OPTICAL MEASUREMENT OF ULTRA FINE LINEWIDTHS USING ARTIFICIAL NEURAL NETWORKS

Richard Smith, MEng.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

August 2006

OPTICAL MEASUREMENT OF ULTRA FINE LINEWIDTHS USING ARTIFICIAL NEURAL NETWORKS I

ABSTRACT V

ACKNOWLEDGEMENTS VI

CONFERENCES & PAPERS VII

1 INTRODUCTION 1

1.1 THE IMPORTANCE OF RESOLUTION 2
1.2 STANDARD MEASUREMENTS 6
1.3 THESIS LAYOUT 11

2 BACKGROUND 13

2.1 OPTICAL MODEL 14
2.2 CALCULATING LINE WIDTH 17

3 LITERATURE REVIEW 28

3.1 OPTICAL SYSTEMS 28
3.1.1 Linewidth measurement systems 29
3.1.2 Commercial profilometers 31
3.1.3 Research Based Profilometers 33
3.1.4 Other optical approaches 34
3.1.5 Contact/ non optical systems 35
3.2 SPECTRAL EXTENSION AND INFORMATION THEORY 36
3.2.1 Analytic continuation and the uniqueness theory. 38
3.3 APPLICATIONS OF SPECTRUM EXTENSION THEORY 43
3.3.1 Sampling theorem in frequency domain [17] 43
3.3.2 Auto-Regressive Models 45
3.3.3 Gerchberg – super resolution through error energy reduction 47
3.3.4 Other Techniques 49
3.4 ARTIFICIAL NEURAL NETWORKS 50

4 ARTIFICIAL NEURAL NETWORKS 54

4.1 INTRODUCTION 54
4.2 FORMING A NETWORK – TOPOGRAPHIES AND APPLICATIONS. 56
4.3 TRAINING 57
4.3.1 Improving training 58

4.4 NETWORK DESIGN 62
4.4.1 Input data and Targets 62
4.4.2 Number of layers 63
4.4.3 Number of Nodes 63
4.4.4 Number of Inputs 64
4.4.5 Training Set Size 64
4.4.6 Improving Training with small data sets 64
4.5 SIMULATION OF SINGLE TRACKS 66
4.5.1 Neural network topography 67
4.5.2 Input format 68
4.5.3 Input / Output Scaling 69
4.5.4 Adding noise to the system 72
4.5.5 Repeatability of training 75
4.5.6 Nodes 76
4.5.7 Number of Inputs 77
4.5.8 Simulation Auto correct 78
4.6 DOUBLE TRACKS SIMULATION 81
4.7 DOUBLE TRACK OR SINGLE TRACK CLASSIFIER 91
4.8 REQUIREMENTS ON THE OPTICAL SYSTEM 93

5 OPTICAL SYSTEMS 95

5.1 ULTRA STABLE COMMON PATH DIFFRACTIVE ELEMENT SCANNING INTERFEROMETER 95
5.1.1 Practical considerations 101
5.2 DSOM – DIFFERENTIAL SCANNING OPTICAL MICROSCOPE 102
5.2.1 Shot Noise 105
5.2.2 Practical considerations 112
5.3 SCANNING NOMARSKI 112
5.3.1 Noise / repeatability / vibration etc / photon noise 116
5.3.2 Repeatability 119
5.3.3 Example scans 120
5.3.4 Practical considerations 124
5.4 COMPARISON TABLES AND COMMENTS 125

6 EXPERIMENTAL RESULTS 127

6.1 1-3 MICRON SAMPLE 127
6.2 TRADITIONAL APPROACH FOR TRACK WIDTH MEASUREMENT 132
6.3 ANALYSIS OF TRAINING RESULTS FOR DIFFERENT OPTICAL SYSTEMS AND SAMPLES 134
6.3.1 Differential Scanning Optical Microscope (DSOM) 135
6.3.2 Scanning Nomarski Microscope 137
6.3.3 Hologram 144
6.3.4 Comparison of training results 147
6.4 USING REDUCED TRAINING SETS 4,3,2, ETC 149
6.5 MISSING TRACKS LEFT OUT AT RANDOM 151
6.6 OUT OF RANGE 152
6.7 INPUT POINTS 154

6.8 AUTO CORRECTION FOR TARGET ERRORS....157
6.9 DOUBLE TRACK EXPERIMENT....160
6.10 SINGLE TRACK / DOUBLE TRACK CLASSIFIER....163
6.11 REPEATABILITY OF TRAINING....165
6.12 OVERALL UNCERTAINTY....166

7 FUTURE WORK....168

7.1 SLOPE SIMULATION....168
7.1.1 Impact of varying sloped tracks on network trained on tracks with fixed slope value....171
7.2 HEIGHT SIMULATION....174
7.2.1 Height / range of heights....175
7.3 ARCHITECTURES....178
7.3.1 Single Track Tree Simulation....180
7.3.2 Single Track Height Classifier....185
7.3.3 Single Track Edge Slope Classifier....189
7.4 PROFILES....192
7.5 INPUT POINTS....195
7.6 NETWORK DEVELOPMENT....196
7.7 FUTURE OF OPTICAL MICROSCOPES....197
7.8 PROVIDING A USER FRIENDLY SYSTEM....198
7.9 SUMMARY....199

8 CONCLUSIONS....201

9 APPENDIX 1 - GENERALISED DELTA RULE AND BACK PROPAGATION....207

10 APPENDIX 2 - CONVERTING PHASE NOISE TO PHOTON NOISE...210

11 APPENDIX 3 – EFFECT OF VIBRATION FOR THE DSOM SYSTEM 214

11.1 RIGHT HAND SIDE WINDOW....215
11.2 FULL WINDOW SIZE....216

12 APPENDIX 4 – DERIVATION OF SYSTEM MATHEMATICS FOR LINEAR INPUT POLARISATION FOR NOMARSKI SYSTEM....218

13 REFERENCES....222

Abstract

Measuring fine track widths with optical instruments has become increasingly difficult as the dimensions of the features of interest have become smaller than the traditional optical resolution limit. This has caused a move to non-optical methods such as scanning electron and atomic force microscopy techniques, or novel optical methods combined with signal processing techniques to provide measurements of these samples. This thesis presents one method to increase the measurement capabilities of an optical system. This is achieved by combining an optical profiler such as a scanning interferometer, with an artificial neural network (ANN). Once trained the ANN can calculate the object parameter for other tracks not contained in the training set. This process works extremely well; with experimental results showing that a 60nm track width can be calculated with a 2nm error using an optical system with a spot size of 2.6 microns. The technique can be extended to obtain other parameters such as height, sidewall slope and for other structures such as double tracks. Various aspects of the ANNs have been investigated, such as the training range, the size of network and the impact of noise etc. These studies show that the technique is extremely robust, and has huge potential for general usage.

Acknowledgements

I would firstly like to thank my supervisors, Dr CW See, Professor MG Somekh and Dr A Yacoot (NPL), for their support and input throughout my research degree.

I would like to thank the EPSRC for providing the funding for this work and the National Physical Laboratory for my CASE studentship.

I would especially like to thank my wife, Catherine, for all of her support over the past 4 years. I could not have done this without your help.

Conferences & Papers

Papers:

RJ Smith CW See MG Somekh A Yacoot,"Optical track width measurements below 100 nm using artificial neural networks", IOP Meas. Sci. Technol. 16 (2005) 2397-2404

RJ Smith CW See MG Somekh A Yacoot, "Optical track width measurements using artificial neural networks ", to be submitted.

Conferences:

SPIE Metrology June 2005, Munich :

R Smith CW See MG Somekh A Yacoot, "sub 0.1um Optical Track Width Measurement", Edited by H Ottevaere P Dewolf D Wiersma, Proc of SPIE vol 5858, 58580M (2005)

EPSRC PREP 2005 Lancaster University 30 March 10th April 2005:

R Smith CW See MG Somekh A Yacoot " Combined Optical Interferometer and ANN for Track width measurement beyond conventional limits", Prep 2005 (EPSRC IEE IEEE),Poster pp11 p.254 2005

SPIE Metrology, inspection and process control for microlithography XXI. Conference 6518, 26 Feb – 1 March 2007:

“Optical line-width measurement below 50 nm”, C. W. See, R. J. Smith, M. G. Somekh, The Univ. of Nottingham (United Kingdom); A. Yacoot, National Physical Lab. (United Kingdom) [6518-49]

Micro & Nano Technology (MNT) Measurement Club: Critical Dimension Metrology using SPM, SEM and related techniques:

“Optical track width measurements below 50nm using artificial neural networks”, C. W. See, R. J. Smith, M. G. Somekh, The Univ. of Nottingham .A. Yacoot, National Physical Lab.

For my wife Catherine and our son Alexander.

1 Introduction

Since Anton Van Leeuwenhoek used the first simple single lens microscope to look at tree bark and insects in 1674 [1], people have been captivated by the microscopic world. Finally being able to see things beyond the capability of the human eye captured the imagination of these early pioneers of optical microscopy. This drove the field forward and in the 18th century lens making procedures improved greatly and led to the reduction of some of the aberrations present in the images obtained with these early systems, thus improving the image quality and opening new areas for observation[2]. In 1830 J. Jackson discovered that combining several weaker lenses instead of using one strong lens reduced chromatic aberration and allowed good magnification without blurring of the object and so the compound microscope was born[3][4]. By 1872 the mathematical theory behind the microscope was beginning to be formulated and Ernst Abbe formulated his 'Abbe sine condition' [5] providing the mathematics to explain the maximum resolution of a particular optical microscope. In 1879 Lord Rayleigh [6] derived the image resolution criteria of optical systems based on the separation of diffraction limited images of point sources. By 1896 H Powell had made very high power objectives[7], corrected for three wavelengths and using oil immersion to obtain a lens with a numerical aperture of approximately 1.50, that enabled objects smaller than a micron to be observed. As the years progressed the advent of more complex and sophisticated systems proliferated, in 1936 Zernike invented the phase contrast microscope[8]opening up the possibility to image a whole range of samples that were previously impossible to observe due to the lack of

intensity variation in these samples. Nomarski patented the differential interference contrast microscope in 1953[9] and M Minksy presented the confocal microscope in 1955. These different system configurations have led to a huge range of applications for optical imaging systems covering the fields of biology, medicine, physics, engineering, archaeology, manufacturing etc

1.1 The Importance of resolution

The ability to resolve fine features is the corner stone of microscopy. As technology has progressed the demand to resolve finer and finer features has also increased. However, the optical microscope has a finite resolution due to the nature of light, which will now be considered.

We can consider the propagation of light as a wave and relate this to resolution. Each point on a wavefront can be considered to be a secondary point source that radiates a spherical wavelet [10]. To calculate the field distribution at some distant point all that is required is to sum the contributions from all of the secondary point sources, taking into account the relative phase and directions of the components. Waves will generally spread out as they propagate, the way in which this spread occurs depends on the source, the medium that the wave is travelling through and the effect of any objects in the beam path. A relevant example is when a plane wavefront is incident upon a circular aperture the field pattern observed in the optical far field is that of the Airy disk [11]. This is also the focal pattern that is formed from an aberration free, diffraction limited optical imaging system and can be described by:

$$I = I_0 \left[\frac{J_1(kNAr)}{kNAr} \right]^2 \qquad \text{Equation 1-1}$$

Where I is the intensity distribution, J_1 is the first order Bessel function of the first kind, $k = 2\pi/\lambda$, r is the radial coordinates of the image plane and NA is the numerical aperture of the system. The equation above describes how a point object in the object plane will appear in the image plane; this is why it is often referred to as the 'point spread function' (PSF). It is clear that the finite sized image of a point object is due to the finite resolution of the optical system. Consider imaging two point objects that are far apart, the two point sources are incoherent with one another, so that in the image plane, the intensities of the two images of the point objects are summed. The resulting image will contain two peaks corresponding to the two objects with the same dimensions. If the two objects are moved closer together eventually the images due to the point objects will begin to overlap ultimately merging so that it is extremely difficult to tell whether one or two objects are present in the recorded image. This is illustrated in Figure 1 where the separation between the point objects is reduced for each plot and the dotted line shows the sum of the images due to the two objects for a partially coherent system.

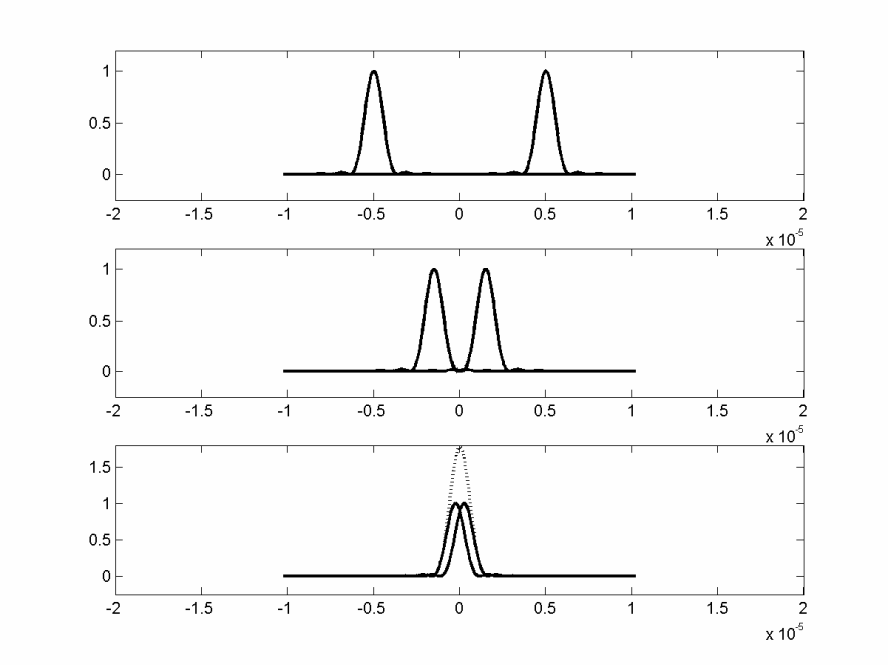


Figure 1 - 2 point objects and corresponding images as distance reduces

In the first two images the sum image hardly differs from the individual images of the point sources. The lower figure shows the two individual images of the point sources and the actual image of the two point objects (dotted line). It is clear that the two objects have not been resolved in this case as only one peak is contained in the final image. The precise separation for two point objects to just be resolvable has been defined by several people. The most common definitions are the Rayleigh and Sparrow criteria.

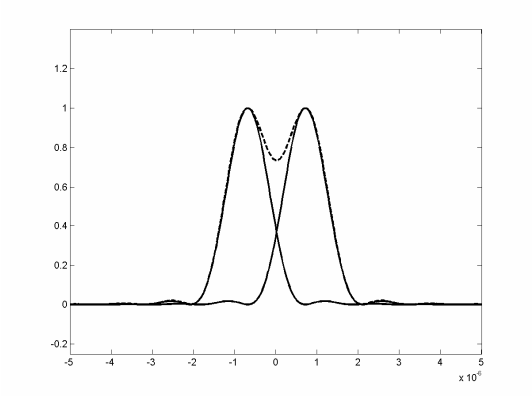


Figure 2 - Rayleigh criteria

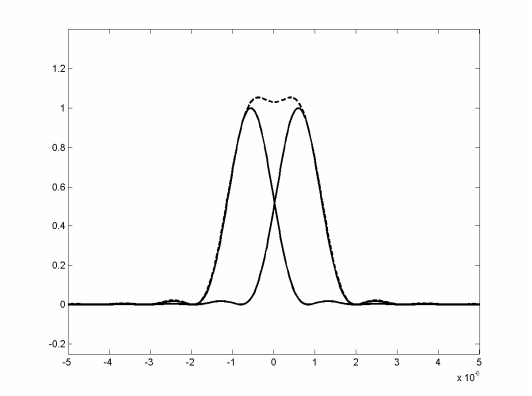


Figure 3 - Sparrow criteria

The Rayleigh [12] criteria states that two point objects are just resolvable when the maximum of the PSF due to one point object overlaps with the first minimum of the PSF due to the other object as shown in Figure 2. This criterion is independent of the coherence of the two point sources. Using the mathematical expression for the PSF for the optical system in Equation 1 this can be expressed as:

$$\delta_R = \frac{0.61\lambda}{NA} \qquad \text{Equation 1-2}$$

Where δ_R is the object separation.

Sparrow [13] on the other hand, defined the point where the objects are just resolvable as when the intensity of the mid point of the image has the same value as the peak value of the PSF as shown in Figure 3. In this case this criterion is more general, as it is also applicable to coherent sources, and is therefore related to the amplitude of the sources. This can be expressed as:

$$\delta s = \frac{0.504\lambda}{NA} \qquad \text{Equation 1-3}$$

As can be seen the sparrow criteria for a system yields an answer 17% smaller than the Rayleigh criteria. This shows that the subject of resolution is somewhat arbitrary, the criteria used to define resolution above is based on being able to simply detect where two objects are present it is not based on some fundamental limit, but rather on the simplicity and ease of use in detecting the presence of two point objects.

The values for the resolution of a system presented above are the best possible assuming that the optical system is perfect in everyway. Obviously in practice this will not be the case, the resolution of the optical system will depend upon the quality of the optical components used, the quality of the light source, the signal to noise ratio and the detector in the imaging system.

From the above equations we can see that using a shorter wavelength or increasing the numerical aperture of the imaging system will increase the resolution of an optical system. However, this does not mean that arbitrary small resolution is available for optical microscopy as the NA of the system is limited to a maximum of 1 in air and the practical range of useable wavelengths for optical imaging is restricted to the visible spectrum. This gives the resolution limit of a conventional optical microscope to be around 250nm and while this resolution is more than adequate for many modern

day applications for others it is not, for example track width measurement in the electronics industry.

The limit to optical resolution has meant that in recent years there has been a move away from optical methods for some kinds of measurements as the features of interest can no longer be easily resolved. The alternative techniques employed are usually scanning probe or scanning electron microscopy; while these techniques yield excellent resolution they are not without their problems. Firstly the systems can be expensive to buy and maintain compared to optical microscopes. Optical techniques are non-contact and are therefore non-destructive. The other techniques are either contact techniques or in the case of scanning electron microscopy the beam powers required are high and so damage can be caused to the sample under observation unless great care is taken. The other techniques require operation by highly skilled users, as these systems are difficult to set up and operate relative to optical systems. They are not suitable for as wide a range of samples as compared to optical systems and often require special sample preparation before they can be imaged.

1.2 Standard measurements

The importance of measurement standards has been recognised for thousands of years. The royal court architects responsible for building the pyramids 5000 years ago faced the death penalty if they forgot or neglected to calibrate their standard unit of length against the Royal Cubit Master at each full moon [14]. Throughout history every civilisation has developed its own measurement systems and calibration processes. More recently, International standardisation began in the electro-technical field in 1906 with the International Electrotechnical Commission, followed in 1926

[15] by International Federation of the National Standardising Associations in the mechanical field [15]. By 1946 25 countries had created International Standards Organisation that began operation in 1947 with the aim of *'facilitating the international coordination and unification of industrial standards'*[15].

Standards are very important for both industry and consumers as 'International standards provide a reference framework or a common technological language betweens suppliers and their customers' [16]. Standards must be traceable to some physical value, which is measurement independent. For example the definition of the meter is the length of the path travelled by light in vacuum during a time interval of 1/299792458 of a second. The speed of light is 299792458 m/s, the second is determined to an uncertainty of 1 part in 10^{14} by the Caesium clock. The iodine stabilised helium-neon laser is the recommended light source for realising the meter, its wavelength is 632.99139822nm. These values are all governed by the underlying physical process that governs the clock and the speed of light in a vacuum which is a universal constant.

This thesis is concerned with the measurement of very fine structures, many times smaller than the point spread function of the optical system. One important application is being able to provide calibrated linewidth standards to industry so this will be used as an illustrative example of the importance of measurement standards.

For semiconductor components, the ability to create precise structures in silicon substrates is crucial. During the fabrication process, for example, a silicon substrate coated with photo resist will be exposed to light through a mask of the circuit. The

exposed material will then be etched away before subsequent processing of the wafer. This fabrication process needs to be monitored to make sure that when, for example, a 350nm structure is created on a substrate its dimensions are correct. The quality control systems that provide this service will produce an answer regarding the dimension of the structure, for example, 365nm width 40 nm high. This in itself does not tell us much as we need to know if the quality control system has been calibrated and to what standard. What is the uncertainty of the quality system? This calibration process is crucial as the actual value for this track could be considerably different from the measured value, which could spell disaster for the components being created. The calibration process could be carried out in a number of ways but the simplest would be to measure a calibrated linewidth standard.

These standards usually provide different structures of varying width values that have the exact traceable widths and the uncertainty of the values supplied with them. This sample is then measured with the quality control system, so that the quality system can be calibrated to a known standard. Then the unknown sample can be measured and its width value established. This will now be traceable to the instrument that measured it, which in turn is traceable to the measured standard sample, which is traceable to the meter through the standards process.

Standards also provide a common language for different groups in an industry to communicate, when one body states the parameters of an object all other parties know exactly what is meant as the standards process defines the terms that the industry uses.

Being able to provide calibrated standards is therefore a vitally important service to industry. Standards institutions such as the National Physical Laboratory in the UK spend a lot of time producing calibrated samples for industry as well as developing new ways to provide these measurements. An important part of the traceability process is the development of an uncertainty budget for a measurement process and so each calibrated sample is provided with a set of quoted parameter values and their associated uncertainty, as measured with that particular system. The uncertainty takes into account any sources of noise or errors that could affect the measurement, data acquisition or signal processing.

A method to quickly and efficiently measure small track structures would be of great interest to NPL, as it is becoming more difficult to keep up with the technological advances of this industry.

However this is becoming more difficult to achieve due to the technological advances of this industry in recent years. In 1974 the typical feature size for semi-conductor components was 6 microns, in 1985 it had reduced to 1.5 microns and in 1993 features sizes were sub-micrometer at 800nm. Over recent years this has rapidly reduced and is now at 90nm for cutting edge processes with 65nm processes not far away. For these processes to operate we need to be able to supply calibrated standards of at least these dimensions and ideally smaller. As mentioned earlier this is very difficult, if not impossible, to achieve with a purely optical system. Because of the advantages that optical systems possess is there anything we can do to improve the optical microscope to enable the measurements we require?

Others [17][18][19][20] have attempted to address this problem by returning to the underlying mathematics that is used to model the interaction between the imaging system and the object itself. They have attempted to reverse the effect of the optical system and retrieve the spectral components lost in the imaging process. This allows the construction of a super-resolved[1] image and enables much smaller object features to be observed with a conventional microscope. This process works to some extent for samples that are simple and well behaved where the optical system is perfect and noiseless, but has serious problems when real, noisy data is used so in practice the resolution enhancement achieved is small. For these techniques to work they usually require detailed knowledge of the response of the optical system, which is often not available.

We have chosen to take a different approach to this problem by attempting to directly measure parameters of objects below the classical resolution limit without trying to increase the system bandwidth. We firstly pose the question: is the measurement limitation of an optical system due to our inability to detect and/or recognise the changes between two images and the differences in the corresponding objects? Can we design a system that can detect the changes and know how these changes relate to the object that was measured thus? This is essentially what this thesis sets out to achieve.

We firstly limit the objects under investigation to track structures as our main application is providing calibrated track width samples to industry. These tracks have

[1] Where *super-resolved* here means that the bandwidth of the final image is greater than the bandwidth of the optical system used to obtain the original image.

various parameters associated with them such as width, height and separation etc. We can then measure a whole range of known objects to build up a set of images. The objects measured are below the classical resolution limit of the optical system. The images are then used to train an artificial neural network that learns how the images are changing as the known object parameters vary. If we then measure another, unknown object, we can apply it to the trained network and find out structural information about this object.

The technique requires the use of a good optical system to measure the small track objects. Even though the objects measured are below the resolution limit the information regarding the objects parameters are contained within the measured profile and using an artificial neural network is one way to extract this information.

It is important to note that this technique is not providing an increase in resolution as the images obtained by the system are still diffraction limited, but the *measurement* range is increased to include the parameters of objects that are well below the conventional limits of the optical system thus providing a way to *measure* very small track structure parameters optically.

1.3 Thesis layout

This thesis is set out as follows**:** Chapter 2 discusses the way in which the track width is measured in current systems and discusses the possibility of track width measurement below the resolution limit with different noise levels.

Chapter 3 introduces relevant research carried out in the field. Firstly discussing optical systems used to provide linewidth measurement as well as commercial and research profilometers that could be used to obtain the optical profiles. The basis of the spectral extension techniques is discussed and several important implementations are presented. Finally the development of ANN and their applications is discussed.

Chapter 4 introduces the ANN in more detail, discussing all of the important factors to obtain a working network. Simulations of single and double track objects are presented to demonstrate the ability of this technique

Chapter 5 presents the optical systems used for the experimental work. Repeatability and noise performance is given as well as considerations of complexity and ease of use of the systems.

Chapter 6 shows the experimental results for various optical systems and samples to demonstrate the practical ability of the technique.

Chapter 7 discusses future work to be carried out, from improvements in the ANNs and optical systems to the extension of the technique to obtain other parameters and the possibility of extracting profiles by using the technique in conjunction with other signal processing techniques.

Chapter 8 contains the conclusions that can be drawn from this thesis with regards to linewidth measurements and the ability of the technique to obtain parameter measurements on extremely small objects optically.

2 Background

The image spectrum obtained from measuring an object is the spectrum of the object modified by the optical system. The finite resolution of an optical microscope is due to the finite pass band of the optical system truncating the spectral components of the object. If the significant spectral components of the object are within the system bandwidth the object parameters can be measured. If they lie outside the pass band then they cannot. This is demonstrated by Figure 4, which shows the image and spectra for two objects (a 10 micron track and a 1 micron track). For the larger object (a) it is easy to measure the track width as the significant components of the object spectra are within the pass band (dotted line) (c). For the small object (b) this is not the case, all of the side lobes of the object lie outside the pass band and are truncated (d) and so the track width cannot be easily measured.

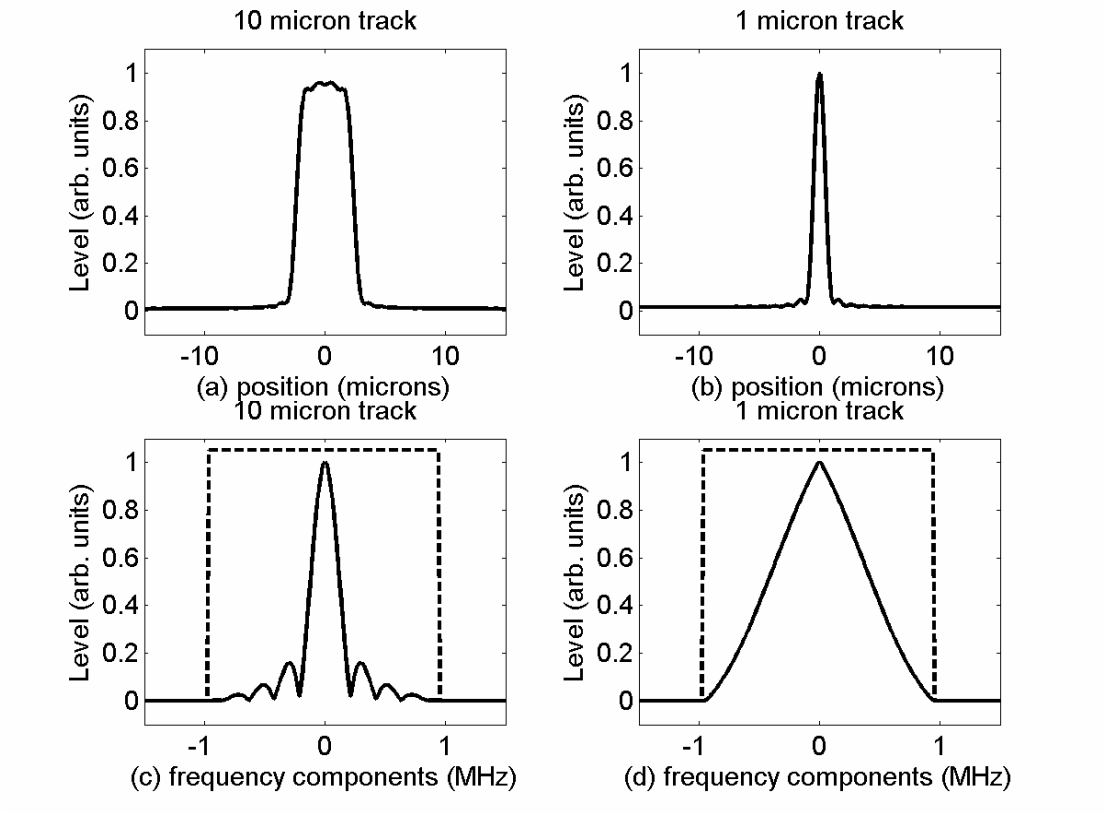


Figure 4 - Effect of optical system on resolution

As track width structures approach the resolution limit the relationship between the width of the image and the actual width is nonlinear, the change in measured width for a given change in actual width reduces, and this change in sensitivity makes it increasingly difficult to measure the track width accurately.

2.1 Optical model

The simulations presented were obtained from a simple scanning microscope model. In this model the PSF due to the system objective is scanned point by point across the object of interest. At each scan location the reflected light is captured by a number of different detectors (Figure 5). The object could be purely reflective (amplitude) or a phase object or mixed. This allows objects such as silicon substrates with tracks etched into them or glass on metal samples to be modelled.

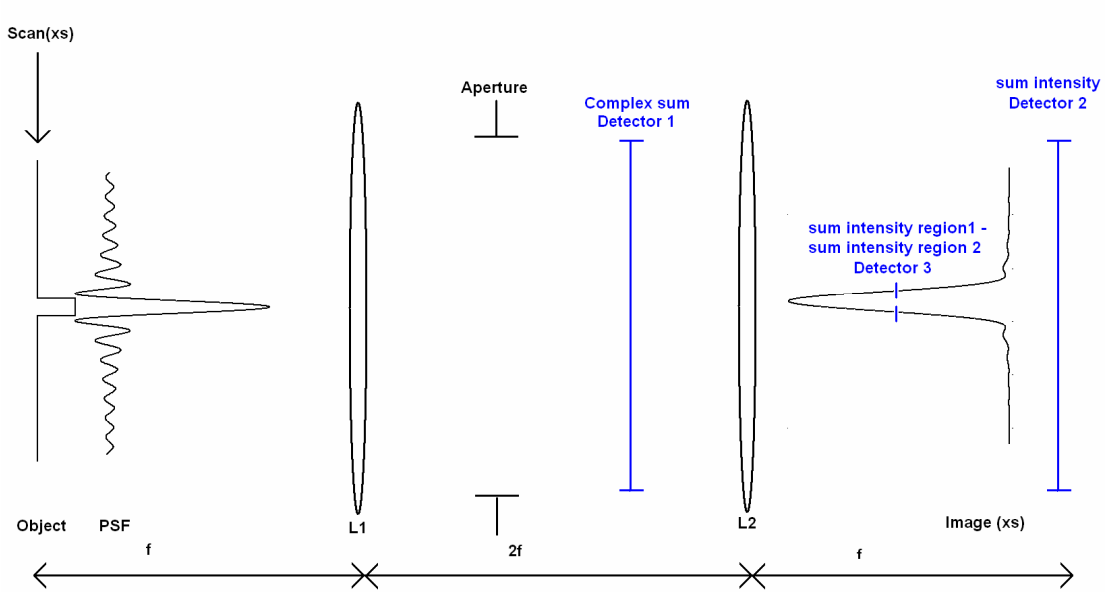


Figure 5 - Schematic of optical model

Detector one: this is at the back focal plane of the objective and is the complex sum of all the Fourier components passed by the system. This detector returns both the

amplitude and phase response of the object. This detector is used when phase profiles are used as the inputs to the signal processing and ANN.

Detector two is in the image plane and returns the intensity at each scan point by summing the absolute of the image of the PSF squared.

Detector 3 is the same as detector two except it uses two small regions for the summation. The difference between these two summed regions forms the signal. These regions are symmetrically offset from the optical axis. This is shown in Figure 6.

