2. An Efficient Digital Algorithm for Envelope Detection

This series is divergent; therefore we may be able to do something with it.

Oliver Heaviside

2.1 Background to White Light Interferometry

Recently there has been much interest shown in the area variously known as white light interferometry[1] (WLI), coherence radar,[2] coherence probe/scanning,[3, 4] correlation microscopy[5-7], interference microscopy,[3, 6, 8] and low coherence interferometry.[9] The main reason for such interest is that the ambiguity present in conventional monochromatic interferometers is not present in white light interferometry. White light interferometers have a virtually unlimited unambiguous range whereas their monochromatic counterparts are usually limited to not more than half a wavelength (slightly more for systems using high aperture microscope objectives.[10-12]) The close parallel between white light interferometry and confocal (as well as conventional) microscopy was noted in the early literature,[3] but has been largely ignored since. Like confocal microscopy, WLI allows surface profiling with high accuracy over a large range, but unlike confocal microscopy WLI allows the entire image field to be captured in one instant without the need for scanning apertures.

Although the objective of WLI can be simply stated: to find the location of peak correlation (or peak fringe visibility), a problem arises because of the large three-dimensional sample datasets and the associated computational burden. A typical system[6] collects images containing 256x256 pixels over a series of 64 equi-spaced sections. If this data were processed using the exact Fourier method[13] to ascertain the peak correlation depth at each pixel then at least 128x6 multiplications must be evaluated at each pixel resulting in about 56 million multiplications. A new algorithm which is many times faster (about eight times faster on the above dataset) than the Fourier method can be developed from a generalised form of the well-known 5-sample phase-shifting algorithm (PSA).[14, 15] The application of *spatial* phase-shifting algorithms to WLI was novel when this work was originally published in 1996.[16] Nevertheless it should be mentioned that the application of *temporal* rather than *spatial* phase-shifting algorithms using 3 full datasets has been used previously.[2] The original aim of this work was to investigate the suitability of phase-shifting algorithms for white light fringe analysis, but as the analysis and simulations progressed it became clear that one particularly neat[17] algorithm combined the properties of simplicity and robust efficiency . The intent of this chapter is to clarify the rather circuitous development of the new algorithm and demonstrate some of its unique properties.

Like most things in science, mathematical algorithms are often discovered and independently rediscovered many times, and the new algorithm is no exception. A final section containing a sleuth's eye view of an algorithm which has been called the energy separation algorithm, phase congruency, "a procrustean technique", and Shank's method to name but a few, is included for completeness.

2.2 Structure of a White Light Interferogram

The light intensity, g, measured in a white light interferometer which is spatially incoherent has the following form (see Chim and Kino,[5] for example):

$$g(x,y,z) = a(x,y) + b(x,y)c[z - 2h(x,y)]\cos[2\pi w_0 z - \alpha(x,y)]. \qquad (2.1)$$

Coordinates x, y correspond to the conventional transverse object and image coordinates, while the coordinate z indicates the axial location or defocus of the object. The quantity $a(x,y)$ is an offset related to the reference and object beam intensity profiles. The reflected beam intensity determines $b(x,y)$. The interferogram envelope function c is related to the spectral profile of the white light source, while the spatial frequency of interference fringes in the z direction, w_0, is related to the mean wavelength of the light. A phase change on reflection due to the complex reflectance of the surface determines the parameter α. Many papers have assumed $\alpha = 0$, although this is generally not the case.[18] Arbitrary control of the fringe phase, α, using the geometric phase is now possible and has been recently demonstrated.[19] I shall use the term correlogram[2] to describe the function $g(z)$ when the emphasis is upon the z variation displaying its characteristic form of fringes within an envelope determined by spectral correlation.

The exact form of the envelope $c = c(z)$ is not critical although it is usually approximated by a Gaussian function to simplify calculations, especially those in the Fourier domain. In interferometer systems without suitably matched reference and object paths $c(z)$ may not be symmetric because of dispersion.

A spatially incoherent source ensures that correlograms can be considered independent of (x,y) location. In a practical system, intensity measurements are performed over a uniform array of x, y and z values. The x and y array values are determined by the pixels of a CCD array. Generally CCD arrays of 256×256 pixels or 512×512 pixels are common with even larger formats becoming popular. The sequence of values of z at which the intensities are sampled is determined by the sequence of positions of a piezoelectric transducer. The sampling in the x, y and z direction must satisfy certain constraints. In the transverse directions these constraints can be summarised in terms of the conventional x, y image bandwidth. Sampling in the z direction is covered in section 2.4.2.

Three-dimensional datasets need significant memory storage capacity. For example, a $256 \times 256 \times 64$ dataset at one byte resolution requires 4 megabytes of memory. Memory cost is rapidly becoming much less significant in digital instrumentation; in fact the renewed interest in white light interferometry has been partly stimulated by these lower costs.

2.3 Ideal Envelope and Phase Detection

The usual purpose of white light interferometry is to determine the profile (characterised by $h(x,y)$) of a surface too steep for monochromatic interferometry. Also of importance is the phase change on reflection, (related to $\alpha(x,y)$) which is determined by the dielectric properties of the surface. The function $\alpha(x,y)$ although often ignored in correlogram analysis shall be considered here because its value inevitably appears in an analysis of the main envelope-detection process.

Figure 2.1 shows a typical correlogram intensity distribution as a function of z. In this particular instance, the envelope is Gaussian. Generally it is possible to derive $h(x,y)$ and $\alpha(x,y)$ from a sequence of samples of the intensity (given by equation (2.1)) over a range of z values.

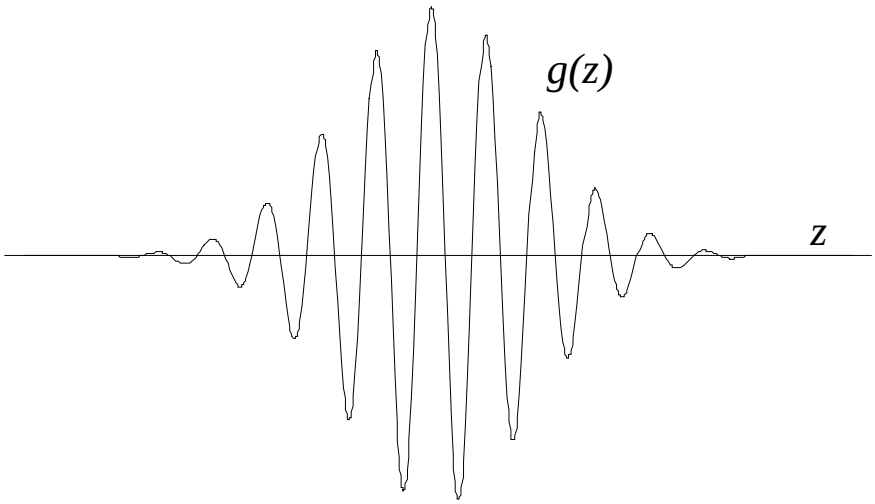


Figure 2.1.
Typical white light interferogram, $g(z)$. The fringe phase offset in this example is $\alpha = \pi/4$.

Some insight can be gained by first considering the continuous Fourier transform of equation (2.1) with respect to the z coordinate.

$$G(x,y,w) = \int_{-\infty}^{\infty} g(x,y,z)\exp(-2\pi i w z)\,dz\,. \tag{2.2}$$

Hence

$$\begin{aligned} &G(x,y,w) = \\ &a(x,y)\delta(w) \\ &+\frac{b(x,y)}{2}\left\{C(w)\exp\left[-4\pi w h(x,y)\right]\right\} \otimes \left[\exp(i\alpha)\delta(w-w_0) + \exp(-i\alpha)\delta(w+w_0)\right]. \end{aligned} \tag{2.3}$$

The symbol $\otimes$ indicates one-dimensional convolution. Spatial frequency in the z direction is denoted by w, and $\delta(w)$ is the Dirac delta function. The Fourier transform of $c(z)$ is $C(w)$. Explicit (x,y) variation can be ignored in the following analysis as long as it is remembered that the calculations are always performed over an array of points (x,y) in the sampled data. Equation (2.3) can be rewritten

$$\begin{aligned} G(w) &= a\delta(w) \\ &+ \frac{b}{2}\exp(+i[\alpha + 4\pi w_0 h])C(w - w_0)\exp(-4\pi i w h) \\ &+ \frac{b}{2}\exp(-i[\alpha + 4\pi w_0 h])C(w + w_0)\exp(-4\pi i w h). \end{aligned} \tag{2.4}$$

In terms of phase and modulus

$$G(w) = |G(w)|\exp(i\phi(w)) \tag{2.5}$$

with

$$|G(w)| \cong a\delta(w) + \frac{b}{2}|C(w - w_0)| + \frac{b}{2}|C(w + w_0)| \tag{2.6}$$

and

$$\begin{cases} \phi(w) \cong -\alpha - 4\pi(w - w_0)h, & w > 0 \\ \phi(w) \cong +\alpha - 4\pi(w + w_0)h, & w < 0 \end{cases}. \tag{2.7}$$

Thus, Equation (2.6) shows that $G(w)$ has an impulse at the origin and sidelobes centered at frequencies $w = \pm w_0$. A typical plot of the modulus $|G(w)|$ is shown in figure 2.2. The approximate equality in the previous equations is achieved when there

is minimal overlap of the sidelobes. Generally the impulse is somewhat spread out by noise and a variation of parameter *a* with *z*. The two lobes in figure 2.2 are well separated; that is to say, the separation is typically greater than the bandwidth of the lobe. The bandwidth is inversely related to the envelope width in the spatial domain.

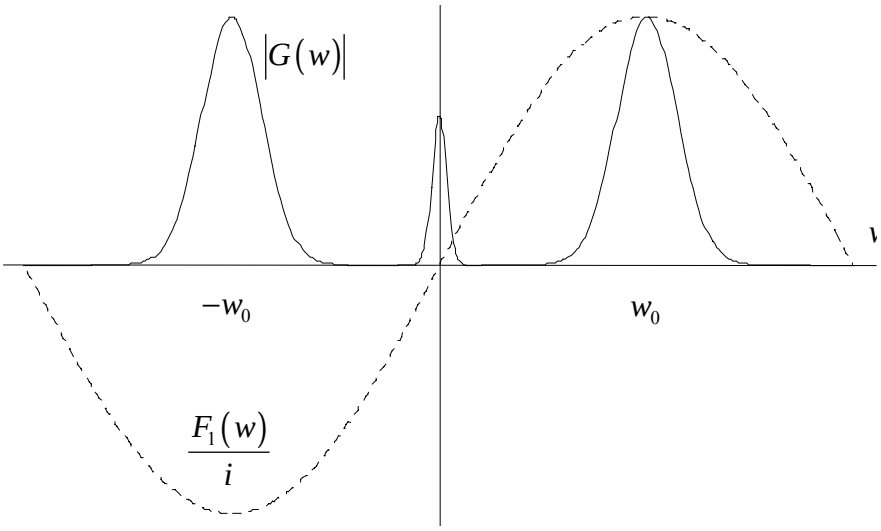


Fig 2.2

Modulus of the Fourier transformed correlogram $|G(w)|$. The normally large DC term has been reduced in magnitude to fit in the chosen scale. Also shown for comparison (dashed) is the FT $F_1(w)$ of the finite difference filter $f_1(z)$ with $\Delta = (4w_0)$.

