zero piston. (If it does not, it can be easily converted to do so by multiplication with a piston-removal matrix.) Furthermore, $\mathbf{M}$ is static and does not adaptively change with conditions (as it would if it depended on the noise variance). This assumption is sufficient for the present analysis of FT reconstructors, though it is not applicable to the more general result that the optimal reconstruction matrix is dependent on the noise distribution.[7]

The error of this estimate is simply the difference between the estimate and the wave-front phase:

$$\epsilon = \mathbf{Mg} - \phi \tag{24}$$

The bias of the estimate is defined as the expectation of the error of the estimate:

$$\mathbf{b} = E(\epsilon). \tag{25}$$

The error variance is then

$$\mathbf{\Lambda}_\epsilon = E[(\epsilon - \mathbf{b})(\epsilon - \mathbf{b})^{\mathrm{T}}]. \tag{26}$$

The mean squared error is also of interest. The mean squared error is a random variable and is the average of the squared error at every point inside the aperture. Therefore we want to deal with its expectation. Recalling that a is the number of actuators, we have

$$\mathrm{mse} = \frac{E(\epsilon^{\mathrm{T}}\epsilon)}{a}. \tag{27}$$

B. Modeling the Wave Front

1. *Nonrandom Parameter Estimation*

The wave front has been frequently modeled as a nonrandom parameter to be estimated. When ϕ is deterministic, the bias and the error variance reduce to

$$\mathbf{b} = (\mathbf{MH} - \mathbf{I})\phi, \tag{28}$$

$$\mathbf{\Lambda}_\epsilon = \mathbf{M\Lambda_n M}^{\mathrm{T}}. \tag{29}$$

The error of the estimate is dependent on the model chosen to describe the wave-front sensor behavior, but the error variance is independent of that model and depends only on the reconstructor and the noise.

In this case, the noise propagator metric results from the error analysis. The noise propagator, called mse_{np}, is defined as the mean squared error divided by the average variance of the noisy slope measurements. The noise propagator is of dual use in analyzing the performance of a reconstruction method. For a single configuration of sensors and actuators, it determines how the reconstruction responds to noise. For a group of various configurations of increasing size, the set of their mse_{np}'s can describe how the system's size affects its response to noise. In the case of white noise of variance σ_n^2, the covariance matrix $\mathbf{\Lambda_n}$ is diagonal with value σ_n^2, which reduces the noise propagator expression to

$$\mathrm{mse}_{np} = \frac{\mathrm{mse}}{\sigma_n^2} = \frac{\mathrm{Trace}(\mathbf{MM}^{\mathrm{T}})}{a} \tag{30}$$

This result agrees with standard derivations.[8]

2. *Random-Vector Estimation*

If the wave front is assumed to be a random vector, with known mean $\mathbf{m}_\phi$ and covariance matrix $\mathbf{\Lambda}_\phi$, the bias and the error variance are given by

$$\mathbf{b} = (\mathbf{MH} - \mathbf{I})\mathbf{m}_\phi, \tag{31}$$

$$\mathbf{\Lambda}_\epsilon = (\mathbf{MH} - \mathbf{I})\mathbf{\Lambda}_\phi(\mathbf{H}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} - \mathbf{I}) + \mathbf{M\Lambda_n M}^{\mathrm{T}}. \tag{32}$$

Assuming that the wave-front phase is zero mean in time, the estimate is unbiased, regardless of the structure of $\mathbf{M}$. This means that, in time, the expected error is zero, though any instance has nonzero error. The mean squared error can be calculated by using Eqs. (27) and (32):

$$\mathrm{mse} = \frac{\mathrm{Trace}(\mathbf{\Lambda}_\epsilon)}{a}. \tag{33}$$

This result depends on both the reconstructor and the method of gradient generation. This performance metric allows comparative evaluation of different reconstruction methods (by varying $\mathbf{M}$) and of different sensor models (by varying $\mathbf{H}$). This equation shows the importance of considering the impact of noise in the system. It is possible to design an $\mathbf{M}$ such that $\mathbf{MH} - \mathbf{I} = \mathbf{0}$. However, this also affects the portion of the error due to the noise. The total error may not be minimal in this case.

