

Scene-based Shack–Hartmann wave-front sensing: analysis and simulation

Lisa A. Poyneer

In many situations it is not possible for an adaptive optics system to use a point source to measure the phase derivative, such as imaging along slant paths through the atmosphere and observation of the earth from space with a lightweight optic. Instead, small subimages of the observed scene can be used in a scene-based wave-front sensing technique. This study presents three important advances in the understanding of this technique. Rigorous analysis shows how slope estimation performance depends precisely on scene content and illumination. Scaling laws for changes in illumination are derived. The technique, when applied to point sources, is more robust to detect size changes and background levels than current methods. © 2003 Optical Society of America

OCIS codes: 010.1080, 010.7350.

1. Problem Description and Background

Adaptive optics (AO) systems work by estimating and correcting for the phase aberration along the optical path. The performance improvements obtained with AO can be quite dramatic.¹ Normally a derivative of the phase is measured, and the phase is reconstructed from those measurements. A point source of light is in most cases used as the reference.

There are many situations, however, when a point source is not available but AO correction is desirable. Of prime interest are two situations. The first is that imaging conducted along horizontal or slant paths through the atmosphere could be improved with AO correction for the turbulent atmosphere. Note that because in this paper we are concerned with how to measure the phase, not how to correct it, we will not discuss whether scintillation or refraction will inhibit the performance of multiwavelength AO along slant paths. The second situation is the observation of the Earth from space with lightweight optics. These optics are less expensive but can suffer from time-varying aberrations due to thermal effects and vibration.² AO could correct these phase aberrations. In both cases, observations of an extended scene do not provide a point source. Instead

of trying to create an artificial point source, the scene itself can be used to make phase estimates.

There are two major ways to use images to estimate phase. The first is called phase retrieval.³ This method uses multiple images, some with a known additional phase aberration, to iteratively determine the phase. This method is commonly used in telescope alignment. In addition, phase-diversity methods can be used to correct for atmospheric aberrations with image post processing.⁴ These methods are normally quite computationally intensive. Instead, we chose to investigate in detail a second approach that is very similar to current point-source-based AO systems. In fact, the only difference is how the slope estimates are computed.

The second option is the use of small images as produced by a Shack–Hartmann sensor array. Instead of forming a spot (an image of the point-source reference), the lenslets form small images of the observed scene. The subimage is at lower resolution because of the smaller size of the subaperture. Each subimage will have a field of view that is restricted by two main considerations. First, the field will need to be large enough to contain scenes capable of being used for slope estimation given the subaperture diffraction limit. Second, the field must be small enough that the subimages will not overlap on the wave-front sensing (WFS) camera. This maximum size is set by the physical separation of the lenslets in the WFS array and the magnification. The exact specifications are a systems-design issue, and most likely will vary depending on the remote imaging scheme. Each of these subimages will be shifted by

The author (e-mail poyneer1@llnl.gov) is with Lawrence Livermore National Lab, Livermore, California.

Received 14 March 2003; revised manuscript received 1 July 2003.

0003-6935/03/295807-09\$15.00/0

© 2003 Optical Society of America

a small amount due to the local phase distortion across each lenslet, just as a point source is. The problem is then how to best estimate the shifts between the images.

This Shack–Hartmann approach, which we refer to as scene-based wave-front sensing (SBWFS), is used successfully at the National Solar Observatory⁵ and other solar observatories. For the case of solar granulation, the scene has very specific characteristics. For the special case of low-contrast solar granulation, the error standard deviation σ_x of the slope was estimated to be proportional to the background noise standard deviation divided by the image contrast.⁵ However, no rigorous performance analysis of wave front slope estimation based on arbitrary scene content and illumination has been published. A sensing system could look at an arbitrary scene, whether it contains buildings in a city or a road in the desert. The illumination, which is due to time of day, angle of observation, and atmospheric characteristics, is highly variable. Intuitively, the performance of a given scene depends on the content of the scene as well as the illumination characteristics. The more features and high-spatial-frequency content and the more light, the better the performance.

This intuitive understanding needs to be rigorously quantified for several reasons. First, a system can be specifically designed to meet performance criteria given expected illumination and scene content. Second, as the system is built and tested, analysis enables confirmation of performance. Finally, real-time knowledge of scene quality could be used to optimize AO system performance by adjusting system parameters, such as frame rate or changing the scene that is used. For all of these reasons, detailed analysis of the scene-based slope estimation problem has been carried out and is presented here. Simulation confirms the analytic predictions, showing that SBWFS is a viable technique that will enable AO to be used in a broad range of new situations.

2. Slope Estimation from a Scene

As described above, multiple subimages of the observed scene are formed for use on the WFS camera, just as multiple images of the point source are. The basic concept of SBWFS is to then compare the image in a given subaperture to a reference image and estimate the shift between the two images. This shift could be determined by cross correlating a reference image with each subaperture image.⁶ However, many different methods exist for aligning images.⁷ Several were analyzed as candidates for obtaining slope information. This section describes the model of the subimages and discusses possible algorithms for slope estimation. Based on both performance with noise and computational simplicity, periodic correlation was chosen as the preferred method.

