INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700  800/521-0600
The use of digital signal processing techniques for the interferometric profiling of rough surfaces

Caber, Paul James, M.S.

The University of Arizona, 1991
THE USE OF DIGITAL SIGNAL PROCESSING TECHNIQUES
FOR THE INTERFEROMETRIC PROFILING OF ROUGH SURFACES

by

Paul James Caber

A Thesis Submitted to the Faculty of the
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
In Partial Fulfillment of the Requirements
For the Degree of
MASTER OF SCIENCE
WITH A MAJOR IN ELECTRICAL ENGINEERING

In the Graduate College
THE UNIVERSITY OF ARIZONA

1991
STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: [Signature]

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

[Signature] [Date]

Bobby R. Hunt
Professor of Electrical and Computer Engineering
ACKNOWLEDGEMENTS

I would like to thank all of the people that contributed to this research project and to my education. Sincere thanks to my graduate advisor, Bobby Hunt, for his intellectual support throughout my graduate studies, and also for his contributions to the theory behind the technology being presented here. I would also like to express my gratitude to Jack Gaskill for serving on my thesis committee. My sincere appreciation to James Wyant for serving on my committee, and also for his understanding, patience, and commitment to the success of this project.

Thanks to WYKO Corporation for funding this project and providing a work environment that encouraged me to pursue my graduate degree. Special thanks to all the members of the TOPO team, especially Donald Cohen, without whose continued enthusiastic support I could never have accomplished so much. I would also like to thank Denis Fournier for his work in developing the software simulation.

My deepest gratitude is saved for my wife Jennifer, who provided me with love and understanding during the course of my work. Without her my life and my career would be without purpose.
TABLE OF CONTENTS

LIST OF FIGURES .................................................. 5
LIST OF TABLES .................................................... 7
ABSTRACT .............................................................. 8

1. INTRODUCTION .................................................... 9
   1.1 Motivation for Optical Surface Roughness Measurement .............. 9
   1.2 Overview of Topics ........................................... 12

2. REVIEW OF OPTICAL SURFACE ROUGHNESS MEASUREMENT ... 14
   2.1 Phase-Shifting Interferometry .................................. 14
   2.2 Optical Probes ............................................... 18
   2.3 Light Scattering Techniques ................................... 20

3. A NEW METHOD FOR PROFILING ROUGH SURFACES ............ 22
   3.1 Basic Theory Behind the Technique ............................. 22
   3.2 Initial Implementations ....................................... 28

4. SIGNAL PROCESSING TECHNIQUES ................................ 35
   4.1 Envelope Demodulation ........................................ 35
   4.2 Digital Filtering Techniques ................................... 44
   4.3 Peak Detection of Envelope ................................... 49

5. IMPLEMENTATION OF THE ALGORITHM ........................... 53
   5.1 Digital Filter Design .......................................... 53
   5.2 Error Analysis ................................................ 67

6. RESULTS ............................................................ 76
   6.1 Physical Implementation ....................................... 76
   6.2 Surface Measurements .......................................... 77
   6.3 Summary and Conclusion ...................................... 85

REFERENCES .......................................................... 86
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Phase shifting interferometer</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>Fringe signal produced by broadband illumination</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>Rectangular source spectrum</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>Irradiance signal for rectangular spectrum</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>Blackbody curve</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>Initial implementation test setup</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>Measurement of roughness comparator (125 μin RA)</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>(a) TOPO-3D measurement of optical flat</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>(b) Optical flat measured using modulation algorithm</td>
<td>34</td>
</tr>
<tr>
<td>9</td>
<td>Power spectrum of amplitude modulated signal</td>
<td>37</td>
</tr>
<tr>
<td>10</td>
<td>Amplitude demodulation by absolute value detection</td>
<td>39</td>
</tr>
<tr>
<td>11</td>
<td>Simple AM demodulator circuit</td>
<td>40</td>
</tr>
<tr>
<td>12</td>
<td>Amplitude demodulation by square-law detection</td>
<td>42</td>
</tr>
<tr>
<td>13</td>
<td>Algorithm block diagram</td>
<td>44</td>
</tr>
<tr>
<td>14</td>
<td>Peak detection of coherence function</td>
<td>51</td>
</tr>
<tr>
<td>15</td>
<td>Test setup for digital filtering algorithm development</td>
<td>54</td>
</tr>
<tr>
<td>16</td>
<td>Vertical distance and frequency domain plots of irradiance signal</td>
<td>55</td>
</tr>
<tr>
<td>17</td>
<td>Frequency plot after square-law detection</td>
<td>56</td>
</tr>
<tr>
<td>18</td>
<td>Irradiance signal using 200X objective</td>
<td>58</td>
</tr>
<tr>
<td>19</td>
<td>Frequency plot of highpass filter response</td>
<td>59</td>
</tr>
<tr>
<td>20</td>
<td>Phase plot for highpass filter</td>
<td>60</td>
</tr>
</tbody>
</table>
Figure 21 Pole-zero plot for highpass filter .................................................. 61
Figure 22 Irradiance signal after highpass filtering ........................................... 62
Figure 23 Frequency, phase, and pole-zero plots of fourth-order IIR lowpass filter ................................................................. 63
Figure 24 Signal after filtering with fourth-order lowpass filter ......................... 64
Figure 25 Frequency, phase, and pole-zero plots of second-order IIR lowpass filter ................................................................. 65
Figure 26 Signal after filtering with second-order lowpass filter ......................... 66
Figure 27 Histogram of simulation output for an input SNR of 30 dB ................. 72
Figure 28 Simulated RMS measurement error versus input noise
    (a) measurement error versus input noise standard deviation .............. 73
    (b) measurement error versus input SNR ............................................. 73
Figure 29 Simulated measurement error versus optical bandwidth for input SNR
    of 40 dB ................................................................. 74
Figure 30 Measurement of optical flat (fringes nulled) .................................. 78
Figure 31 Measurement of optical flat (4 fringes) ....................................... 79
Figure 32 Effects of quantization on measurement error ................................. 80
Figure 33 Measurement of section of roughness comparator with average
    roughness of 8 \( \mu \)in RA .................................................. 83
Figure 34 Measurement of section of roughness comparator with average
    roughness of 32 \( \mu \)in RA .................................................. 83
Figure 35 Measurement of section of roughness comparator with average
    roughness of 125 \( \mu \)in RA .................................................. 84
Figure 36 Measurement of backside of silicon wafer after acid etching process .... 84
Figure 37 Measurement of 43.5 nm step showing overshoot on edges ............... 85
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81</td>
</tr>
</tbody>
</table>

Comparison of measurements of optical flat
Conventional non-contact optical methods of surface profiling are limited in the range of surface heights that can be accurately measured due to phase ambiguity errors on steep local slopes. Instruments that have been developed thus far to avoid the problems with local slope typically suffer from poor measurement height resolution and slow measurement speeds. Contact profilometers such as stylus-based instruments suffer from poor lateral resolution due to the finite radius of the stylus tip, and slow measurement speeds, especially when two-dimensional scans of the surface are required. Stylus tips can also scratch delicate surfaces during the measurement. We propose a new method of optical, non-contact profiling of rough surfaces without the limitations on local slope that other methods suffer from. This new method utilizes interferometric techniques as well as digital signal processing algorithms to produce fast, accurate, and repeatable three-dimensional surface profile measurements on a wide variety of surfaces.
1. INTRODUCTION

1.1 Motivation for Optical Surface Roughness Measurement

Various methods of profiling the surface features of objects have evolved throughout the years. The simplest methods of determining surface roughness characteristics involve tactile comparators that machinists use to evaluate the comparative roughness of specimens on the shop floor. These devices typically are made of strips of electroplated nickel, with various machining operations represented in separate areas of the device, including grinding, polishing, and lapping operations. The operator simply drags a fingernail across the test surface and the gauge, and a relatively accurate assessment of roughness can be made [1].

Contact methods of profilometry typically involve the use of stylus-based instruments to obtain a two-dimensional trace of the surface profile as a sharp probe tip is dragged across the surface. More advanced instruments generate a three-dimensional surface map by scanning the surface being measured in a raster or radial pattern. While the height resolution may be quite good, the lateral resolution is limited by the size of the probe tip, which is typically on the order of 1 micron in diameter. The result of measuring a surface with higher spatial frequencies than a stylus can accurately detect is to effectively lowpass filter the data. Modern instruments typically digitize and store the data in a digital computer that can then perform further filtering, analysis, and display.
Stylus profilometers also require physical contact of the stylus tip with the surface, which makes them impractical for measuring very delicate surfaces, or materials that deform easily. The setup and operation of these instruments also requires a good deal of manual intervention, which makes the measurements difficult to automate. Depending on the setup time required, a typical surface profile can take 10 minutes or longer, making them unsuited for most on-line applications [2].

Optical surface profiling has become popular recently with the introduction of high-speed digital processing equipment. The benefits of optical surface profilometry include the ability to perform non-contact measurements of delicate surfaces, increased spatial resolution over contact probe techniques, and increased measurement speed. Various methods of optical surface profiling have evolved recently, including interferometry, measurement of local slopes using laser techniques, and detection of focus. Optical phase-shifting interferometers can be used to obtain fast, three-dimensional profiles of very smooth surfaces, with accuracies on the order of 1/100 of a wavelength of light. These instruments are typically limited to measurements of smooth, polished, homogeneous surfaces. When measurements of rougher surfaces, or surfaces with dissimilar optical properties are measured, severe errors in measurement arise.