3. *Wave Front as a Random Vector*

To use the performance metrics derived in Subsection 8.B.2, the statistics of the wave front must be known. These statistics will be different for closed-loop and open-loop performance. Obtaining them is a nontrivial problem. These statistics could potentially be derived from theoretical knowledge of the phenomenon that produces the phase aberrations. This could be atmospheric turbulence, or the heating of optics due to high-power lasers. If the theoretical approach is not possible, the statistics could be estimated from observations of the process or by simulation. Wallner[7] has presented a method for assessing Kolmogorov turbulence across an aperture with piston removed. Work by the authors in applying this method is still in progress.

C. Linear Model of Mean Squared Error

The above results can be combined to create a linear model of reconstruction performance. Expanding Eq. (33) produces

$$\mathrm{mse} = \frac{\mathrm{Trace}[(\mathbf{MH} - \mathbf{I})\boldsymbol{\Lambda}_{\phi}(\mathbf{H}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} - \mathbf{I})] + \mathrm{Trace}(\mathbf{M}\boldsymbol{\Lambda}_{\mathbf{n}}\mathbf{M}^{\mathrm{T}})}{a}. \tag{34}$$

The left-hand term of the numerator is the contribution to the mean squared error by the wave-front phase. For a given wave-front phase distribution, this remains fixed. This part of the error will be called mse_{ϕ}, the latent error. The right-hand term is the contribution of the noise. This is exactly what the mean squared error is in the nonrandom-parameter case. This depends entirely on the variance of the noise. This part of the error will be called $\mathrm{mse}_{n}(\boldsymbol{\Lambda}_{\mathbf{n}})$. Note that the assumption in Subsection 8.B of a reconstruction matrix independent of noise allows this simplification. The total error is given by

$$\mathrm{mse} = \mathrm{mse}_{\phi} + \mathrm{mse}_{n}(\boldsymbol{\Lambda}_{\mathbf{n}}). \tag{35}$$

Assuming white noise of variance σ_n^2, this can be reduced further. When we recall Eq. (30) defining the noise propagator, the above expression becomes

$$\mathrm{mse} = \mathrm{mse}_{\phi} + \sigma_n^2\, \mathrm{mse}_{np}\,. \tag{36}$$

This equation describes a line. It says that the expected performance of a reconstructor with white noise on the measured gradients is simply a fixed component (the latent error) and a noise component that grows linearly with the noise variance. This allows easy graphical comparison of the performance of various reconstructors under the same wave-front and noise conditions.

9. PERFORMANCE RESULTS WITH HUDGIN AND FRIED GEOMETRIES

If we use the metrics defined in Section 8, the performance of the Hudgin-FT and Fried-FT methods can be analyzed. The most significant result of this analysis is that for a given aperture size, a trade-off exists between speed and error in FT methods. This is due to the specific power-of-2 grid sizes required by the FFT. This section also presents the noise propagator results for large systems and confirms the linear model of mean squared error.

Note that this performance analysis is done given the discrete model and the specific geometry of each method. In particular, the gradients generated for the simulations are created directly from the Hudgin- and Fried-geometry equations [see Eqs. (1), (2), (11), and (12)].

A. Noise Propagation

The theoretical noise propagator mse_{np} is evaluated for a variety of different aperture sizes for both the Hudgin-FT and Fried-FT methods. The calculation comes directly from Eq. (30). This is tractable for grid widths up to 32 actuators across, allowing a circular aperture of 716 total actuators. Of most concern, however, are systems with thousands to tens of thousands of actuators. To predict performance in this regime, one has to rely on simulation. Simulations to estimate the sample mean of the noise propagator prove to be reasonably accurate compared with what theory predicts (see Fig. 9). Simulation is therefore used to predict mse_{np} for larger apertures.