A. Definition of Terms

Consider two discrete images (made of pixels on the WFS camera) that are formed by distinct subapertures. One image is the reference, $r[m, n]$, which

will be used in comparison with all the other subapertures. The second image is $s[m, n]$. The two signals are shifted slightly from each other by x_0 and y_0 samples, which are not necessarily integer amounts. For now, consider the signals as being infinite in extent, and as sampled versions of the same underlying continuous signal. The sample interval is D . Given the high-resolution base signal $i(x, y)$, the reference signal is

$$r[m, n] = i(mD, nD). \quad (1)$$

The subaperture image is assumed to be simply a shifted version

$$s[m, n] = i(mD - x_0D, nD - y_0D). \quad (2)$$

The shifts we desire to find are x_0, y_0 . Because x_0 and y_0 are in general not integers, the expression

$$s[m, n] = r[m - x_0, n - y_0] \quad (3)$$

has no meaning except in the model described above. We also have the frequency domain counterparts of the above equations. Where $\tilde{I}(f_x, f_y)$ is the continuous-time Fourier transform of $i(x, y)$, we can write expressions for the discrete-time Fourier transform of the other signals (which use frequency variables ϕ_x, ϕ_y)

$$\tilde{R}(\phi_x, \phi_y) = \frac{1}{T^2} \tilde{I}\left(\frac{\phi_x}{D}, \frac{\phi_y}{D}\right), \quad (4)$$

$$\tilde{S}(\phi_x, \phi_y) = \tilde{R}(\phi_x, \phi_y) \exp[-j2\pi(x_0\phi_x + y_0\phi_y)]. \quad (5)$$

In reality, the two images r and s are of small finite extent $N \times N$ pixels and \tilde{I} is not necessarily band limited. Converting the above equation for use with discrete Fourier transform produces

$$\tilde{S}[k, l] = \tilde{R}[k, l] \exp\left[\frac{-j2\pi(x_0k + y_0l)}{N}\right]. \quad (6)$$

B. Derivation of Algorithm

Based on this model, three possible methods to determine x_0 and y_0 were analyzed. They are maximum-likelihood estimation, deconvolution, and correlation. These methods are applied to the finite, discrete signals obtained from the WFS camera. The reference subaperture image $r[m, n]$ and the subaperture image for slope estimation $s[m, n]$ are both $N \times N$ pixels. Note that N may be preferentially a power-of-2 for some of the algorithms.

Because the two images are in fact noisy realizations, we can regard the search for x_0 and y_0 as a nonrandom parameter estimation problem. Given a noise model for both images and the actual scene content, the probability distribution for each could in principle be derived. The maximum-likelihood estimate is then determined. This technique has been applied to point-source WFS.⁸ However, this estimator is directly dependent on scene content, meaning it would have to be recalculated for every scene. Furthermore, since this is not a linear function of the

nonrandom parameters x_0 and y_0 , producing the estimator is nontrivial computationally. This method of estimation was therefore discarded as being too computationally intensive.

The second method is deconvolution. This method exploits the frequency-domain relationship between the two images [Eq. (6)]. With use of the discrete Fourier transform, $\tilde{S}[k, l]$ is known, and if $\tilde{R}[k, l]$ is nonzero at all k, l , division produces

$$\tilde{F}[k, l] = \exp\left[\frac{-j2\pi(x_0 k + y_0 l)}{N}\right]. \quad (7)$$

If x_0, y_0 are integers, the signal $f[m, n]$ (obtained by inverse transforming the above) is simply a unit impulse at x_0, y_0 . If x_0, y_0 are not integers, $f[m, n]$ is a shifted sampling of a unit impulse. Based on this signal, the shifts can be estimated.⁹ This deconvolution method is sensitive to noise, particularly when the spatial frequency content of the images is mainly low frequency. Tests of this method in our application showed it to be highly susceptible to noise.

The final method under consideration is correlation. This method is an implementation of a minimal mean-squared error (MSE) metric. The shift with the least-squared difference between the two images r and s is the best answer. Formulated explicitly, the MSE $e[m, n]$ of the two images at displacement m, n is

$$e[m, n] = \left[\sum \sum (r[i - m, j - n] - s[i, j])^2 \right] \times [(N - m)(N - n)]^{-1}, \quad (8)$$

where the summation is, for $m \geq 0$, from $i = m$ to $i = N - 1$, and for $m \leq 0$, from $i = 0$ to $i = N - 1 - m$. This holds likewise for n with index j . Note that $e[m, n]$ is defined only over the range $-(N - 1) \leq m, n \leq N - 1$. Expanding the terms produces

$$e[m, n] = \left(\sum \sum r[i - m, j - n]^2 + s[i, j]^2 - 2r[i - m, j - n]s[i, j] \right) \times [(N - m)(N - n)]^{-1}. \quad (9)$$

The last term in the summation in Eq. (9) looks like a correlation. But simply calculating the correlation is not a shortcut to calculating this metric. Because of the limits of summation, all the terms are dependent on m and n . Minimizing the MSE is not equivalent to maximizing the correlation between r and s , even when it is calculated with energy normalization.

Instead, the finite signals can be treated as a single period of an infinite periodic signal (just as for the

discrete Fourier transform.) Now the limits of summation are constant for all values of m and n

$$e[m, n] = \left(\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} r[i - m, j - n]^2 + s[i, j]^2 - 2r[i - m, j - n]s[i, j] \right) N^{-2}. \quad (10)$$

Owing to the periodicity of the signals, as one end of the signal moves away, it wraps around from the other side. In this case the formula simplifies dramatically as the two energy terms remain constant. The MSE equation now becomes

$$e[m, n] \propto - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} r[i - m, j - n]s[i, j], \quad (11)$$

which is exactly the correlation of the two signals calculated with periodic convolution. Minimizing the MSE [as given by Eq. (11)] is the same as maximizing the correlation. This correlation can be calculated quickly in the frequency domain by use of conjugation and the correlation theorem.