Optical profilers that rely on local slope, or focus, to determine surface profile must scan the surface being measured and integrate the data to obtain a complete profile. These methods suffer from errors when local slopes are too steep to obtain reflection of the light being measured, and are also limited in spatial resolution by the spot size
of the light on the surface. The speed of these instruments is generally much slower than the speed of phase-shifting methods, and while they can typically measure much rougher surfaces, the vertical resolution is not as good.

Clearly, there exists the need for an instrument that has the benefits of optical profiling (non-contact, fast, accurate), but does not suffer from the limitations of local slope and surface property variations. This paper describes a technique that provides fast, three-dimensional, non-contact measurements of rough surfaces, while preserving good lateral as well as vertical resolution.
1.2 Overview of Topics

A method will be described herein for profiling a surface using a combination of optical interferometry and digital signal processing technologies. The emphasis of this presentation will be put on the signal processing theory as it relates to the algorithms required to process the optical data. The optical theory behind the method will not be investigated in detail, and therefore some assumptions may need to be made, and some details may remain unproven. However, the basic optical theory behind the method will be shown to be valid, and experimental results will be shown to support these assumptions.

In Section 2, classical methods of optical surface profiling will be reviewed, including optical probes, and phase-shifting interferometers. Light scattering systems will also be described, although they are not used to measure surface profile directly. Theory and formulas behind these methods will be discussed, as well as the limitations of each method in profiling rough surfaces. Recent advances and techniques will also be investigated.

Section 3 will discuss the history behind the method being presented. Early attempts to produce an accurate surface profile using this technique will be reviewed. The optical theory behind this technique will also be discussed.

Section 4 will describe the method and algorithms required for interferometric profiling using digital signal processing. Basic background theory in digital signal
processing will be discussed, as well as its relationship to optical profiling. In
Section 5, a detailed description of the digital signal processing algorithm used to
produce the measurement will be presented. A section on error analysis and
prediction will also be given.

Finally, experimental results obtained with this method will be demonstrated in
Section 6, and these results will be shown to support the theory. The results given
will prove that it is indeed possible to produce fast, accurate, non-contact, three-
dimensional measurements of a wide variety of surfaces using a combination of
optical and digital signal processing techniques, without the limitations of earlier
methods. The method will be shown to be well suited for implementation in fast,
digital hardware, and robust enough to be incorporated into an instrument capable of
taking measurements in a wide variety of process control applications.
2. REVIEW OF OPTICAL SURFACE ROUGHNESS MEASUREMENT

2.1 Phase-Shifting Interferometry

Of the many methods of profiling surfaces using optical techniques, phase-shifting interferometry (PSI) has by far become the most popular in recent years. Although the analysis of surface features using interferograms has been done for some time, only recently has the process of digitizing and analyzing the interference fringe data using a digital computer been made possible. In the 1960’s, optical comparators were used to measure distance between interference fringe centers, and then analyzed on a computer. Later, video systems were incorporated to detect the fringe centers, speeding up the technique. These systems were limited to accuracies of no better than 1/10 fringe [3].

Recently, with the development of solid-state detector arrays and high-speed digital computers, a technique has been developed that digitizes the fringe wavefront while a known phase shift is introduced between the sample and the reference beams of the interferometer. Typically multiple (three to five on the average) frames of data are taken during the course of a measurement and the data is then analyzed on the computer in order to obtain the phase at each point on the detector, which then can be related to relative surface height on corresponding points on the test surface. Using this technique, surface height accuracies on the order of 1/100 of the wavelength of the source [3], which is typically on the order of 500 to 700 nm, or better can be obtained.
A typical phase-shifting interferometer setup for the measurement of surface topography is shown in Figure 1. Light from the source, either quasi-monochromatic light from a laser or filtered white light, travels down the illumination path where it reflects off of the first beamsplitter to the microscope objective. The beamsplitter in the interferometer (a Mirau interferometer in this case) divides the light into two beams, one which reflects off of the reference mirror, and the other which reflects off of the sample. These two beams then recombine at the focal plane, where a solid-state detector array images and digitizes the intensity information into discrete frames of video data. If the optical path difference (OPD) between these two beams is small, or on the order of the coherence length of the source, interference fringes will be present at the detector.

Techniques for determining phase typically involve the sequential shifting of the phase of one beam of the interferometer relative to the other beam by known amounts, and measuring the resulting interference pattern irradiance. A modulation calculation is usually done as well at each sampled data point in order to determine validity, with the assumption being that pixels with higher values of modulation have a greater probability of being valid. The phase data is then processed by another "unwrapping" algorithm that removes phase ambiguities between adjacent pixels. The spatial sampling requirements of the detector array are such that there must be at least two detector elements for each fringe, in order for the wavefront to be accurately reconstructed. This critical sampling frequency is commonly referred to as the Nyquist frequency. If the fringe pattern is undersampled, an accurate measurement of phase cannot be made [3].
The intensity at the detector plane can be described with the following standard interference equation:

\[ I(x,y) = I_1 + I_2 + 2 \sqrt{I_1 I_2} \left[ 1 + \cos(\phi(x,y) + \alpha) \right] \]  

where \( I(x,y) \) is the intensity measured at pixel coordinates \((x,y)\), \( I_1 \) and \( I_2 \) represent the intensities of the reference beam and the sample beam, \( \phi(x,y) \) is the phase difference between the two wavefronts, and \( \alpha \) is the introduced phase shift. This equation is often rewritten in the following form:
\[ I(x,y) = I_0[1 + \gamma \cos(\phi(x,y) + \alpha)] \] . \hspace{1cm} (2)

At least three measurements of intensity will be required to determine the phase of the wavefront since there are three unknowns in Equation 2, namely the DC intensity term \( I_0 \), the fringe modulation \( \gamma \), and the phase \( \phi(x,y) \).

Typically, multiple frames of intensity data are recorded, and the interference equation is solved using least-squares techniques. For the simplest case of three frames, the expression for the phase at each point is [3]

\[ \phi(x,y) = \tan^{-1}\left[ \frac{I_3(x,y) - I_2(x,y)}{I_1(x,y) - I_2(x,y)} \right], \hspace{1cm} (3) \]

where \( I_1 \) through \( I_3 \) are the intensities measured at phase shifts of \( 90^\circ \), or \( \pi/2 \). The expression for the intensity modulation is [3]

\[ \gamma(x,y) = \sqrt{\frac{(I_3(x,y) - I_2(x,y))^2 + (I_1(x,y) - I_2(x,y))^2}{2I_0}}. \hspace{1cm} (4) \]

The relative surface height can then be determined from the phase data by the following formula:

\[ h(x,y) = \frac{\lambda}{4\pi} \phi(x,y), \hspace{1cm} (5) \]

where \( \lambda \) is the wavelength of the source. Discontinuities in phase between adjacent pixels due to the discontinuous nature of the arctangent calculation must now be removed, if possible, using some sort of phase unwrapping procedure. This typically involves comparing phase changes between adjacent pixels and adding or subtracting
a multiple of $2\pi$ if the phase difference exceeds $\pi$. For reliable removal of these discontinuities, the phase must not change more than $\pi$, or $\lambda/2$ in optical path difference, between any two adjacent pixels, or errors of $2\pi$ ($2\pi$ ambiguity errors) will result in the integrated surface map [3]. Thus, conventional phase-shifting interferometry is limited to the measurement of fairly smooth, continuous surfaces.

Various techniques have been employed to get around the $2\pi$ ambiguity problem inherent in phase-shifting techniques. Possible solutions would be to use a higher resolution detector array, or longer wavelength sources. A method has also been developed where a longer wavelength source is synthesized by taking multiple measurements using two or more shorter visible wavelengths. These multi-wavelength techniques can greatly extend the range of conventional single-wavelength techniques, but are still limited by the depth of focus of the microscope objective, and are much more susceptible to noise caused by air turbulence and vibrations [4].

2.2 Optical Probes

Optical probe profilometers are characterized by the fact that they must scan the surface being measured, unlike phase-shifting interferometers which can produce a three-dimensional surface map of a complete area in a single step. One type of optical probe profilometer utilizes a dynamic focusing system where a laser spot with a diameter of 1-2 microns is scanned across the surface being profiled. The technique is similar to that used in compact disk technology, where a feedback
system is constructed using a focusing lens attached to a displacement transducer, and an optical detector to provide a reference signal [5]. This method can produce surface height measurements with resolution on the order of a nanometer over a total range of ±10 microns, but measurements are typically much slower than phase-shifting techniques, and integration errors due to the scanning process are likely to be present, especially when measuring surfaces with low reflectivity, steep slopes, or high spatial frequency content.

Another type of optical probe uses a differential laser system to measure local slopes on the surface, and then integrates the local slope errors to obtain a surface profile. These systems are also limited by surface reflectivity and steepness of slope, and are subject to statistical propagation of errors in the measurement and integration of local slopes [6].

Other types of optical probes make use of confocal properties of high numerical aperture microscopic lenses. These confocal microscopes typically have high magnifications of between 100X and 10,000X, and can obtain profiles of surfaces with vertical resolution as low as 0.1 micron [7]. These instruments typically scan a spot derived from either a laser or broadband source across the surface being measured, as the focus of the objective is maintained. This technique has the advantages of being able to measure surfaces with large variations in height, and also the ability to section layers of translucent materials such as biological specimens. The drawbacks of this technique are the limited height resolution compared to other
optical techniques, and the fact that the surface must be scanned, typically resulting in slow measurement speeds.