Careful analysis of the phase function ϕ reveals that the two parameters of interest, *h* and α, can be simply extracted from the slope and intercept of the linear portion of the curve. Nevertheless, care must be taken performing a linear regression on the phase. Only in the regions where $|G(w)|$ has significant, non-zero, values will the phase have meaningful values. A weighted least square fit to the phase using $|G(w)|^2$ as the weight automatically gives good estimates of *h* and α. In this context *h* actually corresponds to the position of the centroid of the envelope as can be shown by a Fourier correspondence theorem.[20] More recently this relationship has be

explored in greater detail by Sheppard and Larkin.[21] Unfortunately, calculation of h and α using the above method is computationally intensive. The fast Fourier transform of a real array alone requires $N\log_2 N$ floating point multiplications followed by N multiplications for an optimized least square fit over the positive frequency region. A weighted fit requires additional $2N$ multiplications. Here N is the total number of samples in the z direction (typically $N = 64$). The calculation is then repeated for each element in the (x, y) array, which is typically 256×256.

Conceptually, perhaps the easiest way to obtain the phase and envelope, is by using the transform technique which has been outlined in several papers.[5, 7, 13] Briefly, the method entails fast Fourier transforming the raw data (typically 64 samples at each x,y location) then removing the negative and zero frequency components. Finally, the transform data are re-centered at the midpoint of the sidelobe and then inverse transformed. In fact the very same technique is better known as the Fourier Transform Method of fringe analysis in interferometry.[22] The technique essentially generates the analytic signal[23] by imposing causality. The signal $s(z)$ that results has the following form

$$s(z) = b.c(z-h).\exp(i[\alpha - 4\pi w_0 h]) \tag{2.8}$$

The modulus of $s(z)$ is the envelope we require and the argument of $s(z)$ contains the phase offset α. The method outlined above is even more computationally intensive than the least-squares fit method because forward and inverse FFTs are required in addition to the squaring operations required to determine the modulus.

More recently an improved, computationally efficient, method of determining the envelope has been developed.[6] The method relies on a real space implementation of the previous double FFT method. The crucial point is the introduction of the Hilbert transform convolution kernel as a digital FIR filter.[24] Computational efficiency is gained by realising that an approximate Hilbert transform can give near perfect results for typical interferograms. A real space implementation of the Hilbert transform has also been proposed for 2-D interferogram analysis by Zweig and Hufnagel.[25] As I shall show in the following sections of this chapter the real space implementation can be greatly simplified, so much so that the final algorithm requires only 2 multiplications per point to obtain the envelope squared at each point. In terms of multiplications alone this is believed to represent a lower limit upon numerical envelope detection using quadrature functions (i.e. a counterexample has not yet been found).

2.4 Approximations to Hilbert Transform Envelope Detection

2.4.1 The Quadrature Property of the Hilbert Transform

The perfect Hilbert transform (HT) can be considered equivalent to a wideband 90° phase-shift operation,[25] but, as we have already seen, a typical white light interferogram is an approximately bandlimited signal. That is to say, the spatial frequency content is limited to a region centered about a "carrier" frequency. The wideband property of the perfect HT is not required for such signals. Thus the constraints upon an approximation to the HT can be relaxed - it only has to have a 90° phase-shift over a limited frequency band. Outside this band, the transform can have

any phase-shift but the modulus of the response should be low, thus suppressing noise present outside the passband. In the case of discrete sampled data with white noise, it is desirable for the transformer to have zero or near zero response at frequencies below and above the pass band. An important practical requirement for a numerical discrete envelope detector is computational efficiency. A standard of efficiency to compare any discrete implementation by is the method of Chim and Kino[6] in which the main computational burden for N samples is due to the $6N$ multiplications and N square roots required to obtain the envelope. All the above requirements can be met by a pair of quadrature filter functions well known to researchers working in the area known as phase-shifting interferometry. The crucial Fourier properties of these functions have been derived[26] and extended,[27] but have not previously been applied to the analysis of white light interferograms. Freischlad and Koliopolous[26] introduced the two filter functions $f_1(t)$ and $f_2(t)$ which are correlated with a generalized interference pattern $g(x,y,t)$ to produce two quadrature functions, the ratio of which gives the tangent of the phase sought by the technique while the envelope (or modulation) is given by the root sum of squares. A well-known algorithm which uses 5 samples[14, 15] has the following discrete filter functions:

$$f_1(t) = 2\big(\delta(t-\Delta) - \delta(t+\Delta)\big) \tag{2.9}$$

and

$$f_2(t) = -\delta(t-2\Delta) + 2\delta(t) - \delta(t+2\Delta). \tag{2.10}$$

Here Δ is the step between samples. Readers familiar with signal processing or digital filtering may recognize these as simple, finite impulse response (FIR) filters.

In the case of white light interferograms, the temporal parameter t is replaced by a spatial coordinate parameter z. In fact f_1 and f_2 are symmetrical and anti-symmetrical linear phase digital filters respectively.[28] The function f_1 has the property that its Fourier transform is an imaginary odd function , whilst that of f_2 is a real even function. Figure 2.3 shows F_1 and F_2, the transforms of f_1 and f_2. The choice of the functions f_1 and f_2 is important. There are many possible choices, with three samples being the minimum number required for such an algorithm.

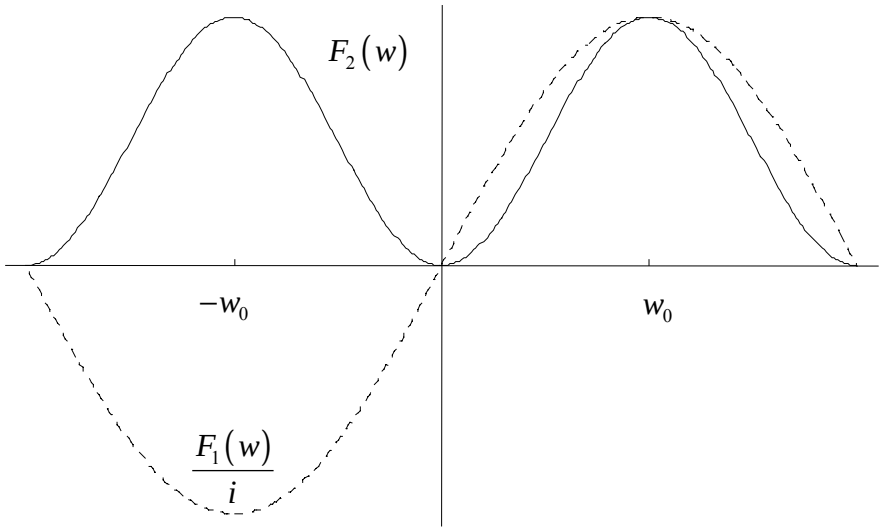


Figure 2.3
The spectral responses of the numerator and denominator filters of the five-sample phase-shifting algorithm.

The 5-sample algorithm, however, has several important properties not possessed by any other algorithms. The most important of these is related to the matched gradients of the Fourier transforms F_1 and F_2 at the fundamental frequency $w = 1/(4\Delta)$.[26] In figure 2.3 the gradients at the fundamental frequency can be seen to be near zero and slowly varying in this region. A consequence of the zero gradient is that both phase

and envelope calculations for the 5-sample algorithm are relatively insensitive to sample spacing (or step size) errors; essentially the filter response is stationary in this region. Conversely, such an algorithm is insensitive to fringe spacing variation when the sampling is fixed.

2.4.2 Sampling Considerations

At this point it is worth considering the sampling requirements for white light interferometry. Most sampling schemes greatly oversample the correlogram data. A typical example, Chim[6] uses about 8 to 10 samples per period. This exceeds the Nyquist criterion by at least a factor of 4. For a typical correlogram this gives a very one-sided distribution of information in the Fourier frequency domain. The transform data are centered on the frequency $1/(8\Delta)$, which is one quarter of the Nyquist frequency. An even distribution of frequency components requires about 4 samples per fringe and keeps the peak frequency midway between DC and the Nyquist frequency $1/(2\Delta)$.

More detailed analysis of the optimum sampling requirements for bandpass signals have been considered elsewhere.[29, 30] It is generally agreed that it is the *bandwidth* of a narrow-band signal which determines the sampling frequency not the *carrier plus bandwidth* as recently suggested by Caber.[31] So we can see that the proposition that $4f_c + 2B$ is the minimum sampling frequency necessary to avoid aliasing when the signal is squared (here f_c is the fringe frequency and B is the envelope bandwidth) fails to account for the fact that the aliasing that results when

sampling at $4f_c$ occurs at $4f_c + 2B$ and consequently has no effect after low pass filtering (with a cut-off below $2f_c$).

A criterion of just 4 samples per fringe is considered an adequate compromise here because it corresponds to the optimum sampling for a 90° step phase-shifting algorithm. A common misconception is that increased sampling frequency gives a proportionate increase in accuracy, but this is not the case. All the following analysis is based on a nominal sampling frequency of 4 samples per fringe (practically, this will cover the range from about 3 to 8 samples per fringe) but is easily modified for other values.

Recently a method has been to sample correlograms at a frequency determined by the bandwidth,[32, 33] although such "undersampling" was previously used in 1991.[9] The technique has been called sub-Nyquist sampling. Instead of sampling the fringe pattern at 4 samples per carrier fringe there is undersampling by an odd integer factor so there are $4/(2L+1)$ samples per fringe, where L is typically 1 or 2. Undersampling of this kind avoids common aliasing problems so long as the envelope bandwidth alone is adequately sampled. An inevitable consequence of undersampling is that the allowable error in the initial prediction of the mean wavelength is inversely proportional to the undersampling factor. Undersampling also increases the effective bandwidth of the bandpass signal and can be expected to degrade the performance of demodulators in the presence of noise. There are some issues regarding the effects of sampling upon the envelope peak detection process which shall be discussed in Section 2.7.

2.4.3 Computational efficiency

Another issue worth mentioning is optimality with regard to computational efficiency. It is generally agreed (see Chim and Kino[6] for example) that mathematical operations such as multiplication, division, and the evaluation of transcendental functions and square roots are much more significant to calculation time than the add or subtract operations (although these distinctions are less important for present day personal computers). So, to estimate the computational burden of a numerical procedure it is convenient to ignore the addition and subtraction operations and just count the other operations. Envelope detection algorithms can be compared in efficiency to an idealized numerical scheme, which cannot be realized in practice but gives an idea of the limits to efficiency. A perfect envelope detection scheme could consist of the following steps:

1) Read in function values g_n and the perfect quadrature function values $\hat{g}_n$ for $n = 1 \rightarrow N$. (The sequence g_n is assumed zero-mean.)

2) Calculate envelope values $e_n = \sqrt{g_n^2 + \hat{g}_n^2}$.

In this idealized scheme evaluation of the envelope function requires $2N$ multiplications and N square root operations. The sequences g_n and $\hat{g}_n$ are considered known. In practice $\hat{g}_n$ has to be derived from the sequence g_n (which has, itself, to be made zero mean) and more computational steps must be involved. As mentioned earlier, the recent computational scheme of Chim and Kino[6] requires $6N$ multiplications and N square roots to obtain the envelope, almost a factor of three below optimality.