Computing the exact MSE for all possible overlaps is $O(N^4)$ in FLOPs, where N is the number of pixels along each dimension of the square image. The periodic correlation method can employ fast Fourier transforms (FFT), so it is $O(N^2 \log_2 N)$. Except for the case where the shifts are known to be small and only a few points of the MSE must be calculated, the periodic correlation method is significantly faster. For a 16 by 16 pixel subaperture image, the periodic correlation requires a factor of 16 times less computation than calculating the MSE for all possible overlaps. However, the periodic correlation technique is susceptible to errors due to the wraparound of pixels. For low-quality images there is a trade-off: MSE provides better estimates but with more computation.

Despite this wraparound effect, the periodic correlation technique can work just as well as the exact MSE method. As shown in Fig. 1, Monte Carlo simulation demonstrates the comparable estimate error and standard deviation for the same image with both the MSE technique and periodic correlation. For high-quality images, this overlap effect is negligible, and both methods perform equally well. For poorer quality images, the non-common image content at the edges has a significant effect.

It is important to note that the answer we want out of this technique is not the whole-pixel shift with the best value of the metric, but rather an estimation of the best sub-pixel shift. In the general case the shift will not be a whole-pixel amount. The correlation is actually just the scaled sampling of the autocorrelation of the base signal sampled at the shift x_0 and y_0

$$C[m, n] = \frac{1}{D^2} C(x, y)|_{x=mD-x_0, y=nD-y_0}. \quad (12)$$

To obtain the true maximum of the function, we simply need to interpolate to estimate the continuous signal $C(x, y)$ and take its maximum value. Para-

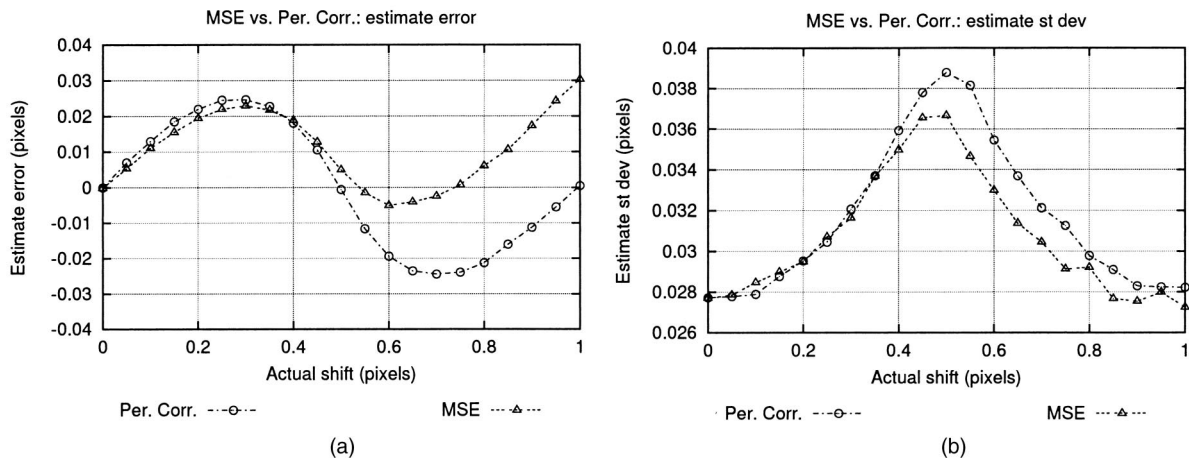


Fig. 1. Periodic correlation has competitive performance with the MSE method for a good image. In this case, Monte Carlo simulation results for slope estimation of a shift in the x direction are shown: (a) Error of estimation in both cases is small, (b) the standard deviation of the estimates are comparable.

bolic interpolation is used, though another curve, such as a Gaussian, could be fit to the peak. Assume that the maximum of the discrete signal $C[m, n]$ is located at integers $[\Delta_x, \Delta_y]$. Then the following equation fits a parabola to the closest points around the peak and gives the shift estimate:

$$\hat{x}_0 = \Delta_x + \frac{0.5(C[\Delta_x - 1, \Delta_y] - C[\Delta_x + 1, \Delta_y])}{C[\Delta_x - 1, \Delta_y] + C[\Delta_x + 1, \Delta_y] - 2C[\Delta_x, \Delta_y]} \quad (13)$$

The estimate of y_0 is obtained in an analogous fashion.

In summary, the best method for estimating the shift between two subimages uses periodic correlation. This is done by calculating the correlation with FFTs and doing parabolic interpolation around the peak location to determine the best sub-pixel shift estimate. We have elected to use a reference subaperture to provide the new reference subimage at every time step. This ensures that any dynamic changes in the scene will be compensated for. There are two complications to this. First, at any time step there is no knowledge of the true tip and tilt, because all slope estimates are measured relative to the offset of a single subaperture. Therefore the drift of this subaperture through time should be followed to obtain tip and tilt information. Second, using a new, noisy reference at each time step increases the amount of noise in the system. Compared with a fixed (nonrandom) reference image, the noise variance when using a new reference every step is two to three times greater. However, if there is any dynamic change in scene content, this fixed reference may provide poor quality estimates because of differences in the images. In the general case we expect changes in the scene. The performance of this method of shift estimation is rigorously analyzed below. Simulation is then used to confirm the analysis.

3. Performance Analysis and Simulation Results

A. Motivation

In AO systems, noise on the wave-front sensor manifests itself as noise on the slope estimation. This error then propagates through to the compensated phase. Scene-based slope estimation can be analyzed by use of this model, where image quality as well as photon noise leads to noise on the slope estimates. Image content will vary to a large degree, with some scenes being full of features, such as buildings or vehicles, and others being relatively uniform such as a barren patch of land. Differences in image content will produce images that are better or worse for estimation. The total amount of light and background scatter will also vary with angle of viewing, time of day, air quality, etc. These characteristics will affect the noise level and change the performance of a specific scene.