2.3 Light Scattering Techniques

Total integrated scatter (TIS) is an optical technique for measuring the average surface roughness over an area. Although instruments that employ this technique do not produce actual surface profiles, they will be discussed here nonetheless. The basic principle behind this approach is to measure the backscatter of laser light reflected off of an opaque sample, and relate the amplitude of the scattered light to RMS surface roughness. TIS instruments typically consist of a laser light source illuminating a sample at normal incidence placed at the focal point of a specularly reflecting hemispherical dome, and an optical point detector placed at the conjugate focus of the sphere used to detect the scattered light [8]. Specularly reflected light is also sensed by a point detector placed along or near the axis of the laser beam. If the RMS surface roughness is smaller than the illuminating wavelength, the relationship between surface roughness and TIS is given by:

$$\text{TIS} = \left( \frac{\text{total scatter}}{\text{specular reflection}} \right) = \left( \frac{4\pi h_{\text{RMS}}}{\lambda} \right)^2,$$

where $h_{\text{RMS}}$ is the RMS surface roughness, and $\lambda$ is the wavelength of the source. The RMS roughness can then be found by measuring the output voltage of the point detectors:
where $V_d$ is the scatter signal, $V_{RS}$ is a scatter reference signal, $r_{RS}$ is a specular reference signal, and $r_s$ is the sample specular signal.

Although the TIS technique is a simple, relatively fast method to obtain surface roughness measurements, it gives no information of the spatial characteristics of the test surface, and since the measurement is taken over a small area equal to the spot size of the laser beam, the results obtained may differ from results taken over a long scan, such as in the case of other types of optical profilers.
3. A NEW METHOD FOR PROFILING ROUGH SURFACES

3.1 Basic Theory Behind the Technique

Another method of extending the range of interferometry that has become popular only very recently is the technique of "coherence scanning", where the modulation, or degree of coherence of an interference signal produced using a broad bandwidth source is used to determine surface characteristics. Systems have been described by Davidson et al. [9] and Kino et al. [10] that use this technique to determine surface characteristics, such as line widths, on integrated circuits. A coherence scanning microscope was described by Lee and Strand [11] that used a similar technique. The basic principle is the same in all of these techniques, what is different however is the method in which the modulation information is extracted from the interference data and how the surface profile is generated using this information. The intent of this development effort was to produce a fast, accurate method of profiling rough surfaces using principles similar to coherence scanning that could eventually be developed into an instrument capable of measuring a wide variety of surfaces in a wide range of applications.

If broad bandwidth light is used as the source in a typical interferometer, such as shown in Figure 1, the modulation, or visibility of the fringes drops off rapidly from its maximum value at minimum OPD, to zero at distances greater than the coherence length of the source. Figure 2 shows a typical fringe signal obtained from a detector in the image plane of an interferometer as the OPD is varied through focus. The
variation in path length could be accomplished by either varying the position of the reference mirror or by translating the sample. If the envelope of this signal, whose width and shape is a function of the spectral characteristics of the source, could be extracted, and the peak, which corresponds to the point of minimum OPD, could be detected, a measurement of relative surface height at that point could be made. If this procedure could be done for each point on the detector plane, a three-dimensional surface map of corresponding points on the test surface could be made. The modulation signal is inherently non-periodic for most broad bandwidth sources, and is therefore not subject to the ambiguity problems that affect phase measuring methods.

The quality of fringes produced by an interferometric system can be expressed in terms of the visibility which is given as [12]:

Figure 2 Fringe signal produced by broadband illumination
\[ V(z) = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} , \]

where \( I_{\text{max}} \) and \( I_{\text{min}} \) are the irradiances of the minimum and maximum of the fringes at position \( z \). Assuming a diffraction approximation, and a spatially-coherent source, the normalized form of the complex degree of mutual coherence of two complex wave fields, \( E_1(x,t) \) and \( E_2(x,t) \) can be expressed as [12]:

\[ \gamma(\tau) = \frac{<E_1^*(t)E_2(t+\tau)>}{\sqrt{<|E_1|^2><|E_1|^2>}} . \]

Now Equation 1 can be rewritten as:

\[ I(\tau) = I_1 + I_2 + 2\sqrt{I_1I_2} \text{Re}\gamma(\tau) , \]

where \( \gamma(\tau) \) is a complex quantity and can be expressed as:

\[ \gamma(\tau) = |\gamma(\tau)|e^{i\phi} , \]

and

\[ \text{Re}\gamma(\tau) = |\gamma(\tau)|\cos(\phi) . \]

Equation 10 then becomes:

\[ I(\tau) = I_1 + I_2 + 2\sqrt{I_1I_2} |\gamma(\tau)|\cos(\phi) . \]

For the case of a completely coherent source, \( |\gamma(\tau)| = 1 \), and for two completely incoherent waves, \( |\gamma(\tau)| = 0 \). For all other cases, the waves are said to be partially coherent, and \( 0 \leq |\gamma(\tau)| \leq 1 \). Therefore, \( |\gamma(\tau)| \) can be thought of as the degree of coherence in an optical interferometer, with the time difference \( \tau \) equal to OPD/c.
For a monochromatic source, the degree of coherence is always unity, for a partially incoherent source, as is the case for broadband illumination, the degree of coherence is unity when the OPD is zero, and drops off to zero as the OPD is varied [13].

Relating this back to the visibility, \( V \), of the fringes, it is clear that the minimum and maximum values of \( I(\tau) \) will occur when the cosine term in equation (13) is at values of +1 and -1. The degree of coherence is therefore equal to the fringe visibility at these points, assuming that \( I_1 = I_2 \), as is often the case.

It can also be shown [13] that the coherence function can be obtained from taking the Fourier transform of the source spectrum. The normalized Fourier transform of the source spectrum, \( G(\nu) \), can be expressed as:

\[
\gamma(\tau) = \frac{\int G(\nu) e^{-2\pi i \tau \nu} d\nu}{\int G(\nu) d\nu} \quad (14)
\]

It is also clear from Equation 14 that when \( \tau \) is equal to zero, the coherence term becomes unity.

We will define modulation, \( m(z) = |\gamma(\nu)| \) with the dimension \( z = \text{OPD}/2 \) representing the variation in the sample position through focus (since we are dealing with double-pass interferometry, a variation in \( z \) represents a two times variation in OPD). The interference equation can be once again rewritten as:

\[
I(z) = I_1 + I_2 + 2\sqrt{I_1 I_2} m(z) \cos(\phi) \quad , (15)
\]
Figure 3 Rectangular source spectrum

or if $I_1 = I_2$:

$$I(z) = I_0 [1 + m(z) \cos(\phi)] .$$  \hspace{1cm} (16)

Although the actual source spectrum is a function of many parameters and often difficult to predict, for the simplified case of a rectangular source spectrum, shown in Figure 3, with bandwidth $\Delta v = c \Delta \lambda / \lambda^2$ and constant irradiance $G(v) = I_0$ over that bandwidth, the coherence function obtained by taking the Fourier transform of $G(v)$ is given by:

$$\gamma(\tau) = \frac{\int G(v) e^{-2\pi i \nu} dv}{\int G(v) dv} = e^{-2\pi i \nu_0} \frac{\sin(\pi \tau \Delta v)}{\pi \tau \Delta v} ,$$  \hspace{1cm} (17)

and
\[ m(z) = \frac{\sin(2\pi \Delta \lambda z/\lambda_0^2)}{2\pi \Delta \lambda z/\lambda_0^2} = \frac{\sin(k\Delta \lambda z/\lambda)}{k\Delta \lambda z/\lambda}. \] (18)

In this case the irradiance \( I(z) \) consists of a cosine function with a frequency of \( 2/\lambda_0 \) multiplied by a sinc function whose width is proportional to the bandwidth of the source spectrum, as shown in Figure 4.

For an actual optical system, the simplification of a rectangular source spectrum will not be valid, and some other function will need to be used. In the interferometric system used for these experiments, for example, a standard tungsten-halogen bulb was used as the source. In this case the spectrum will follow closely the blackbody curve given in Figure 5. The optical transmission and detector response will also need to be taken into account in order to accurately model the behavior of the interference signal. A complete analysis of the optical path is beyond the scope of this paper,
however, and for simplicity we will assume that the source spectrum is either a simple rectangular function, or approximates a Gaussian shape, which is closer to what is measured in practice. In either case, the mathematics are simplified greatly.

![Blackbody curve](image)

**Figure 5** Blackbody curve

3.2 Initial Implementations

Initial attempts at proving the concept behind a coherence-measuring method of profiling rough surfaces made use of a vertical translating stepper-motor stage (Klinger P/N UZ80-PP) with a vertical step resolution of 0.1 micron, to vary the OPD in an interference microscope by moving the sample through focus. The stage was mounted on the base of a standard WYKO TOPO-3D phase-shifting interferometer,
and controlled via an IEEE-488 communications link from the Hewlett-Packard 300 series computer to a programmable indexer (Klinger P/N CC-1.2). A 256 x 256 solid-state detector array was used to produce images of the interference fringes that could be analyzed in the computer. Figure 6 shows a diagram of the arrangement used for the initial implementation.

Several algorithms were developed to measure the degree of coherence at each pixel in the solid-state detector array at each step of the measurement. A typical measurement consisted of the following steps:

1. The stage was positioned so that the highest point of the sample was just below the optimum focus point (point of minimum OPD) of the microscope.
2. The stage was then stepped through focus at a rate of 0.1 microns per step. At each step of the measurement, a measurement of the degree of coherence, or modulation, for each pixel in the array was made.

3. Two arrays of data were stored in the computer, one for the maximum value of modulation measured, and another for the stage position when that value was measured. The modulation calculated for each pixel at each step was compared with the previous maximum value, and if the current value was higher than the previous value, the previous value would be replaced with the current value and the current stage position would also be recorded.