The difficulty of implementing a perfect Hilbert transform using a real space filter function is closely related to the well-known problem in communication theory of implementing a wideband 90° phase shifter for Single Side Band modulation (SSB).[34] One way round this is to start with the original function $g(z)$ and produce two new functions $g_a(z)$ and $g_b(z)$. The two new functions are in quadrature to each other but the phase relationship with $g(z)$ is not constrained. This gives an extra degree of freedom, which allows a practical implementation using realistic available electronic components. For example, instead of requiring a 90° phase-shifter we could consider a +45° and a -45° shifter. In terms of numerical filter functions, it is in principle trivial to derive a -45° shifter once the +45° shifter is known. Consider a real filter function $f(z)$ and its spatial Fourier transform $F(w)$. This can be conventionally represented

$$f_1(z) = f(z) \rightleftharpoons F(w) = \int_{-\infty}^{+\infty} f(z) \exp[-2\pi i w z] dz = |F(w)| \exp[i\chi(w)]. \quad (2.11)$$

The symbol $\rightleftharpoons$ here represents Fourier transformation, and $\chi(w)$ is the FT phase. A filter function with the opposite phase is simply

$$f_2(z) = f(-z) \rightleftharpoons F(-w) = \int_{-\infty}^{+\infty} f(-z) \exp[-2\pi i w z] dz = |F(w)| \exp[-i\chi(w)]. \quad (2.12)$$

In both cases the modulus of the frequency response is the same, $|F|$. The two functions above, f_1 and f_2 are the bases for a whole series of quadrature functions that

are linear combinations of f_1 and f_2. For example, a zero phase function f_o can be defined

$$f_o(z) = f(z) + f(-z) \rightleftharpoons 2|F|\cos\chi \quad . \tag{2.13}$$

In a similar way, a 90° phase function can be defined

$$f_{90}(z) = f(z) - f(-z) \rightleftharpoons 2i|F|\sin\chi \quad . \tag{2.14}$$

In this case, although f_o and f_{90} are exactly 90° in phase, their moduli are no longer necessarily equal at all frequencies. Compare this to f_1 and f_2 which have equal moduli but the phase difference is 2χ which does not necessarily equal 90° at all frequencies. The function pairs f_1 and f_2 or f_o and f_{90} can be used to generate approximate quadrature pairs of functions from the correlogram $g(z)$ simply by correlation (or by convolution with z reversed functions)

$$\left.\begin{aligned} g_1(z) &= f_1(z) \otimes g(z) \\ g_2(z) &= f_2(z) \otimes g(z) \end{aligned}\right\}. \tag{2.15}$$

2.5 Calculation of Phase and Modulation using Phase-Shifting Algorithms

The following analysis will consider continuous functions and continuous Fourier transforms. However, the techniques outlined are applicable to discrete

sampled data, and discrete Fourier transforms. Important differences between the continuous and sampled will be noted as necessary.

The main idea in this section is the application of phase-shifting algorithms for white light interferogram demodulation. Dresel et al[2] first considered the idea in 1992. Typically, such algorithms are used in interferometry to estimate the phase of a wavefront over a two dimensional array of points. In the case of the 5-sample algorithm, 5 interferograms are captured by a digital imaging system. At each pixel location in the image, all 5 intensity values are combined to extract the phase $\phi(x,y)$ at each pixel. The algorithm can be simply defined

$$\phi(x,y) = \tan^{-1}\left(\frac{2\left[I_2(x,y)-I_4(x,y)\right]}{-I_1(x,y)+2I_3(x,y)-I_5(x,y)}\right), \tag{2.16}$$

where I_1 to I_5 are the interferogram intensities. The modulation at each point $M(x,y)$ can also be calculated from

$$M(x,y) = \tfrac{1}{4}\sqrt{4\left(I_2-I_4\right)^2+\left(-I_1+2I_3-I_5\right)^2} \quad . \tag{2.17}$$

Both these formulae are exact when the phase-shift between interferograms is 90°. For phase-shifts in the region near 90° the errors in both ϕ and M are small because the first order error terms cancel.[15] This error compensating property of the 5-sample algorithm has made it popular in many digital interferometer systems. Other error compensating algorithms exist, the best known being that of Carré.[35] The interesting

point about the Carré algorithm is that it compensates *exactly* for step errors and requires only 4 interferograms whereas the 5-sample algorithm compensates only partially (second order residuals are present). The problem with the Carré algorithm is that it is more complex and hence more time consuming to compute. A little known fact is that an exact compensating form of the 5-sample algorithm exists and is somewhat less complex than the Carré algorithm. The essential background to this exact compensating 5-sample algorithm is present in the paper of Hariharan, Oreb, and Eiju,[15] but it is not shown explicitly. It can be shown that the following phase and modulation expressions are exact

$$\tan^2\phi = \frac{4\left(I_2 - I_4\right)^2 - \left(I_1 - I_5\right)^2\left(I_2 - I_4\right)^2}{\left(-I_1 + 2I_4 - I_5\right)^2} \tag{2.18}$$

$$M = \frac{\left(I_2 - I_4\right)^2\left[4\left(I_2 - I_4\right)^2 - \left(I_1 - I_5\right)^2 + \left(-I_1 + 2I_4 - I_5\right)^2\right]^{\frac{1}{2}}}{4\left(I_2 - I_4\right)^2 - \left(I_1 - I_5\right)^2} \tag{2.19}$$

The phase step between interferograms here is ψ which can have any value not equal to an integral multiple of π (this condition also applies to most phase-shifting algorithms including the Carré algorithm). The denominator of equation (2.19) can also be expressed in a form that avoids problematic zero by zero divisions whenever $\sin\psi = 0$. A serendipitous simplification of equation (2.19) leads to

$$M^2\left(\frac{\sin^4\psi}{4}\right) = \left(I_2 - I_4\right)^2 - \left(I_1 - I_3\right)\left(I_3 - I_5\right) \tag{2.20}$$

$$M^2 \propto (I_2 - I_4)^2 - (I_1 - I_3)(I_3 - I_5) \tag{2.21}$$

For values of ψ near 90° (and odd multiples thereof) the sine factor is near unity and so M can be calculated using just two multiplications and one square root operation; precisely the number of operations required for the ideal detection scheme. The division operation in equation (2.19) is then neatly avoided. Both the Carré algorithm and the conventional 5-sample algorithm require an additional multiplication to calculate M. Remarkably the optimum sampling (Nyquist and sub-Nyquist) is predetermined by the maxima of $\sin^4 \psi$ which is entirely in accord with the bandpass signal sampling discussed earlier in this section.

Strictly speaking, the temporal phase-shifting analysis appearing so far in this section only applies to phase-shifting interferometry where the modulation and offset remain constant between interferograms. A technique known as spatial phase detection[36, 37] applies the same algorithms to a single interferogram. In spatial phase-shifting the intensity values I_1 to I_5 represent spatially adjacent pixels instead of phase-shifted intensities recorded at the same pixel position. If the modulation and offset is assumed to vary slowly across the interferogram then the algorithms are approximately correct. Generally, it is necessary to introduce a large number of tilt (or "carrier") fringes into the interferogram. This is because the maximum phase variation detectable is proportional to the mean phase variation, which in turn is related to the total number of fringes. The application of spatial phase detection algorithms to white light correlograms initially appears counter-intuitive because spatial phase techniques normally assume slowly varying offset and modulation whereas white light correlograms, generally, have rapidly varying modulation.

2.6 Application of Spatial Phase-Shift Algorithms to Envelope Detection

Perhaps the easiest way to investigate envelope detection using phase-shift algorithm is by computer simulation.[38] Three representative algorithms are compared in this section. The Carré derived envelope algorithm has been omitted because it requires three multiplications to evaluate (also preliminary simulations indicate poor performance). Two of the (5-sample) envelope algorithms have been mentioned in the previous section. The third is based on the simplest 3-sample algorithm utilizing 90° phase steps.[39] The modulation factor in this case is

$$M(x,y) = \frac{1}{4}\sqrt{(I_3 - I_2)^2 + (I_2 - I_1)^2} \tag{2.22}$$

which is only correct for exact 90° steps. The above algorithm has been selected because it only requires two squaring operations and one square root (both operations are possible using fast look up table computation).

A number of other 3, 4, or 5 sample algorithms (see Creath, for example[40]) could have been chosen with and without error compensating properties. Initial testing has shown that most phase-shifting algorithms have inferior performance to the algorithm defined in equation (2.21). The three algorithms chosen here illustrate the performance of :

i) no error compensation (NEC) [equation (2.22)],

ii) partial (1st order) error compensation (PEC) [equation (2.17)], and

iii) exact compensation (five-sample adaptive or FSA) [equation (2.21)].

For convenience the algorithms shall be denoted NEC, PEC, FSA, respectively. Each algorithm has been applied in turn to a simulated white light correlogram without noise. The correlogram is shown in figure 2.1. The calculated envelopes are shown in figure 2.4 for the case where the carrier, or rather the mean fringe spacing is known precisely. In this case the algorithm step size and sampling step are set at 90° or, equivalently, one quarter of the mean period. Figures 2.1 and 2.4 show continuous functions but it can be easily shown that the discretely sampled case involves samples that occur at points located on the continuous curves.

Often the exact value of the carrier frequency is unknown before a measurement is made. The approximate value can be readily calculated from a knowledge of the spectral profile of the illumination and the numerical aperture of the imaging system (typically a microscope objective). The exact value depends on other system parameters and the spectral reflectance of the sample being measured. Therefore it cannot be known exactly before a measurement is made. Calibration of the system for each sample can be a rather tedious addition to a measurement procedure. A preferable method requires a self-calibrating (or error compensating) algorithm which works effectively over a range of carrier frequencies. To show the effects of carrier frequency variation the demodulated envelopes for two extreme values of the frequency have been calculated. In figure 2.5 the frequency is 0.5 times its nominal value and hence the samples are now 45° apart. Figure 2.6 shows the envelopes calculated when the frequency is 1.5 times its nominal value and the samples are thus 135° apart.