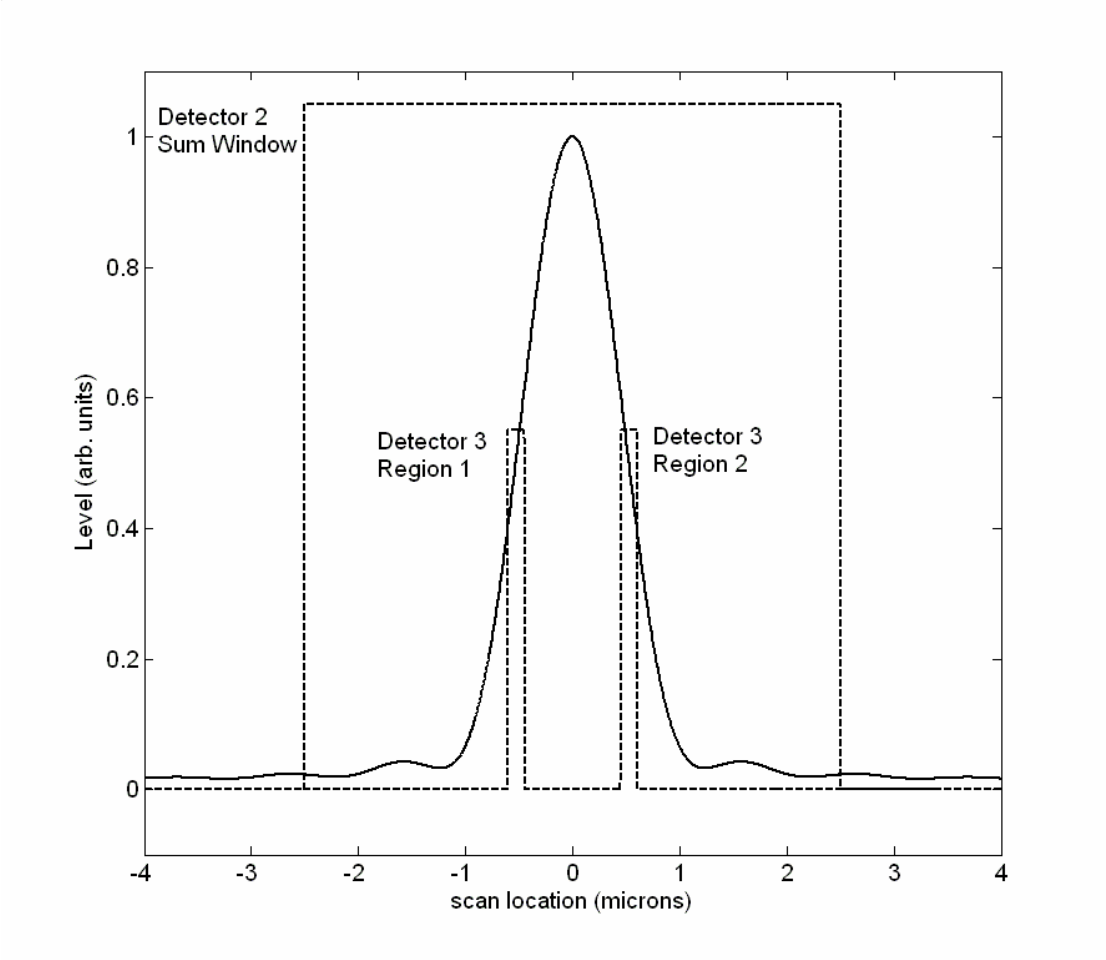


Figure 6 - Summation regions for detector 2 & 3

Noise is added to the profiles after they have been measured. This was to reduce computation time as the sets of tracks took a long time to compute. Photon noise was added by taking the profile of interest and scaling the maximum value to correspond to the maximum value of photons. The square root of the value of a pixel is taken and multiplied by a random number, whose distribution is normal, with zero mean and standard deviation of one. This value is then added to the original pixel value. This is repeated for each pixel in the image. If the phase profile was used then additive phase noise was added instead, where the standard deviation of the phase noise was scaled to correspond to the equivalent level of photon noise used for the amplitude/intensity case (see appendix 2 for more detail).

The signals used for the networks tend to be differential signals as explained later in chapter 4, this means that the signal from detector 3 can be used directly, but for the other detectors the signal needs to be differentiated. This is achieved by taking the difference between the profile and a shifted version of the profile. The shift is performed using the Fourier shift theorem so that the shift distance can be controlled precisely. This process is shown in the equations below.

$$\begin{aligned} P(f) &= \Im\{p(x)\} \\ P_d(f) &= P(f)[1-\exp(-i2\pi fd)] \\ p_d(x) &= \Im^{-1}\{P_d(f)\} \end{aligned}$$ Equation 2-1

Where d is the amount of shift between the two profiles and $\Im$ is the Fourier transform. The differential profile ($p_d(x)$)is then processed and used as the input into the system.

2.2 Calculating line width

There are many methods for deriving the width of a structure from the optical profile, two of the techniques used by the National Physical Laboratory (NPL) for commercial linewidth measurement service will be discussed here. The first is a thresholding technique that is based on the positions of the 25% total intensity positions. The 100% level is obtained by measuring a wide structure and using the value in the central flat region (marked in figure 2). The choice of the threshold value used is a critical one for calculating the linewidth. The level is chosen as according to scalar theory that states that the 25% points correspond to the position of an abrupt edge of an opaque film when spatially coherent illumination is used.[21] [22]

An optical model is required to relate the values at the 25% intensity levels to the actual linewidth. As the model is based on scalar theory it is only valid for a low NA system and coherent illumination In general the optical system used for measuring linewidth will be operating at high NA and with partially coherent illumination.

The second method is based on the total transmitted intensity. It is a measure of the area under the intensity profile and is normalised by the intensity range, as shown in Figure 7. This is a valid method of calculating linewidth, as the amount of light transmitted through a clear line is proportional to the width of that line. This relationship is not perfectly linear due to the effects of diffraction and so again an optical model is required to relate the measured area with the actual linewidth.

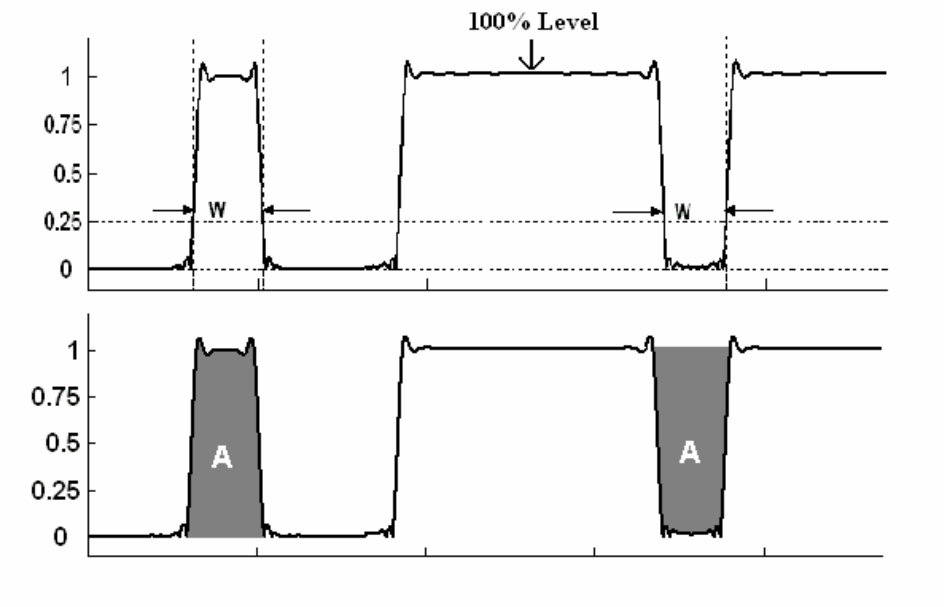


Figure 7 - 25% threshold and area method

This method has two main advantages over the thresholding method. Namely it is less sensitive to focusing errors and can still provide linewidth measurements for narrow objects where the 25% threshold level is never reached. The draw back of this technique is that the method is sensitive to any changes in the overall intensity level so if the intensity is known to have +/-x% error then the linewidth value will also have this error. This means that the threshold method is more suited to larger linewidths and the area method is best suited to smaller structures. In practice anything smaller than 2 microns is measured using the area method.

Figure 8 shows simulated intensity profiles for a chrome on glass object of various widths from 17-0.22 microns, where the optical system had a NA 0.3. The tail off in resolution occurs because the image is a convolution of the actual object with the point spread function of the optical system. Therefore as tracks get increasingly small,

the variation in the image reduces as the images tend towards the point spread function of the system.

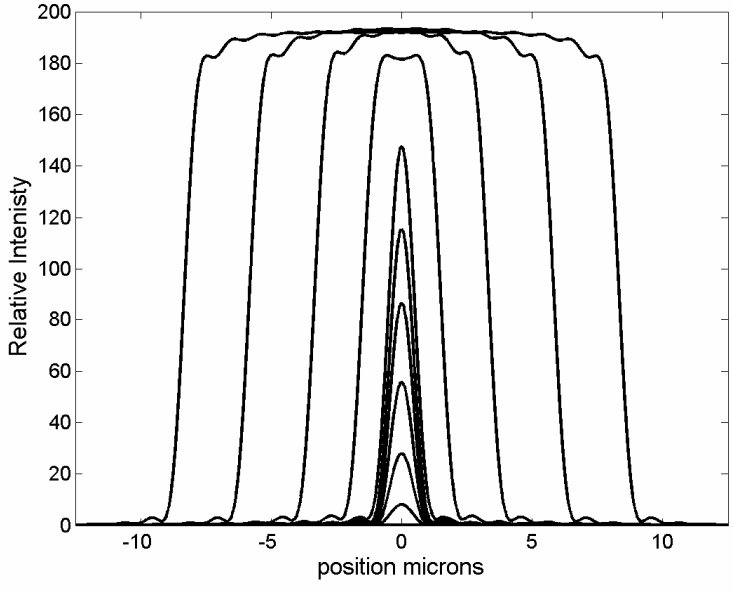


Figure 8 - Intensity profiles as track width reduces from 17 – 0.220 microns

The two methods to calculate the width of an object were used on a set of simulated tracks from 20 microns to 100nm where the NA of the simulated optical system was 0.3. Figure 9 shows the results from using the 25% threshold method for a set of noiseless tracks. This method breaks down for tracks below 750 nm as these tracks no longer reach the 25% threshold level. There is a fairly constant gradient between the actual and measured value, which could be easily removed by fitting a line to the tracks up to approximately 1.5 microns. Figure 10 shows the results from using the total area method. The global scaling factor was calculated by scaling the area value for the 20-micron track to coincide with 20 microns. As can be seen this is acceptable as it produces good agreement between the actual and measured value for the range 3-20 microns. Below 2 microns the graph starts to tail off as expected.

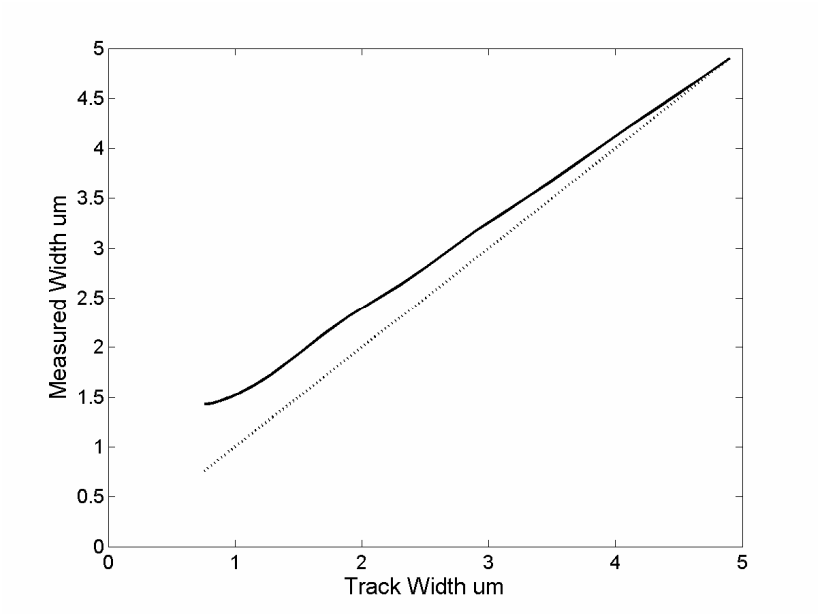


Figure 9 - 25% intensity method

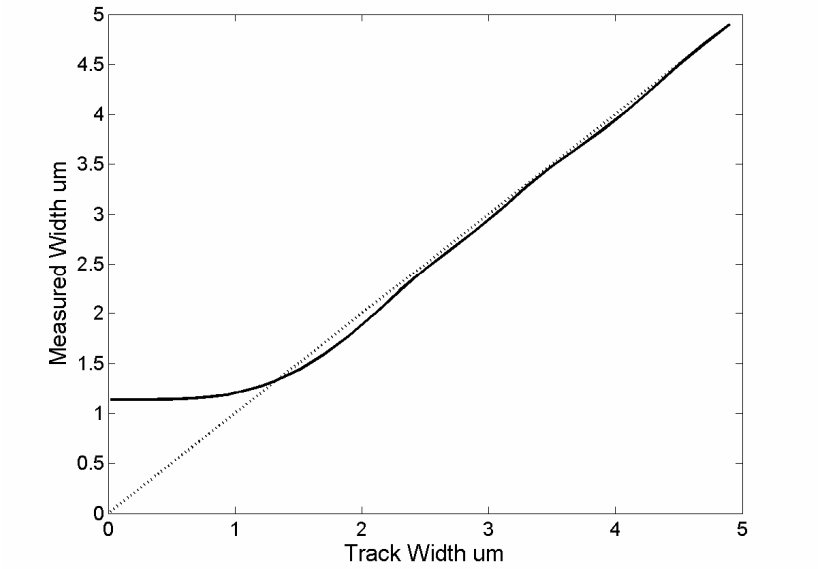


Figure 10 – Normalised total area method

The impact of noise on the two methods has also been investigated. Intuitively one would expect the 25% method to perform poorly as it only uses 2 points out of all of the available data to calculate the width value, whereas the area method uses all of the

available data and will therefore have a lower sensitivity to noise as many more photons are used to derive the width value.

Figure 11-Figure 12 show the results where the peak value of a large track is 100 000 or 10 million photons and the noise is from photon noise only. The range has been reduced to 100nm-1000nm so that the error bars are visible. The error bars are clearly visible in figure 6 where the noise level is high. For the other figure the error bars are harder to see as the noise is having less effect on the measurement of the width value.

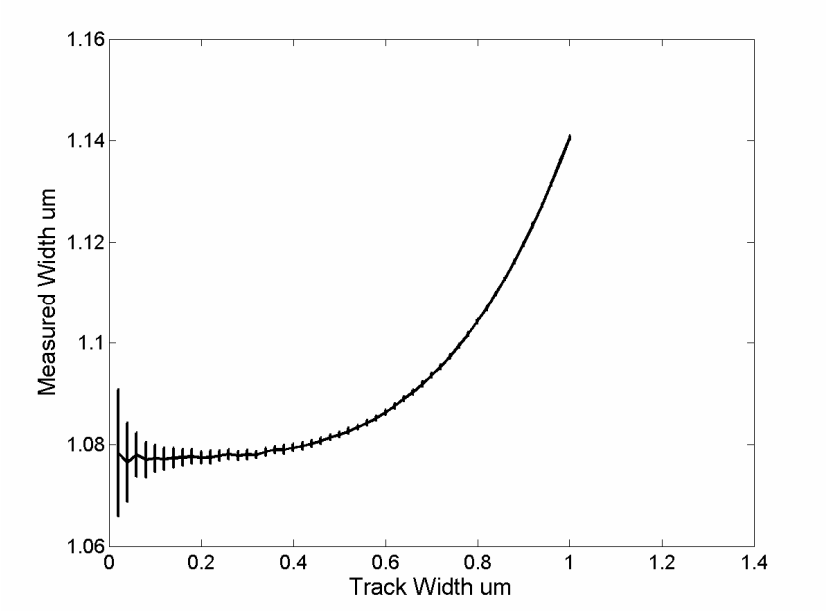


Figure 11 – Close up of smallest tracks (100 000 photons)

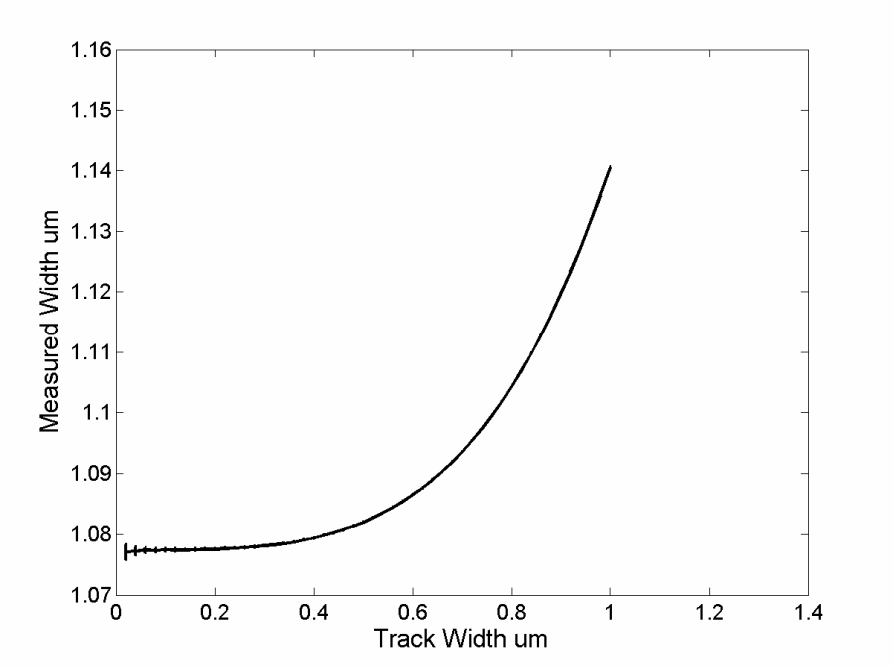


Figure 12 Close up of smallest tracks (10 million photons)

The actual number of photons used to calculate the width in the 25% threshold case is 50 000 and 5 million (25% of the number of photons at the 100% level x2). Whereas for the area method the total number of photons used to calculate the track width is approximately 13000 and 1.3million respectively (for a 1 micron track), however for this case the wider the track the more photons there are unlike the previous case where the number of photons was independent of track width. (There are approximately 500 pixels across the extent of the smallest track width (x increment is 5nm) and 8192 points were summed in total.)

The actual number of photons used to provide the width value for the two methods for different maximum photon levels is given in Table 1and 2.

Table 1 Photons and predicted SNR for area method

Peak photons per pixel	**1e+5**	**1e+6**	**1e+7**	**1e+8**	**1e+9**	**1e+10**	**1e+11**	**1e+12**
SNR 100 nm	13.589	42.973	135.89	429.73	1358.9	4297.3	13589	42973
SNR 1100 nm	121.6	384.7	1216.5	3847	12165	38470	121655	384707

Table 2 Photons and predicted SNR for 25% threshold method

Peak photons per pixel	**1e+5**	**1e+6**	**1e+7**	**1e+8**	**1e+9**	**1e+10**	**1e+11**	**1e+12**
SNR	223.61	707.11	2236.1	7071.1	22361	70711	2.23e+5	7.07e+5

The uncertainty of measuring any specific track width can be calculated for the different noise levels used. This is achieved by taking the standard deviation at a specific track value and calculating the change in actual width for the change in measured width. This was done for the two methods, for several track values and noise levels. The results from this are tabulated in Table 3 and Table 4.

Table 3 - Uncertainty (nm) of 25% method for different track widths and maximum photons

Width µm \ Peak photons	**1e+5**	**1e+6**	**1e+7**	**1e+8**	**1e+9**	**1e+10**	**1e+11**	**1e+12**

0.84	26.0	7.4	3.2	0.75	0.26	0.071	0.024	0.0078
0.94	9.5	3.8	1.20	0.42	0.11	0.046	0.013	0.0038
1.04	8.4	2.6	0.68	0.28	0.076	0.024	0.0080	0.0024
1.5	7.0	2.9	0.53	0.19	0.055	0.018	0.0064	0.0020
2.5	5.3	2.0	0.50	0.16	0.053	0.019	0.0060	0.0018
4.5	5.1	1.82	0.54	0.15	0.059	0.017	0.0054	0.0018

Table 4 Uncertainty (nm) of max intensity method for different track widths and maximum photons

Width µm \ Peak photons	**1e+5**	**1e+6**	**1e+7**	**1e+8**	**1e+9**	**1e+10**	**1e+11**	**1e+12**
0.12	353	111	35	11.17	4.01	1.22	0.32	0.12
0.22	126	38	10.8	3.76	1.25	0.35	0.10	0.026
0.32	30	13.7	4.63	1.40	0.40	0.15	0.046	0.014
0.42	17.2	6.3	2.24	0.62	0.20	0.074	0.023	0.0069
0.52	13.8	3.92	1.16	0.30	0.11	0.044	0.012	0.0036
0.62	7.5	2.67	0.67	0.22	0.068	0.025	0.0076	0.0023
0.84	3.05	0.95	0.37	0.077	0.028	0.012	0.0037	0.00084
1.04	1.39	0.56	0.16	0.070	0.017	0.0064	0.0021	0.00053
3.5	0.80	0.28	0.078	0.030	0.0099	0.0029	0.0010	0.00025

Table 3 shows that the measured value of the track width for the 25% threshold method is affected by the noise level more than the area method as is expected. With 10^8 photons per pixel the error standard deviation is around 1 nm for tracks 840nm and above.

The Area method is less sensitive to the noise level for example using 10^8 photons for the 840nm track the error standard deviation is 0.0077nm for the area method as

opposed to 0.75nm for the 25% method. This is because the area method has an averaging effect on the noise in the signal. Taking measurements with 10 billion photons will produce nanometre uncertainty across the range of 120nm-3500nm track widths.

This shows that even though the classical resolution limit of this optical system is around 1.4 microns it is possible to measure track widths down to 120nm with nanometre uncertainty as long as the signal to noise ratio is high enough and suitable corrections can be applied to the measured track width value. These corrections are usually provided by modelling of the optical system and comparison between the known object profiles and the modelled profiles as for the OPTIMM system used by NPL discussed in chapter 3 section 1.

One of the problems of modelling the optical system is that the model for the system will undoubtedly differ from the real system due to many factors, including, alignment conditions, aberrations and component tolerances. This mismatch between the model and the real system could lead to large errors in the width assigned to each profile. A model for every optical system that may be used for measurements of different sample types would also be required.

Ideally the track width measurement will be independent of the optical system used to make the measurement, allowing different systems to be used for measuring different samples. This will give maximum flexibility to the system as some samples will be purely phase samples e.g. an etched silicon surface, other samples may be chrome on glass and therefore provide intensity profiles. Having a system that increases the

measurement capability but is independent of the optical system is therefore very desirable.

The method we have employed is to train an artificial neural network on input values derived from optical profiles obtained from a known sample of track structures. Once trained, the network can then provide a value for the width of any other track structures in the specified range for that network. This method has the advantage that no model of the optical system is required, however the system has to be trained using a set of track widths calibrated by some other method.

This technique works because the input profiles presented to the network contain information relating to the output targets namely the track width in this case.

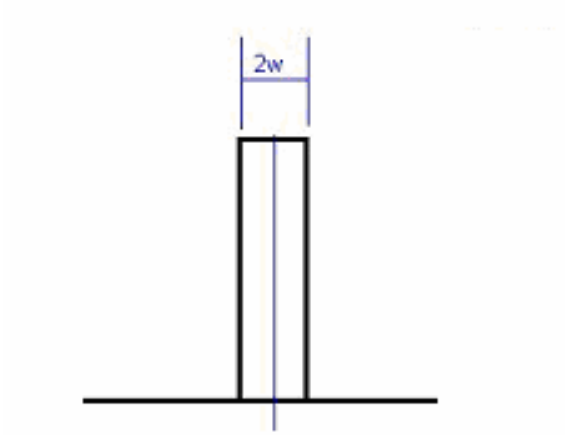


Figure 13 - A simple single track object

In the simple case of a single-track object shown in Figure 13, how the spectrum of the object changes with track width is given below.

$$A = rect\left(\frac{x}{w}\right)$$ Equation 2-2

$$\Im\{A\} = \Im\left\{rect\left(\frac{x}{w}\right)\right\} = w\text{sinc}(f_x w) \qquad \text{Equation 2-3}$$

This shows how the spectrum of the object will vary as the track width *w* decreases. It is this relationship between spectrum shape and track width that the ANN will learn.

This technique is not limited to just single track structures and single parameters, in theory all sorts of information are contained within the spectrum of the profile, for example, for a single track, information regarding the height, the angle of the side walls as well as the width of the structure. For double track structures the width, separation and other parameters should be possible to measure.

The next chapter describes current systems for linewidth measurement and other optical systems suitable for this task. It then goes onto discuss the signal processing techniques that others have applied to attempt to overcome the diffraction limit.

3 Literature review

This chapter examines work related to line width measurement and super resolution. Firstly optical systems are discussed, starting with the requirements of optical systems for these types of measurements before going on to discuss a number of different systems used both commercially and in the research environment. Signal processing techniques and their basis are then discussed in some detail. Information theory and the reasons for the limitations of previous approaches to this problem are described. Finally, a brief overview of the development and applications of ANNs is given, demonstrating their wide-ranging use for many different types of applications.

3.1 Optical systems

Many different types of optical systems are suitable for providing profiles for line width measurement. The main criteria for selecting a suitable system covers many aspects such as, the size of the features to be measured, the types of samples that need to be measured (phase or amplitude). In an ideal world the optical system would provide:

- Precise profiles
- High SNR
- High repeatability
- Low sensitivity to environmental conditions vibrations etc
- Be capable of measuring phase structures
- High lateral resolution
- Easy to use / setup

Not all systems can provide all of the above features. The first system discussed is an actual system used to provide standard measurements at the NPL as part of their measurement service, secondly commercial systems that can be used for linewidth measurement are described before an example of a research instrument is given.

3.1.1 Linewidth measurement systems

OPTIMM - NPL linewidth measurement service

OPTIMM [23] is the current system used by NPL as part of the measurement service for providing calibrated linewidth standards in the form of the BCR standard. The system consists of an optical microscope to obtain profiles of the structures. The linewidth is generated after some processing and use of optical modelling techniques.

The optical system shown schematically below in Figure 14

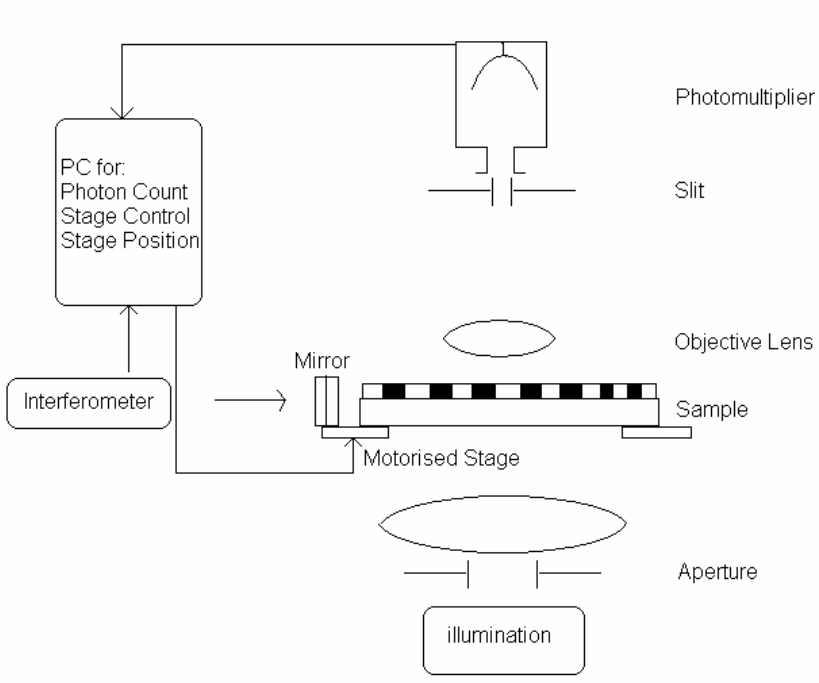


Figure 14 Schematic of the optical setup

The sample is illuminated from the underside. As the lines are transparent the light passes through the sample and is focused by the objective lens onto a slit in the image plane. The light that passes through the slit is collected by a photo multiplier and recorded by the computer. The intensity profile of the object is generated by moving the object so that the image of the structure is scanned across the slit. The system is not 100% confocal as a slit is used as opposed to a pinhole but for line structures this is acceptable. Also the slit reduces background light, which is important for any measurement system.

The scanning location is recorded by an interferometer; this along with the photomultiplier signal produces images of the linewidth under investigation. These images are then processed and with the aid of optical models, the width of the structure is calculated.

The shape of the intensity distribution is dependent upon the object dimensions, the wavelength used, the width of the slit, the NA of the objective lens as well as the NA of the illuminating condenser lens. NAs typically used are 0.1-0.6 for the condenser, 0.9 for the objective. The slit is usually between 80-90nm (actual) and the wavelength is from 502-572nm.

OPTIMM requires the use of optical models to provide corrections to the widths obtained by the threshold and area methods. The models need to take into account the parameters of the optical microscope used to measure the intensity profiles. If any

microscope parameter is changed new models need to be generated to provide the adjustments to the linewidth values.

The main problem with the OPTIMM system is that the types of samples that can be measured are very restrictive. Only samples with opaque sections can be measured. Phase objects are therefore not measurable with the current instrument configuration. The accuracy of the method is dependent upon the model and how well the system is aligned/setup so that the corrections applied are correct. If various operating conditions are required, for example using various different NA objectives then the models and calibration procedure need to be repeated for these new operating conditions. The lower limit of sample that can be measured is approximately 300nm. This is partly due to the technique but also due to the types of sample used. The samples are usually chrome on glass samples and if the track widths get much smaller than 200nm they tracks start to come off the glass so a safe limit of around 300nm is chosen

3.1.2 Commercial profilometers

Two commercial profilometers that could be used to provide calibrated linewidths will now be discussed.

Olympus – Confocal Laser Scanning Microscope - LEXT

LEXT [24]is a commercial optical system suitable for many imaging applications. It has several different operation modes such as confocal, darkfield and DIC Nomarski all as scanning modes using laser illumination. It can also provide color images by illumination by a white light halogen source for wide field images.

The system has a large sample area and like most microscopes it does not require special sample preparation unlike SEM systems. The sample is positioned by xyz translation stages.

The repeatability of the system is important if the system is to be used for linewidth measurements, depth measurement and surface roughness etc. The repeatability is quoted as 0.002Lμm (L=measurement length). The light source used has a wavelength of 408nm and the lateral resolution of the microscope is quoted as 0.12microns (from measurement of a 120nm on 120nm off grating of height 0.01microns). The reliability of the measured data is traceable to international standards set out by Physikalisch-Technische Bundesanstalt (PTB), Japan Quality Assurance organization and the United Kingdom Accreditation Service. Because the system is traceable it would be possible to use it to provide calibrated linewidth standards.

ZYGO Newview 6300

The Newview 6300 [25] is a commercial microscope provided by Zygo for the applications of measuring surface roughness, step heights and critical dimensions etc. Like most optical microscopes it provides fast, non-contact measurements. It uses white light interferometry utilising an LED. The interferometer works as a Mirau or Michelson type and uses objective lenses with an internal beam splitter and reference path that matches the optical path length. The system provides sub nanometre *z* resolution (0.1nm) with sample step heights up to 15mm. The lateral resolution and field of view is dependent upon which objective is used from the objective turret. The best lateral resolution is quoted as 450nm. The step height accuracy is quoted as

<0.75% and the repeatability is better than 0.1%. The system comes with a software package to provide 2d and 3d imaging and image analysis. All systems are certified and traceable to NIST standards.

This system is capable of providing many different types of measurements on different types of samples.

3.1.3 Research Based Profilometers

There are many different types of optical profilometers [26 27 28 29 30 31]for measuring a variety of samples. A heterodyne differential interferometer [32] will be discussed in some detail as it demonstrates several important features of different approaches found in many profilometers.

In this case two beans are produced on the sample surface by passing a collimated laser beam through a Bragg cell. This splits the beam into the zero and first order (either +1 or –1). The zero and first order beams are then focused onto the sample surface by the objective lens. The Bragg cell also imposes a frequency shift to the first order beam of f_2. By modulating the Bragg cell drive signal the beam are amplitude modulated in anti-phase at frequency f_s. Upon reflection from the sample surface the beams are recombined by the Bragg cell and interfere. The signal is captured with a photodiode.

The differential amplitude can be expressed as:

$$A\left(r_1^2 - r_2^2\right)\cos\left(2\omega_s t\right) \qquad \text{Equation 3-1}$$

and is located at $2f_s$. Where A , B and δϕ are constants, r_1 and r_2 are the reflectivities experienced by each beam at the sample surface, $\delta\vartheta$ is the phase difference between the returning beams due to the object. The differential phase is expressed as

$$Br_1r_2\cos\left[2(\omega_1-\omega_2)t+\delta\vartheta+\delta\phi\right] \qquad \text{Equation 3-2}$$

And located at $2f_2$

The system is simultaneously able to measure differential phase and amplitude. By using a heterodyne approach the signal processing is much easier as phase stepping is not required. The differential nature and common path arrangement means that microphonics are reduced substantially.

3.1.4 Other optical approaches

Another set of microscopes are concerned with modifying the point-spread function so that it is smaller than the diffraction limited spot size and so provide improved lateral and axial measurement precision.

The microscopes either employ superposition of beams or use an optical mask to shape the point spread function . The aim is to reduce the FWHM of the PSF and improve lateral resolution. [[33] [34] [35]]

These techniques while interesting have their problems; firstly they waste a lot of light, the effect of the side lobes is increased and the optical transfer function of the system is distorted. Most importantly, these techniques do not increase the overall system bandwidth. They do not provide a large increase in measurement precision and would still not enable the measurement of 100nm tracks optically.

3.1.5 Contact/ non optical systems

There are several different types of non-optical measurement systems capable of measuring very small features. As they are not limited by the nature of light the resolution of these systems can be extremely high. Atomic force microscopes are very common and can provide measurement resolutions in the nm range. The measurement signal is generated by monitoring the Van der Waals force between an ultra fine probe tip on the end of a cantilever and the sample. As the sample changes during a scan the forces on the tip change and causes the cantilever to move and so a measure of the object surface is obtained. The cantilever angle is usually monitored with a laser beam, which is reflected from the cantilever to a position detector. As the cantilever moves the deflected beam moves at the detector and the signal is recorded to produce and image of the surface. The main issues with the AFM regard the relatively small image areas that can be measured (100μmx100μm), the small maximum sample step height of around 1μm and the slow scanning speed. The type of tip used also plays an important role in the profile obtained by the AFM; this effect needs to be taken into account if the AFM is used for critical dimension measurements.