4. After the entire surface had been scanned, or after the lowest point on the surface had gone through focus, the stage position array contained the vertical position in which each pixel had the highest value of modulation, which corresponded to the point at which the OPD was minimum for each corresponding point on the surface. A three-dimensional representation of the relative surface height of each point on the surface could then be plotted. Typically, the maximum modulation value stored for each pixel would be compared to a pre-determined threshold to determine the validity of each point.

The first algorithm used to measure the modulation involved phase-shifting the reference mirror in the interferometer and taking five frames of data for each step of
the measurement. Five-frame phase calculations are commonly used in phase-shifting interferometry due to their decreased sensitivity to system miscalibrations [14]. If we assume that the phase-shift between each frame of intensity is 90°, then five equations of the form given in Equation 2 can be written:

\[
\begin{align*}
I_1 &= I_o + I_o' \cos(\phi - 2\alpha) = I_o - I_o' \cos\phi \\
I_2 &= I_o + I_o' \cos(\phi - \alpha) = I_o + I_o' \sin\phi \\
I_3 &= I_o + I_o' \cos\phi \\
I_4 &= I_o + I_o' \cos(\phi + \alpha) = I_o - I_o' \sin\phi \\
I_5 &= I_o + I_o' \cos(\phi + 2\alpha) = I_o - I_o' \cos\phi
\end{align*}
\] (19)

These five equations can be solved for the fringe modulation, \(\gamma\), yielding:

\[
\gamma = \frac{\sqrt{(2I_3-I_1-I_2)^2 + 4(I_2-I_4)^2}}{4I_0}.
\] (20)

The piezoelectric transducer that translated the reference mirror was then calibrated for a phase shift of approximately 90 degrees between frames, and Equation 20 was used to calculate the modulation, with \(I_o\) assumed constant.

Measurements taken with this technique produced fairly good results on reasonably "rough" surfaces, but suffered from poor resolution on surfaces that were fairly smooth, but with surface roughnesses out of the range of standard phase-shifting methods. The resolution was expected to be as good as the minimum vertical step size of 100 nm, but in practice it was found to be at least five to six times larger than that due to the uncertainty of the modulation calculation when used with broad bandwidth light. Figure 7 shows a measurement taken on a roughness comparator strip (GAR P/N S-22) using this technique.
Various other algorithms were tried using the stepper-motor stage, ranging from simple measurements of maximum intensity (assuming that the "brightest" fringe occurs at the point of minimum OPD), to phase-shifting algorithms using two and three frames of data for the modulation calculation. Although some of these algorithms produced faster measurements, all suffered from the problem of poor resolution on smoother samples. Each algorithm was tested with both unfiltered white light, and filtered red ($\lambda=650\text{nm}$, $\Delta\lambda=20\text{nm}$) light, with the best results obtained with the unfiltered source, as expected, due to the shorter coherence length of the light and more peaked modulation envelope. Figure 8a shows a measurement of an optical flat (mirror) taken on a WYKO TOPO-3D system, with an RMS roughness of 2.63 nm and a peak-to-valley height of 26.7 nm over a $1\text{ mm}^2$ area. Figure 8b shows
the same surface measured using Equation 20 to calculate modulation. The RMS roughness obtained with this technique is now 113 nm, with a peak-to-valley height of 500 nm.

Although these initial implementations did not result in a method that provided the resolution necessary for profiling a wide range of surface textures, it nonetheless proved the feasibility of profiling very rough parts using coherence-measuring techniques.
Figure 8 (a) TOPO-3D measurement of optical flat (b) Optical flat measured using modulation algorithm
4. SIGNAL PROCESSING TECHNIQUES

4.1 Envelope Demodulation

If the irradiance signal produced by a broad bandwidth source in an interferometer is observed as the OPD is varied through focus, as shown in Figure 2, several properties can be noticed. First, the fringe signal is generally sinusoidal with a fairly constant center frequency, implying a bandpass characteristic. Second, the amplitude of the signal is modulated by another lower frequency, non-periodic, envelope function. Third, and this is stated as an assumption at this point, the shape of the modulation envelope does not depend on the relative phase of the fringe signal. These features are very similar to those found in amplitude modulated (AM) communication signals, where the signal to be transmitted (the sound information in an AM radio, for instance) is modulated by a carrier signal with a much higher center frequency than the information signal.

An amplitude modulated signal, \( s(t) \), can be mathematically expressed as:

\[
 s(t) = A [1 + kx(t)] \cos(2\pi f_c t + \theta) ,
\]

(21)

where \( A \) is the amplitude of the carrier signal, \( k \) is the amplitude sensitivity, \( f_c \) is the center frequency of the carrier, \( \theta \) is the carrier phase, and \( x(t) \) is the signal to be transmitted, or the modulating signal [15]. It is clear that this signal has the same form as that given in Equation 16, without the DC offset, or \( I_0 \) term. This therefore implies that the same basic principles that apply to classical communication theory in
the area of amplitude modulation/demodulation may apply in our case, since what we are interested in, after all, is recovering a modulation signal from a carrier, or fringe signal.

In the case of the amplitude modulated signal, the modulated signal, $s(t)$, is equal to the product of the sinewave carrier, $u(t) = A\cos(2\pi f_c t + \theta)$, and the modulating signal, $m(t) = 1 + kx(t)$. In the case of the interference signal, the modulated signal, $I(z)$, is equal to the product of the sinewave fringe signal, $u(z) = I_0\cos(\phi)$, and the modulating signal, $m(z)$, plus the DC term, $I_0$. The main difference between these two cases is that in the first case, the signals are all functions of time, with frequency represented as $f = 1/t$, while in the second case, the signals are functions of the variable $z$, and the frequency in this case must be expressed as $f = 1/z$.

The amplitude modulated signal, $s(t)$, will have a spectral density $S(f)$ with the following form:

$$S(f) = \frac{A}{2} \left[ \delta(f-f_c) + \delta(f+f_c) \right] + \frac{A}{2} \left[ M(f-f_c) + M(f+f_c) \right],$$

where $M(f)$ is the Fourier transform of the modulating signal $m(t)$. The frequency domain representation is shown in Figure 9.

In order to retrieve the signal of interest, the amplitude modulated signal must be demodulated using one of several techniques. Demodulation by synchronous detection involves multiplying the modulated signal by a periodic auxiliary signal of the same frequency as the carrier, and removing the resulting undesirable components.
with a lowpass filter [16]. The multiplication process effectively shifts the carrier information to a higher frequency of $2f_c$ and shifts the modulating signal to a bandwidth centered at zero. With the appropriate choice of lowpass filter, the modulating signal can be successfully extracted. The disadvantage of this technique, of course, is the need to reconstruct a synchronous signal of the same fundamental frequency as the carrier signal. This would be especially difficult when the carrier is a fringe signal, which is only present over a very short interval, and whose frequency characteristics may vary somewhat due to fluctuations in the source spectrum and variations in sample reflectivity.

A more practical technique of demodulation of the irradiance signal is by the method of amplitude demodulation by envelope detection. This method requires no generation of an auxiliary signal, and can be implemented using simple filtering techniques in high speed digital hardware. We start out with an amplitude modulated
input signal, \( s(t) = m(t)\cos(2\pi f_c t + \theta) \). The signal is first rectified by one of several methods, the most common of which involves either taking the absolute value of the signal or by squaring the signal. This also causes a shift in frequency of the carrier and modulating signal, which can then be separated by a suitable lowpass filtering operation.

For the case where the rectification is performed by taking the absolute value of \( s(t) \), the output of this operator, \( y(t) \) can be expressed as:

\[
y(t) = |s(t)| = |m(t)|\cos(2\pi f_c t + \theta) .
\]

Expanding \( y(t) \) in a Fourier series yields [16]:

\[
y(t) = \frac{2|m(t)|}{\pi} \left[ 1 + 2\sum_{n=1}^{\infty} \frac{-1^n}{4n^2-1} \cos[2n(2\pi f_c t + \theta)] \right] .
\]

This results in a shift of the modulating signal to zero frequency, while the carrier is shifted to frequencies at \( (2f_c, 4f_c, 6f_c, \ldots) \) with decreasing amplitude as the frequency order is increased. If a lowpass filter with a cut-off frequency greater than \( B_m \) but less than \( 2f_c-B_m \), where \( B_m \) is the bandwidth of the modulating signal \( m(t) \), is used to filter the output, an effective demodulation of the envelope can be obtained. Figure 10a shows the block diagram of the absolute value envelope demodulator, with the transfer function given in Figure 10b, and the resulting frequency plot and lowpass filter characteristics given in Figure 10c. The absolute value operator is very similar to the case of a detector used in practical AM demodulation circuits, where a simple half-wave diode rectifier is used along with an R-C network to provide the lowpass filtering function, as shown in Figure 11.
Figure 10 Amplitude demodulation by absolute value detection
For simple analog circuit implementations of the envelope demodulator, the simple rectifier works well. However, if the lowpass filtering is to be done using digital techniques, there is the potential problem of aliasing of the higher frequency components produced by the non-linear absolute value operation folding back into the passband of the lowpass filter. In this case, the technique of rectification by squaring, or square-law detection of the modulated signal produces a much more "ideal" separation of frequency components. The output of the squaring operator in this case is:

\[
\begin{align*}
  y(t) &= s^2(t) = m^2(t) \cos^2(2\pi f_c t + \theta) \\
  y(t) &= \frac{1}{2} m^2(t) + \frac{1}{2} m^2(t) \cos(4\pi f_c t + 2\theta)
\end{align*}
\]  

(25)

The resulting frequency spectrum of this signal includes the components resulting from the $1/2m^2(t)$ term centered at zero with bandwidth of $2B_m$ due to the multiplication process, and components centered at $\pm 2f_c$ with bandwidth $2B_m$ on either
Therefore if $f_c$ is greater than $2B_m$, the envelope can be recovered by lowpass filtering with cut-off frequency of greater than $2B_m$ but less than $2f_c - 2B_m$ [16]. The block diagram, transfer function, and frequency plot for the squaring operator is given in Figure 12. Now, with the proper choice of sampling rate and digital filter characteristics, assuming that the frequency of the carrier is sufficiently higher than that of the modulating signal, there should be no possibility of higher order frequency components of the rectification process aliasing into the output waveform. The squaring operator is also a simple operation to accomplish using digital hardware. The resulting output from the lowpass filter would then have the form:

$$z(t) = \frac{1}{2} m(t)^2 .$$  \hspace{1cm} (26)$$

Typically, a square root operation would follow the lowpass filtering operation. This step would not be necessary, however, if the modulating signal did not need to be reconstructed exactly, but some characteristic such as the presence or absence of the signal, or the peak of the signal, was all that was required.