The bias and variance of the slope estimate are of interest. The ideal scene would produce an unbiased estimate with very low (or no) variance. This is of course not true in the general case, and both bias and variance depend on the scene, the noise level, and the amount of shift between the subimage and its reference. This Section analyzes the characteristics of the slope estimation based on image content, illumination and actual image shift. This analysis provides both a measurement of performance for an arbitrary case, as well as scaling laws that describe how performance changes with specific parameters.

Monte Carlo simulations were used as a technique for verifying the analytic results. In these simulations a large diffraction-limited image (obtained from commercial satellite imagery) was used. The sub-pixel shifted image was obtained by convolution. The larger signal was windowed down to simulate the field stop and the real drift of the image within the field. Given a specific illumination profile, image content, and noise model, random realizations of the images were generated. Tens of thousands of trials

on the same set of conditions were conducted, and the results were statistically analyzed.

B. General Case

As presented above, slope estimation involves computing the cross correlation of two subaperture images ($r[m, n]$ and $s[m, n]$), finding the maximum, and then using that maximum value and the two neighboring values each to determine the estimate of the shifts x_0 and y_0 via parabolic interpolation. The random vector $C[m, n]$ represents the cross-correlation function of these two images. Specifically, this periodic correlation

$$C[m, n] = \sum_i \sum_j r[i - m, j - n]s[i, j] \quad (14)$$

can be computed with FFTs. The maximum of the correlation will be (for whole-pixel shifts) at exactly $C[x_0, y_0]$. For sub-pixel shifts, the maximum of this correlation function is at $[\Delta_x, \Delta_y]$. We will assume that the maximum of $C[m, n]$ is within half a pixel of the actual shift. For some degenerate images, it will not be. For a single estimate of the x slope, parabolic interpolation is used. This requires the discrete maximum $C[\Delta_x, \Delta_y]$ (as opposed to the true maximum of the continuous correlation) and the two points bracketing it, $C[\Delta_x - 1, \Delta_y]$ and $C[\Delta_x + 1, \Delta_y]$. For notational simplicity, these three points will be referred to as C_0 for the maximum and C_{-1} and C_1 for the neighbors. The estimate of the shift is then rewritten into Eq. 13 as

$$\hat{x}_0 = \Delta_x + \frac{0.5(C_{-1} - C_1)}{C_{-1} + C_1 - 2C_0}. \quad (15)$$

This expression is very difficult to analyze. Because it involves division of random variables, full knowledge of the probability distributions of C_{-1} , C_0 , and C_1 is required to characterize the resulting random variable. Preferably, knowledge of just the means (m_{-1} , m_0 , m_1), variances (σ_{-1}^2 , σ_0^2 , σ_1^2) and covariances ($\sigma_{-1,0}^2$, $\sigma_{-1,1}^2$, $\sigma_{0,1}^2$) of the random variables C_{-1} , C_0 , C_1 will be adequate to evaluate the performance of the estimator. This is true if the estimator \hat{x}_0 is approximated as a linear combination of C_{-1} , C_0 , C_1 instead of being a quotient. The tangent plane to the function is constructed at the point $[C_{-1} = m_{-1}, C_0 = m_0, C_1 = m_1]$. The function $f(C_{-1}, C_0, C_1)$ is approximated as $\tilde{f}(C_{-1}, C_0, C_1)$. This is done by taking partial derivatives and evaluating them at the correct location. Where the partial derivative of $f(C_{-1}, C_0, C_1)$ with respect to variable C_i is $f_i(C_{-1}, C_0, C_1)$, the approximation is

$$\begin{aligned} \tilde{f}(C_{-1}, C_0, C_1) = & f(m_{-1}, m_0, m_1) \\ & + (C_{-1} - m_{-1})f_{-1}(m_{-1}, m_0, m_1) \\ & + (C_0 - m_0)f_0(m_{-1}, m_0, m_1) \\ & + (C_1 - m_1)f_1(m_{-1}, m_0, m_1). \end{aligned} \quad (16)$$

Evaluating the above expression produces the linear approximation for the slope estimate:

$$\begin{aligned} \hat{x}_0 \approx & \Delta_x + [C_{-1}(m_1 - m_0) + C_0(m_{-1} - m_1) + C_1(m_0 \\ & - m_{-1}) + 0.5(m_{-1} - m_1)(m_{-1} + m_1 - 2m_0)] \\ & \times (m_{-1} + m_1 - 2m_0)^{-2}. \end{aligned} \quad (17)$$

Because this is a linear combination of random variables, the mean and variance can be determined with knowledge of only the means and variances of its components. Most terms cancel to make the expectation simply

$$E[\hat{x}_0] = \Delta_x + \frac{0.5(m_1 - m_{-1})}{m_{-1} + m_1 - 2m_0}. \quad (18)$$

The variance can also be explicitly calculated and after simplification is

$$\begin{aligned} \sigma_x^2 = & [\sigma_{-1}^2(m_1 - m_0)^2 + \sigma_0^2(m_{-1} - m_1)^2 \\ & + \sigma_1^2(m_0 - m_{-1})^2 \\ & + 2(m_1 - m_0)(m_{-1} - m_1)\sigma_{-1,0}^2 \\ & + 2(m_1 - m_0)(m_0 - m_{-1})\sigma_{-1,1}^2 \\ & + 2(m_{-1} - m_1)(m_0 - m_{-1})\sigma_{0,1}^2] \\ & \times (m_{-1} + m_1 - 2m_0)^{-4}. \end{aligned} \quad (19)$$

The means, variances, and covariances (e.g., m_0 , σ_1^2) that appear in the above equations can be easily calculated from the statistical models of the images. The explicit formulas for these terms are given in the Appendix. This long expression is not particularly informative on its own. As the following discussion shows, it can be drastically simplified for specific useful cases. The results provide powerful descriptions of estimation performance.