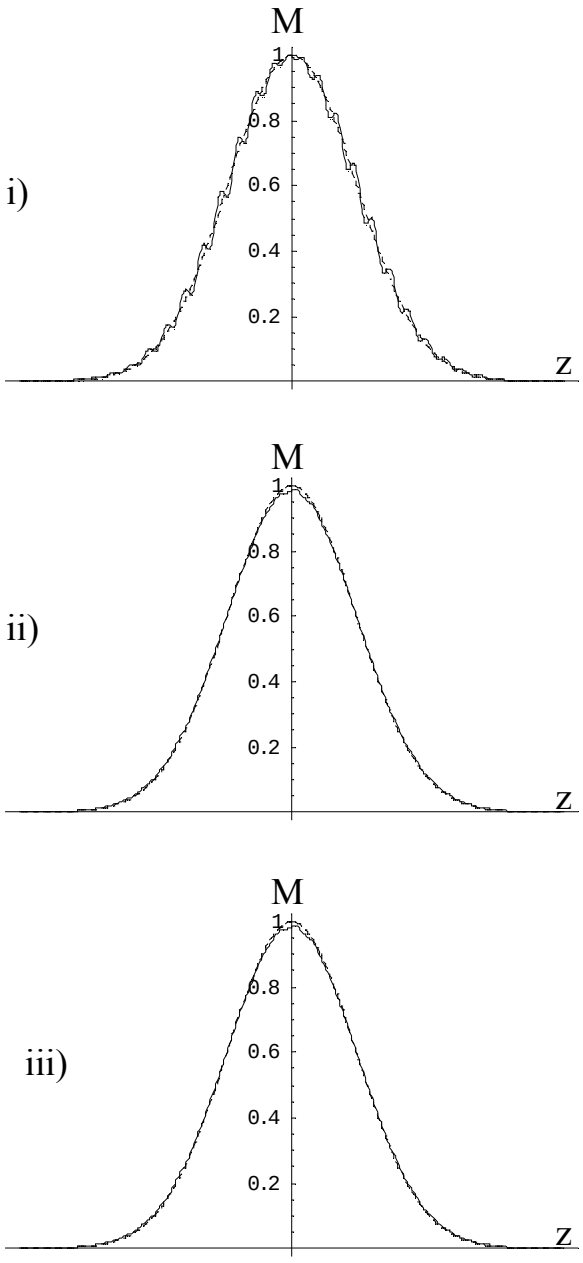


Fig 2.4

Envelope demodulation for all three algorithms using a 90° step size. In this particular case the envelopes produced by ii) and iii) are identical and close to the ideal. Algorithm i) performs less well and has noticeable fringe structure in the envelope.

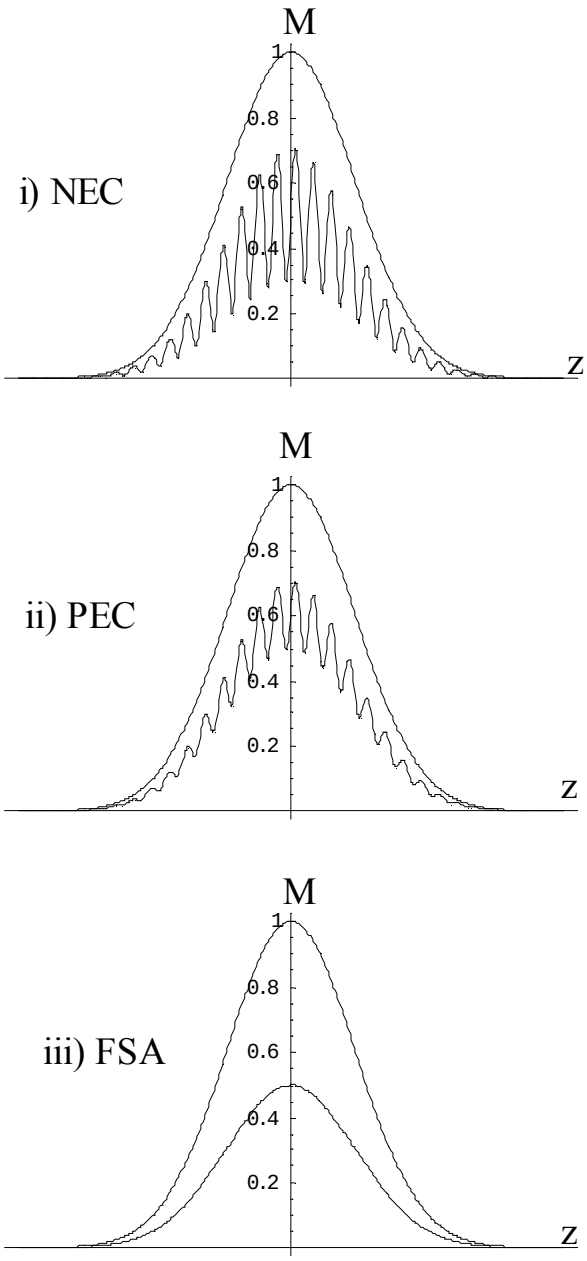


Fig 2.5

Envelope demodulation for all three algorithms using a 45° step size. Both algorithms I) and ii) have significant fringe structure visible. Algorithm iii) performs exceptionally well and has a 50% reduction as predicted.

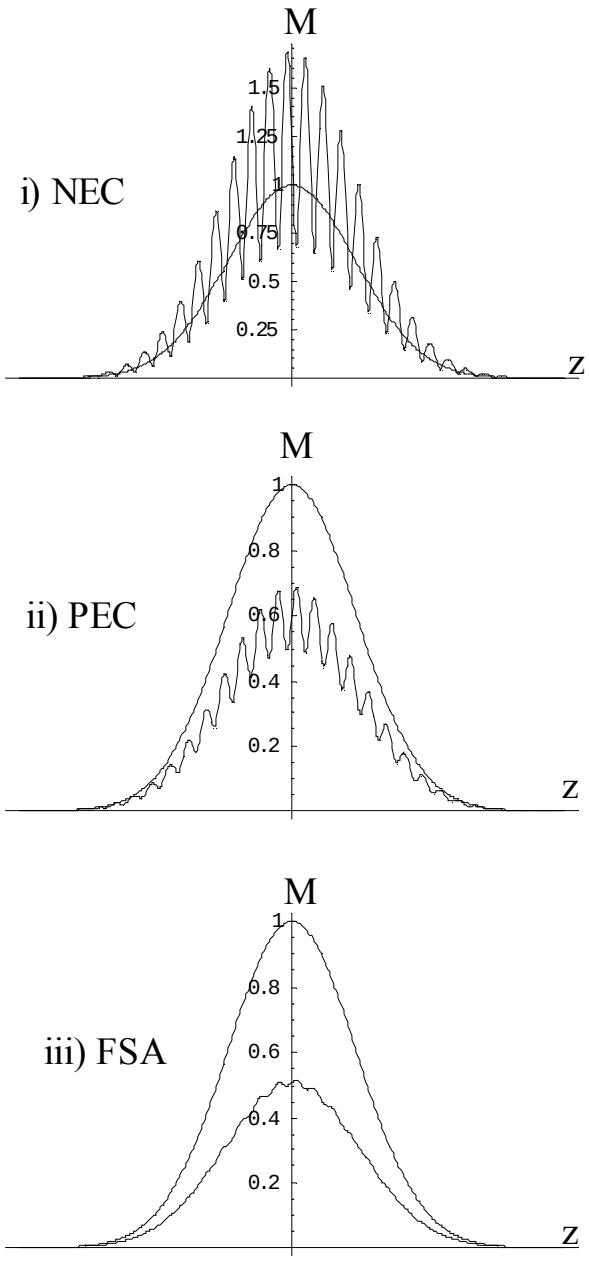


Fig 2.6.

Envelope demodulation for all three algorithms using a 135° step size. Again both algorithms i) and ii) have significant fringe structure visible, whereas algorithm iii) only shows a trace of second harmonic fringe structure.

In a system with white light in the range 400nm to 700nm the extreme frequency variation due to spectral effects alone is in the range 0.72 to 1.28 times nominal. Even then such extremes can only be achieved if the reflected light is very narrow band at either 400nm or 700nm. The range from 0.5 nominal to 1.5 nominal frequency is perhaps rather harsh, but is useful for an algorithm comparison.

A simulation of algorithm performance in the presence of noise is presented in a paper by the author[16] and in section 2.8. Some general observations on the linear filtering properties of all three algorithms[27, 41] as well as the perfect Hilbert transform method can be made. To summarize: NEC is a high pass filter, PEC and FSA are bandpass filters centered on the nominal carrier frequency, and the Hilbert transform is a wide band (all-pass) filter. In the presence of zero-mean Gaussian white noise algorithms PEC and FSA suppress spectral components of noise outside the signal bandwidth and can therefore be expected to perform well when compared to both NEC; which boosts high frequency noise, and the Hilbert transform method; which neither suppresses nor boosts noise.

2.7 Interpretation of Calculated Correlogram Envelopes

In the previous section three algorithms have been used to estimate the envelope of the white light correlogram. In all cases some fringe structure propagates through into the calculated envelopes. This problem does not occur in the Fourier transform method and only to a miniscule degree in the real space Hilbert transform technique. Of the algorithms, FSA has by far the smallest residual of fringe structure over the full range of sampling intervals from 45° to 135°. This factor is important in the process of finding the envelope peak, which is the crucial parameter. In Section

2.2, the height of the sample surface at any point, $h(x,y)$, was directly linked to the ideal envelope peak position $z_p(x,y)$

$$z_p(x,y) = h(x,y). \tag{2.23}$$

Inevitably, the three algorithms tested only approximate the desired envelope. Applying a simple point-to-point peak detection process[42] to the calculated envelope can give significant errors with respect to the ideal peak position and requires significantly more than 4 samples per fringe to work correctly. A better way to find the peak is by using the overall shape of the envelope around the approximate peak position. Simple curve fitting using three points has been proposed in an alternative approach to the envelope detection process.[31,43] In the region of the peak the calculated envelope can be expected to be well approximated by a Gaussian function, $\exp(-\beta z^2)$ where β characterises the ideal envelope. A better estimate of the peak position can thus be obtained from a least-squares fit to this function. For example the nearest and next-nearest neighbour samples can be used for a five point, symmetrical least-squares fit (LSF) to the function's exponent $\left[\beta_o - \beta\left(z - z_p\right)^2\right]$ which is a quadratic in z. The use of a symmetrical LSF greatly simplifies the calculation.[44, 45] The full process is simply implemented by taking the logarithm of the calculated envelope values and computing the peak position from an explicit solution of the symmetric LSF.

The peak detection process can just as easily be applied to the envelope squared as this only produces a factor of two in the envelope exponent. Thus, the

overall computation can be reduced by N square root operations. The increased computational burden of the 5-point LSF is only 19 operations (5 logarithms , 12 multiplications and 2 divisions). Initial analysis indicates that the dominant error in the calculated envelope occurs at the second harmonic of the carrier frequency which is typical of a second order nonlinearity. As a result the conventional LSF gives a significant error in the peak prediction. However, it is possible to define a weighted five-point LSF which is insensitive to second harmonic errors and thus gives much improved peak prediction. Five is the minimum number of points required to satisfy both LSF and harmonic criteria.[46] Equation (2.24) defines the five-point, frequency-selective, LSF peak predictor, where the symbol L_n represents the logarithm of the envelope value I_n and the distance z_p is measured from the middle (third) sample.

$$z_p = 0.4\Delta\left(\frac{L_1 + 3L_2 + 0L_3 - 3L_4 - L_5}{L_1 + 0L_2 - 2L_3 + 0L_4 + L_5}\right) . \qquad (2.24)$$

Figure 2.7 shows the result of applying the aforementioned peak detection process to the FSA algorithm envelope shown in figure 2.4. Again a continuous function analysis has been performed, but the result for a sampled function is just one point on the curves shown. In this particular instance the error of the proposed procedure is less than $1/20$ sample for calculations with initial estimates of peak location within 2 samples of the actual value. The unweighted LSF is an order of magnitude less accurate. Careful scrutiny of figure 2.7 reveals a small bias in the predicted value; a bias related to the non-zero phase change used in the example(in this case $\alpha = \pi/4$).