An algorithm that successfully detects the location of the peak of the coherence function in an interference microscope can now be developed based on these principles. First, the irradiance signal must be sampled at fixed intervals ($\Delta z =$ constant) as the OPD is varied through focus. The sample rate must be chosen such that the signal is not under-sampled and violates the Nyquist theorem. This requires that the sampling frequency be greater than $2f_c + 2B_m$ for square law detection. Then the DC, or $I_0$ term must be removed from the signal if envelope detection is to be successful. There are a number of ways to do this, depending on the signal
Figure 12 Amplitude demodulation by square-law detection
characteristics and computational resources. If the \( I_0 \) term does not vary with OPD, then it can be treated as a constant, and once its value is determined, it can be subtracted from the irradiance signal. If \( I_0 \) does vary with OPD, then it cannot be removed by a simple subtraction, and some other means must be employed, such as bandpass or highpass pre-filtering.

Next the signal must be rectified by one of the above techniques. Assuming the carrier signal, the fringes, are at a much higher frequency than the envelope, which is typically the case, square-law detection should provide the best results, due to the lack of aliasing of high frequency harmonics, provided digital hardware with register lengths long enough to hold the large dynamic range of signals produced by the squaring operation is used. Modern floating point digital signal processing (DSP) hardware is well suited for this purpose. Then the rectified signal must be properly filtered using digital filtering techniques. Since there are many data points that need to be processed in parallel to produce a three dimensional surface profile, a filter form that requires the fewest possible delay terms should be chosen in order to reduce memory requirements.

Finally the peak of the lowpass filter output must be located, and the vertical position that corresponds to the peak be recorded. In order to increase the resolution of the measurement, some form of interpolation between sample points should be used. This operation can be considered to be a correlation of the output signal to the peak of the actual coherence function. A block diagram of the complete algorithm is shown in Figure 13.
4.2 Digital Filtering Techniques

The algorithm described in the previous section requires the use of filters on the irradiance data from the interference microscope. This data is typically discrete, or quantized in nature, resulting from a "snapshot" of the image at the detector plane taken with a solid-state camera whose output is digitized and transmitted to the computing hardware. Each pixel in each $N \times N$ video frame represents a sample of a continuous interference signal, therefore, in order to process an entire array of data, $N^2$ separate signals must be processed in parallel by the hardware. Due to the sampled nature of the signals, and the number of data points involved, analog filters are not practical in this application. Therefore, we must perform the filtering using digital techniques, which can impose massive computational requirements and require large amounts of computer memory. Clearly, a digital filter implementation that can
be executed quickly on large arrays of signals, as well as one that requires the smallest amount of memory for its storage requirements is desired. Of course, the filter must also perform the required function with as little distortion of the signal as possible that could lead to errors in the measurement. Fortunately, the rather predictable nature of the interference signal, as well as the fact that we are only interested in certain features, such as the peak of the modulation envelope, can relax the filter constraints considerably.

Digital signals are made up of digital numbers typically, but not always, produced as a result of sampling and quantizing a continuous, analog process. Digital signals are defined only at the sample points, and have meaningless values between these points. A digital signal can thus be thought of as a sequence of numbers, with the $n$th number in the sequence denoted $x(n)$, where $n$ is an integer multiple of $T$, the sampling period. An arbitrary digital sequence can be expressed as:

$$x(n) = \sum_{k=-\infty}^{\infty} [x(k) \delta(n-k)].$$

where $\delta(n-k)$ is a delayed unit-sample sequence, similar to the unit-impulse function in an analog system [17]. The input sequence to a digital system can therefore be represented as the weighted sum of delayed unit-sample sequences. The irradiance signal as output from the solid-state camera and digitized by an analog-to-digital converter can be thought of as being made up of a two-dimensional array of such sequences, with new samples taken each frame time. The output of a digital system can likewise be characterized by a weighted sum of delayed unit-sample responses $h(n-k)$:
which is analogous to the convolution of two signals in the analog domain. Any digital system that is both linear and shift-invariant can be completely characterized by its unit-sample response \( h(n) \) [17].

The sampling frequency defined as the inverse of the sample period \( T \), is often expressed in terms of a normalized frequency, or angle around the unit circle, with one sample period corresponding to a radian frequency \( \omega=2\pi \). In order to properly reconstruct the output signal from the sampled input, the input frequencies must be limited to those less than \( \omega=\pi \) to avoid aliasing effects. This frequency, \( \omega=\pi \), is commonly referred to as the Nyquist frequency [17].

The convention of the z-transform is often used in describing operations in digital discrete-time systems. This convention is similar to the Laplace transform in an analog system, where \( s=j\omega \) is the complex frequency domain variable. In a digital system, it is convenient to express the system transfer function in terms of the complex variable \( z \). The one-sided or unilateral z-transform of a digital sequence \( x(n) \) is given by:

\[
X(z) = \sum_{n=0}^{\infty} [x(n)z^{-n}] .
\]

The general transfer function of a digital system can be represented by the following relationship:
Taking the inverse $z$-transform, the following difference equation relating input and output of a digital system can be obtained:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}}.$$  \hfill (30)

\[\begin{align*}
    y(n) &= \sum_{k=1}^{N} a_k y(n-k) + \sum_{k=0}^{M} b_k x(n-k) .
\end{align*}\]  \hfill (31)

This is the general difference equation describing a digital filter. The output sequence, $y(n)$ in this case, is simply the sum of the delayed input values $x(n-k)$ multiplied by the coefficients $b_k$, and the delayed output values $y(n-k)$ multiplied by the coefficients $a_k$ [17]. The values chosen for the coefficients determine the type of filter implementation, either lowpass, bandpass, or highpass, as well as the characteristics of the filter response in the passband and the stopband. There are also many forms that the difference equation can take, and still preserve the relationship of input to output. Each form has its own merits in terms of the implementation in fixed register length hardware and the amount of memory space required for the delay terms.

There are two distinct classes of digital filter implementations, classified by their unit-sample response. If the unit-sample response of the filter is of finite duration, the filter is referred to as a finite impulse response (FIR) system, whereas if the unit-sample response is of infinite duration, it is referred to as an infinite impulse response (IIR) system [17]. An IIR filter typically requires the fewest number of delay terms.
and constant multipliers to implement a given set of filter characteristics. FIR filters, on the other hand, has the advantage of linear phase across its passband and are less susceptible to errors due to limited register size in the hardware that they are implemented on. The IIR filter, on the other hand, can have regions of non-linear phase in the passband, and if not designed properly can suffer from instability due to roundoff or truncation errors due to fixed register arithmetic. Usually, however, the IIR filter can be implemented with lower filter order, or number of delay terms than an FIR filter with similar frequency response characteristics. The IIR filter will thus lead to a much simpler implementation and require far less memory space for the storage of the delay terms. Since we need to store $M$ delay terms, where $M$ is the filter order, for each pixel in the two dimensional image array, the IIR realization is really the only practical implementation for use in our system. We must be careful to design the filter so that register length errors do not cause oscillations for the range of inputs expected, and that the frequencies of interest fall within an area of linear phase. Implementing the filter in DSP hardware with long register lengths and floating point arithmetic should help to solve some of these problems. Careful design of the filter with respect to the pole and zero locations is also important. The fact that the inputs to the system are generally slowly varying and predictable also simplifies the process.

Typically, the IIR filter is implemented as a cascaded system composed of second order sections of the "direct-II" form. This implementation generally leads to the most stable design as well as reducing the computational complexity of the algorithm. The general form of a second order direct-II filter section is given by:
\[ d(n) = x(n) + d(n-1)a_1 + d(n-2)a_2 \]
\[ y(n) = d(n)b_0 + d(n-1)b_1 + d(n-2)b_2, \]

where \( d(n), d(n-1), \) and \( d(n-2) \) are the delay terms, \( x(n) \) is the current filter input, and \( y(n) \) is the current output of the filter. The output is, of course, delayed by two sample periods from the input. Assuming that the frequencies of the fringes and the coherence function are separated by a large enough margin, IIR direct-II filter implementations with orders as low as two may be able to provide adequate performance for our purposes.