In the DFT case, the FT methods have reasonable noise propagation, especially when compared with the results for other methods. These comparisons are shown in Fig. 10. It has been shown that minimum-least-squares VMM reconstructors have a $c + d \ln a$ dependence for the noise propagators,[9] where c and d are constants and a is the number of actuators. Based on work by Herrmann,[10] a Hudgin-geometry VMM reconstructor on an $N \times N$ square aperture ($a = N^2$) has a noise propagator

$$\mathrm{mse}_{np} = 0.46 + 0.087 \ln a. \tag{37}$$

The square-aperture FT case has a noise propagator[1]

$$\mathrm{mse}_{np} = 0.09753 + \frac{1}{\pi} \ln N. \tag{38}$$

when a curve is fitted to the data, the Hudgin-FT method over the range 500–50,000 actuators inside the aperture has a noise propagator

$$\mathrm{mse}_{np} = 0.17 + 0.13 \ln a \tag{39}$$

with correlation coefficient of the fit of 0.97. As shown in Fig. 10(a), the Hudgin-FT method has a lower noise propagator than that in the square-aperture FT case but higher than that of a VMM reconstructor.

For the Fried geometry, a VMM reconstructor on an $N \times N$ ($a = N^2$) square aperture has a noise propagator[5]

$$\mathrm{mse}_{np} = 0.6558 + 0.1603 \ln a. \tag{40}$$

The square-aperture FT case has a noise propagator[2]

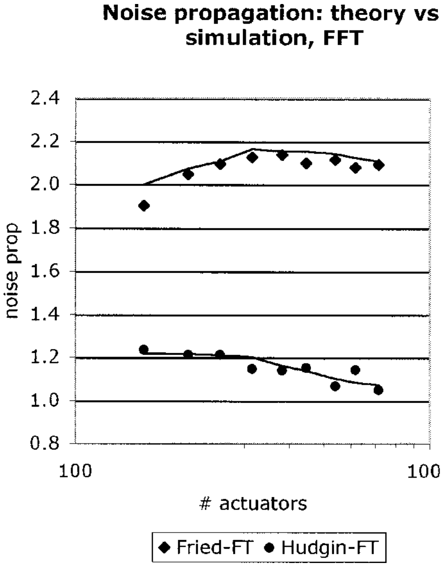
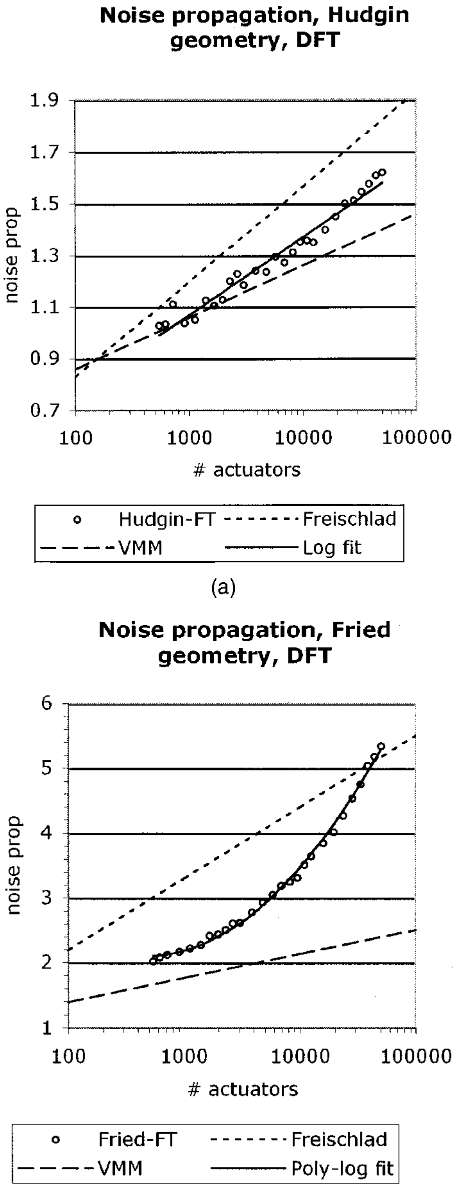


Fig. 9. Theoretical and simulation results of the noise propagation for the Hudgin-FT and Fried-FT methods. Aperture sizes vary on a 32 × 32 grid for the FFT. The solid curves are the noise propagators as determined theoretically. The data points are simulated noise propagator predictions. The simulation converges to the correct solution adequately enough to use it for large numbers of actuators, the theoretical calculations of which are computationally intractable.