Scanning Electron Microscopy (SEM) uses a focused beam of electrons ranging in energy from a few hundred eV to 50keV. The electron beam is then focused to a small spot on the sample this spot can range from approximately 1nm to 5 microns. Where the electrons hit the sample an area of interaction exists and depending on the energy of the beam various different measurements can be made. The resolution of the SEM depends on the size of the volume of interaction as well as the size of the electron spot but is usually in the range of approximately 1-20nm and is very dependant on the

sample being measured. The areas that can be measured are relatively large and many sample types can be imaged compared to other forms of electron microscopy. The main drawbacks to this technique are that the resolution is sample dependant, the beam powers used can also damage samples, they are difficult to use and are generally expensive systems to buy and maintain.

Near field optical techniques can also provide enhanced resolution over far field optical techniques. They work by perturbing the evanescent field that exists in the near field of the sample light interaction. Measurements are made by scanning a very fine fibre tip (around 50nm) extremely closely to the sample surface (tens of nm). The tip perturbs the evanescent field and resolution now becomes a function of the tip/field interaction as opposed to the diffraction limit. Resolutions down to 50nm are routinely quoted. The problems of this technique arise mainly from the fibre tips that are used. The exact size and shape of the tip control the imaging response and resolution of the system. Unfortunately the tip properties vary considerably and as such repeatability and the imaging properties vary between two nominally identical tips. This makes standard measurements difficult to make, as the results are not traceable due to the variation in tips.

3.2 Spectral extension and information theory

The next section discusses the theory behind spectral extension techniques and also considers the spectral extension technique from the point of view of information content theory. Examples of techniques used to provide super resolution and their associated problems are discussed in some detail.

Reversing the effect of the optical system on the object has been the topic of much research in the past. The idea being that if the effect of the optical system can be removed perfect object restoration would be possible providing arbitrary resolution of objects.

In 1955 Toraldo di Francia published a paper showing that under some conditions two different objects would produce identical images[36] . This has implications for super resolving algorithms as without *a priori* information the correct object cannot be reconstructed. In 1963 Harris went on to relate the ambiguous image to its spectral components saying '*... objects can be distinguished one from another as long as the spatial frequency spectra of the two objects are not everywhere identical in the pass band of the optical system'*[17]. He went to propose that the fundamental limit due to diffraction on resolution must be because two or more objects produce identical images. Harris then showed that for a specific case two different objects will never produce the same image. This case is that the object must be bounded, which in practise is the case for most imaging systems. This means that the limitation shown by Toraldo di Francia does not apply to imaging systems where the object is bounded (essentially all practical systems). Harris' theory is underpinned by the fact that the Fourier transform of a bounded structure is analytic. Utilising the uniqueness theory and analytic continuation he showed that in theory arbitrary resolution is possible in a noiseless system.

3.2.1 Analytic continuation and the uniqueness theory.

An analytic function is a function that may be complex and that is infinitely differentiable. These functions have certain properties, which means that '*a function of a complex variable is determined throughout the entire z-plane from a knowledge of its properties within an arbitrary small region of analyticity'* [37][38]] . The uniqueness theorem states that if '*any two functions of a complex variable whose values coincide over an arbitrarily small region of analyticity must have identical values throughout their common region of analyticity and hence be identical'*. [37] These properties of analytic functions imply that if any part of an analytic function is known, then the entire function can also be determined as the known values can only belong to one specific function. This makes possible the reconstruction of the spatial frequencies outside of the pass band of the optical system. An extension on an analytic function will always produce the same answer; the extensions are not ambiguous.

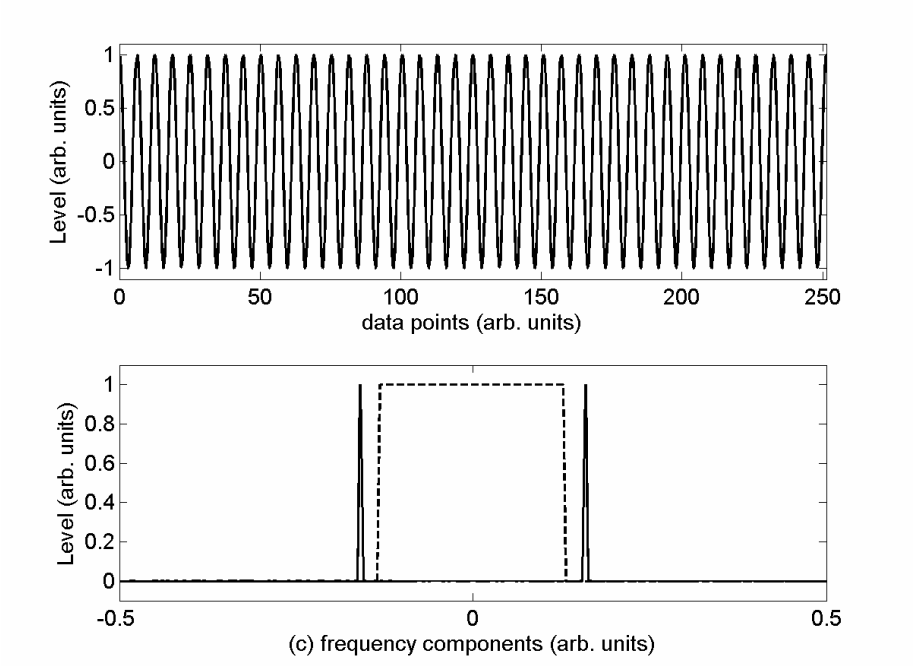


Figure 15 - grating structure (top) and spectrum (bottom)

The importance of the bounded structure property of the object under investigation is discussed here. The example below shows a grating structure whose frequency is outside of the pass band of the optical system. If the object is infinite in extent as in Figure 15 there is no information in the pass band (dashed line) and the spectrum is two delta functions at the grating frequency. As there is no information in the pass band of the system no extension can take place. This demonstrates that the Fourier transform of an object is not necessarily analytic and it is only when the object is truncated will analyticity be ensured.

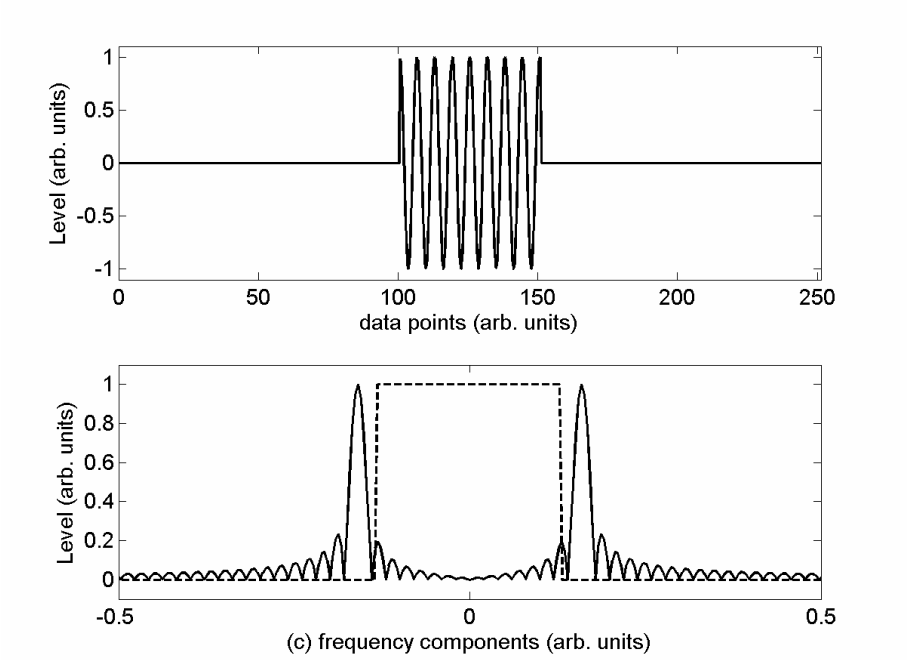


Figure 16 - truncated grating (top) and spectrum (bottom)

If the object is bounded then the spectrum of the grating is convolved with the spectrum of the truncating window. The convolving function is a sinc function, which can be expressed as a power series and by definition, is infinitely differentiable, hence analytic. Thus if the object is bounded as in Figure 16, the pass band of the optical system now contains information about the grating structure and because the spectrum of a bounded object is analytic the known spectrum in the pass band of the optical

system can, in theory, be extended by analytic continuation to acquire the entire spectrum of the object.

Analytic continuation by Taylor expansion

The function *f(x)* can be extended outside of its known range by, for example, a Taylor series expansion, as long as the function is analytic and it is precisely known over some arbitrary region then the function about point x_0 in this known region is:

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0) + \cdots + \frac{(x - x_0)^n}{n!} f^{(n)}(x_0)$$

Equation 3-3

The function $f(x)$ is now known and if enough orders are used then this function is valid for all *x* and can be used to obtain sufficient values of $f(x)$ outside the known extent of the signal.

The success of the extension depends greatly on how well the starting function is known. If there is any noise then the extension will become less and less accurate. This noise can come from inaccuracies in the measurement system but also from digitisation of the function, as using discrete levels introduces uncertainties in the actual value for each specific point. The differentiation in the Taylor expansion will also amplify noise. This will in practice greatly reduce the ability to perform an extension. Another problem arises due to the fact that the response is modified by the optical system transfer function. Decovolution by inversion filtering will magnify the random noise, or if a Wiener filter is used the function is modified and so any extension carried out on these modified signal would be incorrect.

The issues of noise in the system have been the major obstacle for this technique and several other methods [39]. These problems can be considered from the point of view of information theory and will be discussed briefly.

Information theory

The super resolution problem has also been considered from an information theory point of view. This is an attempt to try and establish the theoretical limits on extension and understand the influence of noise on extension techniques. In 1966 Lukosz [40 41 42] suggested an invariance theorem, which states that for an optical system the number of degrees of freedom is fixed not the spatial bandwidth. Cox and Sheppard [43 44] went on to develop this idea and took into account random noise in the system and its effect on the resolution improvement that was possible. The information capacity equation developed is shown in Equation 3-4.

$$N = (2L_zB_z + 1)(2L_yB_y + 1)(2L_xB_x + 1)(2TB_T + 1)\log(1 + s/n). \qquad \text{Equation 3-4}$$

Where N is the degrees of freedom, L_z L_x L_y are the extent of the field of view in the x,y,z directions, B_x B_z B_y are the spatial bandwidths in the x,y,z directions, B_T is the temporal bandwidth, T is the observation time, s is the signal level and n is the noise level.

This is a very useful equation as for any optical system it can be used to calculate the information capacity of the system and how this varies with signal to noise ratio. It should be pointed out that N is the theoretical maximum information capacity available from the system and so in practice the total information carried by the

system may be much less. For example, the object being measured by the system may not vary in time, although the system maybe able to carry time information no actual object information is delivered by the system. This does however mean that if the object is known *a priori* to be restricted in anyway, additional information may be encoded onto the independent and unused parameters of the system.

This invariance theory implies that the degrees of freedom N is fixed and so for example the spatial bandwidth Bx can be increased as long as there is a corresponding decrease in the other terms in the equation to keep N fixed.

Super resolution in optical microscopy was considered by Cox and Sheppard. They showed that the SNR of the super resolution image decreased as the spectrum was extended, if all other parameters were left the same. Using the above equations, for a specified SNR in the final image, the maximum increase in resolution could be calculated. This is because the only thing changing to keep N fixed in Equation 3-4 if the spatial bandwidth is being increased is the SNR of the signal. (The temporal bandwidth is also fixed in most cases.) This explains why there are practical problems with spectral extension methods based on analytic continuation as when they extend the spatial bandwidth the noise in the extended image increases dramatically and unless the original image has an extremely high SNR then the extension yields either very poor extension or very poor SNR in the final image.

Several techniques based on the idea of analytic continuation will now be considered.

3.3 Applications of spectrum extension theory

3.3.1 Sampling theorem in frequency domain [17]

An object of width W in the range +/-W/2 has a spectrum that is exactly determined for all frequencies by specifying the values of the spectrum at discrete frequencies separated by the interval 1/W, this series extends throughout the entire frequency domain. If we have an object of finite dimensions bounded by +/-X/2 then

$$N(x) = \sum_{N=-\infty}^{+\infty} G_n \exp\left[-i2\pi\left(\frac{nx}{X}\right)\right] \qquad \text{Equation 3-5}$$

And

$$G(f_x) = \frac{1}{X}\int_{-\frac{X}{2}}^{\frac{X}{2}} N(x)\exp\left[i2\pi\left(\frac{mx}{X}\right)\right]dx \qquad \text{Equation 3-6}$$

Sampling theorem establishes the relationship between the spectrum $G(f_x)$ and series coefficients G_n by substituting Equation 3-5 into Equation 3-6 and performing the integration yields:

$$G(f_x) = X\sum_{N=-\infty}^{+\infty} G_n \frac{\sin\left(\pi\left(\frac{m}{X} - f_x\right)X\right)}{\pi\left(\frac{m}{X} - f_x\right)X} \qquad \text{Equation 3-7}$$

The method for spectrum extension is to select a number of pairs of f_x and G_n's, and form a set of simultaneous equations that can be solved for G_n values. Once the G_n values are known the use of Equation 3-5 allows the reconstruction of the object.

The equations can be solved as follows.

Equation 3-7 can be re written as

$G_k = SG$ Equation 3-8

Where G_k is the (2N+1) column vector of known frequency components. G is the column vector (2M+1) of unknown coefficients of the G_n's. S is the (2N+1) x (2M+1) *sinc* function matrix.

Be rearranging the equations the unknown coefficients, which contain information beyond the cut off of the system, can be obtained from the known spectrum and *sinc* matrix S.

$G = S^{-1}G_k$ Equation 3-9

Harris presents an example of this, which shows a large increase in resolution, however this approach is only reliable in the absence of noise. Once noise is included in the system the uniqueness theorem no longer applies, as the spectrum of a function with random noise is not necessarily analytic. The limit to resolution extension via this technique is therefore a function of the noise level, and in this case the coefficients need to be known accurately (1 part in 10-1000 billion) so if the noise is greater than this, as is usual for practical signals, the technique breaks down completely. The method is very sensitive to noise as the matrix that is

inverted is near singular, which explains why extreme precision is demanded of the data

3.3.2 Auto-Regressive Models

An Auto regressive (AR) model is one where the current value of the function *x* is based on a linear combination of the previous N weighted values of the function *x*.

For example:

$$x_t = \sum_{i=1}^{N} c_i x_{t-1} \qquad \text{Equation 3-10}$$

where x_t is the series under investigation, c_i are the auto regressive coefficients, N is the order or length.

This type of model can be used to reconstruct the lost spectral components as the known part of the spectrum can be used to obtain the values of the spectrum outside of the bandwidth point by point. This will always yield the same extension in the absence of noise, as the function is analytic. The difficulty of this method is in finding the AR coefficients and this process will now be discussed.

Equation 3-8 can be rewritten in matrix form as follows:

$$X = Cx \qquad \text{Equation 3-11}$$

The AR coefficients are therefore:

$$C = x^{-1}X \qquad \text{Equation 3-12}$$

This can be applied to the spectral extension problem by dividing the known portion of the spectrum into two sets. This means that we know the ***x*** previous values by using the first set and the actual current values for *X* using the second set of known points. This allows the AR coefficients to be calculated. Once we have all of the coefficients for the known spectrum we can use these to calculate the spectrum

outside of the pass band point by point. After each new point the X value is incorporated into the model and its AR coefficient is calculated and the model updated.

Although simple to do in theory the problem becomes more demanding in practice. This is caused by the way the AR coefficients are calculated. Performing the inverse of matrix x is difficult as it is usually ill conditioned. This means that the usual method to calculate the inverse cannot be performed. The method used instead is the general inverse, which is often calculated by singular value decomposition (SVD). The problem with using the general inverse is that when round off errors or noise is involved, small errors in the elements of a matrix lead to larger errors in the general inverse. Performing SVD on the other hand reduces the impact of noise and enables an inverse to be calculated when these other methods fail.

A paper by Minami et al [45] used an auto regressive model with singular value decomposition to obtain super resolution spectra of Fourier Transform Infrared (FTIR) absorption data of benzene and cyclohexane. They showed an increase in spectral extension of 8 times. This agrees with our experience of using this method. Auto regression is particularly suitable to model a simple oscillatory function. However for more complex and non-periodic objects such as double track or triple track structures this method will not perform as well.

3.3.3 Gerchberg – super resolution through error energy reduction

Gerchberg [19] introduces the idea of an 'error energy'. The aim of his technique is to reduce the error energy and therefore increase the resolution beyond the diffraction limit. A pre-requisite for the method is that the extent of the sample must be known.

The error energy is defined as the difference between the energy of a measured function and the actual function. The method works as an iterative process between the spatial and spatial frequency domains. At each step the profile or spectrum is modified by the following rules:

- In spatial domain set all points outside of the known extent of the object to zero.
- In spatial frequency domain replace the spectral components in the pass band with the original spectral components. Leave all others unchanged.
- The error energy is the sum of the signal outside the known extent of the object in the spatial domain.

This process iterates until the change in the 'error energy' level reaches some predefined level. By constraining the spatial extent of the object, and retaining the original spectrum inside the system bandwidth, the total error energy can be shown to reduce for each iteration and will lead to an extension of the image spectrum.

The first thing to note is the conditions in the spatial domain. The extent of the object must be known *a priori* and this is central to the correct working of this method. If the incorrect value is used for this condition it greatly affects the results of the process.

Walker [46] used the Gerchberg algorithm to obtain a 2.5 fold increase in resolution of 3 square apertures imaged with a low Shannon number imaging system. (Where the Shannon number is the ratio of the object extent to the Rayleigh resolution distance). Three images were presented: a high resolution image of the apertures taken at the image plane, A low resolution image taken when a slit was placed in the Fourier plan effectively reducing the bandwidth of the system. With this slit in place the three objects were no longer resolved. A 1D slice from this image was used for the resolution enhance by the Gerchberg method. Firstly the image was square rooted and the phase was reconstructed for this image so that the amplitude profile could be obtained from the intensity profile. This amplitude profile was then subjected to the Gerchberg algorithm. After approximately 2000 iterations the three aperture object was clearly observable in the super resolved image.

The downfall of the Gerchberg method mainly stems from the amount of required knowledge of the sample of interest. The known extent of the object and how much of the extended spectrum is reliable need to be well defined otherwise the reconstruction is very unreliable. If the known extent is over estimated then the reconstructed image remain mainly unchanged, if the known extent is under estimated then the algorithm adds higher frequency components to squashes the data to fit this size.

Another problem with the technique is that there is no feedback as to how good the reconstruction is and how much it can be trusted. It may appear that the algorithm has managed to resolve two points but is the reconstruction correct?

The choice of the band limiting window also has consequences on how well this technique works. Using a top hat function produces a sinc response in the frequency domain and so the added in frequencies tend to follow a sinc like fashion. If the object being reconstructed is top hat like then good results can be obtained, but if the object is not then the reconstruction is poor.

The Gerchberg method can be difficult to use even when the exact form of the object is known so for objects where little information is available the situation is worse. In practical applications where the main aim is to perform a reconstruction to obtain object parameters this technique is becomes too unreliable.

3.3.4 Other Techniques

Several other people have also developed methods for providing super resolution. Barnes' [18] technique consists of trying to solve the imaging equation for the object. By using prolate spheroidal wave functions and the sinc function of PSF, he attempts to remove the effect of the optical system and restore the object information. Barnes presents results for noiseless systems with varying degrees for N, if N → inf then system response is a delta function (infinite resolution). The response of the system is sharpened in the known illumination region; however, outside the known illumination area the response grows to be many orders of magnitude greater than in the known region. In theory this is no problem, as there is no information outside of the illumination area, however practically, great care would need to be taken to exclude stray light from outside the illumination area otherwise the results would be hugely affected. The method has little practical value

when noise is contained in the system, as the errors in the reconstruction become very large beyond the diffraction limit.

Howard [20] developed a non-iterative method for spectral extension based on the Gerchberg approach. It uses the same idea of an additional unknown spectrum that is added to the original signal. The inverse transform of the combined signals produces a new image that is zero outside of the known extent of the object. The idea is that this function should be the negative of the distortion in the image. The method hangs on being able to calculate the coefficients of this signal. The coefficients were found by forming a Fourier series with cosine and sine terms corresponding to the known frequencies leaving the coefficients unknown. The goal is to minimise the squared error between this function and the negative of the distorted image in the regions outside of the true extent of the object. This yields a set of linear equations with unknown coefficients. Once the equations have been solved to obtain the coefficients the negative distortion function can be generated and added to the original spectrum, yielding the new spectrum. The final image can then be obtained by the inverse Fourier transform. This method is much faster then Gerchberg's iterative method but suffers all of the same problems associated with the 'known extent' of the object.

3.4 Artificial Neural Networks

This section briefly describes the development of artificial neural networks (ANNs) and gives examples of the type of different problems to which they have been applied.

The idea of using the brain as a model for computing was developed by Turing in 1936. In 1942 Wiener was formulating ideas about cybernetics, which were dealing

with the ‘control and communication in the animal and machine’, in the same year McCulloch and Pits published the first formal treatment of Artificial Neural Networks and published the Threshold Logic Unit (TLU). –A simple single neuron unit, which accepted the weighted sum of its inputs and the output was determined by a thresholder; if the weighted sum was greater than some value the output would be 1 otherwise it would be zero. In 1949 Hebb developed his learning rule for the human neuron. This rule suggested that synaptic strengths might change to reinforce any simultaneous correspondence of activity levels between the presynaptic and postsynaptic neurons. Simply put the weight of the connection between two nodes will increase the more it is stimulated. The development of ANNs continued with Widrow & Holtt developing models of the ADALINE in 1959. The ADALINE is identical to a TLU except the inputs were +/-1 not 0/1. Several ADALINEs were connected together to form a MADALINE this was the first ANN applied to a real world problem, removing echo on phone lines, and was trained using the delta rule. Other modifications to the TLU produced the Perceptron. This was an enhancement of the TLU where the inputs of the TLU come from a pre-processing association unit, the input pattern was supposed to be Boolean, these pre-processing units can have any Boolean functionality but are fixed, they do not learn. 1962 Rosenblatt, initiated training rules for neural nets, and showed that Perceptron training rule would converge making training for networks of perceptrons possible.

After a period of growth and development Minsky and Papert (1969), showed that a single layer perceptron could not solve non-linearly separable problems (for example XOR logic). The discovery really dented the enthusiasm for neural network and thus followed a period in which very little work was done on neural networks. A few

people continued to work in the area trying to find a way past this problem. During this time different types of networks were developed such as Grossbergs ART (adaptive resonance theory), which were self-organising neural implementations of pattern clustering algorithms. It was not until the early 1980s that interest in neural networks began to gather pace with the Hopfield recurrent network and Parkers rediscovery of back propagation, originally developed in the early 70s. Back propagation lead to the possibility of training multiple layer networks as long as the activation functions were differentiable.

This discovery lead to resurgence in interest in ANNs and since then the field has developed very quickly, and has found many applications (see below). Many different types of networks, nodes and training algorithms have been developed.

Throughout the development of ANNs the practical ability of them to solve very difficult problems has made them very popular. A review paper [47] (and references therein) by Widrow in 1990 summarises these early networks and the applications in a succinct manner. The applications mentioned cover: speech and pattern recognition [1963], weather forecasting [1964], adaptive controls [1987], adaptive filtering and adaptive signal processing [1985] – adaptive antennas [1967], adaptive inverse controls [1986] adaptive noise cancelling [1975], seismic signal processing [1985], Adaptive equalisation in high-speed modems [1965 1968]. Adaptive echo cancellers for long distance telecoms and satellite circuits [1967].

Other recent applications [48] include: mortgage risk evaluator, bomb sniffer, stock market analysis, process monitors for industry, classification problems for the medical world and many others besides.

The field of ANNs is now vast, with people using them in different ways, some simply as tools to solve problems of interest, some work on the underlying theory and node and network development. The next chapter deals more specifically with the neural networks used in this thesis and describes briefly the practical aspects of working with neural networks.

4 Artificial Neural Networks

"A neural network is an interconnected assembly of simple processing units. The functionality of which is based on the human neuron. The processing capabilities of the network are stored in the inter neuron connection weights. These are obtained by adaptation to, or learning from, a training set of patterns".[49]

This chapter introduces the properties of the artificial neuron and network structure. Training and learning of the network are also discussed. Then follows simulations of our system and the performance of the neural networks on the simulated data. The effect of varying different network parameters on the training is also discussed.

4.1 Introduction

Figure 17 shows an artificial neuron that contains the simplified properties of a real neuron. The interconnection weights w_n are simple representations of the synapse, which is multiplied by the input level I_n. The cell body is modelled by the sum and activation function $f(\ )$ and the output representing the axon.

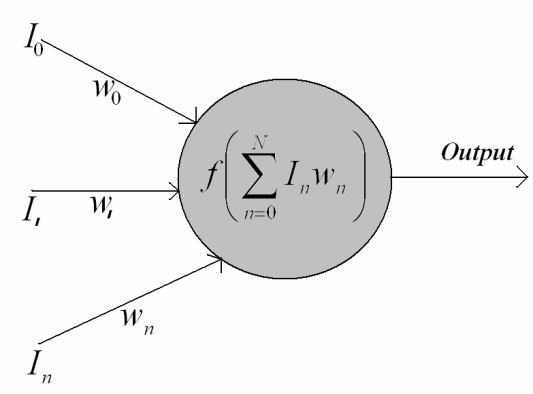


Figure 17 – Simple model of neuron

The central part of the artificial neuron is the processing unit or node. The weight values for the interconnections between the node and the inputs are where the relative impact of each input to that node is stored. The node calculates the weighted sum of its inputs. It is these weight values that are adjusted during the learning process.

The maximum information capacity of the network is governed by the total number of weights in the network, the more weights (interconnections) a network has the higher the information storage or learning capacity.

The output from the node is determined by the nodes activation function. It is these weights, combined with the usually non-linear activation function, which gives the ANN its computational power.

Activation functions are the functions used to calculate the level of the output of the node based on the values from the sum of the weighted inputs to that node. Examples of functions are the Boolean or hard limiter (0 or 1), piecewise linear, squashing functions (sigmoid range +/-inf goes to +/-1) as shown in Figure 18.

The activation function has implications for the training/learning method. Some methods of learning such as back propagation require a continuous derivative of the activation function and so using this technique reduces the choice of activation functions.

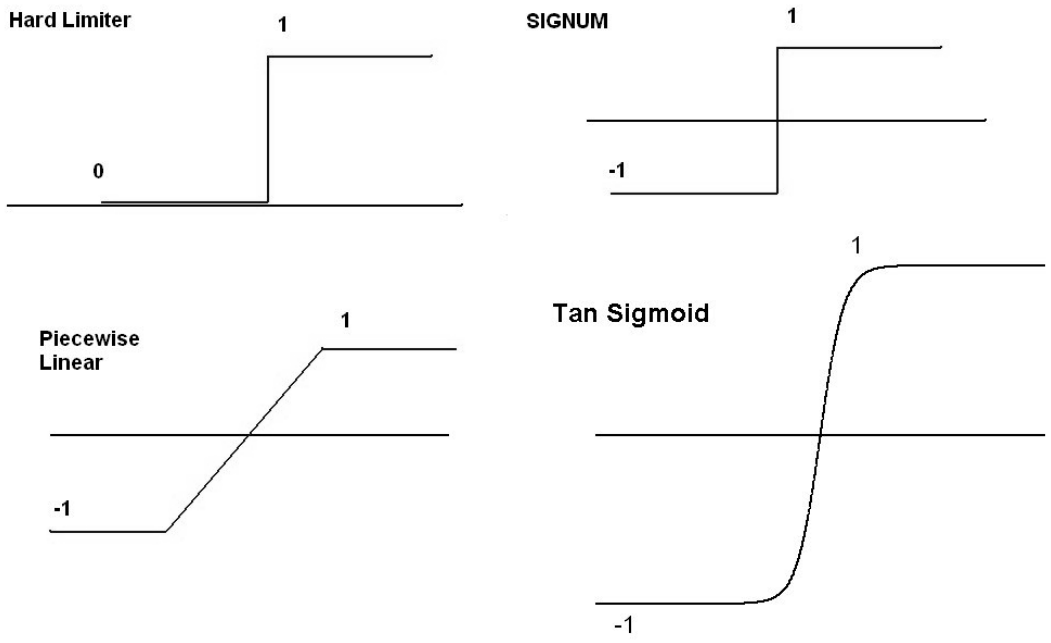


Figure 18 - Activation Functions

It is the choice of activation function that gives a network its power. For example if a linear activation function were used, the network would have the same function as a Perceptron (which has no hidden layer) this is because the linear combination of linear functions is still a linear function. If non-linear activation functions are used then any nonlinear function/operation can be approximated. Activation functions with discontinuities are usually not used, as they are difficult to train as the most common training methods rely on a continuous derivative of the activation function. This will be discussed in more detail later.

4.2 Forming A Network – topographies and applications.

The usefulness of artificial neural networks comes from the ability to combine many nodes into a network. The topography of the network determines its suitability for different applications. There are several broad categories of networks classified by the similarity of their topographical features, these are:

- Feedforward – networks where each successive layer is connected to the next either completely or partly. There are no feedback loops. These networks are typically used for classification and function interpolation [50][51] .
- Recurrent – these networks have feedback loops between the layers and are used for associative memory, noise filtering and content addressable memory [49][52] .
- Competitive – these tend to be self-organising and are used for analysis of topological features and cluster template formation. [53] [52]

4.3 Training

Training is required to generate the required weight values for the network to perform as desired, depending on the network type this training can be either supervised or unsupervised. Supervised training means that the network is given inputs and their corresponding targets. The network is then exposed to these input/output pairs and adjusts the weights to reduce the overall error in a usually least mean squares (LMS) fashion after each pass through the training patterns. This process iterates until some predefined stopping criteria is reached. Unsupervised learning is where the network itself tries to create clusters of similar features; there is no specific target information. After training is complete the weights are usually fixed for normal operation and the network stops learning. [54]

Our network is trained using back propagation, which is an extension of the generalised delta rule originally used to train ADALINES. (Networks with only an input and output layer(no hidden layer)). The mathematics for the training of a multilayer network via back propagation through gradient descent is easy to derive and is given in appendix 1.

Essentially back propagation is the backward pass of error to each internal node within a network; this is then used to calculate weight gradients for that node. Training progresses by alternately propagating forward the activations and propagating backward the instantaneous errors.

The change to the weights $\Delta w_m(i,j)$ of any layer *m* is given by:

$$\Delta w_m(i,j) = \alpha \delta_m(i) a_{m-1}(j) \qquad \text{Equation 4-1}$$

δ_m is the back propagated error of the network for Intermediate layer *m*. Alpha α is the learning rate and *a* is output from the activation functions from the previous layer.

4.3.1 Improving training

The Generalised delta rule (GDR) shown in equation 4-1, works very well but can be very slow to train. There have been several techniques to improve the training speed. Firstly GDR can be modified by adding a momentum term or by having an adaptive learning rate. The algorithm is essentially the same except the static learning rate alpha is either updated after each step or an additional term is added in the momentum case.

One reason why GDR is not usually used even though it is a simple algorithm is because its step size at each iteration is usually the opposite of what we require. For example when the gradient is small the algorithm takes a small step, but if the gradient is small we are moving on a plane and could take larger steps without fear of moving too far through the error plane. If the gradient is large the algorithm takes larger steps and we run the risk of stepping over the minimum completely and could end up stuck in a local minimum having already passed the global minimum, as shown in Figure 19. It is this counterintuitive operation that has lead to the wide spread use of more sophisticated

algorithms. These algorithms are usually second order algorithms and so use the second derivative of the gradient to make an update.

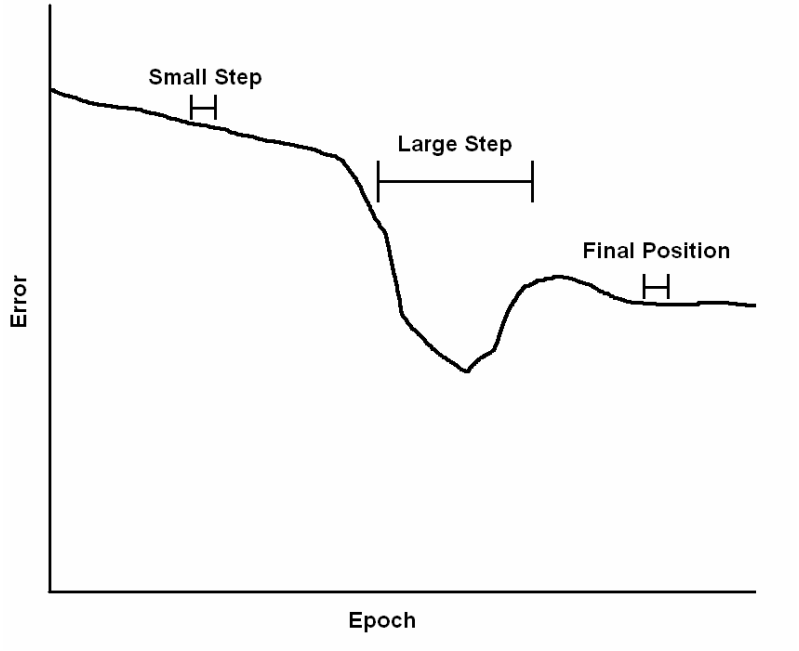


Figure 19 - problem with GDR

The goal of training is to find the global minimum of the error function. Gradient descent does this by moving in the negative gradient direction a small amount after each iteration. The Gauss-Newton method attempts to find the global minimum of the error function in terms of the weights in one step. This means that the Gauss-Newton method takes far fewer iterations than for the gradient descent method.