If the calibration is exact, then the bias error has no effect on the peak location estimate, which is then exact (see middle graph in figure 2.7).

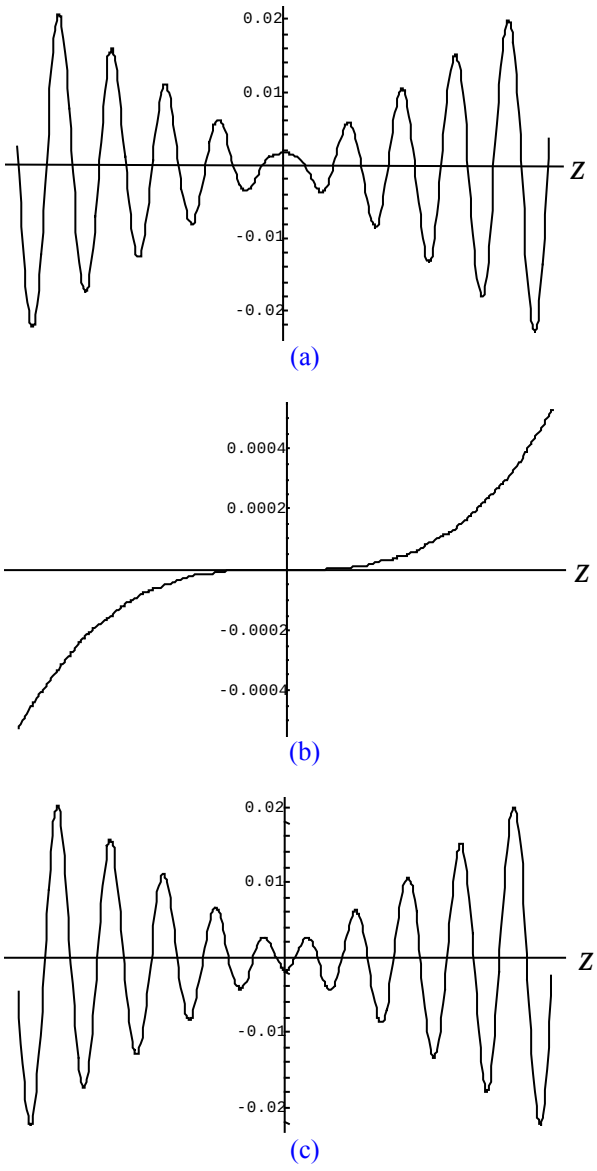


Fig 2.7
Error in the predicted peak position (in units of sample step size) as a function of initial selected peak position. The top graph is for 45° step size, the middle for 90°, and the lower for 135°. Again, a strong second harmonic structure is visible.

Unfortunately figure 2.7 (a) has been misinterpreted in subsequent research. Figure 2.7(a) shows that there is a $\Delta/20$ inherent error for highly detuned envelope

detection using the FSA algorithm. However, $\Delta/20$ is not the accuracy limit for the method. When considering a properly tuned FSA algorithm (figure 2.7 (b)) the inherent (noise free) error is of the order $\Delta/10000$ or less.

Once the peak position of the envelope has been estimated it is then possible to find the phase at that position. The FSA algorithm phase given in equation (2.18) can also be shown to have much smaller errors (due to miscalibration) than the other two algorithms. However, equation (2.18) alone has a sign ambiguity. If we assume that the sampling is not sub-Nyquist, then the phase-step is limited $0 < \psi < \pi$ and the sign ambiguity may be resolved:

$$r_1 = \frac{I_5 - I_1}{I_4 - I_2} = 2\cos\psi \tag{2.25}$$

$$r_2 = \frac{I_5 - 2I_3 - I_1}{I_4 - 2I_3 + I_2} = 2\frac{\sin^2\psi}{\cos\psi - 1} \tag{2.26}$$

$$\tan\psi = \frac{+\sqrt{r_2(r_1 - 2)}}{r_1} \tag{2.27}$$

Equation (2.27) is less noise sensitive than equation (2.25) on its own. Once the step size ψ is known, the unambiguous phase ϕ (modulo 2π) can be evaluated from the Hariharan[15] formula:

$$\tan\phi = \sin\psi \frac{2(I_4 - I_2)}{I_5 - 2I_3 + I_1} \tag{2.28}$$

The preceding four equations may be combined into one equation similar to equation (2.18). In a paper published in 1995 by Servin et al[47] a similar self-calibrating spatial algorithm is developed. In that paper the authors claimed the algorithm to be a 3-sample algorithm, although in a preprocessing step they used a 3-sample differencing operation; effectively defining an overall 5-sample algorithm. The methods are essentially identical, except that the FSA emphasises envelope demodulation, and Servin emphasises phase demodulation.

The phase at the estimated peak must be interpolated from actual calculations of the phase at sample positions on either side of the peak. The expected form of the phase near the peak is linear with respect to z, although in high aperture systems the nonlinear Gouy phase may be expected to have a significant effect.[12] In a low aperture system a two point linear interpolation can give a good estimate of the phase at the peak, in other words an estimate of $\alpha(x,y)$. More points could be used for a LSF estimate and an explicitly nonlinear phase model could be included. However, the main difficulty in estimating the phase is the occurrence of phase discontinuities due to the modulo 2π restriction of the arctangent function. A phase discontinuity in the region of the peak renders simple interpolation useless. To avoid such discontinuities it is necessary to subtract a linear phase component from the calculated phase and re-evaluate module 2π:

$$\phi_1 = \mathrm{mod}_{2\pi}\{\phi - 4\pi w_0 z\} \tag{2.29}$$

The interpolation scheme can then be applied and the mean phase term added afterwards to give α'_p. In principle, a rough estimate of w_0 is sufficient because errors cancel completely, but in practice we have calculated w_0 exactly in equation (2.27):

$$\alpha'_p = \phi_1 + 4\pi w_0 z. \tag{2.30}$$

This process works well in the region of the envelope peak for values of α not equal to $\pm\pi/2$ nor $\pm\pi$. More sophisticated complex methods may be used to evaluate the phase consistently modulo 2π, but they are not considered here. The number of additional significant computational steps for a two-point interpolation is 10. So for a data set with N values of z the total number of operations required to obtain an estimate of $h(x,y)$ and $\alpha(x,y)$ is $2N+29$. This compares favourably to the Hilbert transform method of Chim & Kino[6] which requires $6N$ non-trivial multiplications just to obtain the envelope squared. Typically $N = 64$, which means the phase-shift algorithm method is nearly two and a half times faster than the Hilbert transform kernel method. Essentially the speed gain is due to the remarkable adaptive properties of the fully compensating FSA algorithm of equation (2.21) when applied to envelope detection. Also, by limiting the accurate estimates of h and α to small regions near the estimated peak of the envelope much global calculation has been avoided.

The comparisons in speed are only valid if the methods compared have similar accuracy. Certainly the procedure consisting of the FSA algorithm followed by a weighted least squares peak prediction given by equation (2.24) has a error less than a small fraction of one sample interval. When this work was originally performed in

1995 there appeared to be no published work which considers the performance of any WLI peak prediction schemes in the presence of noise and other degradations. A preliminary analysis of this kind was included as an appendix in the published work[16] to confirm the general principle discussed here. All methods that estimate the envelope can also utilise some form of LSF peak prediction and so, presumably, are capable of sub-sample resolution. In the past not all methods have used such simple curve fitting to such advantage and therefore had a crude resolution limited to half a sample at best.

A recent paper by Caber[31] developed a communication theory approach to the interferogram envelope detection. The well known demodulation process of a bandpass filter followed by a square law nonlinearity followed, in turn, by a low-pass filter is implemented as a sequence of digital filters. Although details are not given, the known computational efficiency of two digital Infinite Impulse Response (IIR) high-pass or low-pass filters[24] in series with a squarer, is lower than the exact error compensating 5-sample algorithm outlined in the preceding sections. The availability of digital signal processing boards with special digital filter hardware may counterbalance the lower efficiency in practice. A minimum of just eight frames of data need to be stored at any moment in the Caber scheme compared to ten frames needed for the FSA algorithm proposed here. An accuracy of $1/25$ sample spacing is claimed for the Caber method.[43] The accuracy of this technique is not tested in the later simulation because details of the IIR filters used have not been disclosed in the open literature.

Some final remarks about the potential accuracy of the sub-Nyquist sampling method of de Groot[32] follow. The method relies upon a best fit to the phase gradient

calculated from the Fourier transform of the sampled data. The phase gradient at the carrier frequency is easily shown to be proportional to the first moment (or centroid) of the envelope by a well-known Fourier correspondence theorem. Similarly the weighted LSF to the phase gradient (weighted by the magnitude squared) is proportional to the centroid of the *square* of the envelope.[20, 21] Several authors have studied the effects of sampling upon centroid estimation[48-50] essentially concluding that sampling must satisfy the bandpass sampling requirements mentioned in section 2.4. The main point however, is that phase gradient estimation (equivalently Fourier transform centroid estimation) is quite distinct from peak detection. In statistics the centroid (mean) is known to be susceptible to noise, that is to say it is not a robust estimator. The effect of noise upon the centroid increases with the interval over which the centroid is evaluated. In contrast the peak prediction schemes outlined earlier only depend upon values of a distribution near the peak. A balanced assessment of the two techniques must compare accuracy versus computational complexity. The emphasis in this and many other publications has been on the detection of the location of the interferogram peak. In many cases a stricter definition of the interferogram properties versus the height parameter may be possible and a truly optimal estimation procedure may be defined. For example, matched (correlation) filtering is appropriate in the case where the exact envelope and carrier properties are known *a priori*. If the correlogram properties are not known beforehand, the FSA or Hilbert algorithms reveal the underlying envelope and phase variation.

A Fourier description of the mechanism defined by Equation (2.21) shows some similarities with the Caber method. The essential difference being that the

Caber method uses two conventional IIR filters to remove low frequency and the second harmonic components produced by the square law nonlinearity. Whereas the FSA algorithm bandpass filters the signal and then shifts it to DC (i.e. demodulates) in one operation. From the point of view of classification the new procedure can be seen as a (nonlinear) second order polynomial (Volterra series) digital filter[51, 52] followed by least squares peak prediction. The filter can be defined in general terms as a finite difference operation followed by a nonlinear difference operation (similar to Servin's method[47]) There are similarities to the quadrature receiver (see for example Whalen,[53] p200), except the sine and cosine modulation terms are derived from the signal itself instead of an external source. Yet another classification known as the bilinear (quadratic with memory) transformation[54] covers such nonlinearities and offers a tractable analysis of noise propagation. The nonlinear filter can be explicitly defined by f_b, where:

$$f_a(z) = g(z+\Delta) - g(z-\Delta) \tag{2.31}$$

is the finite difference operation, and

$$f_b(z) = f_a^{\,2}(z) - f_a(z-\Delta).\, f_a(z+\Delta) \tag{2.32}$$

is the nonlinear difference operation. Such a definition is amenable to Fourier analysis and the following relations can be demonstrated:

$$F_a(w) = 2i\sin(2\pi w\Delta).G(w) \tag{2.33}$$

$$F_b(w) = F_a(w) * F_a(w) - (\exp[-2\pi i w\Delta].F_a(w)) * (\exp[2\pi i w\Delta].F_a(w)). \tag{2.34}$$

The first equation represents bandpass filtering. The second equation represents zero and second harmonic generation (from auto-convolution) with out-of-phase terms canceling at the second harmonic.