C. Zero-Shift Case

A special case worth considering is when the actual shift between the two images is zero. In a closed-loop system, the image shift will be driven toward null. This simplification also enables easier analysis of slope estimation behavior as illumination conditions change. In this zero-shift case, the two subimages have identical distributions. Therefore $m_{-1} = m_1$ and $\sigma_{-1}^2 = \sigma_1^2$. This reduces the approximation of the estimate [Eq. (17)] to be

$$\hat{x}_0 \approx \Delta_x + \frac{C_{-1} - C_1}{4(m_1 - m_0)}. \quad (20)$$

In this special case, the correlation is actually an autocorrelation, so the peak will be at 0 and the means and variances of C_{-1} and C_1 will be equal. Therefore the estimate is unbiased

$$E[\hat{x}_0] = 0. \quad (21)$$

Most of the terms in Eq. (19) cancel, reducing the expression for the variance to

$$\sigma_x^2 = \frac{\sigma_1^2 - \sigma_{-1,1}^2}{8(m_0 - m_1)^2}. \quad (22)$$

The most important term in this equation is the denominator term $(m_0 - m_1)$. As described above, m_0 is the expected value of the maximum of the correlation function (C_0) and m_1 is the expected value one pixel to the side (C_1). The $(m_0 - m_1)$ term is then a measure of the sharpness of the correlation peak. The sharper this peak, the lower the error variance. The correlation function is paired with its Fourier transform partner: the power spectral density. The more impulse-like the correlation function (hence the sharper the peak) the broader the frequency content of the image. This is consistent with the notion that images with more high-frequency content perform better.

In the zero-shift case, all that is needed to calculate the error variance of the estimate is knowledge of the subimage statistics. By use of the parameters that are defined in the Appendix [see Eq. (A5–A7)], the performance of any image can be quickly calculated. When done for a wide range of images, the predicted slope estimate variance σ_x^2 reveals substantial variation in image quality. Some scenes have much lower variance for estimation along one axis than the other by a significant amount. For example, a scene with a road that runs horizontally across the subimage will be a much better estimator in the y axis than in the x axis. This is because the image is self-similar for shifts along the x axis, as one part of the road looks much like another.

The above analysis showed the estimator to be unbiased and the error variance predictable in the zero-shift case. Monte Carlo simulations confirm these predictions. For a wide range of images, the slope estimate mean is extremely small for a wide range of images, confirming the estimator to be unbiased. Fig. 2 shows the predicted error standard deviation σ_x [by Eq. (22)] compared with the results from simulation. For low values of σ_x , the prediction is highly accurate. It starts to under estimate for larger values because of the linear approximation that was used to produce a tractable expression for estimator variance.

D. Illumination Changes

A metric that differentiates between arbitrary images was derived above. It is also of interest how the quality of a specific image varies with changing types and levels of illumination. Exposure time and total amount of light are the first concern. System frame rate is important, as there is a significant design trade-off between the benefits of high speeds (e.g., faster correction results in reduced temporal errors) and the detriments (e.g., reception of fewer photons leads to more noise.)

To capture the effects of changing light levels, the pixel values of each subimage are modelled as inde-

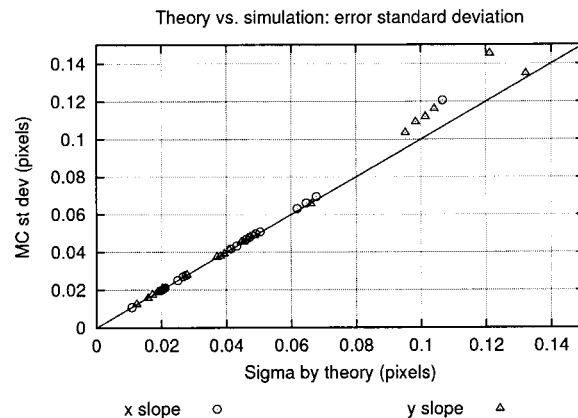


Fig. 2. For the zero-shift case, the bias and standard deviation (st dev) of the estimate can be explicitly calculated from the image itself *a priori*. The predicted standard deviation σ vs. simulation results for a shift of zero. For high-quality images, the prediction is extremely accurate. It starts to under predict for poorer images.

pendent Poisson random variables with parameters $f\tilde{\lambda}_r[m, n]$, $f\tilde{\lambda}_s[m, n]$. The parameters $\tilde{\lambda}$ are normalized from 0 to 1 and multiplied by a scale factor f , which represents the number of photons received by the brightest possible pixel. This is a mathematical convenience, but also could reflect real estimations of the physical object's albedo. The means and variances of C_{-1} , C_0 , C_1 are calculated as given in the Appendix. Insertion of these values shows that the slope estimate expectation is independent of light level and depends only on the image content. This is true regardless of the amount of shift. The new relation for the error variance in the zero-shift case is

$$\sigma_x^2(f) = \frac{\tilde{\sigma}_1^2 - (f-1)f^{-1}\tilde{m}_1 - \tilde{\sigma}_{-1,1}^2}{8f(\tilde{m}_0 - \tilde{m}_1)^2}. \quad (23)$$

This is very close to the basic expression for zero-shift error variance [Eqn. (22)]. For a large number of photons f (i.e., bright light) the $(f-1)f^{-1}$ term in the numerator is approximated as 1, producing a scaling law for the standard deviation of the estimate:

$$\sigma_x(f) = \frac{1}{\sqrt{f}} \frac{(\tilde{\sigma}_1^2 - \tilde{m}_1 - \tilde{\sigma}_{-1,1}^2)^{1/2}}{2\sqrt{2}(\tilde{m}_0 - \tilde{m}_1)}. \quad (24)$$

In the bright-light case, the signal-to-noise ratio (SNR) is simply proportional to \sqrt{f} , where f is the maximum amount of light per pixel. The standard deviation of the estimate follows the same inverse power law to the SNR as quad-cell centroiding with a point source does.¹⁰ Though the constant of the relationship may be different (and is image dependent), this method of wave-front sensing is statistically equivalent to the traditional approach with a point source and centroiding. By use of the same $(f-1)f^{-1} = 1$ approximation, in the general case for shifts of an arbitrary amount, this inverse relationship to the SNR still holds. This means that a change in total amount of light simply results in the

estimate standard deviations for all subapertures to be scaled by a known constant.