The Gauss-Newton method is a technique for solving equations of the form:

$$f(x) = \frac{1}{2}\sum_{j=1}^{m_N} r_j(x)^2 \qquad \text{Equation 4-2}$$

where $x=x_1,x_2 \ldots x_n$ and r_j is a function. In this case r is the difference between the target and ANN outputs this is equivalent to a series of residual errors for patterns $1{:}m$

The equation above can be rewritten as

$$f = \frac{1}{2}\|r(x)\|^2$$ Equation 4-3

where

$$\|r(x)\| = \sqrt{r_1^2 + r_2^2 + r_3^2 .... + r_n^2}$$

The derivatives of f can be written using the Jacobian matrix *J* of *r* w.r.t *x*.

$$J(x) = \frac{\partial r_j}{\partial x_i} \text{ where } \begin{matrix} 1 \le j \le m \\ 1 \le i \le n \end{matrix}$$ where m total number of patterns & n is number of nodes

Equation 4-4

for the general case:

$$\nabla f(x) = \sum_{j=1}^{m} r_j(x) \nabla r_j(x) = J(x)^T r(x)$$ Equation 4-5

$$\nabla^2 f(x) = J(x)^T J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x)$$ Equation 4-6

If the r_j can be approximated by linear functions or are themselves small then the second term in the above equation vanishes, and given the Jacobian it is possible to obtain the Hessian Matrix *H* [55] .

$$H = \nabla^2 f(x) = J(x)^T J(x)$$ Equation 4-7

Near to the global minimum these conditions are met and so this would give good results. Obviously it is very unlikely that a network will start close to the global minimum and so Levenberg and Marquart (LM) [56][57] developed a combination of gradient decent and Gauss-Newton to provide a more useful network update algorithm.

The update rule for gradient descent is simply

$$x_{i+1} = x_i - \lambda \nabla f(x_i)$$ Equation 4-8

For the LM case it is:

$$x_{i+1} = x_i - \left(H + \lambda diag\left[H\right]\right)^{-1} \nabla f\left(x_i\right)$$ Equation 4-9

This uses a second order approximation; all higher orders are neglected. The LM update method works as follows:

If the error increases after an update then λ is increased until the error decreases. This is essentially taking larger and larger steps through the error space until the step locates an area of lower error. If the error decreases after a step then the second order approximation is valid and the influence of the gradient descent is reduced by making λ smaller. Eventually after several iterations the influence of gradient descent is minimal and the weight values for minimum error are accurately established by the Gauss-Newton method.

The gradient descent term is modified by the diagonal of the Hessian matrix. As the Hessian matrix is proportional to the curvature of the error, we will take larger steps in the directions of low curvature and smaller steps where the curvature is high, exactly as desired. The only problem with this method is that it becomes too computationally expensive for larger networks with many parameters (1000's) due to the matrix inversion of the Hessian.

The goal of training is usually to produce a network that is general. The generalisation of a network is the ability of a network to respond correctly to an input that was not part of the training set.

Validation of a trained network allows the generalised nature of the network to be confirmed. Validation is performed by applying a set of input patterns that have not

previously been seen by the network and comparing the known target value with the network response. The deviation from the desired targets should be similar to that of the training set if the network is general. If the errors are much higher then the network has over or under trained, and a new network should be trained.

If the network is not general then it is no more than a look up table of data and for many applications it is completely useless. It is therefore important to ensure that the network remains general when training; this can be controlled to some extent by the network design.

4.4 Network design

The design of the network covers many aspects, from the obvious choices of number of layers, nodes and inputs, the choice of activation functions and training methods, to the choice of scaling of input and output patterns. There are no hard and fast rules that one can use to design a network, but rather general rules that one can apply and modify to get a starting network design, which may need to be modified time and time again as work progresses.

4.4.1 Input data and Targets

The input data is very important and needs careful consideration. Not only does the input pattern have to contain relevant information pertaining to the desired targets but there must not be too much spurious data contained in the inputs otherwise the training will be difficult. Therefore the inputs may require some form of transformation or processing to make them suitable to use.

The input levels also need consideration and usually a training set will be scaled so that the weight values can be kept small relative to the range of the activation function. Having small weight values helps to keep the network general as it reduces the chance of nodes being driven into saturation and producing an unpredictable output function.

If the activation function used at the output has a specific range, for example the tan-sigmoids output range is +/-1, then the target outputs must also fall within this range as a network cannot produce an output larger than the range of its output activation function. It is important to note that the largest and smallest values should not actually be +/-1 as these are the saturation values of the tan-sigmoid function and may lead to infinite connection weights, which in turn can cause network instability and poor training results.

4.4.2 Number of layers

In theory only one hidden layer is required to produce any nonlinear function as long as the network is of sufficient size [58]. If the network is very large due to the complex nature of the problem being tackled then it can be easier to have several hidden layers, as the training can be faster [59]. For our application only one hidden layer is required as the networks used are relatively small.

4.4.3 Number of Nodes

The number of nodes in total or per layer also has to be considered as this partly defines the learning capacity of the network and needs to be of a suitable size for the task being undertaken. If there are too few hidden nodes then the network does not have sufficient capacity to learn the required relationship and training will be poor. If on the other hand the network is too large, problems can also arise especially if real, noisy data is used, as

the network will rapidly overfit the data and lead to poor generalisation. The best practice is always to use a network that is large enough to do the job and no larger. Usually the number of nodes is established through trial and error on simulated or experimental data, starting with a very simple network and increasing the complexity until the performance is acceptable.

4.4.4 Number of Inputs

The number of inputs has a bearing on the information capacity of the networks as the number of inputs along with the number of nodes determines the total number of weights in the network. If the relationship is complex then there may need to be many input points for the network to be able to learn the relationship.

4.4.5 Training Set Size

The size of the training set or the number of patterns is also dependent on many factors. If the relationship is complex many patterns will be required. If too few patterns are used in a larger network then the network can memorise or over fit the data producing unsuccessful training. There are general rules of thumb for the amount of training data required. For example the number of patterns required should be 30 times the number of weights in the network, however there is no mathematical basis for these rules and so they do not guarantee successful training.

4.4.6 Improving Training with small data sets

If there is an insufficient number of training patterns then the network may have problems generalising. However, there are several techniques that can be employed to improve

network generalisation when there is insufficient input data and these will be discussed briefly.

Early stopping techniques are very successful at producing generalised networks where the input data set is small. The usual method involves splitting the input data into two sets, a training set and a validation set. The network is usually ‘large’, with small initial random values for the weights. The validation set is not used in any way to update the weight values, but the error from the forward pass of these patterns through the network after each iteration is monitored. If this error continues to decrease along with the training set then training is continued. If the validation error starts to increase then the network is beginning to over train and memorise the training data. If this occurs then the training is stopped. It should be pointed out that sometimes the error will fluctuate so there is normally a condition on the ‘increasing error’ rule. Often the error in the validation set is allowed to increase but only for x iterations if it does not start to decrease again then the network training is stopped.

Jittering is training with added noise and works because the input output relationship that we wish the network to learn is usually continuously smooth. In this situation very similar inputs will have very similar outputs. If we add noise to an input pattern, then the input pattern will be slightly different but as long as the noise is not too large then the output value will be essentially the same. This increases the number of available training patterns and helps the network to generalise.

4.5 Simulation of single tracks

This approach has been simulated to see how the ANN performs and is illustrated in Figure 20. Firstly the optical system was simulated, this comprised of a simple 1D scanning microscope where at each scan location the complex amplitude of the output signal is given by the complex sum of all the spatial frequency components in the back focal plane of the objective. This allows both the amplitude and phase profile of the object to be obtained. A series of track structures were scanned in this simulation and the amplitude and phase profiles stored. The tracks then underwent some signal conditioning, including a Fourier transformation. The processed spectra were sampled to obtain inputs for the ANN training. Once the inputs are obtained from the profiles the measured tracks are then split into two sets, one set with its known targets is used to train the ANN and the other set is used to test the final network. .

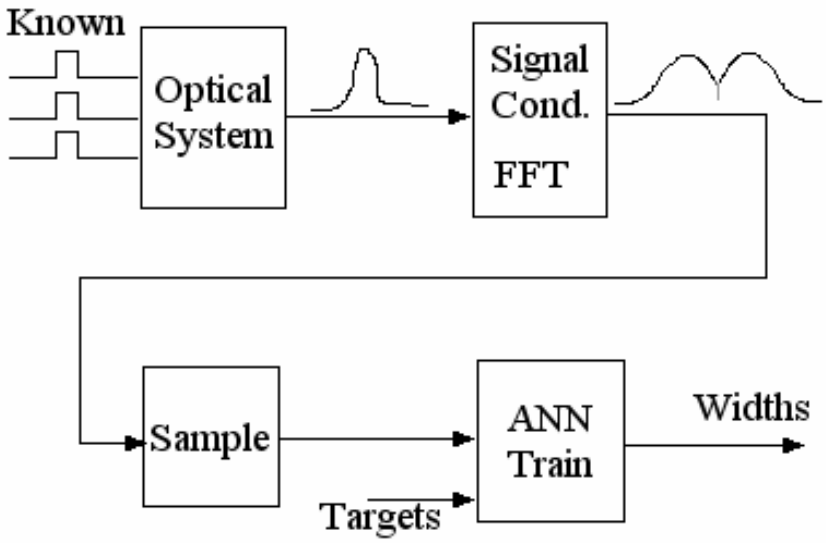


Figure 20 - the whole training process step by step

4.5.1 Neural network topography

The ANNS used both in the simulations and the experimental work all have the same structure (unless otherwise stated). The general form of the network is an 8-5-1 feed forward network. The layers are fully connected and a schematic of the network is shown in Figure 21.

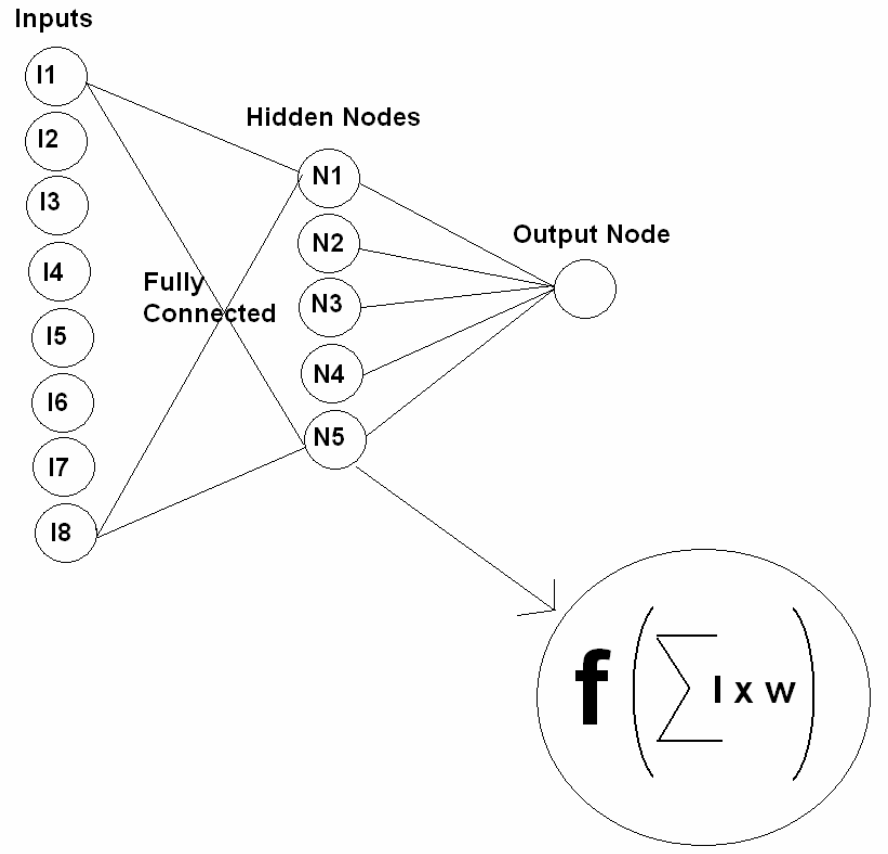


Figure 21 - topography of networks used

There are 8 inputs derived from the object. The input layer is fully connected to the hidden layer, which contains 5 nodes. The activation functions are tansigmoids. The hidden layer is fully connected to a single output node, which corresponds to the track parameter. The output targets are scaled so that the largest target has the value of 0.8. Which is chosen as it is below the saturation value for the tansigmoid function.

The training algorithm employed is based on the method by Levenberg-Marquart described earlier. The networks used for the double tracks contained 16 inputs, 8 hidden nodes and either 1 or 2 output nodes.

4.5.2 Input format

One of the most important stages above is choosing the input format, which combines both the signal conditioning and the Fourier transform stage. The input format could take one of several forms as illustrated in Figure 22:

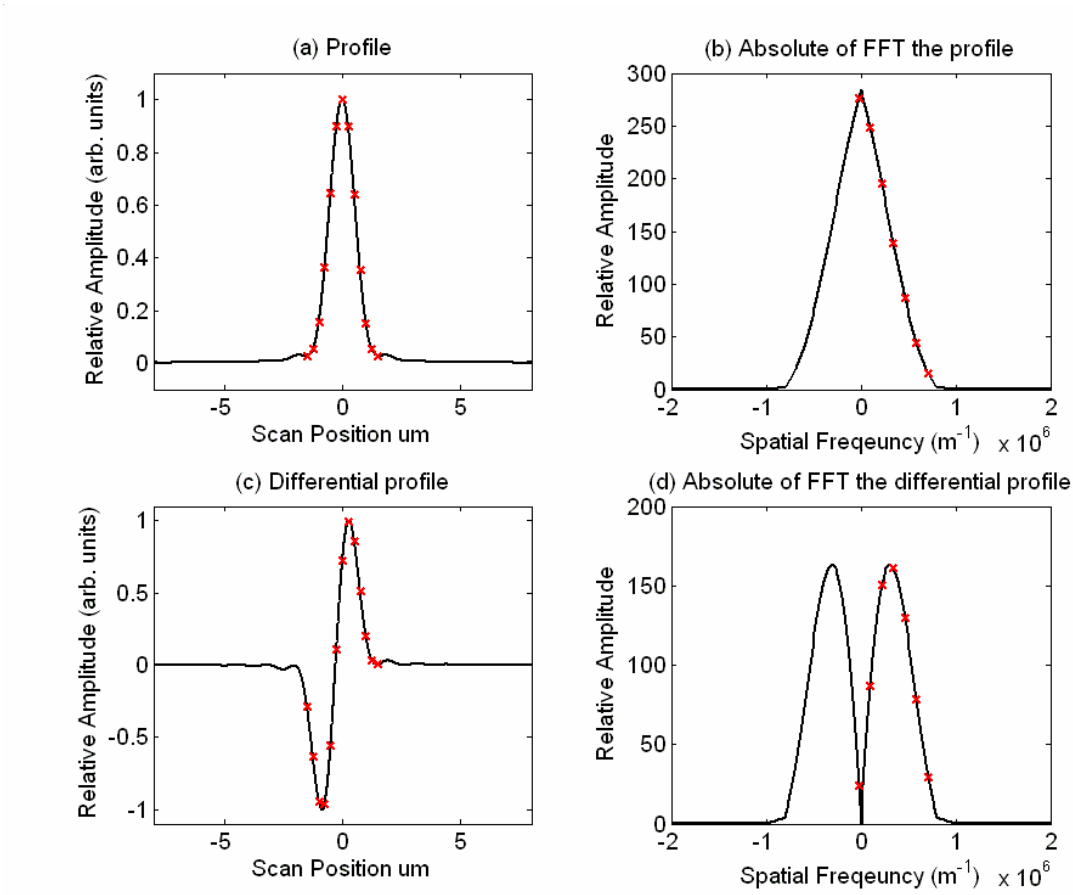


Figure 22 - Input formats

The red crosses indicate possible choices for the input points used for training. Using the profiles or spectrum directly does not produce good training results. After differentiating

the profile and then using the spectrum the system performed well. The differentiation process has the effect of suppressing the low frequency components and these are the most similar components for each track as it is the mid-high frequency components which contain the most information regarding the change of track shape with track width. So essentially the information now being presented to the network is the most relevant and this is why performance is improved.

4.5.3 Input / Output Scaling

Before training the last stage is to scale the input and output points into suitable ranges. The input points are scaled by the largest value in the whole set. This keeps the inputs in the +/-1 range and means that the weight values should be kept small, which helps with the stability of the training. The targets also need to be scaled so that they fall within the output range of the output node activation function. In this case the outputs must be in the range +/-1 for the tanh activation function. However for a target to be exactly 1 this would call for an infinite connection weight into the output node which would make the training unstable and so practically the outputs should be less than +/-1, we typically use +/-0.8.

To recap, we have two data sets that have been processed as below in Table 5:

Table 5 - Input Processing

	Signal conditioning	**Fourier transform**	**Input range**	**Target range**
Training set	Differentiate	Yes	+/-1	+/-0.8
Testing set	Differentiate	Yes	+/-1	+/-0.8

The simulation was carried out to compare to an experimental situation. The range of tracks used was 40-480nm and there are 4 copies of each width so that noise can be applied if desired to allow jittering. The tracks were purely phase objects of 45nm height. The phase profiles from the optical system were used as the input data. An additional set of data was also generated, which have completely different track values from the training and testing data but are within the same range, this extra data is just to demonstrate that the network response is general and that the network is not just behaving as look up table. The simulation of the optical system had the following parameters:

- NA = 0.3
- wavelength = 688nm

The training results in Figure 23 are for the case where there was no noise in the system. The plotted error is the difference between the network output and the known target value. The standard deviation (std) of this error is a measure of how the network responds across the network for different track widths. The training and testing set contain identical data in this case, as there is no noise so in the graph the crosses and stars overlap. For this reason an extra set of data with different width values not contained in the training or testing sets was also plotted on the graphs as the test only set.

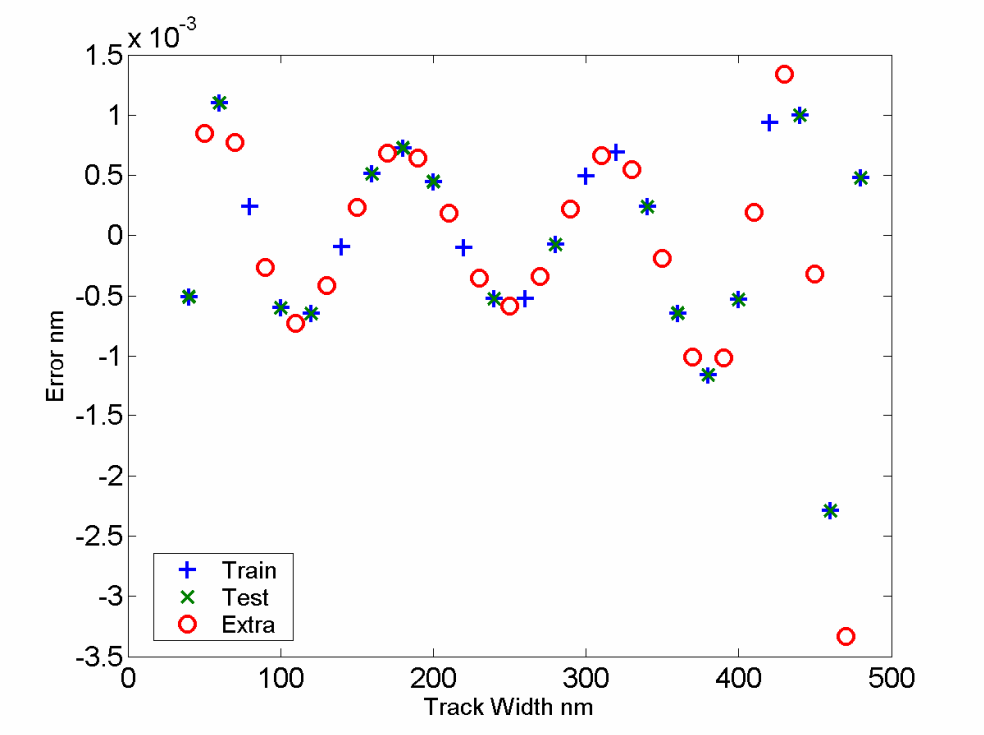


Figure 23 - Results for train set (*), Test (x) and testing only set (o)

The graph above shows that the training was successful. The training and testing set have a continuous response showing that the result is general, if the network had over trained then the error for the testing set would be much larger. The values for the standard deviation are given in the Table 6 below.

Table 6 - Training Errors

	Training set	Testing set	Test Only Set
Standard deviation of error (nm)	0.00073	0.00093	0.00094

Without any noise in the system the standard deviation of the error is better than 0.001nm across the range of 40-480nm track widths. The smallest track here is $1/70^{th}$ of the optical spot size, which is a huge increase in measurement range for this optical system when combined with the ANN.

4.5.4 Adding noise to the system

The noise is additive phase noise, which is added to the phase profiles before they are processed. The level of noise is set by scaling the random noise so that the standard deviation of the noise is $10^{-6} \rightarrow 10^{-3}$ radians. The additive phase noise can be related to shot noise as errors in the amplitude signal will cause errors in the phase signal (a more detail explanation of this process is given in appendix 2). Table 7 presents typical noise values for this system as discussed later in chapter 5.

Table 7 - Conversion between phase noise and shot noise

Maximum photons per pixel	Total number of photons (128x50 window)	Standard deviation phase radians	SNR
2.76E+00	1.00E+04	0.01807	100
2.76E+01	1.00E+05	0.005563	316
2.76E+02	1.00E+06	0.001824	1000
2.76E+03	1.00E+07	0.000552	3162
2.76E+04	1.00E+08	0.000174	10000
2.76E+05	1.00E+09	0.000055	31623
2.76E+06	1.00E+10	0.000017	100000
2.76E+07	1.00E+11	0.000005	316228

As can be seen below in Figure 24, the error distribution is now random across the range due to the added noise; the phase noise in this case had a standard deviation of 0.1milli-radians.

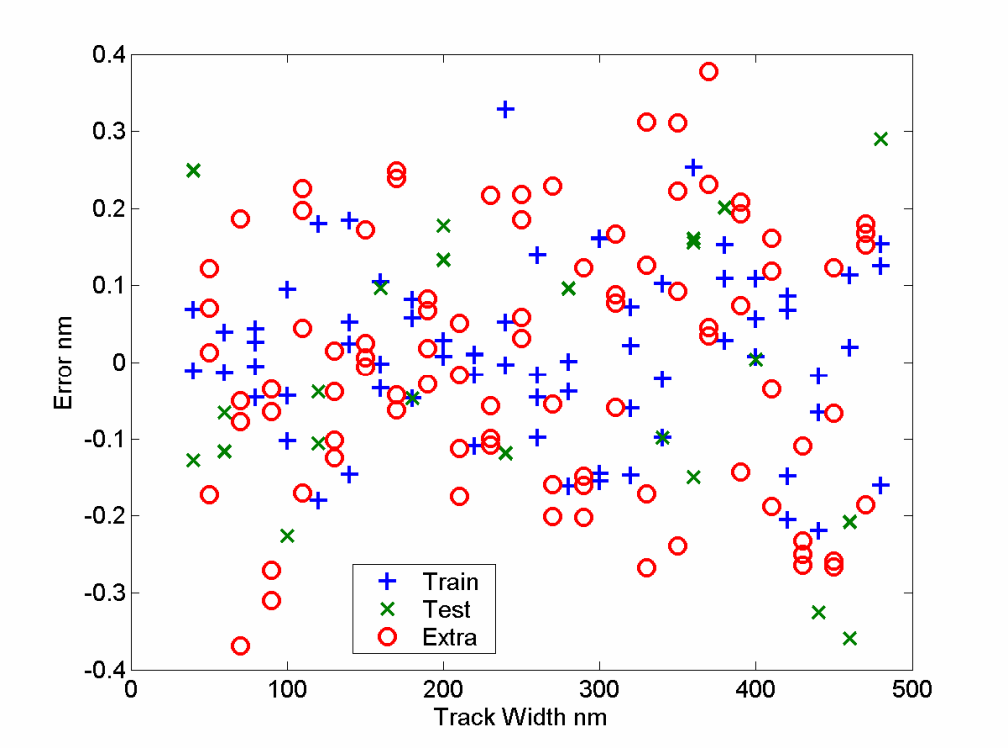


Figure 24 - Training result with phase noise. train (*) ,test (x) ,test only(o)

The standard deviation of the training and testing errors are shown in Table 8 where the phase noise has a standard deviation of 10^{-4} radians. The noise was different for each data set but the standard deviation of the noise was the same. Typical phase noise values for the experimental setup are between 0.5-1.2 milliradians.

Table 8 - Training results for noisy input data

	Training set	**Testing set**	**Test Only Set**
Standard deviation error (nm)	0.11021	0.18114	0.16876

The training error is often slightly smaller than the testing sets and this is probably due to the early stopping algorithm employed. The error for the validation set is allowed to increase for several iterations before the training is stopped and may therefore be slightly higher. This is because the errors often increase slightly before decreasing again and if the network were very strict in stopping on a single iteration where the error increased it

would often fail to train. Also the testing set usually has far fewer examples in it and so the standard deviation is less accurate.

The standard deviation of the testing set is plotted versus phase noise level (log scale) in Figure 25 for simulated data. The relationship is fairly linear, in that a 10-fold increase in noise gives an approximately 10-fold increase in training error.

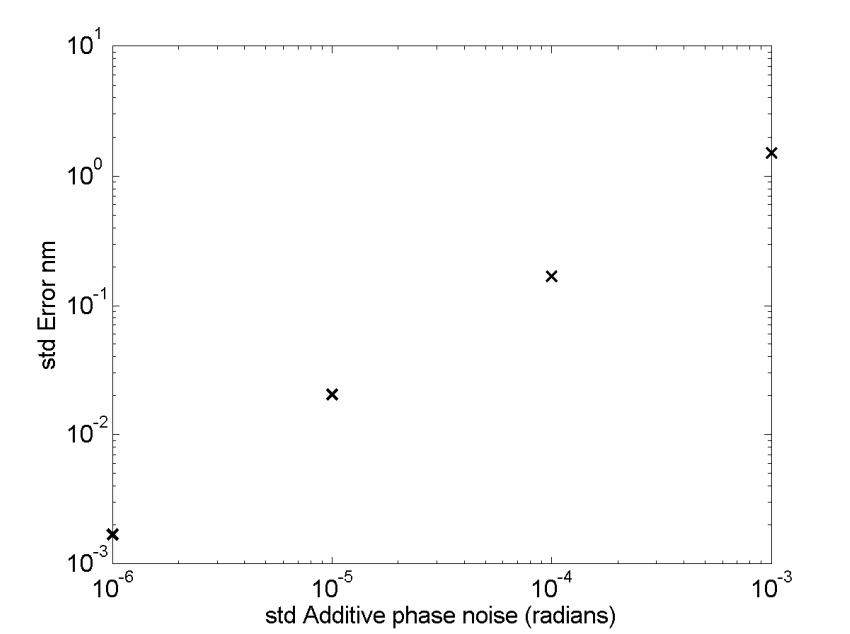


Figure 25 - standard deviation of testing set with increase in noise level

The information in figure 8 above is summarised in Table 9 below:

Table 9 - Training with noise

Photons	Standard deviation of Phase (Radians)	SNR	Error (nm)
INF	0	INF	0.000943
$2x10^{12}$	1e-006	$1.4x10^{6}$	0.001695
$2x10^{10}$	1e-005	$1.4x10^{5}$	0.020377
$2x10^{8}$	0.0001	$1.4x10^{4}$	0.16876
$2x10^{6}$	0.001	$1.4x10^{3}$	1.5037

This relationship will dictate the required SNR of the optical system for a particular training error. If the optical system is the same as the one simulated and has 1mrad phase noise than the standard deviation of the training result would be expected to be around 1.5 nm over the range of tracks used.

4.5.5 Repeatability of training

Figure 26 shows the test only set plotted for three training runs where the training data remained the same. As can be seen the errors for each track are very similar and so the training is very repeatable.

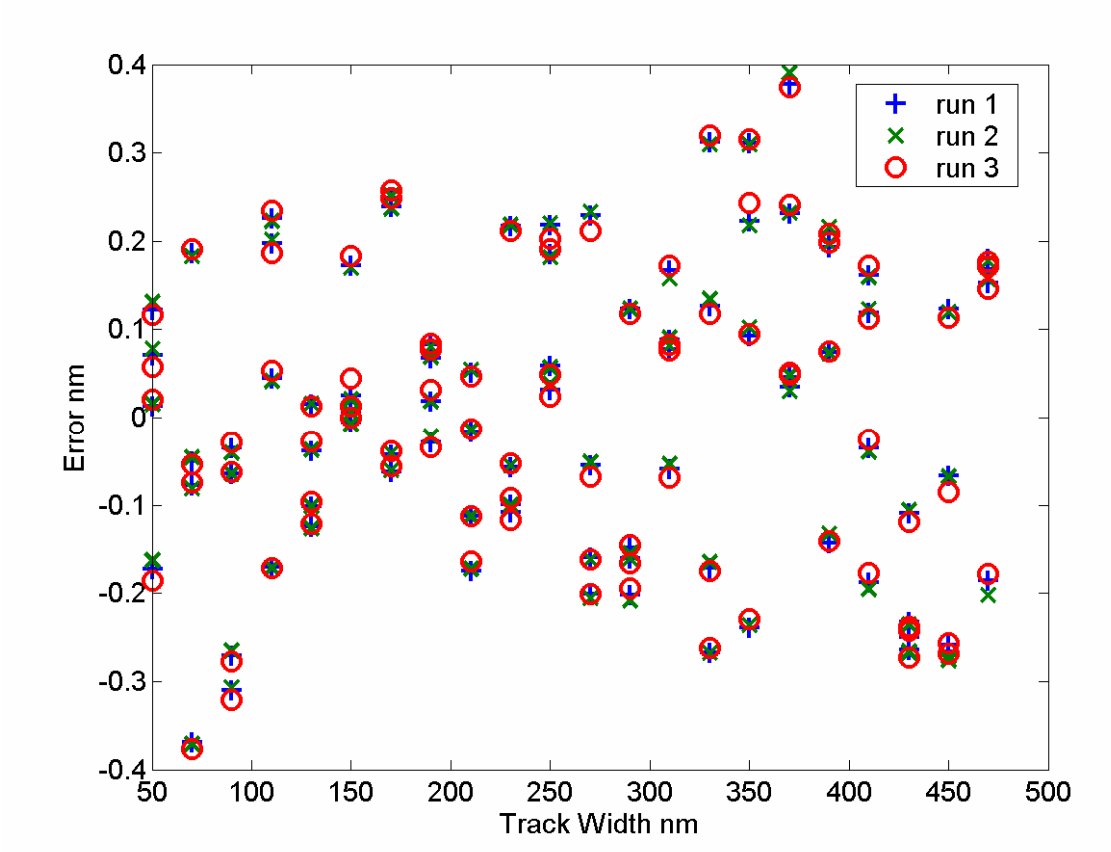


Figure 26 - Repeatability of training.

Table 10 below shows how the values of the track width have changed across the whole range for the different runs compared to the first run.

Table 10 - repeatability of training

	Run1 – Run2	**Run1 – Run3**
Mean (Difference) (nm)	-0.00017542	-0.00073926
Std (Difference) (nm)	0.005412	0.0082584

This shows that the training is very repeatable, for three runs the standard deviation between the values is 0.008nm, compared to the standard deviation of the same track width due to noise being 0.16nm. The repeatability error is 20 times smaller than the error due to the phase noise. This variation in training is caused by the training not necessarily stopping at the same place in the error space, as it may not be at the exact global minimum just very close to it. Each training run starts at different, random location in the error space so the route to minimum during training will be different and therefore can end up in a different place but still very close to the global minimum. These differences give rise to variations in the training errors for successive runs.

The networks above contained 8 inputs and 5 hidden nodes; the following sections will discuss the impact on the training results of varying these parameters.

4.5.6 Nodes

The number of hidden nodes in a network is a key factor in the ability of the network to learn the required relationship. With a fixed set of inputs, the number of nodes was altered and the errors of the trained networks analysed. A table of the training results is given in Table 11.

Table 11 - Training Results for different hidden node number

Nodes	1	2	5	10	15
Training Standard Deviation (nm)	6.4007	0.65021	0.51478	0.48196	0.91102
Testing Standard Deviation (nm)	6.3297	0.85218	0.64239	0.63049	0.8864

As can be seen from Table 11, if there are only a few nodes the training is less accurate, as the node number increases the training error decreases. As the number of nodes gets larger the error no longer reduces. In our case when the number of nodes is around 5-10 the training is reliable and the overall error is low. We have used 5 nodes in the previous training examples we could use more nodes but this would only increase the training time and not improve performance.

4.5.7 Number of Inputs

The total number of inputs presented to the network is also an important parameter as too many inputs may not provide much extra information but will increase the number of weights in the network and therefore have an impact in training times and the complexity of the network. Too few inputs and there will not be enough information contained in the input patterns for the network to learn the required relationship.

Table 12 - Changing the number of input points and training results

Number of Inputs	2	4	8	16	32
Training Error (nm)	0.165	0.059	0.044	0.039	0.035
Testing Error (nm)	0.124	0.052	0.046	0.042	0.046

Several networks were trained with a fixed number of nodes (5) and variable number of inputs; the results of this training are given in the Table 12. The inputs always covered the same region of the spectrum i.e. the first and last points were always the same just the

number of points between the start and end points was varied. Once the number of inputs has increase to over 4 the training level is fairly flat, the testing standard deviations are very similar.