2.8 Numerical Simulation of Algorithms

The FSA algorithm developed in the previous sections has only been compared with two other algorithms based on simple phase-shifting algorithms. In this section the FSA algorithm is compared with three algorithms which represent the best algorithms available in 1995. The simulated data are available as 512x64 pixel images of 1 byte resolution.

2.8.1 Simulated data

The parameters of the simulated data closely resemble the experimental data shown in a number of papers by Chim and Kino.[5-7, 13] Essentially the correlogram is sampled at 64 locations in depth *z*. The sampling occurs at a sample spacing of one eighth the mean wavelength in the full sampling case, and three-eighths the mean wavelength in the undersampling case. These correspond to phase steps of 90° and 270° respectively. The envelope chosen corresponds approximately to a spectral range from 400nm to 700nm. In order that sufficient data exist for useful statistical inferences to be made, there are 512 independent measurements of the correlogram.

To eliminate certain systematic errors, the correlogram shifts *z* position progressively over the full 512 range. The total shift is one sample over the full 512 range. Mathematically the image files used can be defined as

$$g(x,z) = \mathrm{INT}\left[128 + 100\exp\left(z_s^2/\sigma^2\right)\cos\left(4\pi z_s/\lambda_m\right) + n(x,z)\right] \qquad (2.35)$$

The INT() function outputs the nearest integer to the argument input. The *z* sample locations are defined by $z_s = z - 32\Delta - x/512$. The sample spacing is defined by $\Delta = \lambda_m/8$ or $\Delta = 3\lambda_m/8$ in the undersampling case. The coordinates *x* and *z* are defined at the following integer values

$$\left.\begin{aligned} x &= l\Delta_x \\ z &= m\Delta \end{aligned}\right\} \quad \left.\begin{aligned} 0 \le l \le 512 \\ 0 \le m \le 64 \end{aligned}\right\} \qquad (2.36)$$

The selected spectrum has $\sigma = 3.85\Delta$. The noise $n(x,z)$ added to the interferogram is zero-mean Gaussian distributed random noise with a standard deviation (rms) value specified in the range 0% to 8% of the modulation value. The peak modulation is set to 100. Note that even in the case of zero noise the quantisation introduces some systematic (i.e. correlated) noise. The actual noise characteristics of WLI are rather complex, being a combination of such factors as vibration, photon noise, and quantisation, to name just a few. A full analysis requires a multidimensional statistical procedure. Gaussian noise has been chosen as a simple and well-defined starting point for inter-comparisons of algorithms.

All the algorithms tested were configured to predict the *z* peak position at all 512 values of *x*. The ideal results lie upon a straight line in the *x*-*z* plane. The distribution of actual values around the best-fit (least-squares) line is computed in each case, and the standard deviation of the error is tabulated in tables 2.1 and 2.2.

Four algorithms were tested:

1) The FSA algorithm of equation (2.21) in conjunction with the specialised five-point peak detector of equation (2.24). The peak detector is applied twice if the first estimate is more than half a sample from the raw data peak. This iteration removes the systematic error visible in figure 2.7.

2) The Fourier-Hilbert method of Chim and Kino[5, 7, 13] is used to generate the envelope, and a simple three-point peak detector is used.

3) The envelope is predicted by the preceding Fourier-Hilbert method. The centroid of the envelope is then calculated. The method is equivalent to evaluation of the instantaneous phase derivative as proposed by de Groot.

4) The square of the envelope is predicted by the Fourier-Hilbert method. The centroid of the squared envelope is then calculated. The method is equivalent to evaluation of the weighted least-squares phase derivative method of de Groot.

2.8.2 Simulation Results

The results for ideal sampling and 3 time undersampling are shown in tables 2.1 and 2.2, respectively. It is interesting to note that the FSA algorithm gives the best results in the undersampling case. For the full sampling case the FSA algorithm is superior for noise levels above 2%.

The simulation only covers a small sample of the experimental parameters possible, although the sample is chosen to be representative of a typical interferometer. It is quite possible that the superiority of other algorithms may become apparent when the parameters are changed.

RMS noise (%)	FSA Algorithm	Fourier-Hilbert Algorithm	Centroid of Envelope	Centroid of Squared Envelope
0	0.010	0.003	0.000	0.000
1	0.034	0.056	0.175	0.027
2	0.064	0.112	0.322	0.061
4	0.126	0.233	0.553	0.160
8	0.248	0.479	0.856	0.476

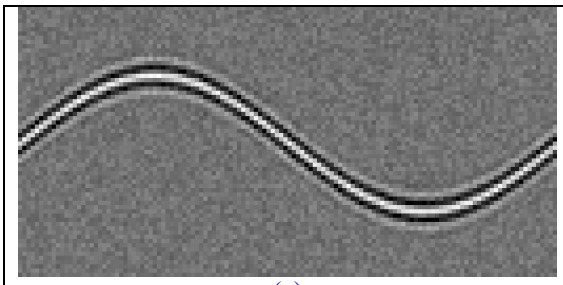
Table 2.1
RMS peak location error for various algorithms for optimal sampling. The error is in units of one sample spacing.

RMS noise (%)	FSA Algorithm	Fourier-Hilbert Algorithm	Centroid of Envelope	Centroid of Squared Envelope
0	0.055	0.166	0.305	0.170
1	0.058	0.166	0.439	0.175
2	0.065	0.168	0.625	0.204
4	0.094	0.176	0.909	0.396
8	0.172	0.206	1.190	1.046

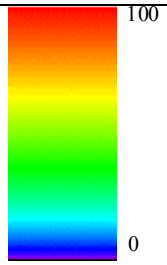
Table 2.2
RMS peak location error for various algorithms for 3 times undersampling. The error is in units of one sample spacing.

2.8.3 Visual Performance of Envelope Demodulation Algorithms

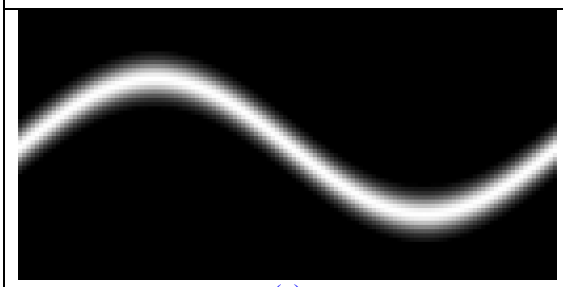
Figure 2.8 shows an *x*-*z* image of a simulated noisy correlogram (8% rms noise). The underlying profile in this case is sinusoidal. The Hilbert algorithm is an all-pass filter and allows more noise through than the bandpass FSA algorithm. This effect is clear in the differences between figures 2.8 (e) and (g). Figures 2.8 (f) and (h) use a pseudocolour scale to help emphasise the difference in noise levels between the envelope detectors.



(a)
Noisy correlogram

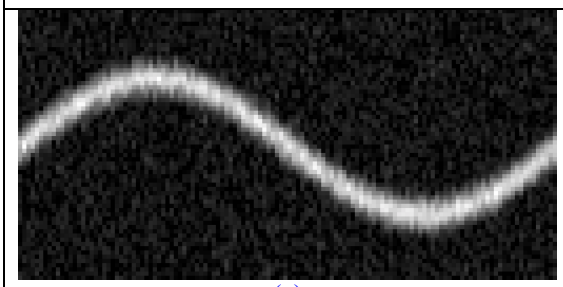


(b)
Pseudocolour scale

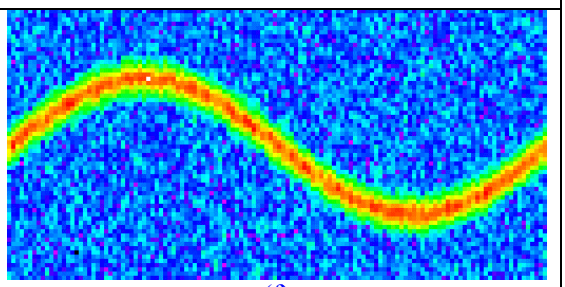


(c)
Ideal (noise free) correlogram envelope.
Linear greyscale display.

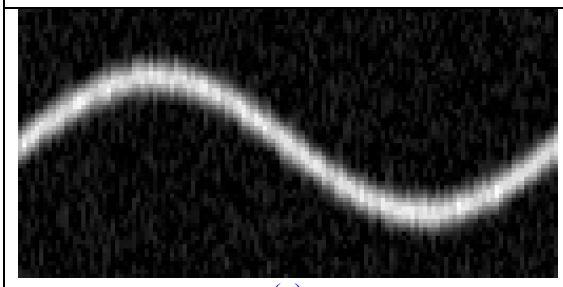
(d)
Ideal (noise free) correlogram envelope.
Pseudocolour display.



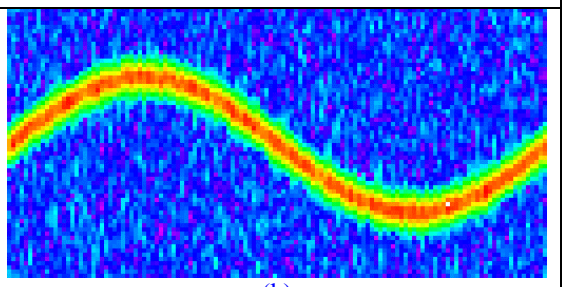
(e)
Hilbert transform envelope of noisy correlogram.
Linear greyscale display.



(f)
Hilbert transform envelope of noisy correlogram.
Pseudocolour display.



(g)
FSA algorithm envelope of noisy correlogram.
Linear greyscale display.



(h)
FSA algorithm envelope of noisy correlogram.
Pseudocolour display.

Figure 2.8
Visual comparison of Hilbert demodulation versus FSA envelope detection. In each case the scanning direction *z* is from bottom to top; the *x* direction is from left to right.

2.9 Conclusion

A simple but highly effective method for envelope detection in white light correlograms has been introduced and demonstrated. The speed due to increased computational efficiency is between two and three times that of the real space Hilbert transform technique. Combined with a new procedure for peak prediction (using a weighted LSF algorithm which removes the residual second harmonic error in the envelope) the fully compensating five-sample envelope detection algorithm is remarkably simple yet effective over a wide range of carrier frequencies. In terms of multiplication operations the new algorithm has been shown to be near the ideal limit of two multiplications per sample suggesting that any further speed improvements from other methods can only be marginal. In situations where the bandwidth is small enough the proposed algorithm can be combined with sub-Nyquist sampling to further improve efficacy.

The method is not limited to white light interferometry and is applicable to any bandpass signals where either the envelope or the phase, or both, need to be detected. Optical measurement techniques such as confocal interferometry and spatial carrier phase-shifting interferometry could benefit from such a method.

The inherent accuracy limit, previously believed to be $\Delta/20$ has been shown to be only true for a severely detuned algorithm. The underlying (noise free) limit for a properly tuned algorithm is actually of the order $\Delta/10000$, which means that the algorithm should not be discounted from use in high-resolution systems.