(a)

(b)

Fig. 10. Simulation results for noise propagation in comparison with the VMM and Freischlad's square-aperture FT methods. In this DFT case, the reconstructions were done on the smallest size grid possible to correctly hold the aperture. (a) Hudgin geometry: The Hudgin-FT case lies between the square-aperture cases. (b) Fried geometry: The Fried-FT case is closer to the VMM case for smaller apertures but reaches square-aperture levels by 35,000 actuators.

$$\mathrm{mse}_{np} = c + \frac{3}{\pi}\ln(N - 1), \tag{41}$$

where c was unspecified and is assumed to be zero here. The Fried-FT method deviates from the linear model. Over the range 500–50,000 actuators inside the aperture, it is best fitted by a second-order polynomial in $\ln a$:

$$\mathrm{mse}_{np} = 0.1456 \ln^2 a - 1.7922 \ln a + 7.6175 \tag{42}$$

with correlation coefficient of the fit of 0.997. As shown in Fig. 10(b), the Fried-FT method has a lower noise propagator than that for the square-aperture FT case for systems with 35,000 or fewer actuators. The noise propagation is worse than that of the VMM method for all cases but has a reasonably close value for fewer than 3000 actuators.

The performance of the Fried-FT method is worse than the Hudgin-FT method's for two main reasons. First, the transformation to the two grids increases the noise variance (as was recognized by Freischlad[2]). Second, as the aperture size increases, it becomes harder to remove all the wafflelike components.

The preceding results are for the DFT case. But as Fig. 11 shows, performance varies significantly between the DFT and FFT cases. There can be significant performance loss when the FFT is used. This is because the FFT grids are of fixed power-of-2 sizes. If an aperture is 34 points across, it will not fit into a 32 × 32 grid but must be reconstructed on a 64 × 64 grid. If the DFT is used, it could be reconstructed on a 36 × 36 grid instead. As the surrounding grid gets bigger, the noise propagation increases. This observed behavior is confirmed by analysis of the small-aperture case. For an illustration of this for the Hudgin-FT and Fried-FT methods on a 112-actuator aperture, see Fig. 12. The increase in noise propagation can be explained as part of the same behavior that causes noise propagation to increase with aperture size. Though the number of noisy gradient measurements is fixed, the gradients are duplicated more and more as the grid size is increased. A DFT is a linear combination of these points, which means that as more are added, the weighting to the noise is increased.

This speed versus error performance trade-off has important design implications. Ideally, the aperture size will be just right so as to fit into a power-of-2 grid. If this is not the case, and the system is of a size that cannot be changed, a choice must be made. For example, a system with 38,000 actuators could be calculated on a 220

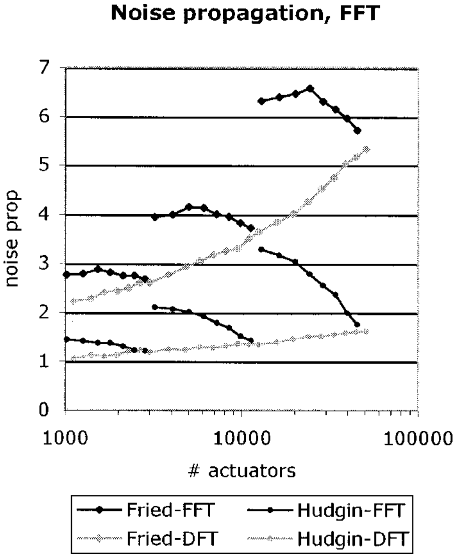


Fig. 11. Comparison of simulation results for noise propagation in the DFT and FFT cases. The DFT case is reconstructed on the smallest grid possible, while the FFT case uses power-of-2-sized grids. For both the Hudgin-FT and Fried-FT cases, there is a clear performance loss when the aperture diameter is small compared with the power-of-2-sized grid. Only the largest size apertures in a given power-of-2 grid approach ideal DFT results.