A second very useful way to look at the problem is to break the subimage down into two components: the actual image and the uniform background. This model is appropriate for paths through the atmosphere, when scattering increases the background level. This can be due to long paths or high particulate levels.¹¹ In this case the image pixels are parameterized by $b + f\lambda[m, n]$, where b is the level of background light. Again, for any shift, the estimate expectation is independent of background level and illumination amount. After some algebraic simplification, the estimate error standard deviation in the zero-shift case is

$$\sigma_x(b, f) = \frac{\{Nb^2 + 2bf^2(\tilde{m}_0 - \tilde{m}_2 + \tilde{t}f^{-1}) + f^3[\tilde{\sigma}_1^2 - (f-1)f^{-1}\tilde{m}_1 - \tilde{\sigma}_{-1,1}^2]\}^{1/2}}{2\sqrt{2f^2(\tilde{m}_0 - \tilde{m}_1)}}, \quad (25)$$

where N is the total number of pixels in a single subimage and t is another image statistic (see Appendix A.) The sharpness of the correlation peak is still the dominant term in this expression.

The advantage of these scaling laws is that, based on a single copy of the subimage at a known light level, the behavior of that image over a broad range of conditions can be predicted. Figure 3 shows simulation results for a specific image over a range of total light and amount of background. The top panel shows the effect of reducing the exposure time; the bottom shows the effect of increasing amounts of background. In both cases, the predictions [using Eqs. (23) and (25)] based on a single copy of the image are a very good fit to the actual performance. The

scaling approximation [Eq. (24)] has significant error at very low light levels, which is to be expected due to the $(f-1)f^{-1} = 1$ approximation.

These laws aid in system design. In terms of choosing the frame rate, performance falls off with the inverse of the SNR. Background light can be very detrimental. When the amount of background light begins to exceed the image signal, the standard deviation increases rapidly, though the estimate mean remains the same. In the case shown in Fig. 3, similar performance is obtained for a no-background image with a maximum of 100 photons per pixel as for a background level of 250 photons per pixel and an image level of another 250 photons. Five times as much total light is required in this case

when the background scatter is equivalent to the image content.

4. Computational Cost Estimates

Better AO correction drives system design to more actuators and faster rates, which are limited by the total amount of light available. But more pixels and more subapertures lead to an increased computational burden. A balance must be struck between computational costs and desired correction levels. As an FFT-based algorithm, SBWFS computation scales as $O(n_p^2 \log n_p)$, where the number of pixels n_p across the subaperture is a power of 2. This computation must be done for each of the n_s subapertures, making the total cost $O(n_s n_p^2 \log n_p)$. These slope

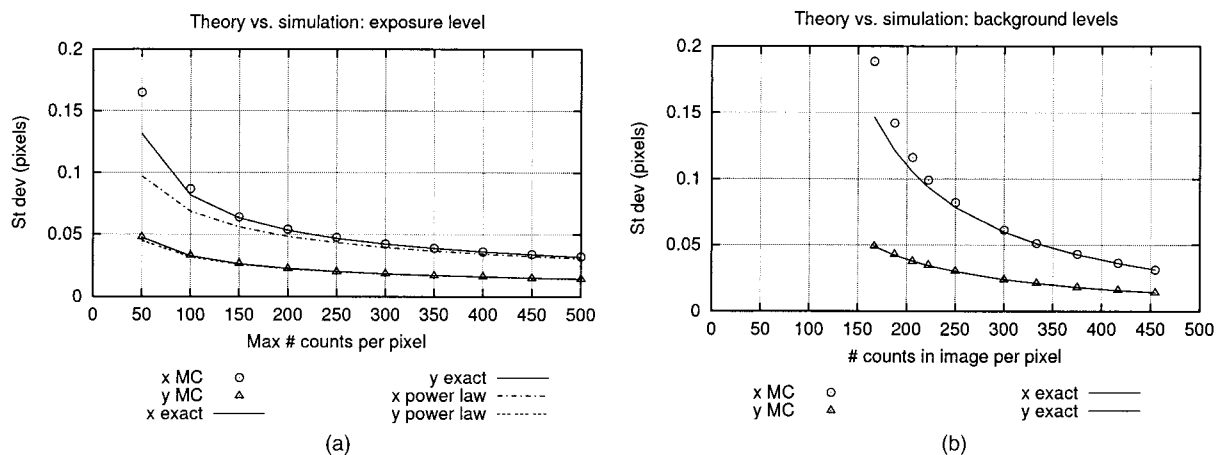


Fig. 3. Scene performance varies with the total amount and type of illumination. Simulation results are shown for the x- and y-slope estimate in the zero-shift case. In both panels the open symbols are simulation results. The solid curves are exact calculations of scene quality. The dashed curves are predictions based on the scaling law [Eq. (24)] and a single copy of the image: (a) Variation in total amount of light. Reduction in the total amount of light causes a fall-off inversely proportional to the SNR. (b) Increasing background and constant total amount of light. Performance falls off rapidly once the amount of background light is equal to or greater than image light. For this image, 100 photons per pixel of image and no background gets comparable performance to 250 photons background and 250 photons image. St dev is standard deviation.