2.9.1 Limitations of this chapter

Although the FSA algorithm has been shown to combine both computational efficiency and accuracy – properties that are often considered to be mutually exclusive – a theoretical lower limit for the peak position error has not been established. A number of nonlinearities in the estimation procedures have prevented the use of simple linear estimation theory on the problem. If we consider the case where the expected interferogram shape (including the width) is known, then we can use maximum likelihood estimation theory to find the best estimate of peak position. In reality the exact shape (and width) is not known *a priori* and such methods are inapplicable.

The analysis of the interferograms has followed convention by ignoring some of the real diffraction effects expected from a surface illuminated by a white light source. In reality the diffracted field will depend to some extent upon the gradient and curvature (and higher derivatives) of the surface.

2.10 Connections

This work presented in this chapter was originally submitted as a manuscript to The Journal of the Optical Society of America, A, and accepted in 1995. The demodulation algorithm I had accidentally stumbled across in my search for self-calibrating algorithms concealed a wealth of connections in a number of different disciplines. A subsequent review of the literature revealed that the algorithm has arisen quite independently in a number of guises over the last half century. My findings are presented briefly in the following subsections.

2.10.1 The analytic signal, the envelope, and the instantaneous frequency

The definition of the analytic signal (as defined by Gabor in 1947[23]) seems to lead directly to the notion of instantaneous frequency (or equivalently the phase derivative). In practice the envelope and phase cannot always be separated unambiguously. There has been some controversy about "instantaneous frequency" (IF) since 1952 when Shekel called the term "erroneous".[55] In 1972 Mandel[56] clarified the term using spectral moments (following work by Ville[57]). More recently the debate has re-emerged because of the centrality of the concept in areas such as time-frequency analysis, AM-FM demodulation, and the fractional Fourier transform. In the work of Loughlin[58] four conditions necessary for the IF to be unambiguously defined are presented. In my work on discrete envelope detection algorithms I have derived specific formula for both the instantaneous phase ϕ, in equation (2.18), and the instantaneous spatial frequency $\psi/(2\pi\Delta)$, in equation (2.27).

$$\left.\begin{aligned} f &= b(x)\cos\left[\varphi(x)\right] \\ \hat{f} &= -b(x)\sin\left[\varphi(x)\right] \\ f_a &= f - i\hat{f} = b(x)\exp\left[i\varphi(x)\right] \end{aligned}\right\}. \tag{2.37}$$

The modulus of the analytic signal is then

$$\left|f_a\right|^2 = f^2 + \hat{f}^2 = \left|b\right|^2. \tag{2.38}$$

2.10.2 The Energy Operator

Closely related to the concept of instantaneous frequency is the idea of the so-called energy operator. The idea of the energy operator in signal analysis seems to have been first suggested by Teager.[59] The idea is that the value of a harmonic signal can be compared to the position of a simple harmonic oscillator (SHO). The analogy then allows energy to be associated with the signal and the total energy is the sum of the kinetic and potential energy of the SHO. The concept was formalised by Kaiser in 1990[60] and subsequently popularised in a number of publications by Maragos[61], Bovik[62] and Quatieri[63], to name but a small selection. The derivation is for an AM signal $f(x)$ is as follows:

$$\left.\begin{aligned} f &= b(x)\cos[2\pi u x] \\ f' &= \frac{\partial f(x)}{\partial x} \approx -2\pi u b \sin[2\pi u x] \\ f'' &= \frac{\partial^2 f(x)}{\partial x^2} \approx -(2\pi u)^2 b \cos[2\pi u x] \end{aligned}\right\}. \tag{2.39}$$

The energy is given by an equation analogous to that for the modulus of the analytic signal

$$\Psi\{f\} = (f')^2 - ff'' \approx (2\pi u b)^2. \tag{2.40}$$

The WLI envelope detector derived in equation (2.21) is, in fact, the discrete form of the energy operator defined in equation (2.40). Most published works on discrete implementations of the energy operator start from equation (2.40) and use discrete derivative approximations. One advantage of starting from equation (2.21) is that the

aliasing properties of the algorithm emerge naturally,[64, 65] whereas research using equation (2.40) fails to reveal the sub-Nyquist effects.

It should be noted that the energy operator is a point operator because each of the component derivatives is evaluated at a point.

2.10.3 Phase Congruency

The local energy model suggests that the features of an image occur in regions where the Fourier phase components are in synch (or, rather, maximally congruent). Phase congruency[66-68] corresponds directly with the modulus of the analytic signal as presented in 2.10.1 above:

$$\left| f_a \right|^2 = f^2 + \hat{f}^2 = \left| b \right|^2 . \tag{2.38}$$

The concept of phase congruency and local energy[69] seems so closely related to the energy operator and the Hilbert transform[70] that it is surprising to find that each of the two main groups research and publish apparently oblivious to each other. I shall revisit some of the statements from both groups with respect to the Hilbert transform in 2-D in Chapter 4.

Ironically the local energy is a non-local operator (unlike the energy operator) because the quadrature (Hilbert transform) component depends on the signal in a non-local manner. Another way of saying this is that the Hilbert transform (of a signal) at a certain point depends upon the value the signal at all points.

2.10.4 Accelerated Convergence of Series

It turns out that both the energy operator and the efficient envelope detector I have developed are quite similar to an algorithm developed some time ago for accelerating the convergence of series. In 1955 Daniel Shanks[71] proved some results in the theory of transformations which accelerate converge of sequences. He gives the classic example of Liebnitz series for π (based on the arctangent series expansion):

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots \tag{2.41}$$

According to Knopp[72] the series is "beautiful, but for numerical purposes – practically valueless" Summing the first ten terms gives an estimate correct to just one figure. Using a transformation that utilizes the partial sums of a sequence

$$S_n = \sum_{m=0}^{n} a_m \tag{2.42}$$

where the transform $\mathrm{T}\{\}$ is defined

$$\mathrm{T}\{S_n\} = \frac{S_{n+1}S_{n-1} - S_n^2}{S_{n+1} - 2S_n + S_{n-1}} \tag{2.43}$$

the convergence of the Liebnitz sequence is so improved that just using the first 4 terms an estimate accurate to 3 figures (2 decimal places) is obtained. Equation (2.43) can be immediately recognised as a nonlinear phase-shifting algorithm. The

numerator is, in fact, the discrete energy operator (if we associate *S* with a finite difference operator).

The theory of such transformations was extended to iteratively applied transforms in the wonderfully entitled "On a Procrustean technique for the numerical transformation of slowly convergent sequences and series" work of Wynn.[73] One way to view the connections here is to notice that if we plot the sequences of Wynn and Shanks we see an oscillating structure contained within an envelope, and having a slowly varying background. The structure is isomorphic with our sampled interferograms and we can expect similar methods to extract the underlying parameters. The conclusion is that nonlinear PSAs date back to 1955 or earlier!

2.10.5 The Ambiguity Function

Recently Hamila et al[74] showed that the energy operator is equal to the Fourier transform of the second derivative of the ambiguity function (AF). The relationship corresponds closely with the second moments of the Wigner distribution function[21] and the concept of instantaneous frequency defined by spectral moments (see section 2.9.1 above). The least squares fit to the phase derivative proposed by de Groot[75] corresponds to the envelope centroid, whereas a weighted LSF corresponds to the centroid of the squared envelope, or, equivalently a first moment of the Wigner distribution function (WDF). Chen et al[76] considered several centroid based techniques including both fringe and envelope centroids. The conclusion is that phase-space (AF and WDF) methods may be useful in white light interferogram analysis, even if the computation (related to the higher dimensionality of phase-space) is substantial.

2.10.6 Error Correcting Phase-Shifting Algorithms

Following the publication of this work in 1996, a number of new error correcting algorithms have been derived.[77-82] The design procedures for linear algorithms involve repeated averaging,[83] the analysis of *z*-transform zeros,[81] and the solution of linear equations.[79] The methods are essentially equivalent and give rise to similar algorithms once certain symmetry constraints are enforced.[80, 82]

At present the only well-known nonlinear phase-shifting algorithms are the four-sample Carré algorithm and the five-sample Servin algorithm (which is equivalent to the nonlinear algorithm proposed in this chapter, equation (2.18)).

2.11 Acknowledgements

John Quartel kindly provided a suit of C programs for Fourier transforming datasets used for the simulation in section 2.8. The algorithms were then implemented and evaluated by the author.

2.12 References and Notes

1 P. A. Flourney, R. W. McClure, and G. Wyntjes, "White-light interferometric thickness gauge", Appl. Opt. **11**, (9), 1907-1915, (1972).

2 T. Dresel, G. Häusler, and H. Venzke, "Three dimensional sensing of rough surfaces by coherence radar", App. Opt. **31**, (7), 919-925, (1992).

3 M. Davidson, K. Kaufman, I. Mazor, and F. Cohen, "An application of interference microscopy to integrated circuit inspection and metrology," Integrated circuit metrology, inspection, and process control, Proc. SPIE **775**, (1987), 233-247.

4 B. S. Lee, and T. C. Strand, "Profilometry with a coherence scanning microscope", App. Opt. **29**, (26), 3784-3788, (1990).

5 S. S. C. Chim, and G. S. Kino, "Phase measurements using the Mirau correlation microscope", App. Opt. **30**, (16), 2197-2201, (1991).

6 S. S. C. Chim, and G. S. Kino, "Three-dimensional image realization in interference microscopy", App. Opt. **31**, (14), 2550-2553, (1992).

7 G. S. Kino, and S. S. C. Chim, "Mirau correlation microscope", App. Opt. **29**, (26), 3775-3783, (1990).

8 M. Davidson, "Method and apparatus for using a two beam interference microscope for inspection of integrated circuits and the like", US patent # 4818110, 1989.

9 B. L. Danielson, and C. Y. Boisrobert, "Absolute optical ranging using low coherence interferometry", App. Opt. **30**, (21), 2975-2979, (1991).

10 K. Creath, "Calibration of numerical aperture effects in interferometric microscope objectives.", App. Opt. **28**, (15), 3333-3338, (1989).

11 G. Schulz, and K.-E. Elssner, "Errors in phase-measurement interferometry with high numerical apertures", App. Opt. **30**, (31), 4500-4506, (1991).

12 C. J. R. Sheppard, and K. G. Larkin, "Effect of numerical aperture on interference fringe spacing", App. Opt. **34**, (22), 4731-4734, (1995).

13 S. S. C. Chim, and G. S. Kino, "Correlation microscope", Opt. Lett. **15**, (10), 579-581, (1990).

14 J. Schwider, R. Burow, K.-E. Elssner, J. Grzana, et al., "Digital wavefront measuring interferometry: some systematic error sources", App. Opt. **22**, 3421-3432, (1983).

15 P. Hariharan, B. F. Oreb, and T. Eiju, "Digital phase-shifting interferometer: a simple error-compensating phase calculation algorithm", App. Opt. **26**, (13), 2504-2506, (1987).

16 K. G. Larkin, "Efficient nonlinear algorithm for envelope detection in white light interferometry", J. Opt. Soc. Am., A **13**, (4), 832-843, (1996).

17 The word *neat* is considered slang by North Americans, although its dictionary definition is quite apposite in this instance:

"**Neat**…4. Cleverly effective in character or execution: a neat scheme…"

The Macquarie Dictionary, Macquarie University, Sydney, 1991.