4.5.8 Simulation Auto correct

The neural network learns the underlying relationship between the inputs and targets, this means that the network can be used to detect errors during training in any of the input/target pairings used to train the network. For example if one track has an incorrect target value the input/target relationship for this track will not fit with the underlying relationship learnt by the network and will therefore have a much larger error as shown in Figure 27, where the target for the 260nm tracks were increased by 3 percent. The mean error for this track is -5.57nm compared with the mean error for the other tracks in the training set of 0.22 nm. The standard deviations are 0.5nm and 0.25nm respectively.

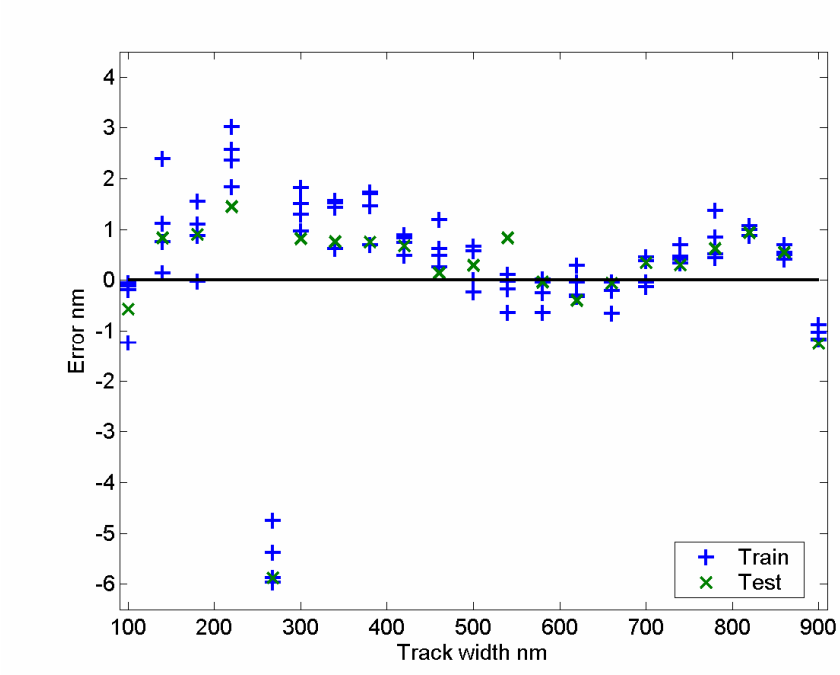


Figure 27 - Effect of incorrect target value on 260nm track

We can use this effect to correct for this target error. The network can look at the mean error for each track/target pair, if one pair has a significantly higher error the network can generate a new target for this pair by subtracting some portion of the mean error. The network can then be retrained and the process can be repeated until all of the tracks have a similar mean error. This is illustrated in Figure 28.

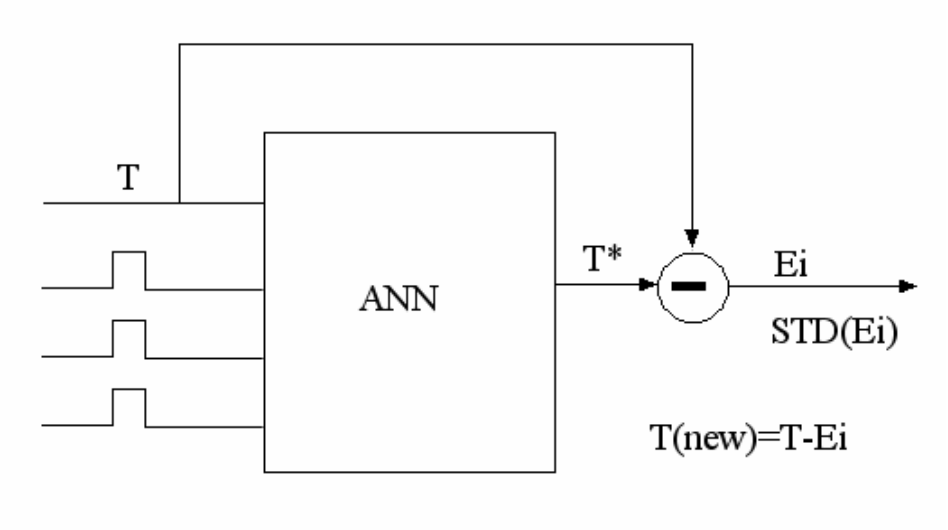


Figure 28 - The auto correction process

This process was carried out on simulated data and the training results are shown for three of the five iterations of the training procedure in Figure 29

After the first training iteration the tracks around the 260 nm track have a mean error greater than zero due to the effect of the incorrect target value for the 260nm track as the network tries to bring the incorrect track into the model. The overall training level is much worse than usual.

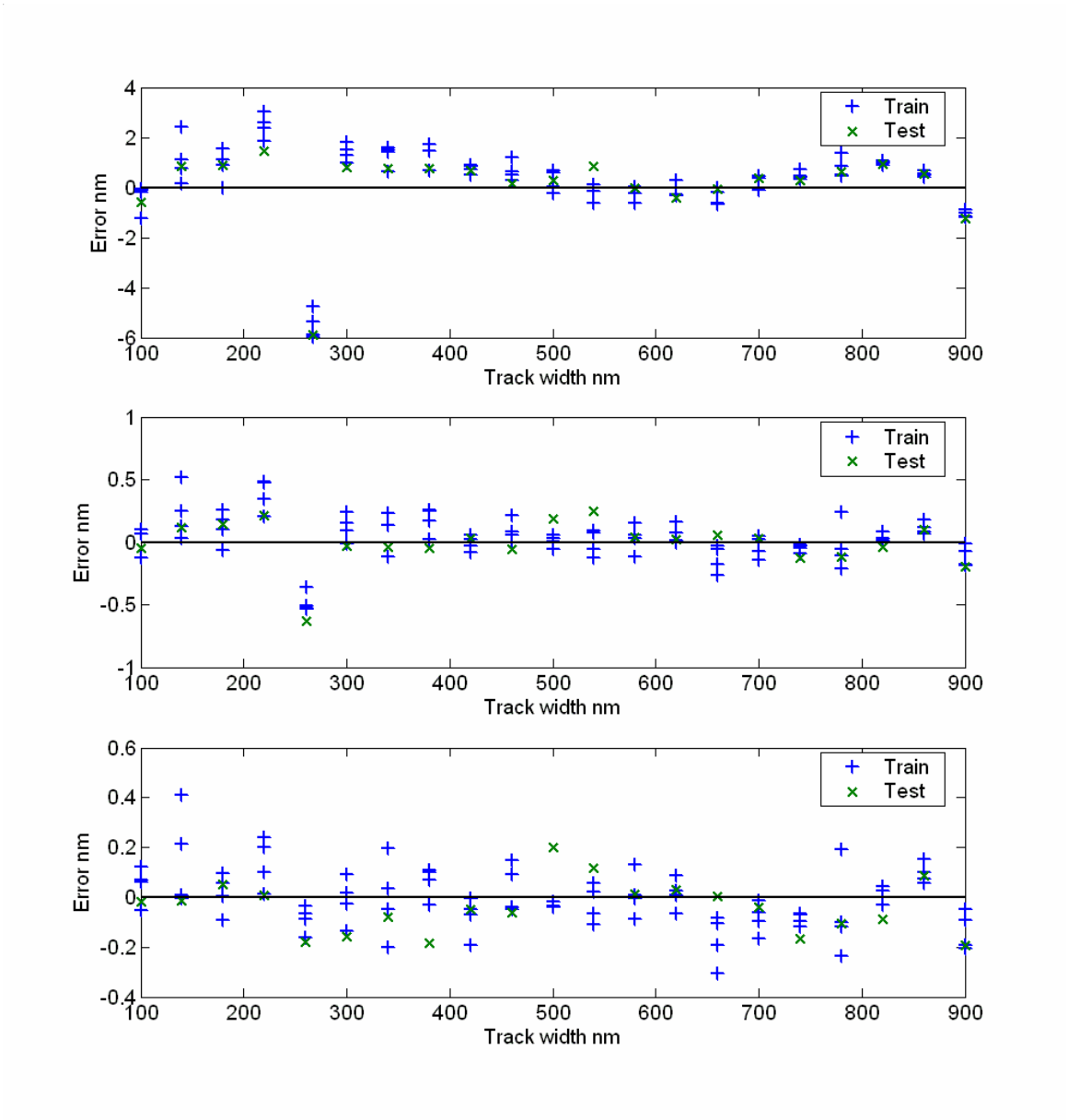


Figure 29 – Run 1, 3&5 of the autocorrect process

After the 3rd iteration the overall error has reduced and the effect of the 260nm is less pronounced on the neighbouring tracks. By the 5th iteration the 260nm track is comparable to all of the other tracks, in terms of mean and standard deviation of the error. The peak error has also reduced to the level expected for this noise level (0.11mrads).

The original track width and the updates after each pass are shown in Table 13 below. The original tracks had a target of 260nm a 3% error was applied to the target values making the starting target 267.92nm.

Table 13 - Auto correct results

	Run1	Run2	Run3	Run4	Run5
Target (nm)	267.92	262.35	260.88	260.37	260.24
Update (nm)	-5.57	-1.46	-0.51	-0.13	-0.11
New target (nm)	262.35	260.88	260.37	260.24	260.13
Overall Error standard deviation (nm)	1.50	0.43	0.18	0.12	0.11

As can be seen after 5 iterations the new target value is correct to 0.13nm and the network has successfully corrected for the training target error. This technique could prove to be very useful when training the ANN. It will allow the identification of any target values that have been specified incorrectly. The ability of the network to cope with large and multiple errors is yet to be established.

4.6 Double tracks simulation

After the highly successful performance with respect to calculating the track widths for single tracks, this technique has also been applied to double track structures with the aim of calculating both the width and the separation of the tracks.

The spectrum of a double track object (Figure 30) has several components, which can be derived as follows.

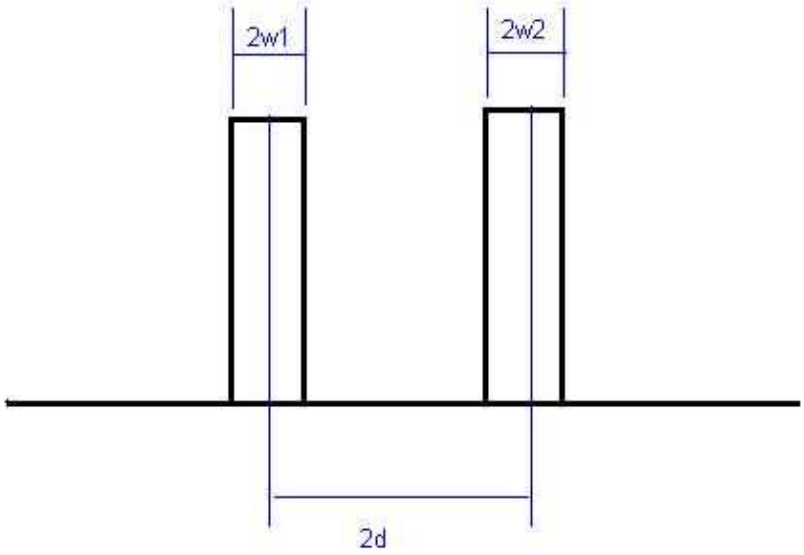


Figure 30 – Simple double track object

The first track is A and the second track is B so the Fourier transform of two-track sample is the sum of the transform of A plus the transform of B (linearity theorem). The transform of A and B are sinc functions based on width w1 & w2, that are shifted by the amount 'd' using the shift theorem. If the widths are the same w1=w2 then the spectrum is:

$$\Im\{i\} = 2w_1 \mathrm{sinc}(f_x w_1)\cos(f_x d) \qquad \text{Equation 4-10}$$

In the simple case where the two tracks are the same width the spectrum consists of two components a sinc term due to the widths of the tracks multiplied by a cosine term due to the separation. This means that the input patterns will contain the relevant information relating to both the separation and the width of the double track structure. After passing through the optical system these components will also be modified by the system transfer function.

A simulation of two simple cases was performed. The optical system had the following parameters: NA = 0.3 wavelength = 688nm and the data was noiseless and both of the double tracks had the same width value. In the first case the separation between the two tracks was kept constant (760nm) and the width was varied from 10 to 740nm. The input patterns (abs differential spectra) are shown below in Figure 31.

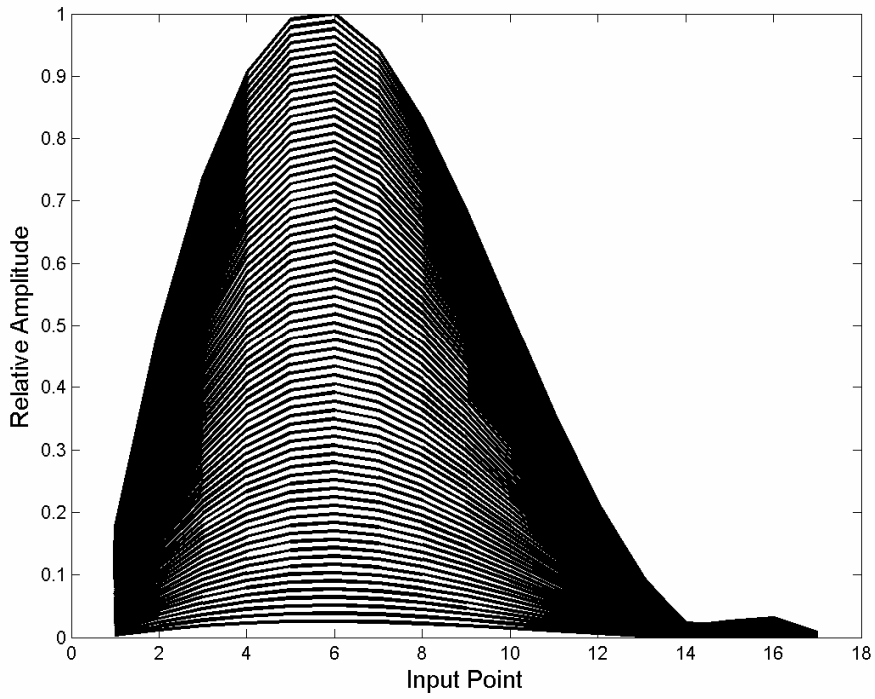


Figure 31 - Width varied separation constant

The network was then trained to calculate the width value of the double tracks and the result of this training is shown in Figure 32.

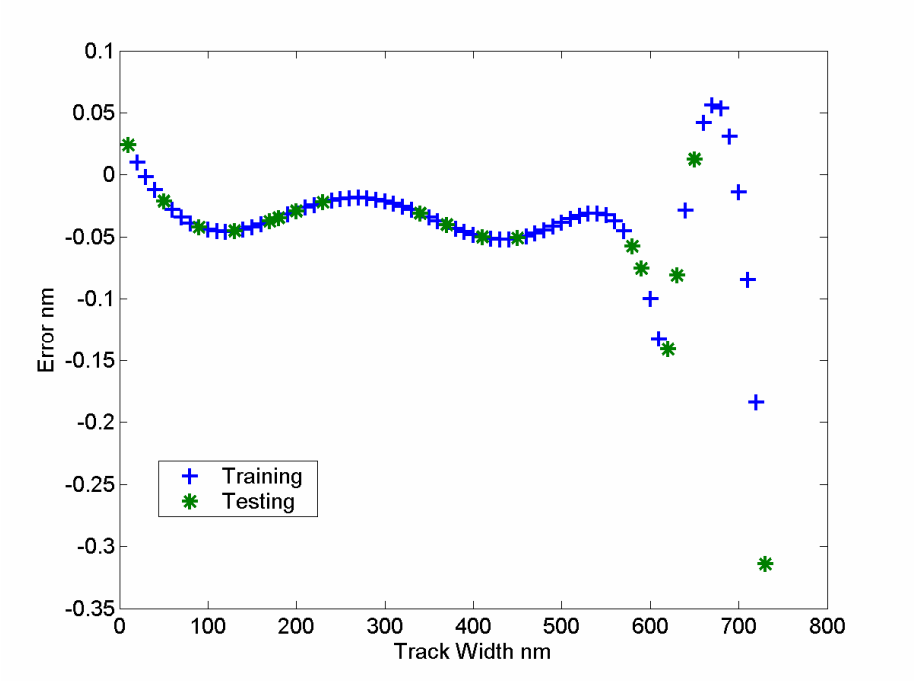


Figure 32 – Results width varied separation constant

The standard deviation of the error in the value of the width of the tracks is 0.001nm when training and 0.003nm when testing.

In the second case the width was kept constant and the double track separation was varied. Again the input profiles are shown in Figure 33. And the network training results are shown in Figure 34.

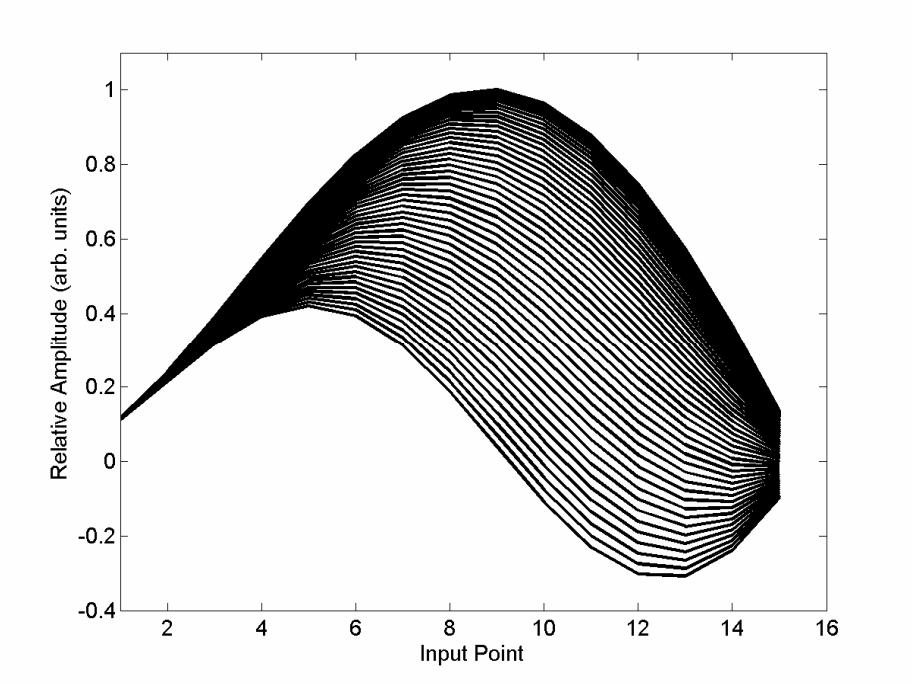


Figure 33 - Constant width variable separation

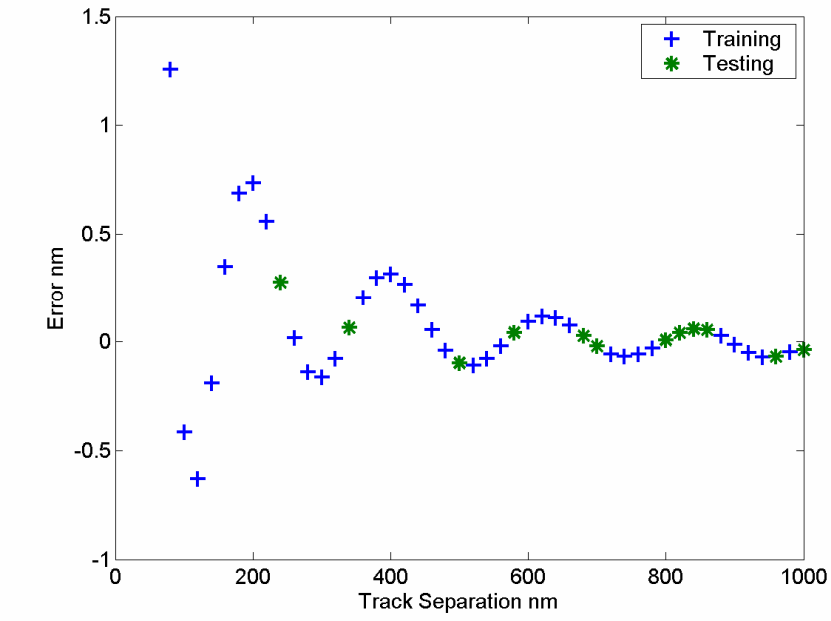


Figure 34 - Results constant width variable separation

The standard deviation of the error in the calculated value of the separation across the whole range is 0.34nm when training data is used and 0.094nm with the testing set. The training set error is higher because the error for the smallest separations was higher and

none of these examples were in the testing set. This shows that the network performs very well for this simple case.

The same process was repeated with the inclusion of noise in the system. The table below shows the error for the width and separation for the same conditions as the above example for different noise levels.

Table 14 - Constant width variable separation training with additive phase noise

Phase noise / radians	Train error (nm)	Test error (nm)
0.01807	29.99	38.59
0.005563	9.96	15.61
0.001824	6.92	8.72
0.000552	1.29	2.75
0.000174	1.32	2.02
0	0.29	0.68

The error decreases with noise, as expected, even with reasonable SNR ratios the training results are acceptable. The errors are worse than for the single track case (for example, 0.0001 phase noise produced track width errors in the region of 0.17 nm) but this is to be expected as this situation is much more complicated.

Table 15 - Constant separation variable width training with additive phase noise

Phase noise / radians	Train error (nm)	Test error (nm)
0.01807	21.05	32.44
0.005563	11.19	14.91
0.001824	2.93	6.63
0.000552	1.49	2.89
0.000174	2.07	2.69
0	0.43	1.49

The same can be said for constant separation and variable width. As the system performed well for this simple situation a more complex task was simulated. In this case both the widths and separations are varied. The simulation details are as follows: A set of 2400 double tracks of height 45nm ranging in width from 10nm→1000nm and with various separations of 40nm-1000nm has been simulated. The processing for the tracks was the same for the single track the only difference being that the number of inputs has been increased, as the spectrum of these objects is more complex, and also the number of hidden nodes has also been increased.

There are two options available for the training of the network as there will be two outputs for the network. We could have two networks, one for the width and one for the separation, or we could have one network with two outputs. Results from these networks will be presented.

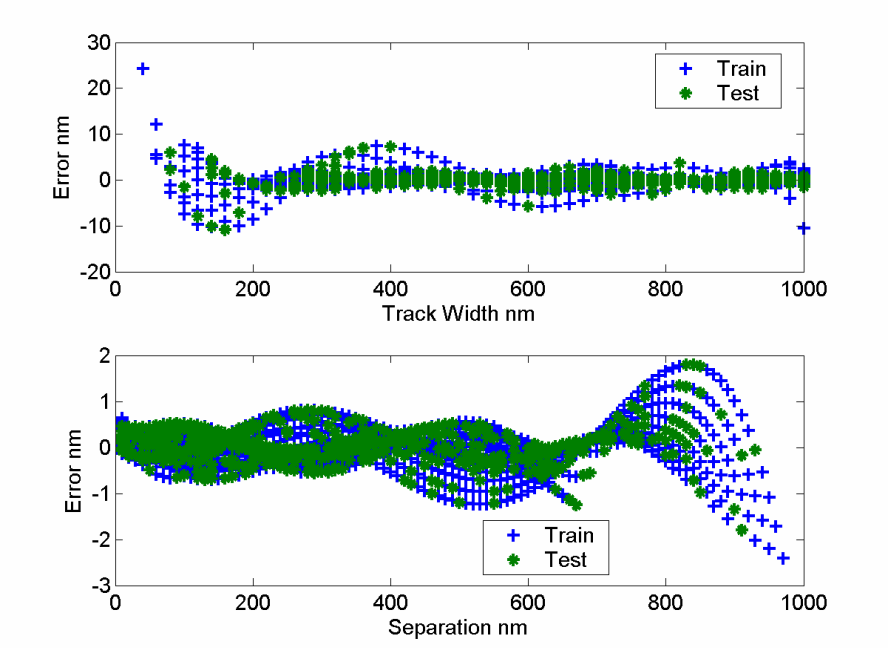


Figure 35 - Results for width and separation in one network

Using one network to calculate both the track width and separation produced better results for the separation than the width as shown in Figure 35. The width values get less accurate as the width reduces, but the separation values get worse as they get larger. The errors are considerably higher than for the simple case discussed previously. The standard deviation of the training and testing sets is presented in Table 16.

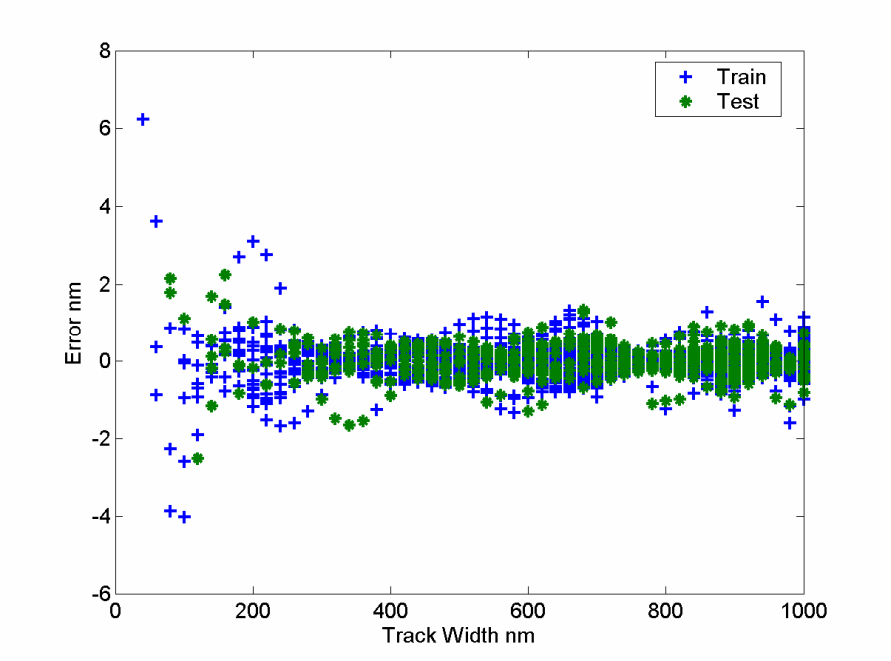


Figure 36 - results for width only network

Figure 36 shows the results where the network has only one output – the track width. The errors are greatly reduced with most cases being sub nanometre. This performs better as the network is only learning one relationship for the track width not two relationships as for the previous case. This makes the task that the ANN is performing much easier and so the training is improved.

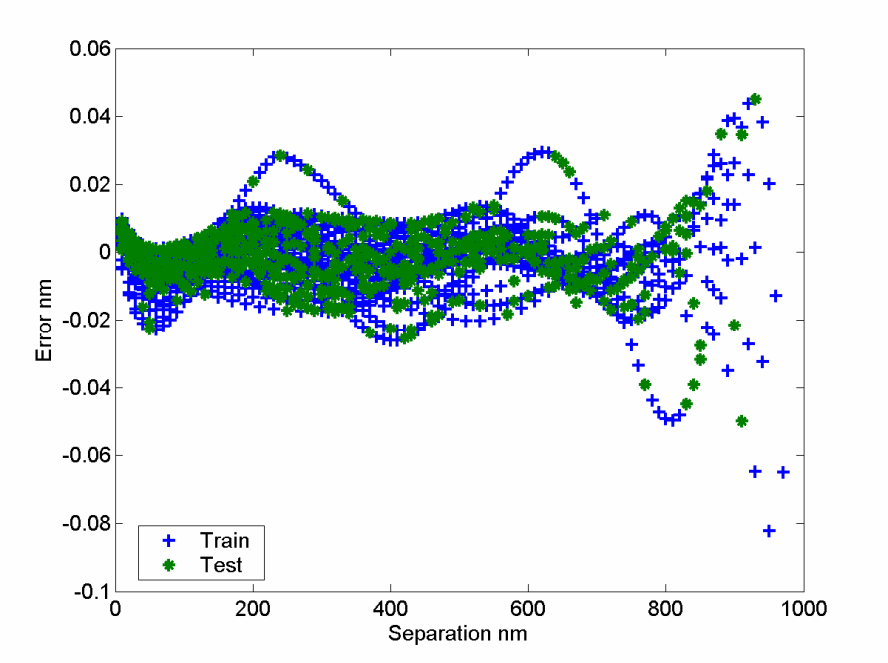


Figure 37 - results for separation only network

Figure 37 is for the case when the network just has to learn the separation and this is performed extremely well and so this relationship must be much easier for the network to learn. All of the training results are summarised in Table 16. Where 'std' is the standard deviation.

Table 16 - Double track training for 3 networks

	Width Network (nm)	**Separation Network (nm)**	**Width &Separation Network (nm)**
Std Train – Width	0.46082	-	1.4806
Std Train – Sep	-	0.0091847	0.39055
Std Test - Width	0.44414	-	1.3908
Std Test - Sep	-	0.0093107	0.39552
Mean Train - Width	0.00056169	-	-0.084009
Mean Train - Sep	-	-0.0015481	0.041172
Mean Test - Width	0.02463	-	-0.030349
Mean Test - Width	-	-0.0016513	0.034281

The mean values are all around zero showing that there is no offset in the errors. The individual networks are superior to the combined network by approximately a factor of three for the width value and a factor of forty for the separation.

The work present above was repeated with noise included in the system and Table 17 shows the results for the testing sets of the networks for various noise levels.

Table 17 - Double track training with noise

	Training Set Std Error (nm)					**Testing Set Std Error (nm)**				
Phase noise (mrads)	**18.1**	**5.56**	**1.82**	**0.55**	**0.174**	**18.1**	**5.56**	**1.82**	**0.55**	**0.174**
Width network	47.74	15.88	6.34	3.29	1.29	55.31	21.49	8.50	4.40	1.85
Separation network	2.21	0.69	0.23	0.07	0.02	2.32	0.71	0.24	0.07	0.02
Dual network width	46.15	16.17	6.53	3.02	2.07	56.22	21.39	8.66	4.03	2.57
Dual network separation	3.64	1.60	0.80	0.73	0.59	3.64	1.64	0.81	0.73	0.60

Again in the presence of noise the separation was much more accurately obtained and the error decreased with noise. The performance of all of the networks could be improved by reducing the range of widths and separations used. For example the widths network could be split into two networks, one that dealt with all tracks in the range 50-500nm and one for 500nm-1 micron this would improve the performance of the lower end of track widths an example of splitting the desired parameter of interest into smaller ranges is given in chapter 7 section 3. There is little improvement in the errors for the width network when going from the two-parameter network to the width only network. This appears to be

caused by the width value being more sensitive to the noise level. When the data is noisy the network has much more freedom in producing a solution as the problem is less constrained by the data points but by the noise. As the noise reduces the solution for the network gets constrained and which is why for the noiseless case there is a marked improvement going to the split networks.

4.7 Double Track or Single Track classifier

As the double track structures get closer together there comes a point at which they look very similar to single track objects. So far we have used *a priori* information to send them to a double track network or a single track network. However it is possible to train a network to classify the tracks into two sets. This is because even though the double and single tracks may look very similar there are still subtle variations in the spectrum that the network can use to classify the tracks into double or single tracks.

A classifier network was constructed to perform this task. The outputs were coded as shown in Table 18. The data set comprised all of the single and double tracks of height 45nm, slope 2nm. In total there were 2400 double tracks and 100 single tracks 75% of these were picked at random to train the network the others were used to test the network.

Table 18 - Output encoding for classifier

	Target1 (Arb. Units)	**Target2 (Arb. Units)**
Single Track	0.8	0
Double track	0	0.8

The output is encoded onto two outputs one for each track type, so that the classification errors are reduced. For this simple case this is not much of an issue as there are only two categories.

Figure 38 shows the outputs from the training. The raw outputs are thresholded and this can be done in two ways. Either everything above 0.4 is set to 0.8 and everything below set to 0, or a confidence zone can be used instead. For example everything above 0.5 is set to 0.8 and everything below 0.3 is set to zero, and outputs in the range 0.3→0.5 are flagged as 'unsure'. This then gives some idea of the confidence of the classification and anything that is borderline can be examined more closely. After the threshold the network value and target should be identical if not then it has been misclassified and is flagged as a failure.

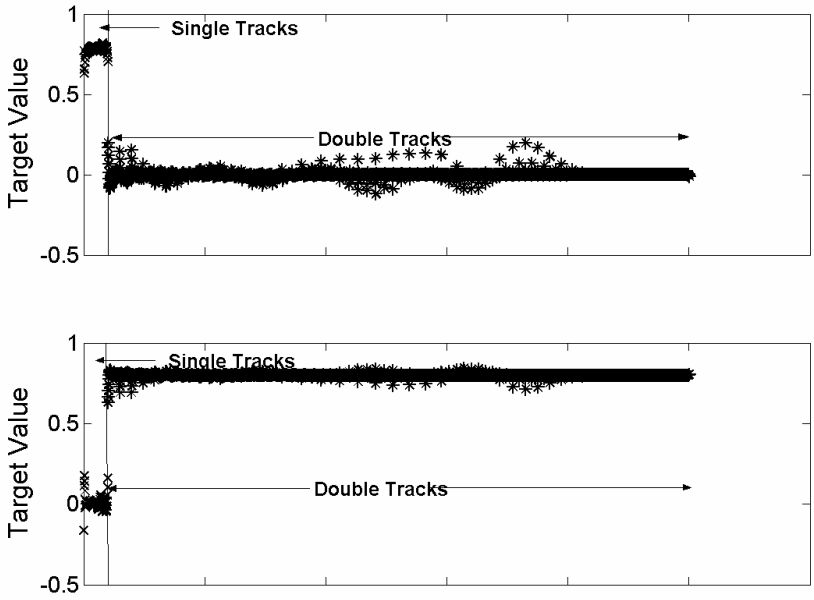


Figure 38 – Single and double track classifier results. top:target1 bottom:target2

As can be seen in Figure 38 none of the outputs would be in our borderline range for this level and so every track was classified correctly as shown in Figure 39.