Table 1. Total kFLOPs^a per Cycle for Slope Estimation and Phase Reconstruction

Step	156 Subaps. ^b	716 Subaps. ^b	3024 Subaps. ^b
Slope 8×8	300	1,370	5,810
Slope 16×16	1,600	7,330	31,000
Slope 32×32	7,990	36,700	155,000
Phase recon ^c	61.4	102	492

^akFLOPs, kilo-floating point operations.

^bSubaps., subapertures.

^cRecon, reconstruction.

estimates are then reconstructed to obtain the phase. For small numbers of actuators n_a , the vector-matrix-multiply method is fine. But it scales as $O(n_a^2)$. For larger systems a faster algorithm, such as Fourier transform reconstruction (FTR),¹² is preferable. FTR scales as $O[(n_a/2)\log n_a]$.

The dominant computational term is the slope estimation, which becomes enormous for large n_s . Exact FLOPs calculations are given in Table 1 for a range of reasonable schemes using FTR. For high control rates on large systems, parallel processing of the slopes will be necessary. However, a low-order AO system, with 14 subapertures across the pupil diameter (156 total), running at 10 Hz with 16×16 pixel subapertures would require a 16 MFLOPs (mega-floating point operations) per second for sensing and reconstruction. This is a very reasonable computational burden.

5. Scene-Based Wave-Front Sensing for Point Sources

The preceding sections have shown that SBWFS is a robust method. It can be applied to arbitrary scenes, with both estimate mean and variance predictable *a priori*. Operating on null, it exhibits no bias and the same inverse-power-law performance the same as using a point source. For an arbitrary shift, the slope estimate mean is insensitive to background light. Why not use it for wave-front sensing with a point source? Current methods calculate the center of mass of the spot formed by the point source.¹³ SBWFS has a couple of particular advantages to these approaches.

In the general case, many AO systems can have background light on the WFS. This can be due to the internal WFS camera characteristics or simply to scattered light. It may be very difficult to characterize and correctly remove the background. Even low levels of background can cause a centroiding algorithm to produce very poor answers. As shown above, any uniform background level is irrelevant to the slope estimate expectation in SBWFS. Therefore no detailed background subtraction is necessary, and a time-varying level of background light will not affect system performance.

SBWFS could address problems specific to astronomical AO systems with the quad-cell (2×2 pixel subaperture) arrangement. The SBWFS would have to use 4×4 pixels to have enough points to

make a correlation peak to interpolate. There are two well-documented situations where the quad-cell method suffers from extra error, both having to do with the size of the spot. As the size of the spot changes (which it does dynamically owing to changing atmospheric conditions) the gain of the WFS measurement also changes, when it is calculated with the quad-cell centroid algorithm. This matter, and how to compensate for it, has been studied.¹⁴ Spots, no matter their size, are good scenes. Analysis shows that within the linear range of the WFS, the SBWFS algorithm is insensitive to the spot size changes. Note that if the spot is too small, the WFS is not linear, and neither method will work well.

Second, and somewhat more subtly, AO systems can be run off null with the use of reference centroids. Once calibrated, a specific subaperture is driven toward a non-zero shift. If the spot size changes, the centroiding algorithm will no longer accurately measure this offset, causing problems with AO operation. This has occurred when Uranus was used instead of a guide star in the Keck AO system.¹⁵ Because SBWFS is insensitive to this spot size change it would enable this kind of operation.

It must be noted that a disadvantage of using SBWFS is the extra computational cost. In the astronomical case SBWFS requires 4 times more FLOPs than a 4×4 centroider and 30 times more FLOPs than the quad-cell algorithm. If the AO system has limited computational resources, the extra computation for SBWFS may cause a reduction in system rate and hence greater temporal errors.

SBWFS has the potential to be an alternative slope estimation algorithm for point sources. Provided a suitable reference is available, SBWFS could produce accurate, consistent slopes as conditions, such as spot size and background level change, unlike current centroiding methods. The author is currently undertaking a detailed comparison of centroiding and SBWFS for point-source sensing in astronomical and vision AO systems.

6. Conclusions

This paper has presented several important advances in the understanding of SBWFS. Scene performance as a slope estimator is shown to be entirely dependent on the scene content and illumination type. In the general case, slope estimate bias is fully predictable *a priori* and is independent of scene illumination and camera characteristics. With a Poisson statistics model, the slope estimation is shown to follow the same inverse power law in SNR for error standard deviation as point sources do. The SBWFS technique can be applied to point sources, potentially providing a WFS method that is more robust to spot shape, size, and background levels. The material presented here enables the design and use of AO systems that can use an arbitrary scene to do Shack-Hartmann-based wave-front sensing instead of a point source.