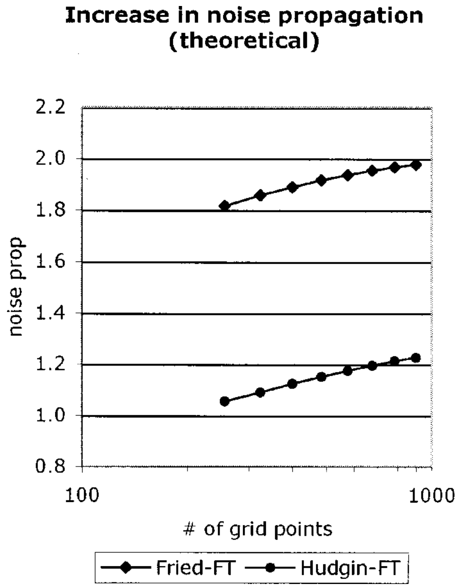


Fig. 12. Theoretical results for the increase in noise propagation of a fixed 112-actuator system as the surrounding grid is increased in size. For both methods, increasing the grid size increases the total noise propagation in a regular manner.

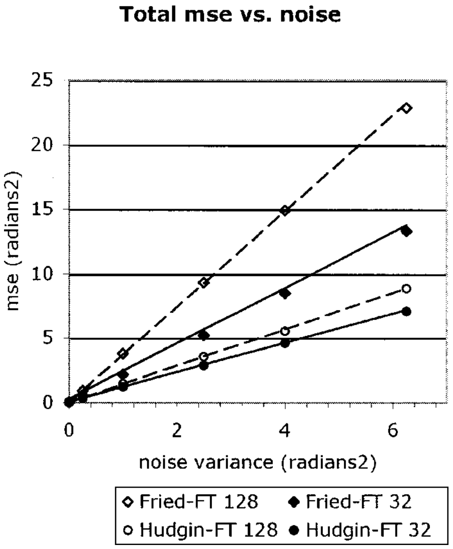


Fig. 13. Total mean squared error versus noise variance for two systems, a 448-actuator aperture on a 32 × 32 grid and an 11,304-actuator aperture on a 128 × 128 grid. Results for both the Hudgin-FT and Fried-FT methods are shown. The lines are the predicted performance, based on either theoretical or experimental noise propagation and an experimentally determined latent error. The data points are the results of simulation at various levels of noise. The effect of the noise propagator is clearly demonstrated in the differing slopes.

× 220 grid by using the DFT. This would have a noise propagation of approximately 1.5 in the Hudgin-FT case and 5.0 in the Fried-FT case. But this DFT is slower than a 256 × 256 FFT, which could also be used. For example, in IDL this DFT takes twice as long to compute as the FFT. Though the FFT is faster, the resulting noise propagation would be increased to approximately 2.0 in the Hudgin-FT case and 5.9 in the Fried-FT case (see Fig. 11).

For the DFT implementation, or for suitably sized apertures with use of the FFT, the FT methods have favorable noise propagation. The noise propagation for the Hudgin-FT method fits the linear dependence on the logarithm of system size. The Fried-FT method has noise propagation that grows more quickly than the logarithm, but it starts at a reasonably low level.

B. Total Mean Squared Error

The second part of the total error is the latent error. As discussed above, using the extension and boundary methods with the FT filtering produces perfect reconstruction of all sensed wave-front phase modes in the absence of noise. For the Hudgin-FT method, this means that the latent error is essentially zero. For the Fried-FT method, the missed waffle mode is extremely small, compared with the large errors that occur without proper processing, as shown in Fig. 2.