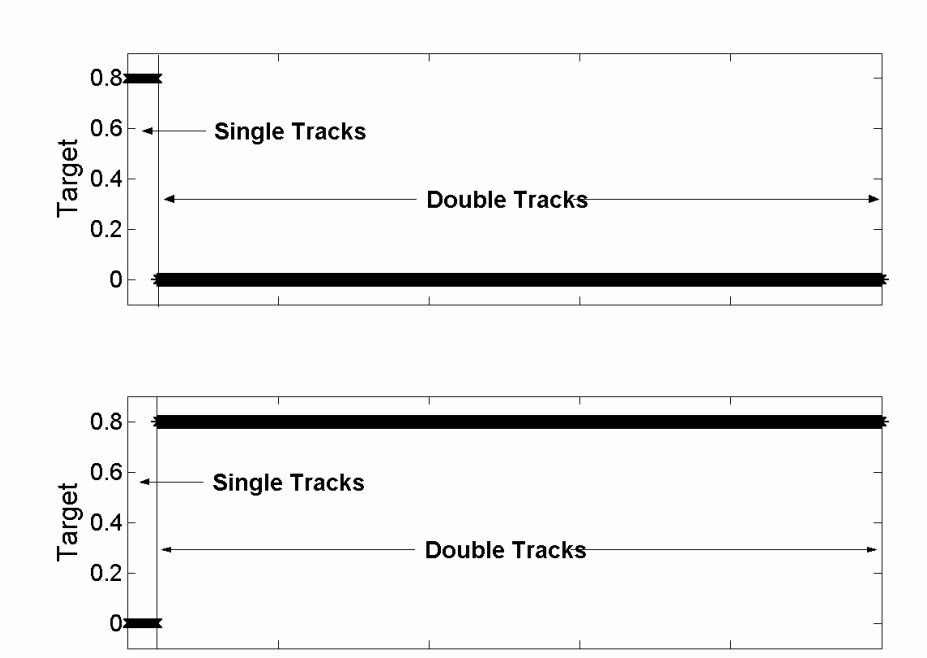


Figure 39 threshold results for single and double track classifier. top: target 1 bottom:target2

4.8 Requirements on the optical system

The simulations presented show that this approach is very good at extending the measurement capability of an optical system. Systems with 0.3NA have been shown to be able to measure track widths substantially below 100nm for both single and double track structures. The most important feature for the optical system is that it provides very repeatable measurements with high signal to noise ratio as this the limiting factor in the final training error. The optical system will ideally be able to measure phase profiles, as many of the samples of interest will be phase objects. One other consideration for the optical system is its ease of use as this system will eventually be used to provide standard measurements in a non-research environment, so the system has to be easy to align and to operate.

The optical systems used for the experimental will now be discussed in some detail, example scans of several samples will demonstrate the suitability of the systems for use with this technique.

5 Optical Systems

Our technique requires the use of an optical system to measure the samples of interest. The actual system used is not important as long as the key requirements of the technique are met as discussed at the end of the last chapter. Three optical systems have been used to obtain profiles that can be used for training. The first is an ultra stable common path diffractive element scanning interferometer [60], the second a differential scanning optical microscope [61] and the third is a scanning Nomarski system [9] that has various modes of operation depending on the configuration. These systems will now be discussed in turn.

5.1 Ultra Stable Common path diffractive element scanning interferometer

The first system used to obtain surface profiles is an ultra stable common path scanning optical interferometer [60]. Because of the common path nature of the system, effects of microphonics due to background vibrations and thermal gradients are greatly reduced, thus allowing the system to perform close to the shot noise limit.

The system uses a computer generated holographic (CGH) diffractive element as the beamsplitter. The arrangement between the objective lens and the hologram (zone plate) is shown in Figure 40. The CGH creates two output beams from a collimated input beam. The first is an unaltered zero order which is focused onto the sample by the objective, this acts as the sample probe beam. The second is a first order beam, converging to the back focal plane of the objective. The objective then collimates the beam onto the sample

surface at some angle depending on the lateral offset of the hologram with respect to the optical axis, and this beam serves as the reference.

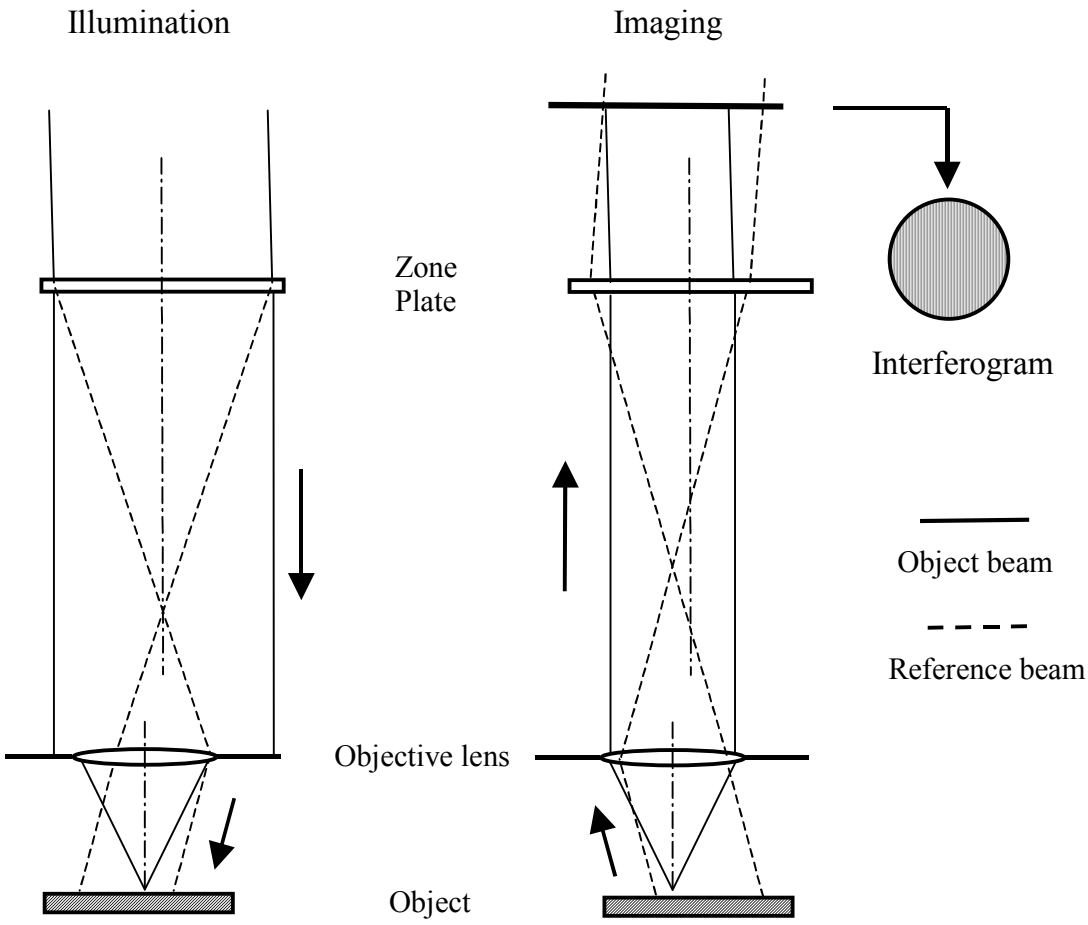


Figure 40 - Hologram & objective alignment

The two returning beams are recombined by the hologram and interfere to form straight fringes, the frequency of which is set by the angle of incidence of the collimated beam at the sample surface. Local surface height variations will change the phase of the probe beam, whereas the average phase of the reference will remain essentially unchanged.

The phase and amplitude profiles are obtained by recording the complex amplitude of the Fourier component due to the fringe frequency at each scan point. It should be noted that the two light beams traverse the optical system through similar paths, and the effects of

microphonics will largely be cancelled when the two beams interfere. This will improve the stability of the system, and allows the system to perform close to its fundamental limits.

Figure 41(a) shows the interferogram recorded by a CCD camera, (b) shows the signal for a single row from the detector (c) shows the Fourier transform of a single row. The amplitude and phase profiles are produced by recording the complex amplitude of the spectral component at the fringe frequency for each scan location. The image was taken using a wavelength of 633nm and an objective with 0.3 NA.

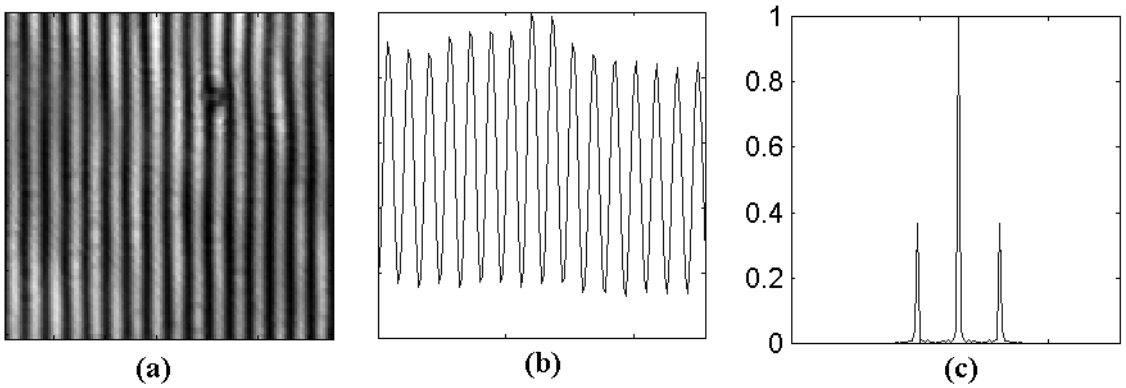


Figure 41 The interferogram and spectrum

The fringe contrast is approximately 0.8 and as can be seen the peak due to the fringe frequency is very sharp showing that the fringe pattern is very uniform. This is achieved by spatial filtering in the imaging arm. This is required because the hologram not only produces the +1 order we require but also the -1 order and the +/-3 three orders as well. These additional orders also interfere and therefore affect the fringe pattern. They are, however, easy to remove as a set of spatial filters at the Fourier plane in the imaging

optics can remove the higher orders and produce the clean fringe pattern shown in Figure 41a.

An example of the use of the system is shown in Figure 42, where it was used to measure a 100nm high and 40 micron pitch phase grating. The phase profile was obtained by scanning 100x80 microns across the grating, which took approximately 20 minutes to obtain.

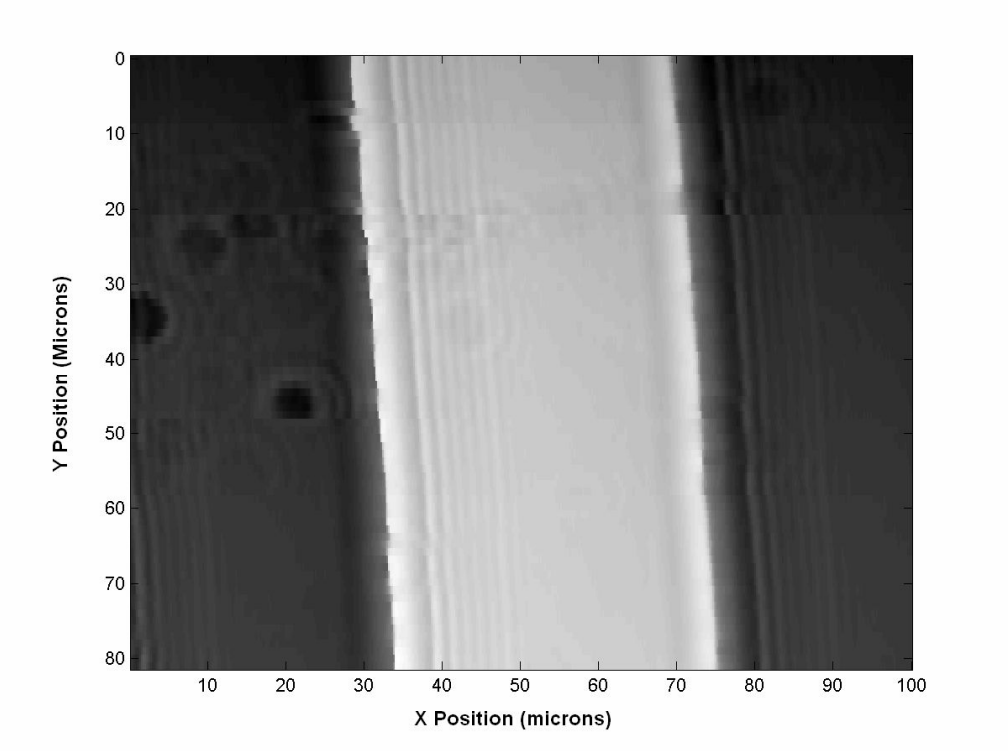


Figure 42 - 2D scan 40 micron pitch 100nm high sample

The dimensions of the second grating were identical to the first except that the grating step height was 17nm. This system successfully measured this sample and a 2d scan of the sample is shown in Figure 43.

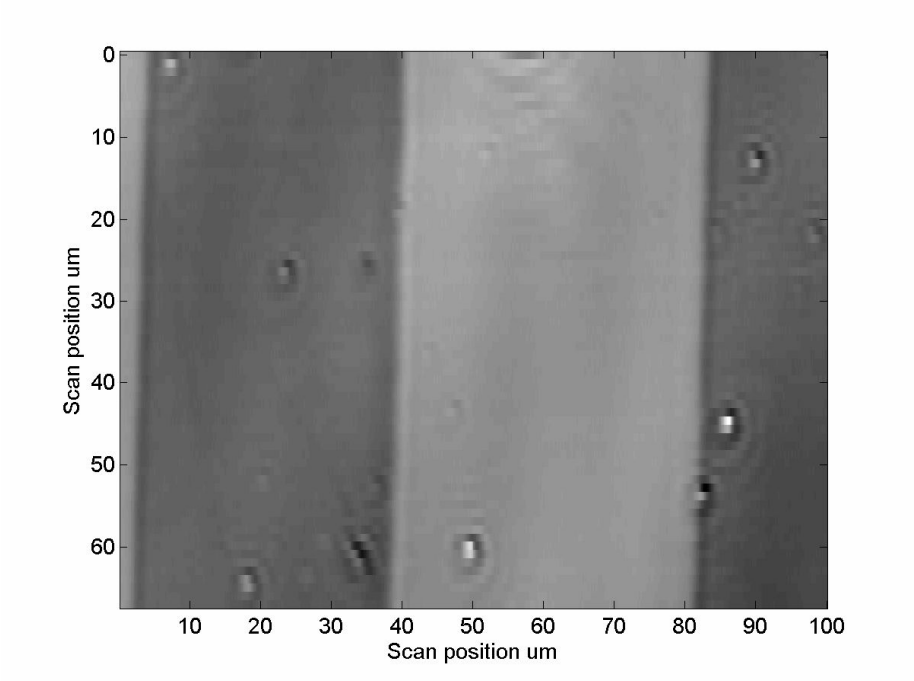


Figure 43 - 2D scan 40 micron pitch 17nm high sample

The signal to noise ratio for the amplitude signal is around 3000:1 and the phase noise has a standard deviation of around 0.5 mrad. These were obtained by recording the complex amplitude while the system was stationary (no scanning took place). The mean value and standard deviations of the amplitude over 1000 data points was used to produce the SNR and the phase standard deviation was calculated for the same 1000 data points.

The system is very repeatable as demonstrated in Figure 44. Four scans of the same location on the sample are plotted on the same graph and the difference between two runs is shown in Figure 44. All four scans have very similar responses showing that the system is stable.

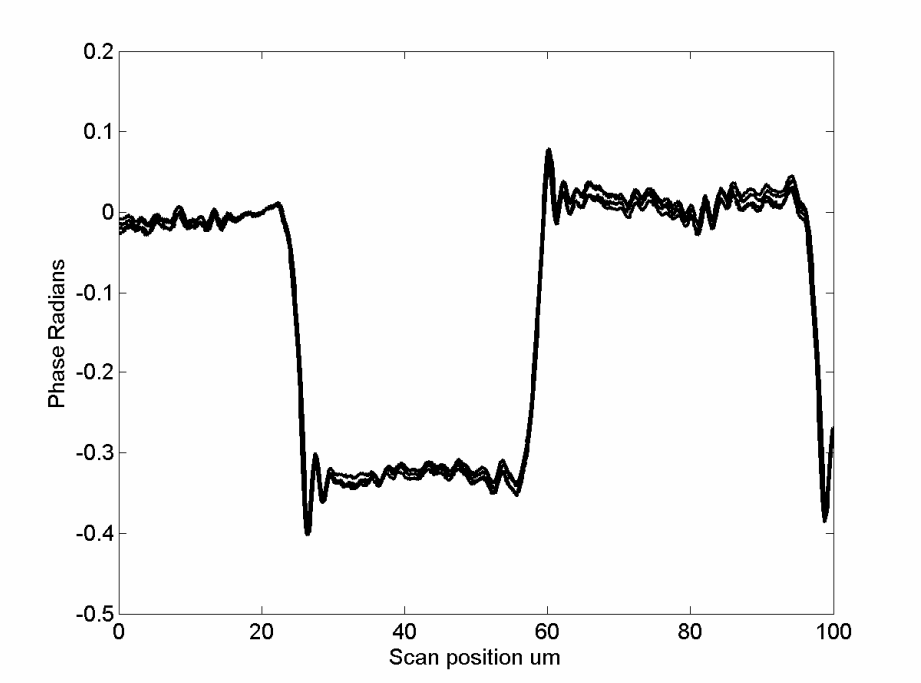


Figure 44 - Repeatability of system

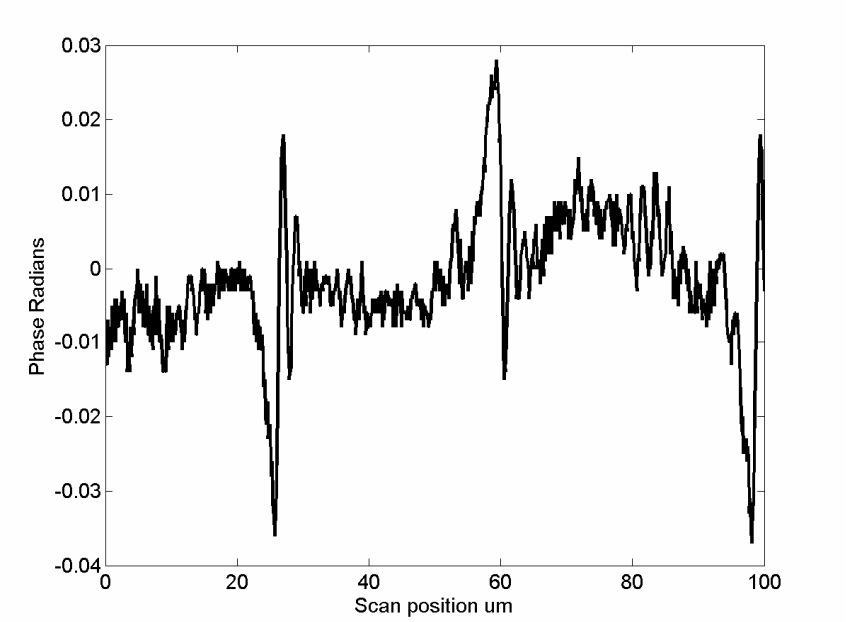


Figure 45 - Difference between two runs

The standard deviation between two runs of the system was 0.008 radians, as can be seen in Figure 45 most of this comes from difference at the transitions between the two phase levels of the sample which could be due to the sample moving laterally between scans.

This could be have been caused by the stage not returning to exactly the same location or thermal drift causing expansion of the sample stages.

5.1.1 Practical considerations

This system can measure all types of objects as it records both the phase and amplitude profiles as well as the intensity signal. The only issue regarding the types of samples that can be measured with this system is the way in which the reference is affected by features on the sample. Ideally there should be no nearby large structures, as this will significantly change the phase profile of the reference beam. Instead of the reference beam being a flat average phase beam, which interferes to produce straight fringes, it will instead have some structure and this will cause the fringe pattern to lose uniformity.

One consideration for the optical system used is the ease of use of the system, as ultimately this system will be used by people making standard measurements of samples and the reliability of the results will depend on their ability to set up, align and be able to spot any problems arising with the optical system.

This system is rather complex to set up and align and also has the disadvantage that precise focusing is difficult as the assumption is that in focus operation is achieved when the fringes are perfectly parallel. Not only does this rely on the system being correctly aligned but also it can be difficult to spot slight curvature on the fringes, especially when the fringe frequency is high. This would lead to operation when the system is defocused. However, ideally in the final system there should be no need for the end user to have to make adjustments to the system.

With robust mechanics regarding the system construction some of these can be easily removed. The main source of error on the fringes will come from misalignment of the hologram and objective but by using specially made holders set at exactly the correct distance this error can be reduced. Checking the focus should be done when the lateral offset of the hologram is small as the curvature on low fringe frequency fringe patterns is much easier to see and adjust for. Using these ideas the system would be much more user friendly.

5.2 DSOM – differential scanning optical microscope

The DSOM [61] is a simple scanning optical microscope where the sample of interest is scanned with a focused beam. The imaging arm magnifies this beam greatly so that an image of the point spread function on the sample is obtained at the CCD camera. The differential signal is obtained by using two regions offset from the centre of the point spread function. The difference between these two regions forms the differential signal. The system is confocal with two displaced pinholes in the image place. A schematic of the system is shown in Figure 46.

This system allows differentiation in any arbitrary direction depending on the choice of the location of the regions used. Usually the region is parallel to the scan direction and perpendicular to the object track. Figure 47 shows the location of two such regions on the point spread function at the image plane

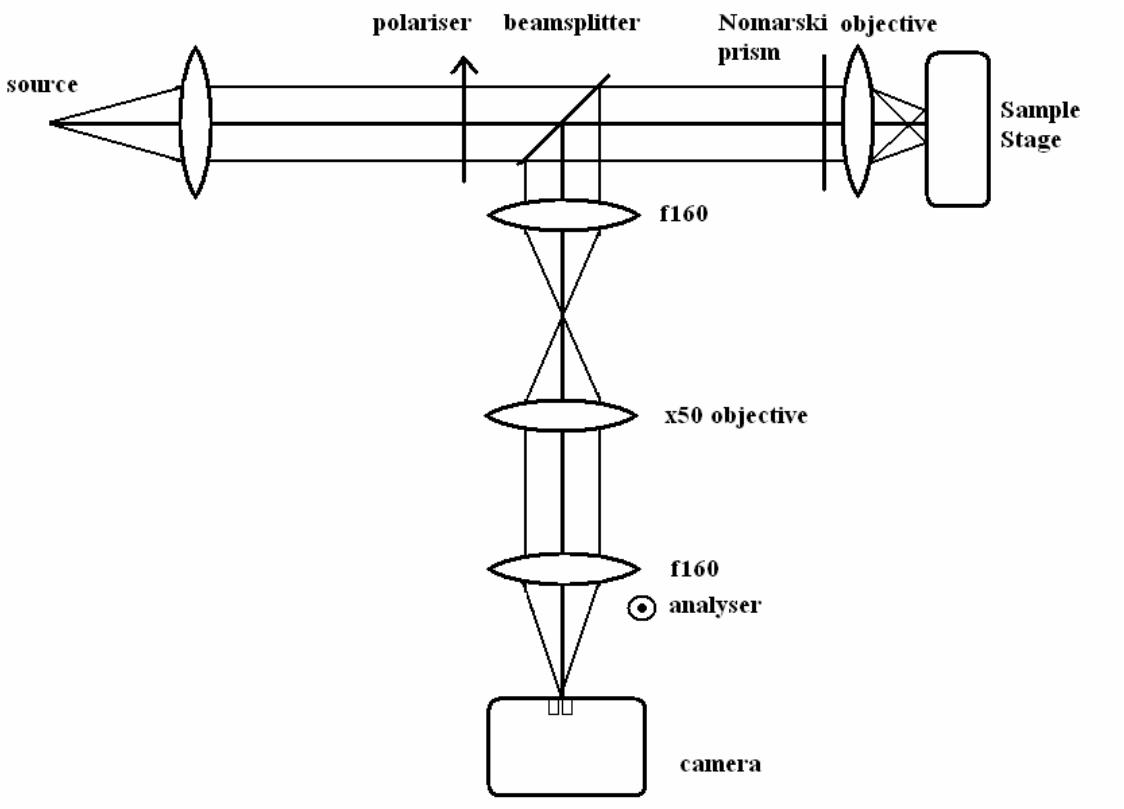


Figure 46 - DSOM setup

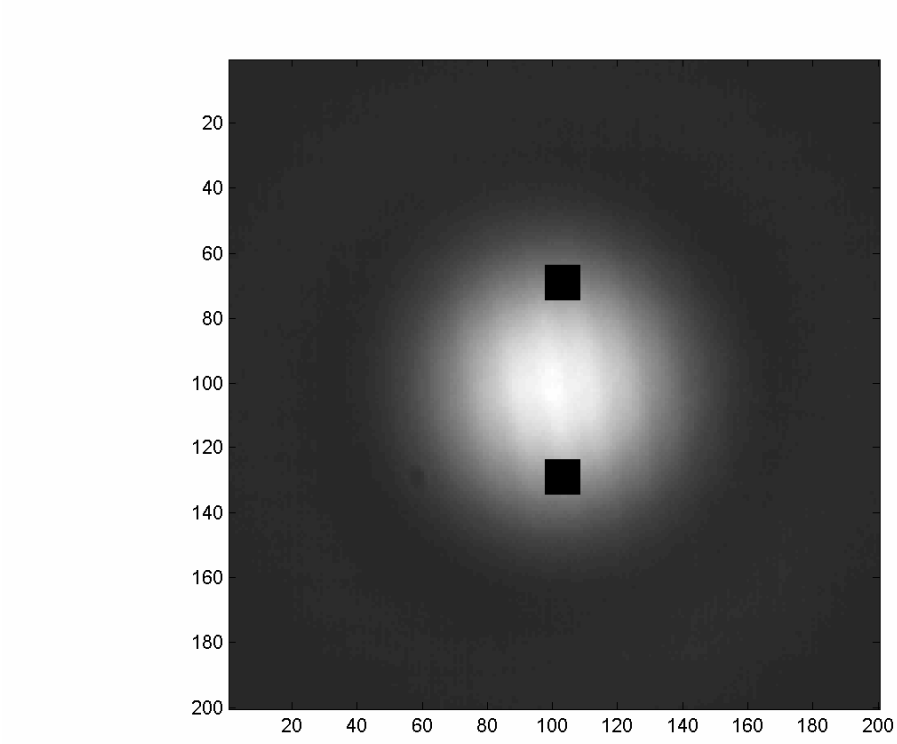


Figure 47 - Location of windows on psf

The point spread function is approximately 140 pixels wide on the CCD camera. The size of the small windows used for the differential intensity measurements are usually between 5x5 and 11x11 pixels. Using larger windows increases the signal level and so reduces the impact of noise. However if the window is too large then sensitivity becomes a problem. The location of the two windows is very important. They must be exactly the same distance from the centre of the PSF otherwise they do not cancel completely when no object signal is present. This makes the scans of tracks asymmetrical and makes the data harder to use for training an ANN.

An intensity profile is also measured by integrating a 150x150 or 200x200 window centred on the PSF. This can be used to monitor the laser output during drift measurements or to provide non-differential profiles for training ANNs. The whole camera field is not used as there is noise on each pixel of the camera even if there is no signal. So using a larger window adds noise to the result without increasing the signal level.

Imaging Equation:

The intensity and differential intensity profiles are obtained by scanning the sample under the focused beam.

The focus beam is simply the Fourier transform of the aperture (P) of the objective lens

$$\Im\{P\} = PSF \qquad \text{Equation 5-1}$$

At the sample service, for scan location x_s this is modified by the object under measurement

$$PSF \times obj(x_s) \qquad \text{Equation 5-2}$$

Finally the object plane is re-imaged to the CCD camera (magnification assumed 1 here for simplicity)

$$\Im\{\Im\{PSF \times obj(x_s)\} \times P\} \qquad \text{Equation 5-3}$$

The intensity for scan location x_s is obtained by integrating the entire camera field. The differential intensity is obtained by subtracting the intensity value for I_1 and I_2. Where I_1 and I_2 are different sub-regions of the camera field as shown in equations 5-4→5-7 below.

$$I(x_s) = \iint \Im\{\Im\{PSF \times obj(x_s)\} \times P\}\, dxdy$$
$$I_1(x_s) = \iint_{region1} \Im\{\Im\{PSF \times obj(x_s)\} \times P\}\, dxdy$$
$$I_2(x_s) = \iint_{region2} \Im\{\Im\{PSF \times obj(x_s)\} \times P\}\, dxdy$$
$$dI(x_s) = I_1(x_s) - I_2(x_s)$$

Equations 5-4→5-7

5.2.1 Shot Noise

The expected levels of photon noise are calculated and actual noise levels are presented along with repeatability measurements to show the stability of the system.

5.2.1.1 Photon Noise Simulation

The level of photon noise for the reference window for different sized windows is given in the table below. (Based on a saturation level of 100000 photons per CCD pixel) this was calculated by simulating the point spread function at the CCD camera and scaling the image in terms of maximum photons. The detector was located offset to the right of the centre of the point spread function by one quarter of the optical spot size.

Table 19 - Photon noise and window Size

Window size	Photons	SNR
3x3 pixels	2.17×10^5	465
5x5 pixels	6.03×10^5	776
7x7 pixels	1.18×10^6	1086

9x9 pixels	1.96×10^6	1400
11x11 pixels	2.93×10^6	1711

The noise level for the 9x9 or 11x11 window is more than adequate for this system. The drift that affects the system is usually fairly linear across the measurement interval.

There are several sources of error that could be contained in the differential signal, one source is the optical power fluctuations in the laser and this is monitored by integrating the entire field of the PSF on the CCD camera and monitoring the overall value during scanning. Another is that the point spread function could be moving with respect to the detector window either due to thermal effects or vibrations.

This drift is shown in Figure 48 for the larger window (150x150 pixels) and Figure 49 for the small right hand side window. The graph is approximately 30 minutes of data; the drift is fairly linear and is probably due to thermal effects there are several features on the data these are probably due environment changes during the scan (for example doors opening etc.)

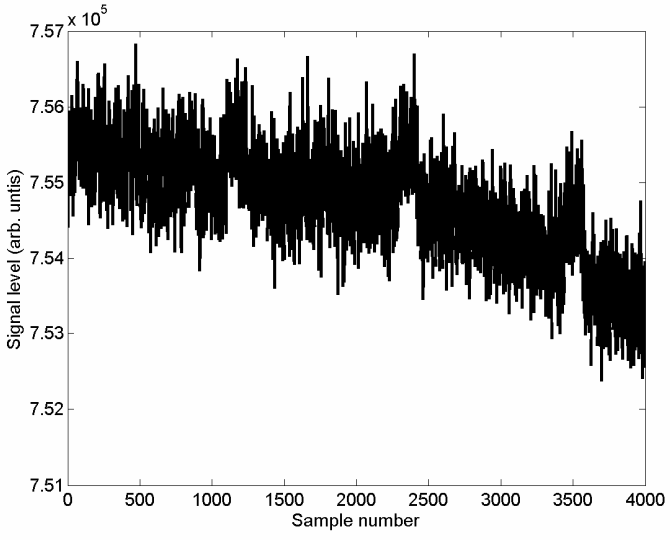


Figure 48 Noise on intensity signal

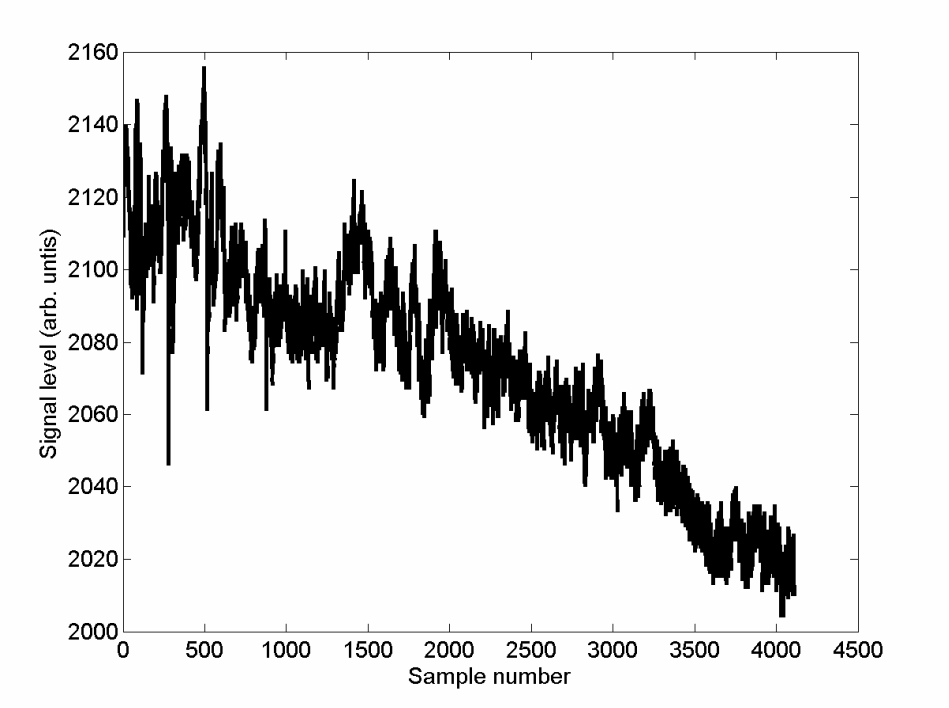


Figure 49 noise on right hand signal

The data is divided into 1000 point sections and the mean and standard deviation is calculated both before and after a linear fit is removed from each section of the data. The camera offset of around 30 levels per pixel has been removed.