Appendix A: Calculation of the Statistics of the Correlation Function

To determine the mean and variance of the estimator, all that is needed is knowledge of the means and variances of the individual pixels of the WFS camera themselves. In the most general case, both images are modelled as independent vectors of independent random variables. For the two vectors $r[i, j]$, $s[i, j]$, each element has mean $m_r[i, j]$ or $m_s[i, j]$ and variance $\sigma_r^2[i, j]$ or $\sigma_s^2[i, j]$. Let C_k be the product of the two vectors at offset k in the x direction, namely:

$$C_k = \sum_i \sum_j r[i - k, j]s[i, j]. \quad (A1)$$

The mean value is

$$E[C_k] = m_k = \sum_i \sum_j m_r[i - k, j]m_s[i, j] \quad (A2)$$

and the variance is

$$\begin{aligned} \text{Var}(C_k) = \sigma_k^2 = & \sum_i \sum_j (\sigma_r^2[i - k, j]\sigma_s^2[i, j] \\ & + \sigma_r^2[i - k, j]m_s^2[i, j] \\ & + m_r^2[i - k, j]\sigma_s^2[i, j]). \end{aligned} \quad (A3)$$

The covariance of any two values of the correlation function is

$$\begin{aligned} \text{Covar}(C_k, C_l) = \sigma_{k,l}^2 = & \sum_i \sum_j (\sigma_s^2[i, j]m_r[i - k, j] \\ & \times m_r[i - l, j] + \sigma_r^2[i, j]m_s[i \\ & + k, j]m_s[i + l, j]). \end{aligned} \quad (A4)$$

If the reference is fixed (nonrandom), the above analysis still applies. The variance $\sigma_r^2[i, j]$ now simply equals zero in the above formulas.

Given that the images are formed from light received by an optical system, the appropriate choice for the random variables is that they have Poisson distributions. This means that the expectation and variance of each random pixel are equal, namely $m_r[i, j] = \sigma_r^2[i, j] = \lambda_r[i, j]$. Plugging these terms into the above equations generates simple summations of λ parameters to obtain the statistics. The mean value of C_k is

$$m_k = \sum_i \sum_j \lambda_r[i - k, j]\lambda_s[i, j], \quad (A5)$$

the variance is

$$\begin{aligned} \sigma_k^2 = & \sum_i \sum_j \lambda_r[i - k, j]\lambda_s[i, j] \\ & \times (1 + \lambda_s[i, j] + \lambda_r[i - k, j]). \end{aligned} \quad (A6)$$

The covariance of C_k and C_l is

$$\begin{aligned} \sigma_{k,l}^2 = & \sum_i \sum_j (\lambda_s[i, j]\lambda_r[i - k, j]\lambda_r[i - l, j] \\ & + \lambda_r[i, j]\lambda_s[i + k, j]\lambda_s[i + l, j]). \end{aligned} \quad (A7)$$

One other term that will be used is t , which is the sum of the parameters $\lambda[i, j]$. This model can be easily expanded to account for gain factors, read noise, and background light.

The work detailed herein is part of a larger program, and the author has benefitted from interactions with her colleagues J. Brase, L. Flath, D. Gavel, E. Johansson, K. LaFortune, R. Sawvel, and C. Thompson. 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 No. W-7405-Eng-48. The document number is UCRL-JC-151642.

References

1. P. Wizinowich, D. S. Acton, C. Shelton, P. Stomski, J. Gathright, K. Ho, W. Lupton, K. Tsubota, O. Lai, C. Max, J. Brase, J. An, K. Avicola, S. Olivier, D. Gavel, B. Macintosh, A. Ghez, and J. Larkin, "First light adaptive optics images from the Keck II telescope: a new era of high angular resolution imagery," *Publ. Astron. Soc. Pac.* **112**, 315–319 (2000).
2. P. C. Chen, C. W. Bowers, D. A. Content, M. Marzouk, and R. C. Romeo, "Advances in very lightweight composite mirror technology," *Opt. Eng.* **39**, 2320–2329 (2000).
3. M. G. Lofdahl and G. B. Scharmer, "Wavefront sensing and image restoration from focused and defocused solar images," *Astron. Astrophys. Suppl. Ser.* **107**, 243–264 (1994).
4. R. G. Paxman, J. H. Seldin, M. G. Lofdahl, G. B. Scharmer, and C. U. Keller, "Evaluation of phase-diversity techniques for solar-image restoration," *Astrophys. J.* **466**, 1087–1099 (1996).
5. T. R. Rimmele, "Solar adaptive optics," in *Adaptive Optical Systems Technology*, P. L. Wizinowich, ed., *Proc. SPIE* **4007**, 218–231 (2000).
6. T. R. Rimmele, O. von der Luehe, P. H. Wiborg, A. L. Widener, R. B. Dunn, and G. Spence, "Solar feature correlation tracker," in *Active and Adaptive Optical Systems*, M. A. Ealey, ed., *Proc. SPIE* **1542**, 186–193 (1991).
7. L. G. Brown, "A survey of image registration techniques," *ACM Comp. Surveys* **24**, 325–376 (1992).
8. M. A. van Dam and R. G. Lane, "Wave-front slope estimation," *J. Opt. Soc. Am. A* **17**, 1319–1324 (2000).
9. H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Trans. Image Process.* **11**, 188–200 (2002).
10. G. A. Tyler and D. L. Fried, "Image-position error associated with a quadrant detector," *J. Opt. Soc. Am. A* **72**, 804–808 (1982).
11. W. L. Wolfe and G. J. Zissis, eds., *The Infrared Handbook*, revised ed. (Office of Naval Research, Washington, D.C., 1978).
12. L. A. Poyneer, D. T. Gavel, and J. M. Brase, "Fast wave-front reconstruction in large adaptive optics systems with use of the Fourier transform," *J. Opt. Soc. Am. A* **19**, 2100–2111 (2002).
13. J. W. Hardy, *Adaptive Optics for Astronomical Telescopes* (Oxford University, New York, 1998).
14. J.-P. Veran and G. Herriot, "Centroid gain compensation in Shack–Hartmann adaptive optics systems with natural or laser guide star," *J. Opt. Soc. Am. A* **17**, 1430–1439 (2000).
15. I. de Pater, S. G. Gibbard, B. A. Macintosh, H. G. Roe, D. T. Gavel, and C. E. Max, "Keck adaptive optics images of Uranus and its rings," *Icarus* **160**, 359–374 (2002).