Simulations confirm the linear model of mean squared error presented in Eq. (36). Simulation and estimation were used to determine both Λ_ϕ (which leads to mse_ϕ) and the sample mean of mse_ϕ. Figure 13 shows the results of Monte Carlo simulations with random realizations of the same wave-front phase profile and varying amounts of white noise on the measurements. The simulator used for the atmospheric phase screen generation correctly simulates the low-frequency components.[11] For both small and large apertures, this linear model of performance was confirmed, with simulation results agreeing closely with predicted values.

10. CONCLUSIONS

The problem of reconstructing wave-front phase on circular apertures with a square grid has been identified and solved. Though large errors result from zero padding, the methods presented in this paper permit accurate reconstruction of all sensed modes when no noise is present. These FT methods have been shown to work on the Fried geometry in addition to the Hudgin geometry. The extra processing steps were shown to not increase the order of growth of the FT method execution time. Detailed performance analysis produced a linear model for the mean squared error of reconstruction. The noise propagation of FT methods is reasonable for apertures that nearly fill the square grid, though there exists a trade-off between speed of performance and reconstruction error when the FFT is used. The above results have been presented for large systems, up to 50,000 actuators. Based on the results in this paper, a reconstruction method for a 10,000-actuator system could be realistically implemented by using current technology and with adequate performance.

These results are for discrete models that are based on the Hudgin and Fried geometries. How these methods perform in a more continuous domain with SH wave-front sensors and a DM is being studied by the authors. The interaction of reconstruction noise with the DM, further filtering approaches, and the latent error of each method when applied to data from realistic SH sensor models, will be addressed in an upcoming paper.

ACKNOWLEDGMENTS

Special thanks go to Carmen Carrano for all her assistance. The authors thank the reviewers, whose comments led to improvement in both the clarity and the content of this paper. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory, under contract W-7405-Eng-48. The document number is UCRL JC-147058. This work has been supported by the National Science Foundation Science and Technology Center for Adaptive Optics, managed by the University of California at Santa Cruz under cooperative agreement AST-9876783.

The corresponding author, Lisa A. Poyneer, can be reached at poyneer1@llnl.gov.

REFERENCES

1. K. Freischlad and C. L. Koliopoulos, “Wavefront reconstruction from noisy slope or difference data using the discrete Fourier transform,” in *Adaptive Optics*, J. E. Ludman, ed., Proc. SPIE **551**, 74–80 (1985).
2. K. Freischlad and C. L. Koliopoulos, “Modal estimation of a wave front from difference measurements using the discrete Fourier transform,” J. Opt. Soc. Am. A **3**, 1852–1861 (1986).
3. K. Freischlad, “Wavefront integration from difference data,” in *Interferometry: Techniques and Analysis*, G. M. Brown, O. Y. Kwon, M. Kujawinska, and G. T. Reid, eds., Proc. SPIE **1755**, 212–218 (1992).
4. R. H. Hudgin, “Wave-front reconstruction for compensated imaging,” J. Opt. Soc. Am. **67**, 375–378 (1977).
5. D. L. Fried, “Least-square fitting a wave-front distortion estimate to an array of phase-difference measurements,” J. Opt. Soc. Am. **67**, 370–375 (1977).
6. R. J. Noll, “Zernike polynomials and atmospheric turbulence,” J. Opt. Soc. Am. **66**, 207–211 (1976).
7. E. P. Wallner, “Optimal wave-front correction using slope measurements,” J. Opt. Soc. Am. **73**, 1771–1776 (1983).
8. J. Hardy, *Adaptive Optics for Astronomical Telescopes* (Oxford U. Press, New York, 1998).
9. R. J. Noll, “Phase estimates from slope-type wave-front sensors,” J. Opt. Soc. Am. **68**, 139–140 (1978).
10. J. Herrmann, “Least-squares wave front errors of minimum norm,” J. Opt. Soc. Am. **70**, 28–35 (1980).
11. E. M. Johansson and D. T. Gavel, “Simulation of stellar speckle imaging,” in *Amplitude and Intensity Spatial Interferometry II*, J. B. Breckinridge, ed., Proc. SPIE **2200**, 372–383 (1994).