4.3 Peak Detection of Envelope

Assuming that the OPD is minimum for all points when the coherence function is maximum, a relative surface height for each data point in the array can be found by determining the vertical position \( z \) at which each point in the array maximizes the output of the lowpass filter. We must assume that the irradiance signal is sampled at a constant interval, and that we know, by some calibration procedure that is done beforehand, the variation in \( z \) for each interval. A coarse measurement of the peak of the coherence function would be to simply compare each filter output to a previous maximum to find the output sample that was maximum for each data point throughout the measurement, and record the value of \( z \) (the sample number) for each maximum. This would result in a measurement resolution in surface height of no greater than the sample interval, which would correspond to a resolution of 50nm to 100nm in most cases. In order to provide an increased resolution in surface height,
we must perform some kind of interpolation on the coherence function output in order to more accurately predict the point at which the function peaks.

There are a number of ways in which to accomplish this interpolation, but we will be limited to those methods that can be implemented using a minimum amount of hardware resources due to the large amount of data that must be processed. If the algorithm is designed so that the peak of the coherence function is detected as the measurement is in progress, so that for each new output of the lowpass filter a comparison to detect the peak would be made, then the amount of data that needed to be saved in order to perform an interpolation may be limited to just a few data points around the peak. If these data points are then saved along with the value of $z$ that occurred when the actual maximum was detected, the interpolation step could be postponed until after the measurement scanning was complete, so that the calculations would not need to be performed at each step of the measurement.

A common method of interpolating points along a curve when discrete sample points are known is to fit a polynomial to the curve of a maximum degree of $n$ with a minimum of $n+1$ data points. If there are more than $n+1$ data points, then least-squares techniques can be used to fit the curve to a higher degree of accuracy. If we assume that the coherence function is continuous near the peak, and that the behavior of the signal is fairly well behaved around this peak, then a polynomial can be fit to the function around this peak and the peak located by finding the value of $z$ where the first derivative is minimum. A quadratic fit is a good start for this purpose, since the coherence function closely approximates a quadratic along an area close to the
peak, and there are only three points required to fit a quadratic function. The value of the peak would be saved in each case, along with the two points on either side of the peak. At the completion of the measurement these points along with the vertical position, \( z \), recorded at the peak would then be processed to perform the interpolation.

![Figure 14 Peak detection of coherence function](image)

This technique is illustrated in Figure 14, where the peak value of a function

\[
y = a_0 + a_1 x + a_2 x^2
\]

is to be determined from three samples taken around the peak, labeled \( y_1 \), \( y_2 \), and \( y_3 \).

We arbitrarily set the value of \( x \) in which the corresponding value of \( y \) is largest, i.e. the peak value, to be \( x_2 = 0 \), with the points on the left and right side of this peak to be -1 and 1 respectively in order to simplify the mathematics. The value of \( x \) that
corresponds to the true peak of the function is simply an offset $-1 \leq x \leq +1$. The first derivative of Equation 33 is given by:

$$\frac{dy}{dx} = a_1 + 2a_2 x \, .$$

(34)

Setting Equation 34 equal to zero and solving for $x$ gives:

$$x_p = \frac{-a_1}{2a_2} \, .$$

(35)

Solving the three equations for the unknowns $a_0$, $a_1$, and $a_2$ and substituting into Equation 35 yields the following expression for the interpolated peak:

$$x_p = \frac{y_1 - y_3}{2(y_1 - 2y_2 - y_3)} \, .$$

(36)

This value is then added to the value of $z$ obtained for the peak modulation ($x_2$ for this example) to obtain the true position of the peak. The resulting value of position is then multiplied by a scaling factor in order to determine the true surface height for that point.
5. IMPLEMENTATION OF THE ALGORITHM

5.1 Digital Filter Design

In order to determine the filter characteristics required for the demodulation algorithm, the frequency characteristics of the input signal must be investigated. A test setup was developed, using a standard WYKO TOPO-3D interferometer as in the initial implementation shown in Figure 6, but instead of using a stepper-motor stage to provide the vertical translation, a piezoelectric transducer with approximately 80 microns of travel was mounted to the microscope base and driven by a precision analog ramp generator as shown in Figure 15. Encoder feedback was also incorporated in the transducer drive electronics to linearize the inherent non-linear motion of the piezoelectric device. A feature was added to the ramp generator that allowed the slope of the ramp to be varied, thus varying the effective sample period. This arrangement allowed unlimited resolution on the vertical positioning of the sample as it moves through focus. As long as the solid-state camera captured data at constant time intervals, and the piezoelectric transducer moved at a linear rate, the sample period would also remain constant.

The system was calibrated such that the sample period was approximately 70 nm when measuring a flat optical mirror with an unfiltered white light source (a sample period of 70 nm was found experimentally to be close to optimum for most situations. It was found that there is actually a fairly wide tolerance of at least ± 10 nm on this value). Figure 16 shows the vertical distance z, and frequency domain...
plots of a single pixel taken with an optical magnification of 20X (The frequency-domain variable has units micron\(^{-1}\), which for a sampling period of 70 nm has a value of 7.14 \(\mu m^{-1}\) at half the sampling frequency of 14.29 \(\mu m^{-1}\)). Note that there is very little variation in the overall DC level of the profile as a function of z. Therefore, for this test at least, there is no need to pre-filter the data to remove unwanted variations. The signal can now be rectified using square-law detection to separate the frequency components of the carrier and coherence envelope. Figure 17 shows a frequency plot of the same signal after this rectification. Note that there is a very distinct separation of frequencies, with the components of the envelope shifted to an area close to zero, and the components of the fringes shifted by a factor of two in the frequency domain.
It is apparent that a simple lowpass filter should be adequate to provide the demodulation of the coherence function. The filter characteristics should be chosen such that the cutoff frequency is beyond the bandwidth of the squared coherence envelope, and the response is down by several orders of magnitude for the squared fringe frequencies. The phase of the filter should also be as linear as possible in the
passband to reduce the possibility of errors due to non-linear phase delays across the data array. The actual performance of the filter will be best evaluated by performing actual measurements, however.

There are many commercially available software packages on the market today to aid in the design and development of digital filters. A package that was found to work well for these purposes was the "Hypersignal-Plus" DSP software package from Hyperception, Inc. [18]. This package provides the means to interactively design various types of digital filters, input and filter waveforms, display both the time domain and frequency domain results, and produce pole-zero and phase plots of the filter characteristics. A function is also provided that allows code to be generated for various DSP devices. All filters described in this paper were designed using this package.
The irradiance signal shown in Figure 16 did not show much variation in $I_0$ as a function of $z$, and a simple mean subtraction was all that was required to remove this term. In some cases, such as when measuring samples with stray reflections due to surface irregularities and when using high-power, short depth-of-focus microscope objectives, $I_0$ shows a strong variation through focus, and therefore must be eliminated by some other method. Figure 18 shows a measurement of the irradiance signal taken with a 200X objective on a flat optical mirror. Note the variation in $I_0$ as the OPD is varied. In order to effectively remove this term, some form of bandpass or highpass filtering must be done on the signal before the rectification process. The simplest method of doing this is to perform a digital highpass filtering operation on the signal, in a similar fashion as the lowpass filtering. The filter type would be IIR for the same reasoning as the lowpass filter, and its cutoff frequency would need to be lower than the lower bandwidth of the modulated fringe frequencies shown in Figure 16.

The nature of the frequency plot of the irradiance signal in Figure 16 suggests that a highpass filter with a relatively smooth cutoff characteristic, i.e. a lower order filter, could be used to remove the DC component. Indeed, a suitable filter was constructed with the following characteristics:

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frequency</td>
<td>14.29 $\mu m^{-1}$</td>
</tr>
<tr>
<td>Center Frequency</td>
<td>7.14 $\mu m^{-1}$</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>6.43 $\mu m^{-1}$</td>
</tr>
<tr>
<td>Stopband attenuation</td>
<td>30 dB</td>
</tr>
<tr>
<td>Filter type</td>
<td>IIR second-order Chebyshev</td>
</tr>
</tbody>
</table>
Passband ripple: 1 dB

A frequency domain plot of the filter characteristics is given in Figure 19. Figure 20 shows the corresponding phase plot while Figure 21 contains the pole-zero plot. The coefficients for this filter for a direct-II realization in the form given in Equation 32 are as follows:

\[ a_1 = 1.655717 \]
\[ a_2 = -0.7328169 \]
\[ b_0 = 0.8471335 \]
\[ b_1 = -1.694267 \]
\[ b_2 = 0.8471335 \]

The phase of the filter is fairly linear in the region of interest and should not cause any errors in measurement due to variations in delay as a function of frequency. The
Figure 19 Frequency plot of highpass filter response

pole-zero plot shows that the locations of the poles are well separated, and do not lay too close to the unit circle, so that the possibility of instability should be minimized. Figure 22 shows the irradiance signal from Figure 18 after filtering with the highpass filter. The slow variation in $I_0$ has now been eliminated, and the overall shape of the signal remained intact, indicating that the filter did not introduce any large degree of distortion to the signal.