Table 20 (a) & (b) results for Right hand and Sum for sections 1-4

Sum Window (a)

	Std with gradient	**Std without gradient**	**Mean signal level**	**SNR with gradient**	**SNR without gradient**
1	449	435	755266	1680.8	1736.2
2	501	467	755023	1506.1	1613.4
3	530	481	754727	1422.1	1566.0
4	571	476	753909	1320.2	1581.7
all	768	512	754687	982.1	1472.2

Right hand detector (b)

	Std with gradient	Std without gradient	Mean signal level	SNR with gradient	SNR without gradient
1	14.2	14.2	2117	148.5	148.5
2	14.5	11.3	2101	144.6	186.0
3	7.5	6.6	2092	278.8	313.2
4	8.8	7.4	2078	235.0	279.8
all	19.0	10.9	2096	110.3	192.1

Typical values for the noise level on the measurements were:

- For the intensity data with a window of 200x200 SNR typically 1 in 1600
- For the right or left signal with 5x5 window typically 1 in 200.

The noise level is much worse than the photon noise limit and this is mainly caused by vibrations in the system. A detailed look at the effects of vibration is shown in appendix 3

A SNR of 1600 for the whole window corresponds to a vibration level of much more the 25% of a pixel. A SNR of 1 in 200 for the right hand detector corresponds to a vibration level of between 5-10% of a pixel, which is equivalent to 550-1100nm of camera vibration or 1.2-2.4nm vibration of components before the magnification arm such as the main objective or the lens used to collimate laser light. Any vibration before the imaging arm has a greater impact on the noise due to the high magnification of the optical system. The bench top design of the system means that it is more susceptible to thermal drift as the components making up the system are isolated (they are not mechanically fixed together); this could be reduced with a better mechanical design of the system. The environment in which the system operates is also far from ideal.

The whole window data is much noisier than it should be for the above to be consistent, the reason for this is that the large window size is much bigger than point spread function and there is a background on the CCD of around 30 levels. This means that each of the pixels outside of the range of the point spread function will have noise associated with it due to the background but no signal. This will reduce the signal to noise ratio considerably. Therefore for better noise performance a smaller window is better for the intensity signal.

The repeatability of the DSOM system is very good. The images below show 4 scans of a track of 2.2 microns. Figure 50 is 4 intensity profiles, Figure 51 is 4 differential intensity profiles and Figure 52 is the difference between two of the intensity profiles.

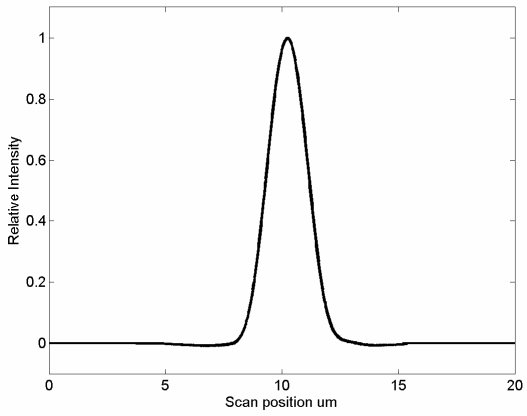


Figure 50 - Repeatability of Intensity profiles - 4 Scans

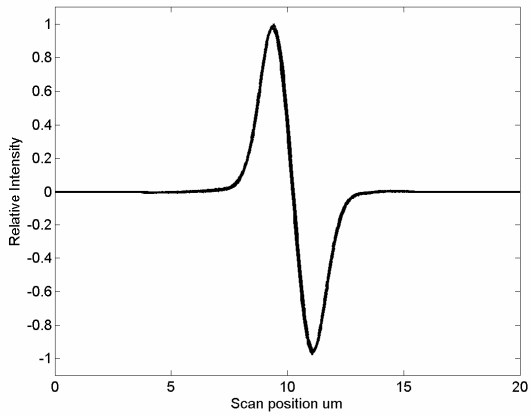


Figure 51 - Repeatability of Differential Intensity profiles - 4 scans

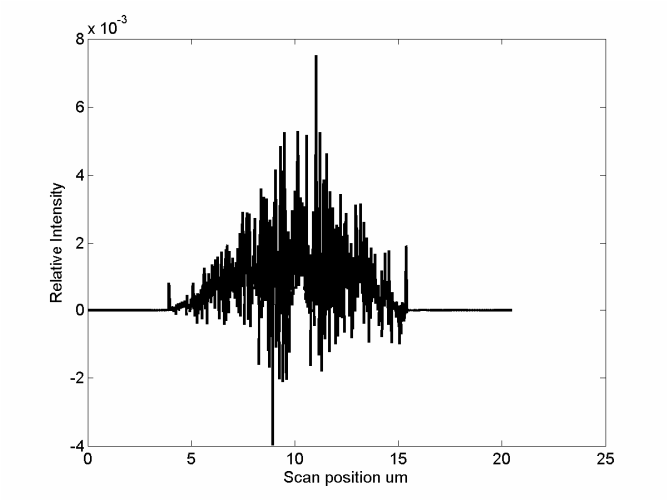


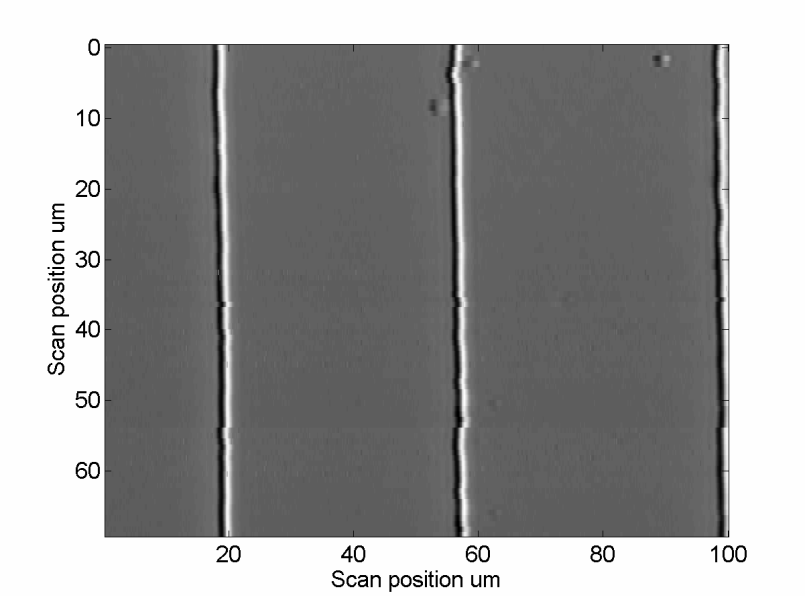
Figure 52 - Difference between two Intensity profiles

Figure 50 & Figure 51 show very little variation and the repeatability of the measurements is clearly excellent. The standard deviation of the difference between two scans is approximately 450. The peak signal level is around $4.5x10^5$ the change between two runs is therefore very small.

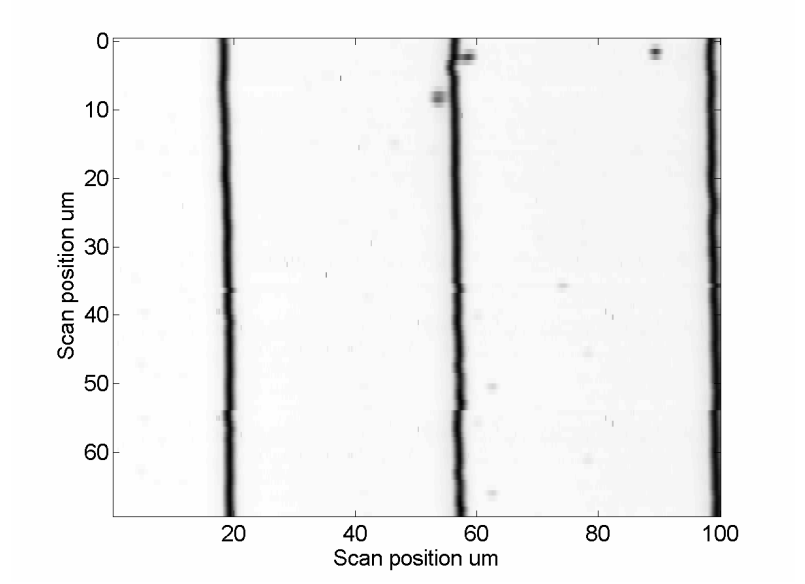
The types of sample that this microscope is suitable for is restricted because the system only measures intensity or differential intensity. Samples are therefore limited to ones with variation in reflectivity or large phase objects where the scattering is significant.

However, this system is can be converted to measure differential phase either as a homodyne using phase stepping techniques or a heterodyne interferometer. The advantages of this are that the system design is still relative simple. This system is also very flexible because the system response can be modified by changing the separation of the pinhole detectors.

As for the previous system a sample has been measured to demonstrate the system operation. The first sample is 100nm high 40 micron pitch phase grating. This is a purely phase structure so the signal obtained from this will be due to scattering alone.



(a) differential intensity



(b) intensity of phase grating

Figure 53 2d Scan of grating sample h=100nm

The images are 90x100 microns. The jagged edges or the gaps in the images could be caused by environmental changes during the scan or the peizo stage not returning to exactly the same location for each line of the image. The differential intensity image shows a dark then bright line at each phase transition, which is the differential signal due to scattering at the grating edges. The intensity grating image (b) shows dark lines just due to scattering. The sample is relatively clean and the surface seems fairly uniform from these images.

5.2.2 Practical considerations

The DSOM system is very simple, making it relatively easy to set up and align. The most critical component is the x50 objective in the imaging arm, as incorrect position can lead to large aberrations of the point spread function image at the camera. Correct focusing is easy to maintain as the point spread function is imaged onto the CCD camera. It is therefore relatively simple to create an auto focus system to keep the system conditions the same for each track scan.

5.3 Scanning Nomarski

Scanning Nomarski [9] system has been developed to provide differentiation on the surface of an object sample. The system setup is shown in Figure 52.

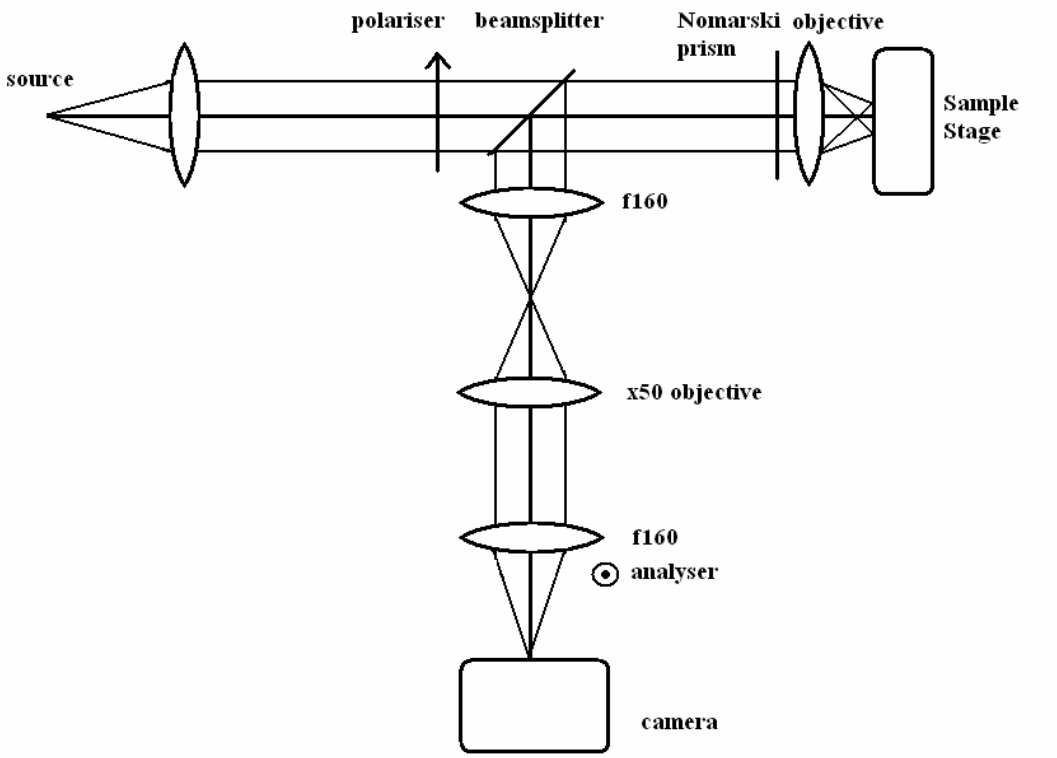


Figure 54 - Nomarski setup

The input light (polarised at 45 degrees) is split into two orthogonally polarised beams by the Nomarski prism. They propagate at an angle and are then focused by the objective onto the sample surface. This produces two point spread functions separated by a small distance related to the Nomarski angle and the focal length of the objective lens. The beams reflect off the sample and are recombined by the Nomarski prism. The analyser is used to mix the two beams and depending on the orientation different modes of operation are possible. The use of the CCD camera for the detector is for convenience, a photodiode could equally be used.

The system can operate in several distinct modes:

- Bright field differential interference microscope
- Dark field differential interference microscope
- A scanning microscope with either s/p polarisations

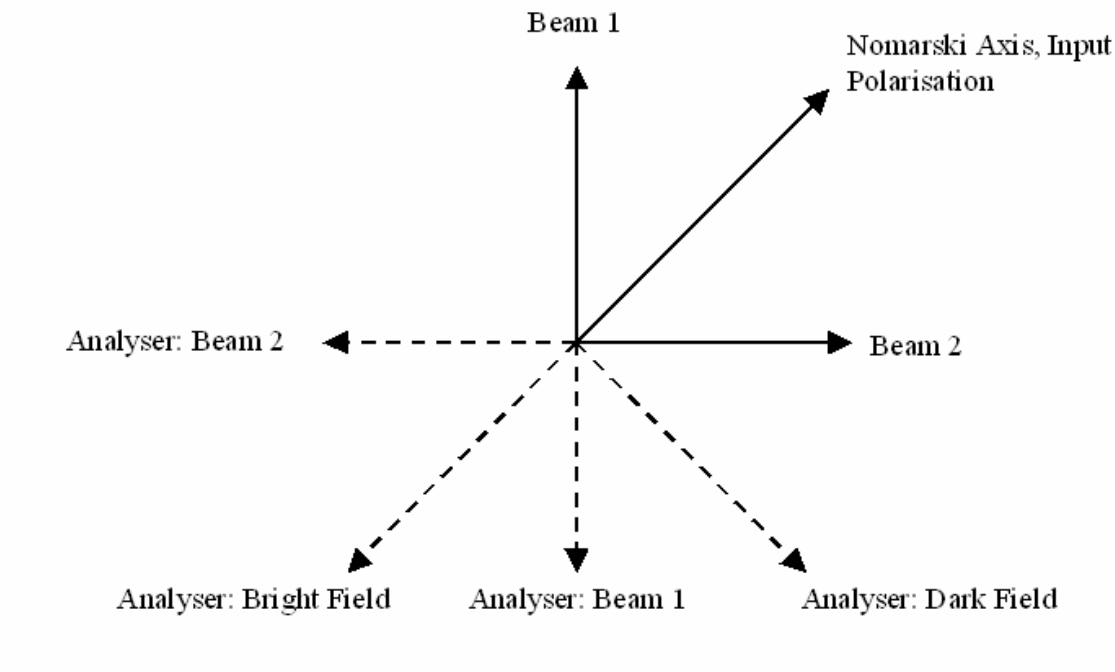


Figure 55 - Nomarski System Polarisation and Analyser Angles

The mathematics of the system and modes of operation are derived in appendix 4 and summarised in Table 21 below. The third column is the intensity signal if a uniform intensity, flat region of sample is observed such that the phase and reflectivity are the same for both probe beams. ϕ is the angle of the analyser with respect to the Nomarski prism axis (see Figure 54), E_1 and E_2 are the Amplitude of the object seen by the two beams respectively, ϑ_1 and ϑ_2 are the phase of the object seen by the two beams respectively.

Table 21 - Operating modes for scanning Nomarski system

Analyser Position (ϕ) degrees	**Intensity (I)**	**If $\vartheta_1 = \vartheta_2$ and $E_1 = E_2$**
90	$I = \frac{1}{2}E_1^2 + \frac{1}{2}E_2^2 - E_1E_2\cos(\vartheta_1 - \vartheta_2)$	I = 0 (dark field)

0	$I = \frac{1}{2}E_1^2 + \frac{1}{2}E_2^2 + E_1E_2\cos(\vartheta_1 - \vartheta_2)$	$I = 2E_1^2$ (bright field)
45	$I = 0 + E_2^2 + 0$	$I = E_2^2$ (SOM *Horizontal* polarisation)
-45	$I = E_1^2 + 0 + 0$	$I = E_1^2$ (SOM *Vertical* polarisation)

The beam separation on the sample surface can be easily calculated. This was achieved by capturing the interference fringes of the two beams after passing through the Nomarski prism. The objective had been removed and an analyser was placed in front of the camera. The wavelength of the fringes is a measure of the angle between the two beams

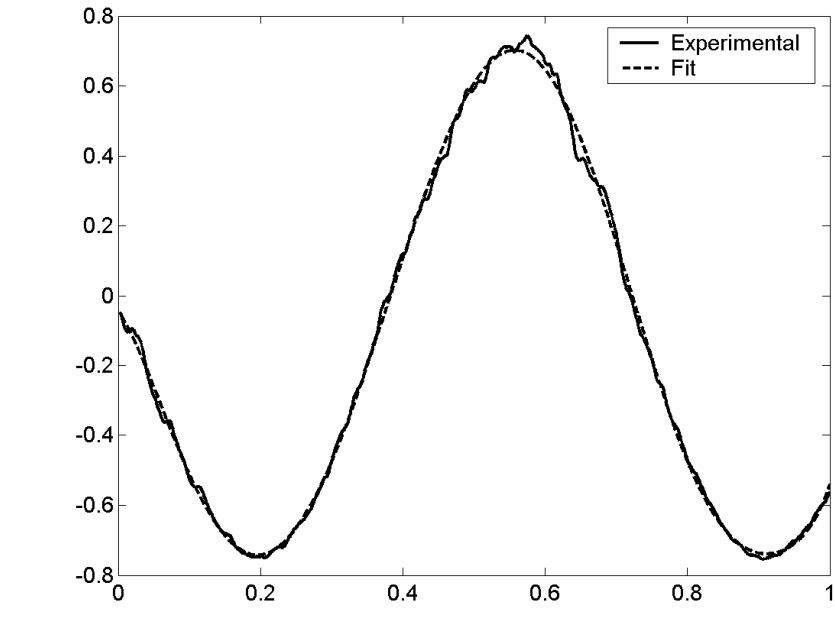


Figure 56 - Slice through fringes due to Nomarski prism

Figure 56 shows a slice through the fringes that were captured. The wavelength is approximately 430 pixels, which corresponds to 4.7 mm, as each camera pixel is 11 microns. This distance corresponds to a path length difference of one wavelength between the two beams. This leads to an angle between the beams of :

$$\sin\phi = \left(\frac{0.688}{4700}\right) = 1.454\text{x}10^{-4}\, radians$$

When the objective is replaced this leads to a beam separation on the sample surface of :

$$Separation = 18\text{x}10^{-3}\sin(\phi) = 2.62\text{x}10^{-6}\, m$$

For this setup the PSF is 2.79um and so the separation of the beams is of the order of the size of the PSF, which is not ideal as the edge response is approximately twice the width of the PSF, but we have no control over this as the angle introduced by the Nomarski prism is fixed.

5.3.1 Noise / repeatability / vibration etc / photon noise

A set of noise and drift experiments were carried out for different signal levels. The window size was 200x200. The intensities were recorded for the bright field condition when no scanning took place (Table 22). The camera offset was removed and the SNR calculated. This was repeated when the final analyser was removed (Table 23). This had the effect of leaving the system unchanged but the signal strength increased as all of the energy was integrated by the camera as opposed to a portion being blocked by the analyser. Finally the same process was repeated for the dark field position (Table 24), as there was no object information the mean signal level here should have been zero as the two beams should cancel completely and so the noise level should also have been zero. A linear fit was removed from the noise profiles to

remove the effects due to thermal drift, which was divided into 1000-point sections. The mean value and standard deviations are given for both sections and the whole file before and after the gradient was removed.

Table 22Noise measurements bright field

Data points	mean	std	SNR	std no gradient	SNR no gradient
0-1000	589431	700	842	549	1074
1000-2000	588532	541	1089	524	1124
2000-3000	587540	583	1009	503	1168
all	588284	1045	563	546	1078

Table 23 - Noise measurements no analyser

Data points	mean	std	SNR	std no gradient	SNR no gradient
0-1000	826148	542	1525	523	1580
1000-2000	825562	494	1672	494	1673
all	825739	607	1361	520	1587

Table 24 - Noise measurements dark field

Data points	mean	std	SNR	std no gradient	SNR no gradient
0-1000	9756	627	15.6	627	15.6
1000-2000	9843	626	15.7	625	15.8
all	9777	633	15.5	633	15. 5

The SNR for the no analyser option is higher because the mean value is higher, not because the standard deviation is any better. For example the mean of the standard deviations for the three Bright field sections is 608, the no analyser case is 518, 15%

smaller. Where as the no analysers mean signal level is 40% bigger than the bright field case. This implies that the noise variation is not due to just shot noise alone otherwise this would increase with signal level.

The noise in the dark field experiment above should be very small as the dark current is small compared to the shot noise; however, due to the camera offset and associated pixel noise, the noise level is much higher. This gives an indication of the impact of the camera offset on the noise level.

For bright field (BF) operation the SNR is good. For the dark field (DF) it is much more complicated to work out but it can be much better than for the bright field situation, as the signal strength is given by the difference between the two beams.

For the bright field case the signal on the camera is:

$$\mathrm{BF}: B_1 + B_2 \le C_{sat}$$

Where C_{sat} is the saturation level of the camera. For the dark field case it is:

$$\mathrm{DF}: B_1 - B_2 \le C_{sat}$$

Beam 2 is the same as beam one unless the object changes, it is this change δ that is of interest to us and so we wish to maximise the strength of this signal.

$$\mathrm{As}: B_1 = B_2 + \delta$$

So for the bright field case we have the signal of interest on the CCD but also a larger DC component due to the sum of the two beams. As the beams are much larger than the change δ the signal strength is limited by the DC power incident on the camera and the signal of interest has relatively little power.

$$\mathrm{BF}: 2B_1 + \delta \le C_{sat}$$

For the dark field case the entire signal recorded at the CCD is the signal of interest and so the power for this signal can be maximised

$$\mathrm{DF}: \delta \leq C_{sat}$$

This makes dark field very attractive for this application, as we will be measuring very small tracks and hence only receiving very small signals.

5.3.2 Repeatability

Not only is the noise performance of this system good but also the repeatability is excellent. Five scans of the same location were performed and the dark field profiles and the spectra are given in Figure 57.

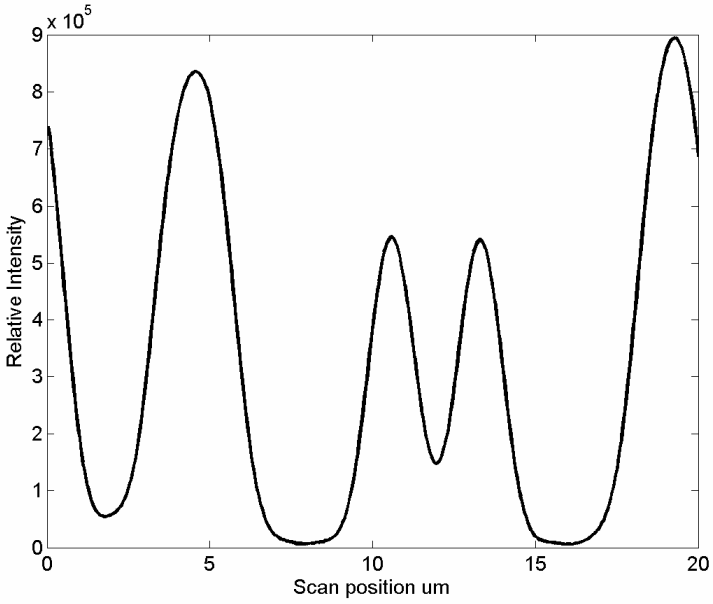


Figure 57 – repeatability - 5 scans

There is very little variation in the profiles, which shows how stable the system is.

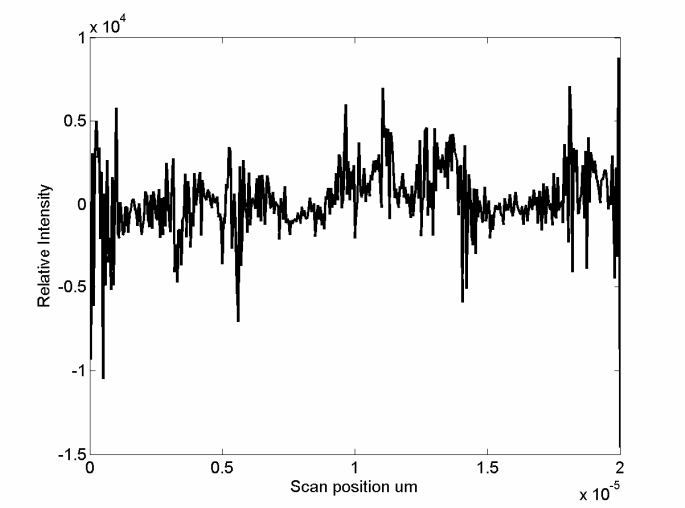
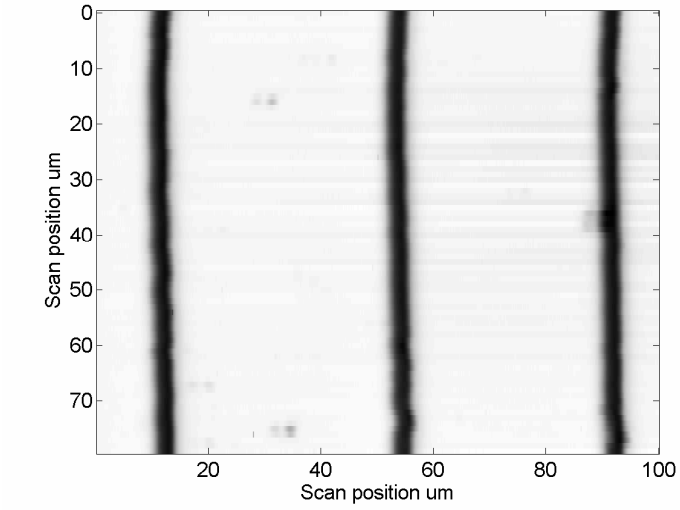


Figure 58 - Difference between two scans

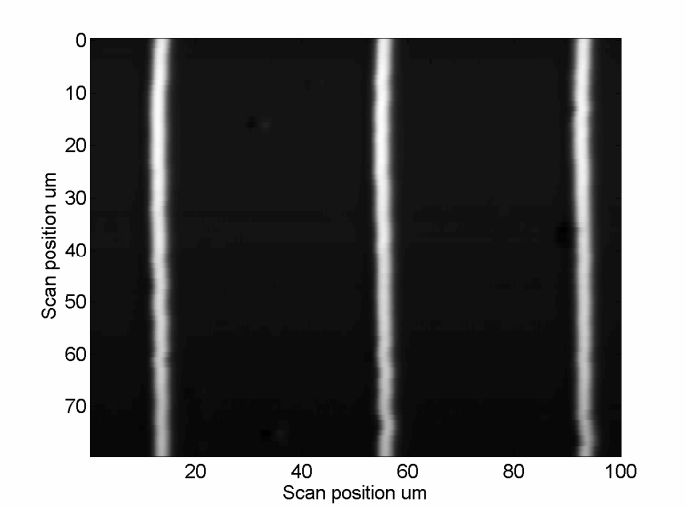
The repeatability is excellent; Figure 58 shows the difference between two scans, the standard deviation of the difference is 2281. The peak of the dark field signal is 894281.

5.3.3 Example scans

The same two samples have been measured with the scanning Nomarski system in both dark and bright field operation, shown in Figure 59.



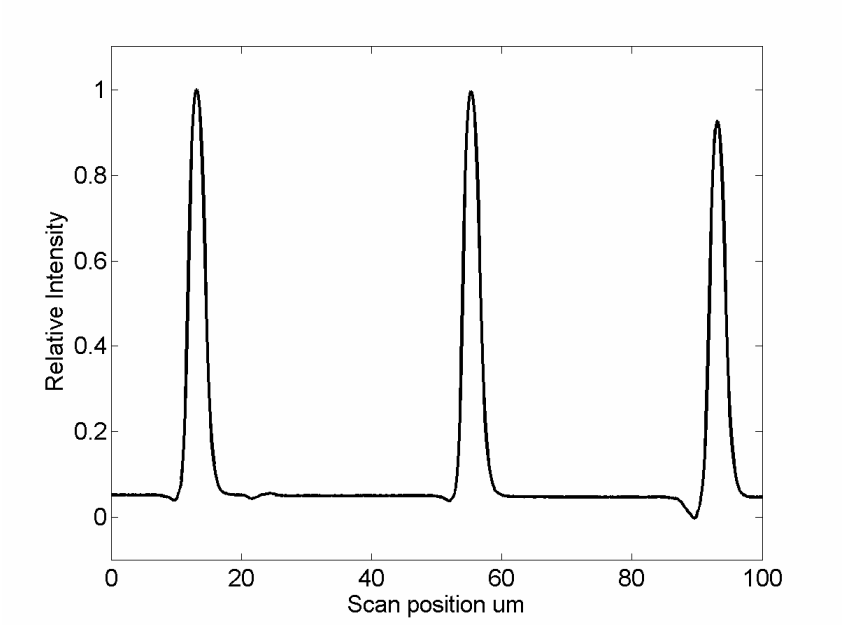
(a) Bright Field



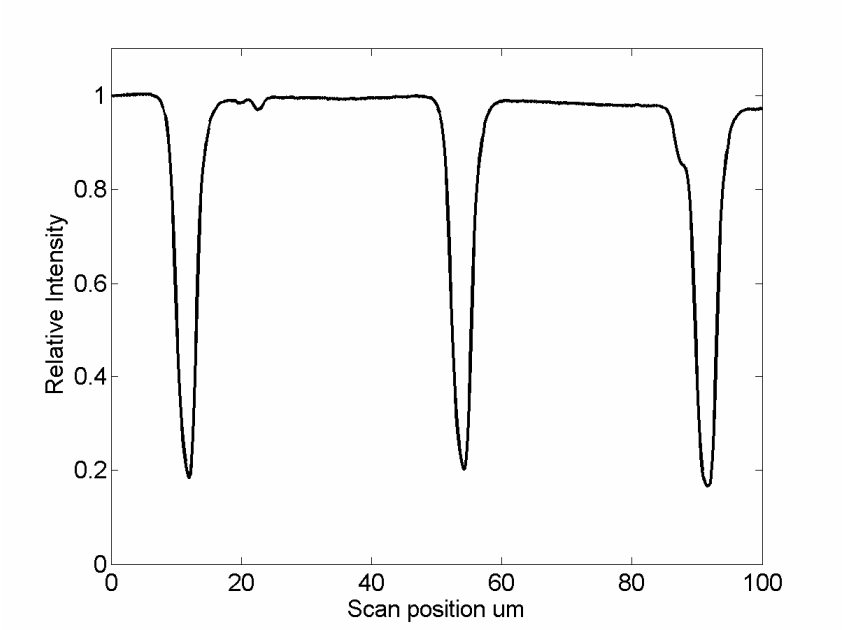
(b) Dark Field

Figure 59 - 2ds scan of 40 micron pitch 100nm high sample

Both the bright and dark field given fairly uniform images. A single line from the above images is given in Figure 60.