18 J. F. Biegen, "Determination of the phase change on reflection from two-beam interference", App. Opt. **19**, (21), 1690-1692, (1994).

19 P. Hariharan, K. G. Larkin, and M. Roy, "The geometric phase: interferometric observations with white light", J. Mod. Opt. **41**, (4), 663-667, (1994).

20 S. C. Pohlig, "Signal duration and the Fourier transform", Proceeding of the IEEE **68**, (5), 629-630, (1980).

21 C. J. R. Sheppard, and K. G. Larkin, "Focal shift, optical transfer function, and phase space representations", J. Opt. Soc. Am., A 17, (4), 772-779, (2000).

22 M. Takeda, H. Ina, and S. Kobayashi, "Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry", J. Opt. Soc. Am. **72**, (1), 156-160, (1982).

23 D. Gabor, "Theory of communications", Journal of the Institution of Electrical Engineers, 93, 429-457, (1947).

24 L. R. Rabiner, and B. Gold, Theory and application of digital signal processing, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1975.

25 D. A. Zweig, and R. E. Hufnagel, "A Hilbert transform algorithm for fringe-pattern analysis," Advanced Optical Manufacturing and Testing, San Diego, (1990), 295-302.

26 K. Freischlad, and C. L. Koliopoulos, "Fourier description of digital phase-measuring interferometry", J. Opt. Soc. Am. A **7**, (4), 542-551, (1990).

27 K. G. Larkin, and B. F. Oreb, "Design and assessment of Symmetrical Phase-Shifting Algorithms", J. Opt. Soc. Am., A **9**, (10), 1740-1748, (1992).

28 R. E. Bogner, and A. G. Constantinides, Introduction to digital filtering, John Wiley & Sons, 1975.

29 O. Brigham, The fast Fourier transform, Second edition, Prentice Hall, New Jersey, 1988.

30 J. D. Gaskill, Linear systems, Fourier transforms, and Optics, John Wiley & Sons, New York, 1978.

31 P. J. Caber, "Interferometric profiler for rough surfaces", Applied Optics **32**, (19), 3438-3441, (1993).

32 P. de Groot, and L. Deck, "Three-dimensional imaging by sub-Nyquist sampling of white-light interferograms", Opt. Lett. **18**, (17), 1462-1464, (1993).

33 ZYGO, "New View 100. 3D Imaging Surface Structure Analyzer", Zygo Corporation, (1993).

34 F. G. Stremler, Introduction to communication systems, Addison-Wesley, 1982.

35 P. Carre, "Installation et utilisation du comparateur photoelectrique et interferentiel du Bureau International des Poids et Mesures.", Metrologia **2**, (1), 13-23, (1966).

36 K. H. Womack, "Interferometric phase measurement using spatial synchronous detection", Opt. Eng. **23**, (4), 391-395, (1984).

37 M. Kujawinska, "Spatial phase measurement methods", in Interferogram analysis: digital fringe pattern measurement techniques, ed. Robinson, D. W., and Reid, G. T. (Bristol: Institute of Physics, 1993).

38 The exact compensating 5-sample algorithm can be shown to be the solution of the linear modulation-demodulation problem $a+(b+cx)\cos(2\pi ux+\alpha)$ using 5 measurements to solve for 5 unknowns, whereas the other algorithms assume $c=0$ and the NEC algorithm assumes $\alpha=90^{\circ}$; the PEC algorithm assumes $\alpha\sim 90^{\circ}$.

39 B. Bushan, J. C. Wyant, and C. L. Koliopolous, "Measurement of surface topography of magnetic tapes by Mirau interferometry", App. Opt. **24**, (10), 1489-1497, (1985).

40 K. Creath, "Phase-shifting holographic interferometry", in Holographic interferometry: principles and methods, ed. Rastogi, P. K. (Berlin: Springer-Verlag, 1994).

41 The nonlinear filter defined by Equation (2.21) can be represented by a bandpass filter followed by a quadratic Volterra series filter.

42 P. Sandoz, and G. Tribillon, "Profilometry by zero-order interference fringe identification", J. Mod. Opt. **40**, (9), 1691-1700, (1993).

43 D. K. Cohen, P. J. Caber, and C. P. Brophy, "Rough surface profiler and method",
US patent # 5133601, (1992).

44 Closely related to Savitsky-Golay, or Digital Smoothing Polynomial Filters mentioned in the next reference.

45 W. H. Press, S. A. Teulolsky, W. T. Vetterling, and B. P. Flannery, Numerical Recipes in Fortran:The Art of Scientific Computing, Second, Cambridge University Press, Cambridge, 1992.

46 Typical peak detection algorithms involve a combination of samples in both numerator and denominator of a quotient resembling equation (2.24). Linear combinations are equivalent to correlation or convolution with a kernel function. In the Fourier domain the transformed signal is multiplied by the kernel transform. If the kernel is suitably chosen, yet still satisfies the LSF criteria, then zeros (of the transform) can occur at certain frequencies, and these frequencies are thus removed from the signal. In essence peak detection and spectral filtering have been combined into one operation. The LSF quotient is similar to the PSA quotient, with corresponding Fourier properties.

47 M. Servin, and F. J. Cuevas, "A novel algorithm for spatial phase-shifting interferometry", J. Mod. Opt. **42**, (9), 1853-1862, (1995).

48 N. Bareket, "Undersampling errors in measuring the moments of images aberrated by turbulence", App. Opt. **18**, (17), 3064-3069, (1979).

49 R. P. Loce, and R. E. Jodoin, "Sampling theorem for geometric moment determination and its application to a laser beam position detector", App. Opt. **29**, (26), 3835-3843, (1990).

50 B. F. Alexander, and K. C. Ng, "Elimination of systematic error in subpixel accuracy centroid estimation", Opt. Eng. **30**, (9), 1320-1331, (1991).

51 B. Picinbono, "Quadratic filters," Proc IEEE Int. Conf. Acoust. Speech, Signal Processing, (1982), 298-295.

52 I. Pitas, and A. N. Venetsanopoulos, Nonlinear Digital Filters: Principles and Applications., Kluwer Academic Publishers, Boston, 1990.

53 A. D. Whalen, Detection of signals in noise, Academic Press, New York, 1971.

54 B. E. A. Saleh, "Optical bilinear transformations: general properties", Optica Acta **26**, (6), (1979).

55 J. Shekel, "Instantaneous frequency", Proc. IRE **41**, 548, (1953).

56 L. Mandel, "Interpretation of instantaneous frequency", Am. J. Phys. **42**, 840-846, (1974).

57 J. Ville, "Theorie de applications de la notion de signal analytique", Cables et transmissions **2A**, (1), 61-74, (1948).

58 P. J. Loughlin, and B. Tacer, "On the Amplitude- and Frequency-Modulation Decomposition of Signals", Journal of the Acoustical Society of America **100**, (3), 1594-1601, (1996).

59 H. M. Teager, and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract", Speech production and speech modelling, ed. Hardcastle, W. J., and Marchal, A. (France: NATO Advanced Study Institute, Series D, 1989) p55.

60 J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," Proc IEEE Int. Conf. Acoust. Speech, Signal Processing, Albuquerque, NM, (1990), 381-384.

61 P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators", IEEE Trans. Signal Process. **41**, (4), 1532-1550, (1993).

62 P. Maragos, and A. C. Bovik, "Image Demodulation Using Multidimensional Energy Separation", J. Opt. Soc. Am., A **12**, (9), 1867-1876, (1995).

63 A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators", IEEE Trans. Signal Process. **41**, (12), 3245-3265, (1993).

64 G. Deng, " A study of sampling frequency for efficient demodulation of AM signals," Proceedings of 1997 IEEE International Symposium on ISCAS, (1997), 2677 -2680.

65 K. G. Larkin, "Efficient Demodulator for Bandpass Sampled AM Signals", Electronics Letters **32**, (2), 101-102, (1996).

66 M. C. Morrone, and D. C. Burr, "Feature detection in human vision: a phase-dependent energy model", Proceedings of the Royal Society of London, B **235**, 221-245, (1988).

67 C. Pudney, and M. Robbins, "Surface extraction from 3D images via local energy and ridge tracing," DICTA, (1995), 240-245.

68 Y. K. Aw, R. Owens, and J. Ross, "An analysis of local energy and phase congruency models in visual feature detection", The Journal of the Australian Mathematical Society, Series B **40**, 97-122, (1998).

69 S. Venkatesh, "A study of energy based models for the detection and classification of image features", The University of Western Australia, 1990.

70 A. Potamianos, and P. Maragos, "A Comparison of the Energy Operator and the Hilbert Transform Approach to Signal and Speech Demodulation", Sig. Proc. **37**, (1), 95-120, (1994).

71 D. Shanks, "Non-linear transformations of divergent and slowly convergent sequences", J. Math. and Phys. **34**, 1-42, (1955).

72 K. Knopp, Theory and application of Infinite Series, Blackie & Son, London, 1928.

73 P. Wynn, "On a Procrustean technique for the numerical transformation of slowly convergent sequences and series", Proc. Camb. Phil. Soc. **52**, 663-671, (1956).

74 R. Hamila, J. Astola, M. A. Cheikh, M. Gabbouj, et al., "Teager energy and the ambiguity function", IEEE Trans. Signal Process. **47**, 260-262, (1999).

75 P. de Groot, and L. Deck, "Surface profiling by analysis of white-light inteferograms in the spatial frequency domain", J. Mod. Opt. **42**, (2), 389-400, (1995).

76 S. Chen, A. W. Palmer, K. T. V. Grattan, and B.T.Meggitt, "Digital signal-processing techniques for electronically scanned optical-fiber white-light interferometry", App. Opt. **31**, (28), 6003-6010, (1992).

77 K. Hibino, B. F. Oreb, D. I. Farrant, and K. G. Larkin, "Phase-shifting interferometry for non-sinusoidal waveforms with phase-shift errors", J. Opt. Soc. Am., A **12**, (4), 761-768, (1995).

78 K. Hibino, B. F. Oreb, D. I. Farrant, and K. G. Larkin, "Phase-shifting algorithms for nonsinusoidal signals compensating nonlinear phase-shift errors and their susceptibility to random noise.," ICO, Taejon, Korea, (1996).

79 K. Hibino, B. F. Oreb, D. I. Farrant, and K. G. Larkin, "Phase-shifting algorithms for nonlinear and spatially nonuniform phase shifts", J. Opt. Soc. Am., A **14**, (4), 918-930, (1997).

80 K. Hibino, K. G. Larkin, B. F. Oreb, and D. I. Farrant, "Phase-shifting algorithms for nonlinear and spatially nonuniform phase shifts: reply to comment", J. Opt. Soc. Am., A **15**, 1234-1235, (1998).

81 Y. Surrel, "Design of algorithms for phase measurements by the use of phase stepping", App. Opt. **35**, (1), 51-60, (1996).

82 Y. Surrel, "Phase-shifting algorithms for nonlinear and spatially nonuniform phase shifts: comment", J. Opt. Soc. Am., A **15**, 1227-1233, (1997).

83 J. Schmit, and K. Creath, "Extended averaging technique for derivation of error-compensating algorithms in phase-shifting interferometry.", App. Opt. **34**, (19), 3610-3619, (1995).