The design of the lowpass filter follows along the same lines as the design of the highpass filter. The filter must preserve the low frequency components corresponding to the coherence signal, while filtering out the higher frequency components corresponding to the fringes. From the frequency plot of the square-law detected signal in Figure 17, it is apparent that this operation may be possible using a fairly
low-order IIR filter. Since the relative amplitude of the fringe frequency components is quite large, the attenuation in the stopband will need to be fairly high, however, if all residual fringes are to be removed from the output signal. A first attempt at producing a suitable lowpass filter resulted in a fourth-order IIR Chebyshev implementation with frequency, phase, and pole-zero plots given in Figure 23. Figure 24 shows the result of filtering the rectified signal, with a fairly clean envelope of the irradiance signal remaining. The characteristics of the fourth-order lowpass filter are as follows:

- Sampling frequency: 14.29 μm⁻¹
- Center Frequency: 7.14 μm⁻¹
- Bandwidth: 1.43 μm⁻¹
- Stopband attenuation: 40 dB
If the characteristics of the lowpass filter can be relaxed, however, the speed of the measurement could be increased and the amount of memory required to store the delay terms would be reduced. There are a total of eleven multiplications and eight additions required to process one step of a fourth-order IIR cascaded direct-II type filter, compared to only five multiplications and four additions for a second-order implementation. A fourth-order design also requires four delay terms to be stored in memory as opposed to two for a second-order filter. The characteristics of a second-order IIR lowpass filter appropriate for this purpose are as follows:

Filter type: IIR fourth-order Chebyshev
Passband ripple: 1 dB

Sampling frequency: 14.29 μm⁻¹
Center Frequency 7.14 μm⁻¹
Bandwidth: 4.29 μm⁻¹
Stopband attenuation: 34 dB
Filter type: IIR second-order Chebyshev
Passband ripple: 1 dB

The frequency, phase, and pole-zero plots of this filter are shown in Figure 25. It is apparent that the filter has a less steep cutoff characteristic in the stopband compared to the fourth-order filter, but the phase response appears to be more linear in the passband, which is a desirable characteristic. There is, however, the possibility that the ripple introduced into the output may become excessive if the sampling period decreases to the point that the fringe frequencies move too close to the filter cutoff.
Figure 23 Frequency, phase, and pole-zero plots of fourth-order IIR lowpass filter
Careful calibration of the stage motion should help to minimize this problem. Figure 26 shows the result of filtering the rectified irradiance signal using the second-order filter, with no noticeable ripple. The coefficients for a direct-II realization are as follows:

\[ a_1 = 1.778255 \]
\[ a_2 = -0.813641 \]
\[ b_0 = 0.008846 \]
\[ b_1 = 0.017693 \]
\[ b_2 = 0.008846 \]

The practical implementation of these techniques requires the use of high-speed digital signal processing hardware. For maximum accuracy in the digital filter
Figure 25 Frequency, phase, and pole-zero plots of second-order IIR lowpass filter
computations, large register sizes should be used to minimize the effects of errors due to roundoff and truncation in the multiplication and addition operations. Also, the memory to store the delay terms should be of a large enough word size so that the recursive effects of roundoff and truncation of these terms do not cause undesirable oscillations in the filter output. Since the signals input to these filters typically have a very large dynamic range due to the roughness of the samples being measured, these effects can become problematic. In the present implementation of this system, a floating point digital signal processor with 96 bit internal registers for arithmetic computations, and 32 bit wide memory has been successfully employed for this purpose. Due to the large dynamic range of the irradiance signal, a solid-state camera with a high signal-to-noise ratio is also desirable.
5.2 Error Analysis

Errors in the measurement of surface height arise from the inability to accurately measure the location of the peak of the coherence signal produced by the demodulation algorithm. The errors result from a number of causes, including but not limited to the following:

1. Noise on the irradiance signal and electrical noise in the input amplifiers.
2. Roundoff or truncation errors in the hardware.
3. Incomplete filtering of frequency components due to the non-ideal response characteristics of the digital filters or aliasing of unwanted high frequency components into the passband.
4. Variations in the shape or position of the coherence envelope due to variations in the optical system, such as phase changes on reflection, diffraction of light on reflection off of rough surfaces, and variations in the source spectrum.
5. Systematic errors such as vibration, stage non-linearities, and thermal drift during the course of a measurement.

Sources two and three can be controlled by careful design of the digital filters and the use of large register length, floating point digital hardware. The input video signal should also be digitized with as many significant bits as possible. Effects due to optical variations are difficult to predict due to the large number of parameters involved (see Reference 21 for a more detailed explanation). Systematic errors can be reduced by careful design of the mechanics. Vibration isolation platforms are
essential to producing good results, and smooth, linear motion of the translating stage is critical. For these purposes, we will assume that the optical parameters are constant and the mechanical motion of the system is constant during the course of a measurement. The major sources of error in the measurement will therefore be noise on the input signal and imperfections in the filter responses.

Current noise on the input signal is made up of a number of sources, including shot noise in the photodetectors $i_{sn}$, thermal noise or Johnson noise $i_{jn}$ in the resistive devices used in signal amplification, input noise in the amplifiers $i_{in}$, and quantization noise in the analog-to-digital converter $i_{qn}$. If these noise sources are all uncorrelated in nature and are the result of Gaussian processes, then the effective noise powers add in quadrature [15]:

$$i_t^2 = i_{sn}^2 + i_{jn}^2 + i_{qn}^2 + \ldots . \tag{37}$$

The resulting noise can be expressed as a variance $\sigma_t^2$ on the input signal. The signal-to-noise ratio (SNR) can be expressed as the ratio of total signal power to total noise power over a bandwidth $B$:

$$\text{SNR} = \frac{\int \Phi_s(f) \, df}{\int \Phi_n(f) \, df} = \frac{\alpha_s^2}{\alpha_n^2}, \tag{38}$$

where $\Phi_s(f)df$ and $\Phi_n(f)$ are the power spectral densities of the signal and noise respectively [16].

We will assume that the noise variance is present on the demodulated envelope function, and that it is this variance that causes the uncertainty in detecting the peak.
The Cramer-Rao bound [19] can be used to determine the best-case estimate of the time delay $\tau$ of a band-limited signal such as a radar pulse. If the SNR of the signal is sufficiently high such that a coarse estimate of the peak of the signal is already made (i.e. $\tau \approx 0$), then using the Cramer-Rao bound for an estimate, the variance on the time of arrival is given by:

$$\sigma_{\tau}^2 \approx \frac{1}{\left(\frac{E}{N_0 B_f}\right)^2} ,$$

(39)

where $E$ and $N_0/2$ are the respective energies of the signal and noise, and $B_f$ is the bandwidth of the signal in radians. It is apparent from this expression that our ability to estimate the peak of the coherence signal will be fundamentally limited by the SNR of the signal, with higher signal-to-noise ratios giving lower $\sigma_{\tau}^2$. It is also evident from this expression that the ability to detect the peak of the function is proportional to the bandwidth of the signal, indicating that the higher the bandwidth of the pulse, the more sharply peaked it tends to be. From Equation 14 we can see that the bandwidth of the coherence signal is inversely proportional to the source bandwidth, and therefore we wish to have as broad of a bandwidth for the source as possible to reduce errors in detecting the peak. On the other hand, we may wish to have a narrow bandwidth source to reduce spread in the frequency domain of the modulation signal and thus facilitate the filter design.

In order to understand how noise on the input signal affects measurement accuracy, we must somehow relate the input noise variance to a variance on the output parameter we are trying to measure, in this case the position. We already have an
expression for the best-case estimate given in Equation 39. What we now need is a way to relate this to actual measurement accuracies. Since the input noise will be mainly dominated by electronic noise in our camera system, a reasonable assumption to make is that the noise follows a Gaussian distribution with a probability density of the form:

$$P_x(x) = \frac{1}{(2\pi \sigma_x^2)^{1/2}} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right],$$

where $x$ is the random variable (the input signal in our case), $\mu_x$ is the mean of the distribution and $\sigma_x^2$ is the variance.

For a given function $y = f(x)$, the probability function $P_y(y)$ for a random variable $y$ given the probability function $P_x(x)$ for random variable $x$ can be expressed as:

$$P_y(y) = \frac{1}{|dy/dx|} P_x[x=f^{-1}(y)],$$

where $f^{-1}(y)$ is the inverse function of $x$ with respect to $y$ [20]. For simple functions this transformation can be derived analytically without too much trouble. In our case, where we want to estimate the variance on the measurement given a variance on the input signal, where there is no simple function relating input to output due to the complexity of the digital filtering steps involved, an analytical derivation of the measurement variance using Equation 41 would be extremely difficult if not impossible. Therefore, we must use computer simulation techniques to estimate the output probability distribution.
A software model of the system was constructed \cite{21} that takes into account the various optical and electrical parameters involved, including source spectral characteristics, optical transmission properties, and detector response. Data is generated for two separate data points and then input to the digital filtering algorithm described in the previous section to simulate the measurement of a step. Random Gaussian noise with a distribution given in Equation 40 generated using the techniques described in Reference 20 is then added to the normalized detector output signal to simulate the effects of input noise on measurement accuracy. The variance, or standard deviation, of the Gaussian noise is input to the simulation program and all other parameters are kept constant. Multiple iterations of the simulation is then performed for each value of standard deviation in order to get good statistical results. It was found that performing the simulation for a total of 1000 iterations was a good compromise between obtaining reasonable statistical distributions and practical computer execution times. A histogram of the output step heights obtained for a normalized input standard deviation \( \sigma \) of 0.02, corresponding to an approximate input RMS SNR of 30 dB, is given in Figure 27. Since the distribution clearly approximates the Gaussian distribution indicated by the dotted curve, a measure of the standard deviation should give a good approximation of the expected RMS output measurement error.

A series of simulations were performed and the resulting RMS measurement noise (the standard deviation of the output distribution) is plotted for various values of input standard deviation and SNR in Figure 28. The system parameters were selected for this simulation to closely match the actual parameters measured in practice and
Figure 27 Histogram of simulation output for an input SNR of 30 dB

are given as follows:

1. Source follows approximate blackbody curve for $T = 3270$ K.
2. Optical transmission of unity (100 %) over a 200 nm bandwidth centered at $\lambda = 700$ nm, with a rolloff following a Gaussian curve with the response down to 1% at 100 nm on either side of the passband.
3. Linear detector response from 200 nm to 800 nm falling to zero at $\lambda = 1100$ nm.
4. A simulated step height of 0.50 $\mu$m with no tilt present.
5. A sample width of 0.070 $\mu$m.
6. Measurement algorithm consisting of second-order highpass filter, square law detection, and second-order lowpass filter as described in
Figure 28 Simulated RMS measurement error versus input noise. (a) measurement error versus input noise standard deviation. (b) measurement error versus input SNR.

7. Three point quadratic interpolation as described in Section 4.3.

The effect of spectral bandwidth on measurement accuracy can also be estimated using these techniques. A series of simulations were performed using a nominal SNR
of 40 dB and the same system parameters as given above, but with a rectangular spectrum centered at \( \lambda = 700 \text{ nm} \). The optical bandwidth was then input for each simulation, and simulations performed over a range of bandwidths from 60 nm to 400 nm. A plot of measurement error versus optical bandwidth is given in Figure 29. Note that the optimum measurement accuracy occurs at the higher optical bandwidths, as expected.

![Figure 29 Simulated measurement error versus optical bandwidth for input SNR of 40 dB](image)

The results of the simulation thus follow what was predicted by the Cramer-Rao estimate given in Equation 39: the measurement error is inversely proportional to the signal-to-noise ratio of the signal being detected (and indeed follows a linear relationship), and is also is inversely proportional to the signal bandwidth (which is proportional to the optical bandwidth in our case). The simulated error for the
estimated signal-to-noise ratio of the camera used in the physical implementation of the algorithm, as described in Section 6.1, is also in line with the RMS error obtained in an actual measurement. For example, the estimated SNR of the camera system is approximately 50 dB, and the RMS measurement error obtained in practice is typically less than 2 nm.
6. RESULTS

6.1 Physical Implementation

For the purposes of evaluating the quality of measurements obtained on actual rough surfaces with this technique, the following algorithm to demodulate the coherence envelope was implemented:

1. The irradiance data was filtered using the second-order highpass filter described in Section 5.1.
2. The output of the highpass filter was then rectified by square-law detection.
3. The rectified signal was then demodulated by lowpass filtering using the second-order lowpass filter described in Section 5.1.
4. Three points around the peak of the lowpass output were saved for interpolation using the quadratic fit technique described in Section 4.3.
5. At the completion of the vertical scan, the maximum output obtained for each pixel was compared to a threshold to determine validity. Data points that fell below this threshold, due to limits in surface slope or reflectivity, were flagged as bad and not used in the calculations or display.

The test setup described in Section 5.1 was used for all measurements, and no averaging or spatial filtering was done on the data post-measurement. The array size of the video data was 256 x 256. Calibration was done on the system by performing
measurements of a step height standard and adjusting the translating stage velocity until the sample interval was a near-optimum 70 nm per step. A VLSI 13.44 micron step height standard was used for this purpose.

6.2 Surface Measurements

In order to obtain an estimate of the lower limit of resolution in surface height measurement obtainable with this technique, a series of measurements were performed on a smooth optical flat, or mirror (Newport 10Z40ER.1). The flat was first measured using a WYKO TOPO-3D phase-shifting interferometric microscope at a magnification of 40X to obtain an average RMS roughness for the surface. A "reference-subtraction" technique was used in the measurements to eliminate errors introduced by the reference mirror and optical aberrations. Multiple measurements are taken of a smooth surface and an average of these measurements is obtained that corresponds to the total system error, which is then subtracted from each subsequent measurement.

Ten measurements were performed using the TOPO-3D system at different locations on the surface of the flat, and the RMS surface heights recorded for each. The same sample was then measured on the rough surface test system using the coherence measuring algorithm, again with reference subtracted. As indicated in Table 1, the TOPO-3D measurement produced an average RMS roughness of 0.314 nm, while the coherence measurement technique produced an average RMS roughness of 0.887 nm.
for no tilt (fringes nulled), and 2.047 nm when there were approximately four fringes in the field of view. Three dimensional isometric plots of the surface as measured using the coherence measurement technique are shown in Figures 30 through 31.

A study was performed to determine the effects of quantization noise on the measurement. The twelve bit data from the camera was truncated in software prior to processing in order to simulate coarser levels of quantization. A series of measurements were then made on both a mirror and a smooth piece of glass using various numbers of effective bits of quantization, with the results plotted in Figure 32. Assuming the input signal is sufficiently random and uncorrelated in nature, quantization noise can be expressed as:
Figure 31 Measurement of optical flat (4 fringes)

\[ \sigma_{\text{opt}}^2 = \frac{2^{-b}}{12} , \quad (42) \]

where \( b \) represents the number of bits in the quantized data [17]. The SNR for an input signal power \( \sigma_s^2 \) is given by:

\[ \text{SNR(dB)} = 10\log_{10}(\sigma_s^2/\sigma_{\text{opt}}^2) = 6.02b + 10.79 + 10\log_{10}(\sigma_s^2) . \quad (43) \]

The SNR is therefore proportional to the number of bits in the quantized data word, increasing approximately 6 dB for each additional bit.

The results of this study follow what was proven in Section 5.2, that measurement error is inversely proportional to input SNR. Since the relative reflectivity of glass with respect to the reference surface is lower than that of a mirror, the fringe visibility is decreased, and hence the SNR decreases as well, explaining the
noticeably higher measurement error for glass. It was found that in order to obtain reasonable results when measuring surfaces with high reflectivity such as a mirror, a minimum of seven to eight bits were necessary. For surfaces with lower relative reflectivity, higher levels of quantization may be required.

The algorithm was tested on a wide variety of surfaces and was found to be very tolerant of variations in surface reflectivity, magnification, source illumination levels, and surface discontinuities. Several measurements taken on the roughness comparator strip (GAR P/N S-22) are shown in Figures 33 through 35. Figure 36 shows a measurement of the back side of a silicon wafer after an etch process taken at a magnification of 100X.
Measurements of sharp steps or discontinuities in the surface produced some
overshoot at the edge of the step. This overshoot was typically on the order of 20 to
40 nm. The exact causes of this problem are not known, but is speculated to be the
result of diffraction of light reflecting off of these edges. Figure 37 shows a
measurement of a 43.5 nm VLSI step height standard exhibiting these effects. The
overshoot was not noticeable on the majority of surfaces measured, where surface
slope are limited to a few degrees. On rougher surfaces with sharp discontinuities,
the errors introduced by these overshoots will be small compared to the overall
roughness of the surface.

<table>
<thead>
<tr>
<th>measurement number</th>
<th>TOPO-3D measurement</th>
<th>rough surface measurement (fringes nulled)</th>
<th>rough surface measurement (4 fringes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.399</td>
<td>0.806</td>
<td>2.163</td>
</tr>
<tr>
<td>2</td>
<td>0.327</td>
<td>0.970</td>
<td>1.577</td>
</tr>
<tr>
<td>3</td>
<td>0.307</td>
<td>0.950</td>
<td>2.568</td>
</tr>
<tr>
<td>4</td>
<td>0.272</td>
<td>0.833</td>
<td>1.495</td>
</tr>
<tr>
<td>5</td>
<td>0.312</td>
<td>0.839</td>
<td>2.871</td>
</tr>
<tr>
<td>6</td>
<td>0.300</td>
<td>0.801</td>
<td>2.588</td>
</tr>
<tr>
<td>7</td>
<td>0.356</td>
<td>0.692</td>
<td>2.074</td>
</tr>
<tr>
<td>8</td>
<td>0.272</td>
<td>0.853</td>
<td>1.974</td>
</tr>
<tr>
<td>9</td>
<td>0.284</td>
<td>0.876</td>
<td>1.740</td>
</tr>
<tr>
<td>10</td>
<td>0.309</td>
<td>1.014</td>
<td>1.421</td>
</tr>
<tr>
<td>mean:</td>
<td>0.314</td>
<td>0.887</td>
<td>2.047</td>
</tr>
<tr>
<td>std dev:</td>
<td>0.039</td>
<td>0.072</td>
<td>0.503</td>
</tr>
</tbody>
</table>

Table 1 Comparison of measurements of optical flat

The speed of the measurements is typically limited by the processing speed of the
digital hardware performing the demodulation algorithm. In our case, this resulted in
a measurement speed of approximately 200 milliseconds per sample step, or an
average measurement time of 10 to 20 seconds for most samples. The speed of the
measurement is, of course, a function of the peak-to-valley surface height that must be scanned. The theoretical speed of a measurement is limited by the camera frame rate, which in our case was approximately 100 milliseconds per frame. Higher frame rate cameras could produce faster measurements if the hardware could process the data at that rate.

6.3 Summary and Conclusion

The goal of this paper was to demonstrate a technique for producing fast, non-contact measurements of a wide variety of rough surfaces without the problems associated with other optical techniques. A method of demodulating the peak of the coherence function of a white light interference signal using digital signal processing methods was described that could be implemented on two-dimensional arrays of data from a video camera, processed at high speeds by fast DSP hardware, and did not require a large amount of computer memory for its calculations. This technique was used to measure the surface profile of a variety of surfaces with results superior to those obtained in earlier implementations. The robust nature of this technique should prove it capable of realization in an instrument capable of measuring a wide variety of surfaces in both industrial as well as scientific applications.
**Figure 33** Measurement of section of roughness comparator with average roughness of 8 μin RA

**Figure 34** Measurement of section of roughness comparator with average roughness of 32 μin RA
Figure 35 Measurement of section of roughness comparator with average roughness of 125 μin RA

Figure 36 Measurement of backside of silicon wafer after acid etching process
Figure 37 Measurement of 43.5 nm step showing overshoot on edges. (a) isometric surface plot (b) two-dimensional profile
REFERENCES