(a) dark field



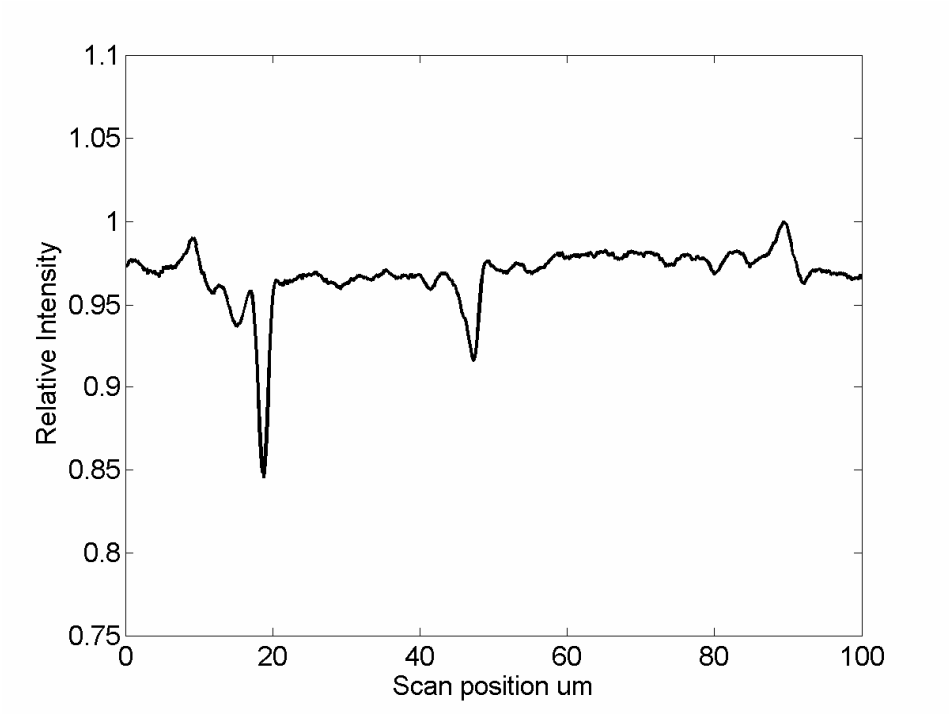
(b) bright field

Figure 60 line scan of the 40 micron pitch 100nm high sample

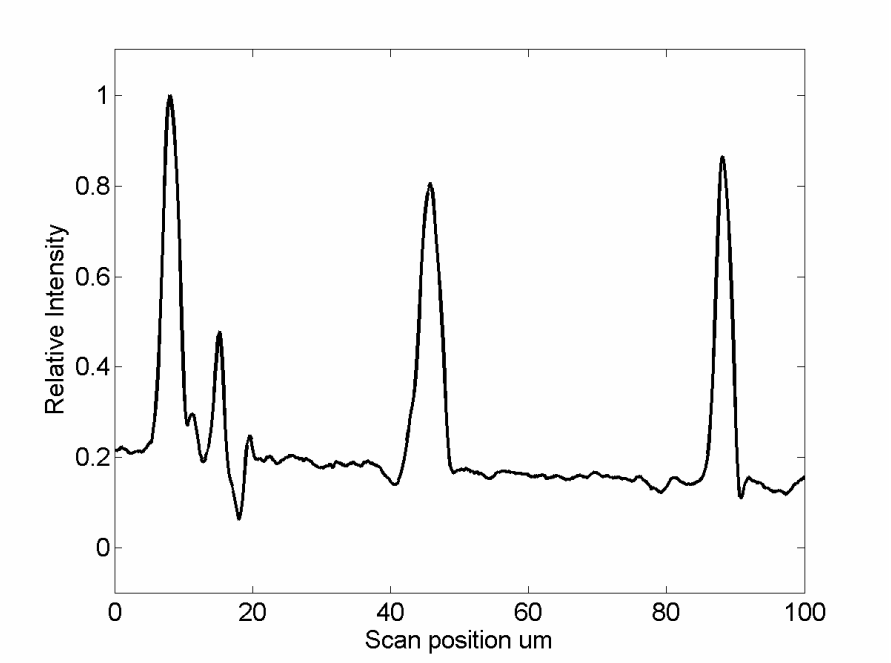
The dip/peak width for bright field and dark field is approximately twice the width of the point spread function. This is due to the two beams being separated by 2.62 microns as describes earlier.

The dark field minimum signal level in this case is also not completely zero because the beams did not completely cancel. This can be caused by several things, mismatch in beam intensity due to misalignment of Nomarski prism or input polariser. Also if there is a slight tilt on the sample or Nomarski prism, there will be a difference in the phase between the two beams and so there will always be a residual signal. The intensity level for the dark field signal was approximately 10 times higher than for the bright field case.

The 17nm high sample was also measured and line scans are presented in Figure 61.



(a) bright field



(b) dark field

Figure 61 - line scans of 40 micron pitch 17nm high sample

The bright field signal is very poor. The scattering due to the edges is very small and so the signal is hard to see. The variations in the sample surface are also visible. For the dark field case the light level was increased considerably. This had the effect of increasing the offset due to incomplete cancellation, but the signal to background strength is much better and there is much less noise on the signal. However dirt on the sample and/or surface scratches are still visible.

5.3.4 Practical considerations.

The alignment of this system is rather complicated as there is no easy way to get a reference for the input polarisation. The input polarisation must be at 0 degrees to the Nomarski prism axis, which in turn must be at 90 degrees to the scan direction otherwise the beams will be focused apart in both the scan direction and perpendicularly giving a differentiation angle of greater than zero degrees which leads to a reduction in the differential signal. Firstly the Nomarski axis is fixed and then the polarisation angle is fixed. The optical system setup is very similar to the DSOM just with added polarisation optics and the prism so the same considerations apply to this system except that the polarisation angles are also very important.

Once aligned the system is easy to use. Keeping the same focus is relatively easy in bright field mode but more difficult for dark field as the point spread function is not visible.

The type of samples that this system can measure is varied but there are some restrictions. While the system can measure phase objects, a phase profile is not obtained. If there are both phase and reflectivity variations then these two signals are

mixed and it is not possible to separate them. So samples must be purely phase or purely reflectivity objects.

The system can however be modified to perform phase stepping so that multiple scans can be used to obtain both the phase and amplitude profiles. This can be achieved by including a wave plate in the imaging arm that alters the phase of one beam with respect to the other. By scanning a multiple of times with varying phase difference between the two beams the phase and amplitude profiles can be recovered. The only draw back to this approach is the 4-fold increase in data acquisition time and the precise control of the phase step angle that is required.

5.4 Comparison tables and comments

The three systems all have advantages and disadvantages. They are suitable for a variety of sample measurements and all have reasonable to good signal to noise ratios. All of the systems have excellent repeatability. Table 25 compares the main features of the three systems discussed in this chapter.

Table 25 - Comparison of optical systems

	Hologram	**DSOM**	**Nomarski**
Sample types	All with some limitations	Reflectivity larger phase structures	All with some limitations
Amp	✓		
Phase	✓		
Intensity	✓	✓*	✓**
Practical SNR	Amplitude 1 in 3000 Phase	Intensity 1 in 1500 Differential	Bright field 1 in 1100 Dark field

	0.5mrad	intensity 1in 3-400	better
Complexity	Medium	Low	Medium
Ease of use	Medium	High	Medium

* and/or differential intensity

** and/or differential interference contrast

DSOM is easy to convert to homo/heterodyne differential phase and amplitude interferometer. Nomarski can measure phase by use of quarter wave plate before the analyser and using a phase stepping algorithm.

The systems described in this chapter have been used to measure a variety of samples to demonstrate the effectives of the combined ANN and optical system approach to measurement enhancement. Chapter 6 will present these results.

6 Experimental Results

Our method for extending the capabilities of an optical system has been tested experimentally using the optical systems described in the previous chapter. These experimental results will be presented in this chapter; firstly a detailed step-by-step example of all of the processing steps is given, before discussing the main results from the optical systems. Then follows several more specific options regarding training and other capabilities of the method. Finally results are presented for a double track object to obtain multiple parameters.

6.1 1-3 micron Sample

The following is a step-by-step example of the training process. The example used is the chrome on glass sample that has tracks ranging from 1-3 microns in 0.1 micron increments. The inputs are taken from the intensity profiles.

The sample was measured with the differential scanning optical microscope described in chapter 5, which produces differential and absolute intensity profiles. The objective used was a x50 Zeiss Epiplan with 0.7 NA, and the wavelength was 688nm. An aperture was placed at the back focal plane of the objective, which would allow the system to operate with an NA of 0.7 or 0.18. The sample was firstly positioned so that the tracks were perpendicular to the scan direction. This was achieved by approximately setting the direction of the sample. Fine adjustments were made by scanning the sample in x and changing the vertical (y axis) position by a known amount, for example 10 microns. By comparing a selection of scans with different y positions (10, 20 30 micron offsets) the shift in the x direction can be calculated. This

can be used with the known y separation to calculate the angle of the sample and then the sample can then be correctly orientated. If the sample is aligned correctly then moving the sample vertically should have no effect on the x position of the track.

It should be mentioned that for this case the targets for the ANN were experimentally obtained. At each scan location the sample was measured with the 0.18NA setup to obtain the profile and also the 0.7NA setup so that a high resolution image could be obtained so the actual width value could be calculated. This had to be carried out as the quality of the sample was such that the variation down the track length was considerable compared to the nominal value.

Each of the 20 tracks was measured 4 times at the same location to give a total set of 80 measurements. The multiple scans allow training with jittering, which improves the training performance when there are relatively few distinct input patterns. The scan increment was 40nm and the scan length was 20 microns. The example below will show the processing steps used for the intensity profiles. An intensity profile obtained with the DSOM is given in Figure 62 for a 1-micron track.

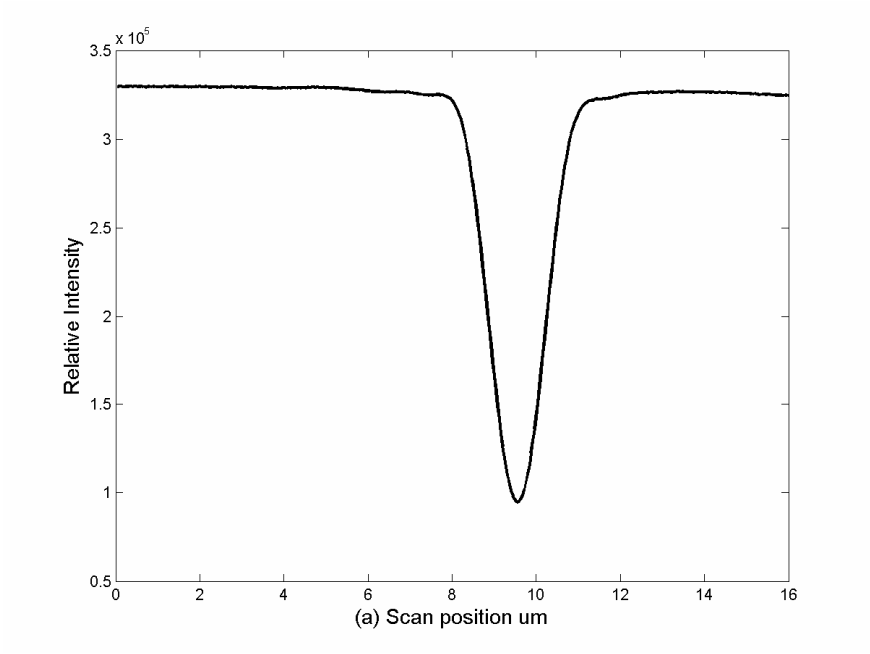


Figure 62 – Intensity profiles of 1 micron track

The profiles are normalised by the mean intensity level of a flat region to remove effects due to possible fluctuations in the laser light level in between scans. Each track is then centred into a padded file that is 2048 long. The padding value has the value of 1 to match the normalised intensity level for the intensity profile as shown in Figure 63.

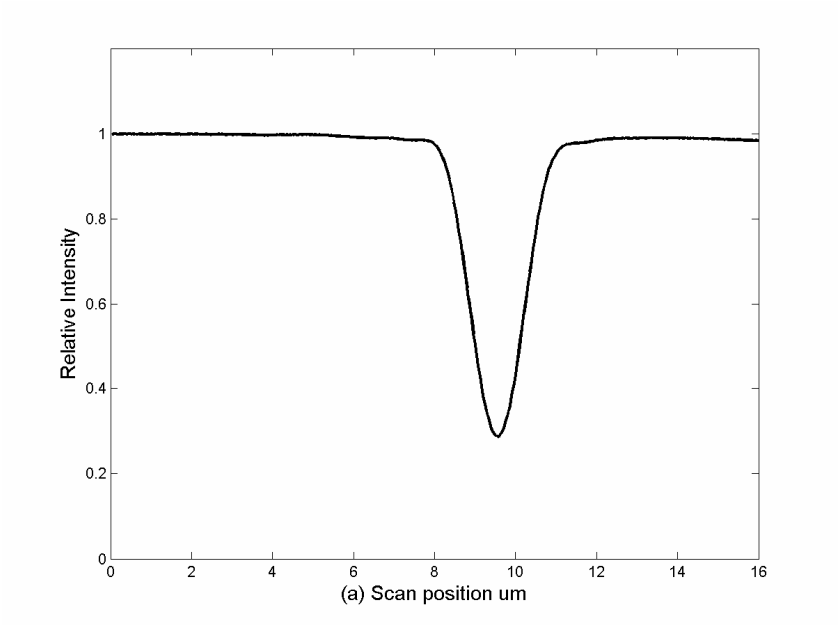


Figure 63 - normalised profiles intensity

The intensity profiles are differentiated in the frequency domain, using the Fourier shift theorem so that the difference distance can be controlled (see section 2.1). The distance is usually one quarter of the optical spot size as this is most appropriate for the ANN. The differential image obtained from the intensity profile is shown in Figure 64.

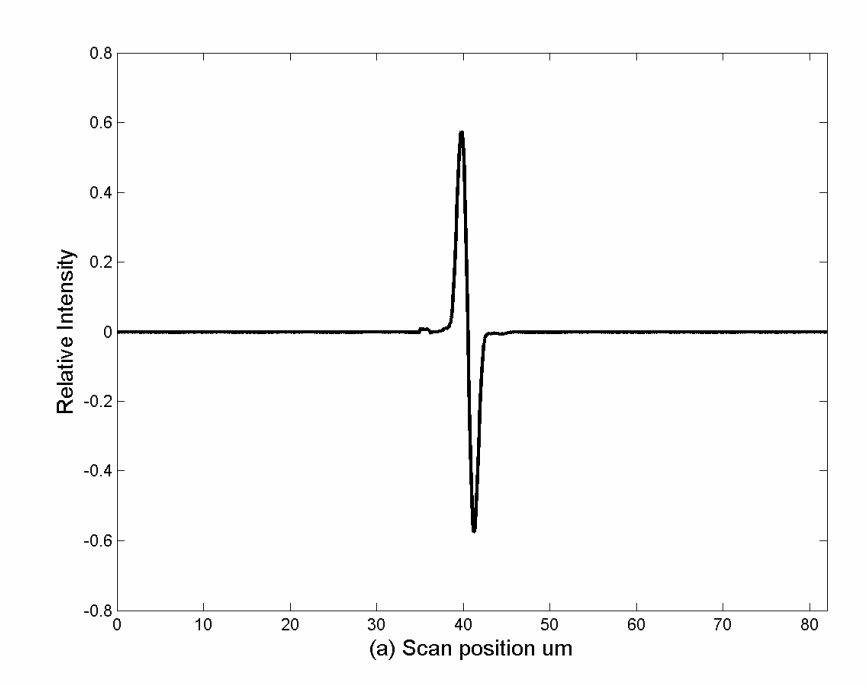


Figure 64 - difference image obtained by from the intensity profile

The magnitude of the spectrum for the difference image is shown below in Figure 65 for experimentally obtained profiles.

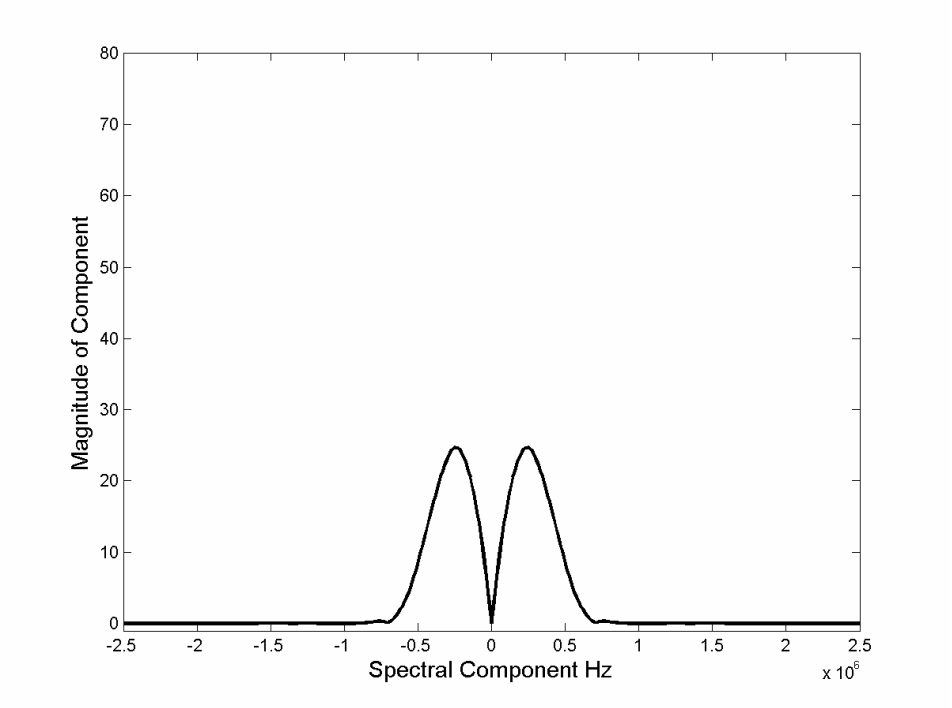


Figure 65 - spectra of differential image,

The positive half of the spectra is then sampled to give 8 input points. Once the samples have been obtained the whole set is scaled to the range 0- 0.8 this helps to keep the training stable as the weight values are kept small refer to chapter 4. The output targets corresponding to each track are also scaled to the range 0-0.8. The maximum output values used have to be less than one, as this is the saturation value of the tanh function used as the output activation function.

The data is then split into two sets. One set used for training and one set used to test the trained network. The split is usually 75% for training and 25% for the testing and the split is done randomly.

The network is then trained and once finished the difference between the target values and the network response (error) is calculated and the scaling removed to give the

answers in nanometres. This is done for both the training set and testing set. The results from this training can be seen in Figure 66.

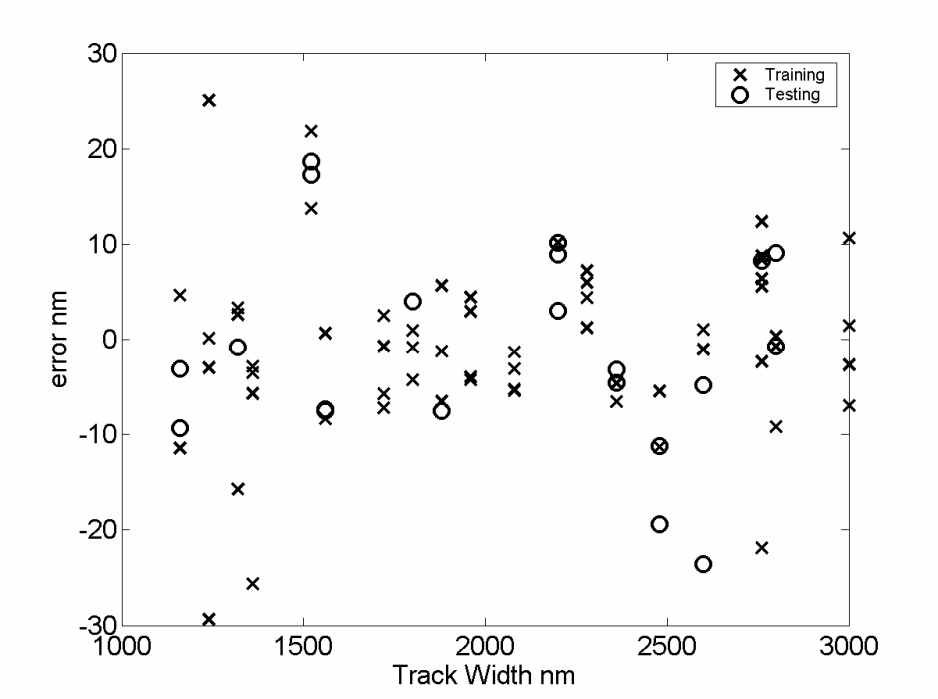


Figure 66 - Training results of 1-3 micron tracks

The standard deviation of the training set is 9.37nm and for the testing set 10.86nm and the means are -0.92 nm and 1.13nm respectively. This corresponds to an error standard deviation of 0.6% of the track width across the entire track range.

6.2 Traditional Approach for Track Width Measurement

Two traditional approaches used for calculating the track width of a profile have been used on intensity profiles obtained from the DSOM microscope for the BCR sample. The first approach was to look for the 25% intensity crossing for the tracks. The reference 100% level was taken as the intensity level from the centre of a reference pad. (See sample diagram figure 8). The 25% crossing points were then calculated and the widths determined. The value for the largest track was then scaled to correspond with the value measured with the 0.7NA objective and the results are plotted in Figure 67.

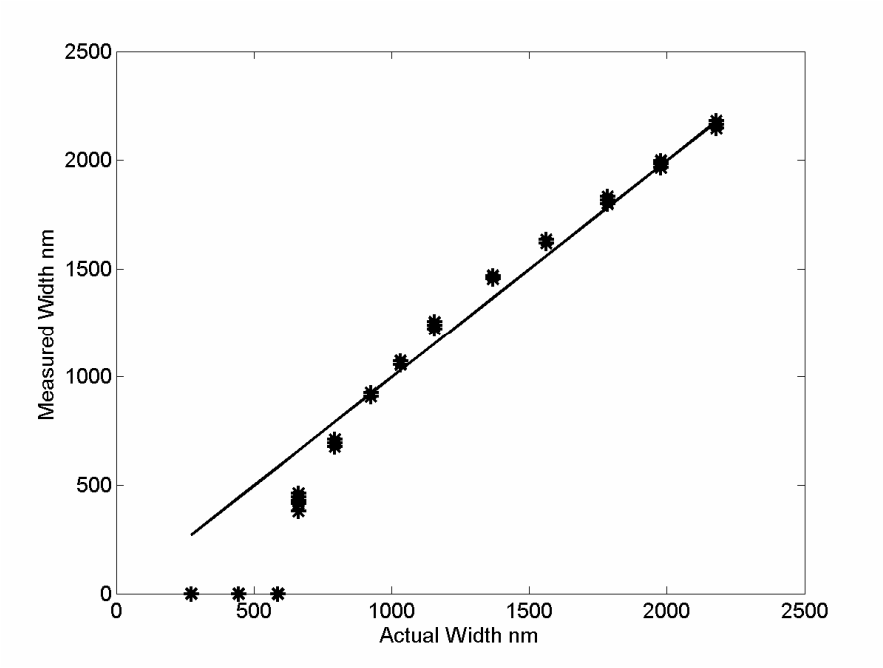


Figure 67 - 25% threshold method for track width calculation

Figure 67 shows that for tracks smaller than 660nm this method breaks down as the tracks no longer cross the 25% intensity level. The other tracks are approximately correct but the spread for successive tracks is quite large and the values are not linearly decreasing with track width.

The second approach was to integrate the profile to get a measure of the area under the track. This approach does not have a lower cut off like the previous method. The mean errors and the standard deviations for the width value are calculated and tabulated in Table 26 below. The standard deviation is for the 6 profiles for each track measured.

Table 26 - Track width errors (nm) for the 25 % and Area method for the BCR sample

Track width µm	**2.18**	**1.98**	**1.79**	**1.56**	**1.37**	**1.16**	**1.03**	**0.93**	**0.79**	**0.66**	**0.59**	**0.44**	**0.27**
std	13.5	12.4	12.4	9.0	8.5	12.4	8.5	8.5	13.5	28.4	NA	NA	NA

25%													
mean 25%	22.0	-6.6	-28.9	-64.7	-96.3	-85.4	-36.0	5.7	105.9	234.4	NA	NA	NA
std area	21.7	33.5	21.1	22.7	28.8	38.0	34.7	24.3	30.8	46.8	18.8	17.9	14.3
mean area	28.4	81.4	126.2	130.5	166.6	306.8	374.3	381.9	369.2	313.8	180.7	234.6	141.4

The mean errors are very large but they could be reduced with better correction methods calculated from optical models of the system as opposed to just a simple scaling factor as was used in this case. The next part of this chapter shows how well the artificial neural network performs in measuring the track widths for various samples and optical systems.

6.3 Analysis of training results for different optical systems and samples

This section examines the ANNs performance at increasing the measurement capability of three optical systems measuring two additional samples. The two samples that have been examined are the BCR standard produced by NPL and a sample made from silicon (Figure 68). Each optical system will be discussed in turn describing the measurements that were taken and the training results. At the end of this section comparisons between the systems will be drawn.

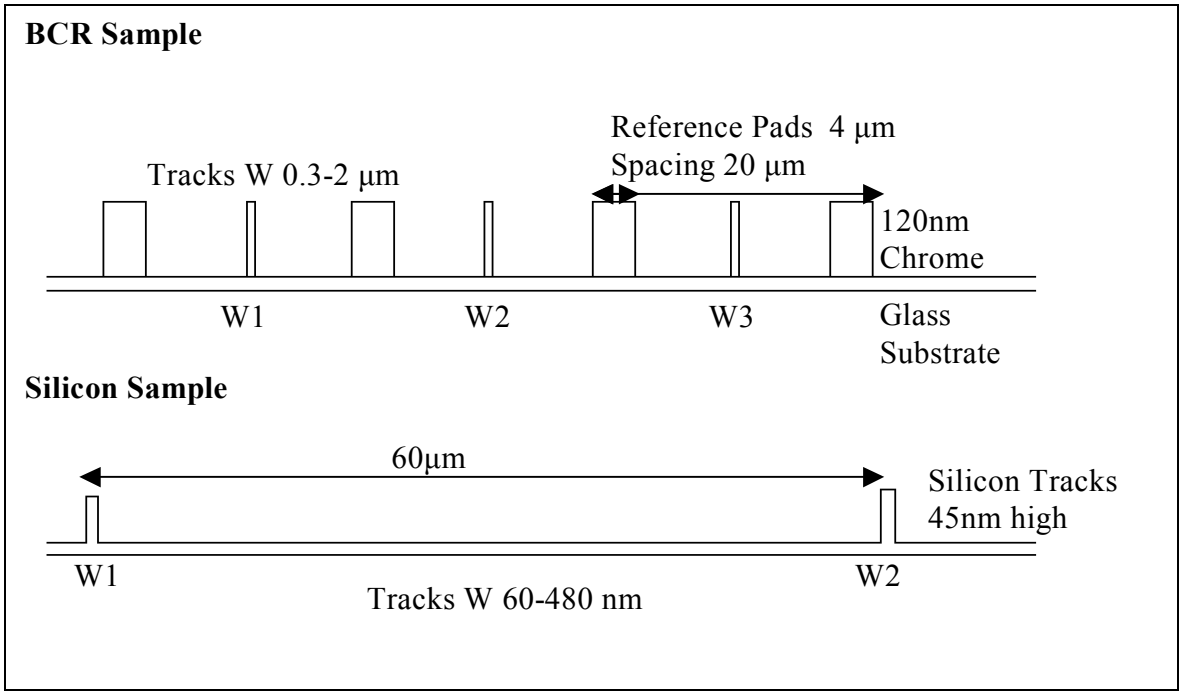


Figure 68 - Layout of samples

The BCR is a chrome on glass sample containing tracks of various widths. The set of tracks used for our experiments were nominally 0.3-2 microns. There were 13 tracks in this range and the target values used were the calibrated values provided by the NPL after the sample had been measured with the OPTIMM system.

The silicon sample contained a row of tracks that were 45nm high and ranged from 40 nm up to 480 nm in width. The target values in this case were the nominal values, as the sample has not been calibrated due to the very small widths of the features. Nominal values can be used to demonstrate the measurement precision of the technique.

6.3.1 Differential Scanning Optical Microscope (DSOM)

The DSOM measured the 1-3 micron sample and these results were used as the example given earlier in this chapter.

The BCR was measured with the DSOM microscope. The objective used was an Olympus x10 0.3NA and the wavelength used was 0.688nm. An example of a differential intensity profile obtained with the DSOM microscope is shown in Figure 69 for a 2.1 micron track

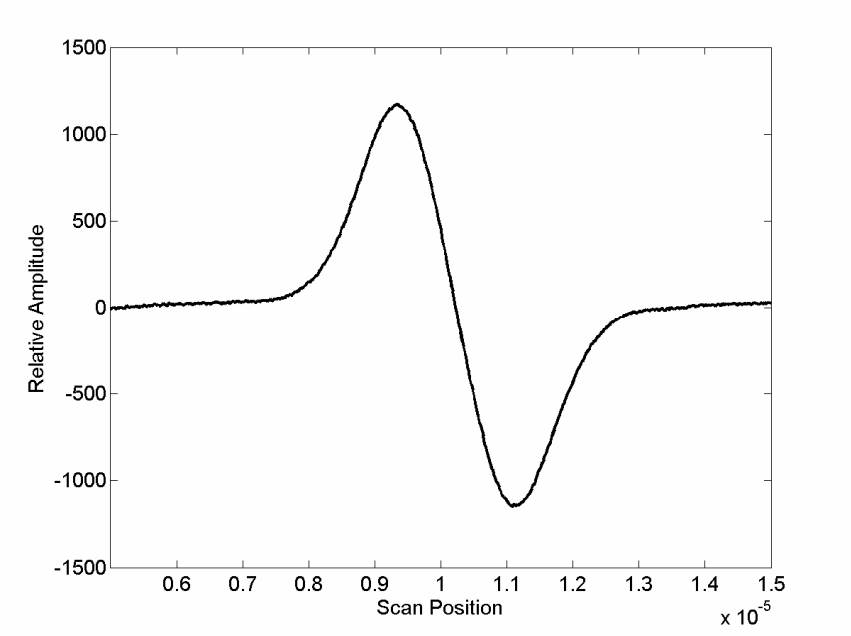


Figure 69 - Differential profile of 2.1 micron track

Each of the 13 tracks were scanned six times to build up a set of 78 measurements. The scan length was 15 μm and 750 samples were taken at 20nm intervals. The PSF width of the DSOM system was 2.8 microns and tracks in the range of 0.272 - 2.1 μm were measured. Figure 70 shows the training results of the ANN on this data.

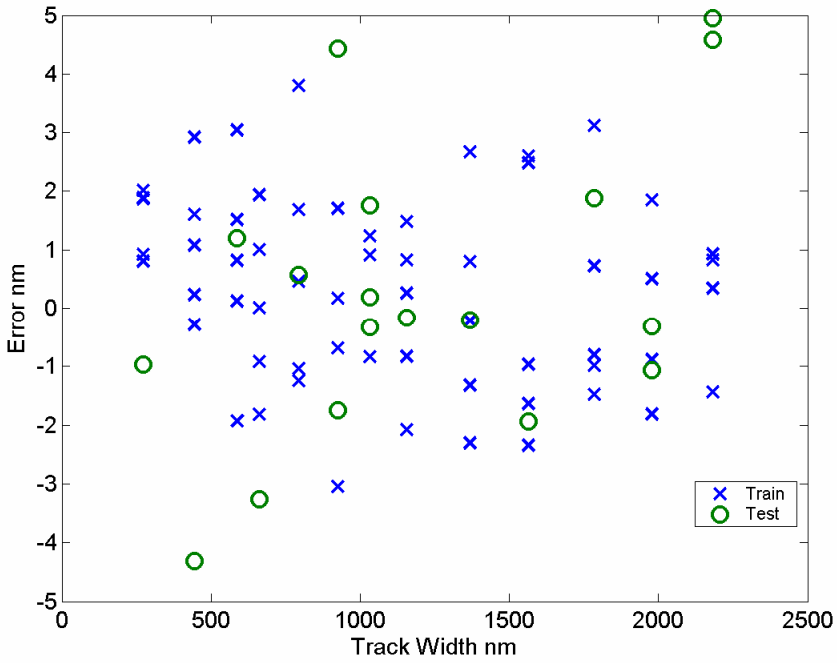


Figure 70 - Training results for 1-3 micron sample

The standard deviation of the training set is 1.59nm and for the testing set 2.55nm. Combined with the ANN the system successfully measured a track width down to 273nm (approximately one tenth of the optical spot size).

6.3.2 Scanning Nomarski Microscope

The Nomarski microscope was used to measure the BCR sample using dark field mode the Nomarski objective used was an Olympus x10 0.3NA with Nomarski prism. As for the DSOM case each track was measured six times, an example profile for a track is presented in Figure 71.

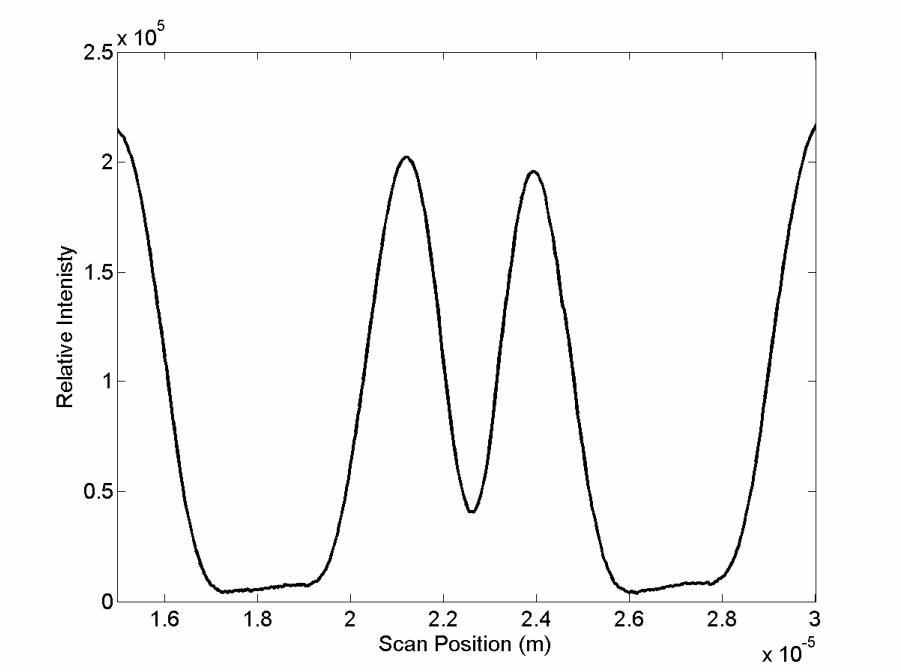


Figure 71 - Dark field scanning Nomarski profile

The reference pads (the two wings) were removed from the profiles, any offset subtracted. The processed profiles were then used to train an ANN. The training was much poorer than for the DSOM system as shown in Figure 72.

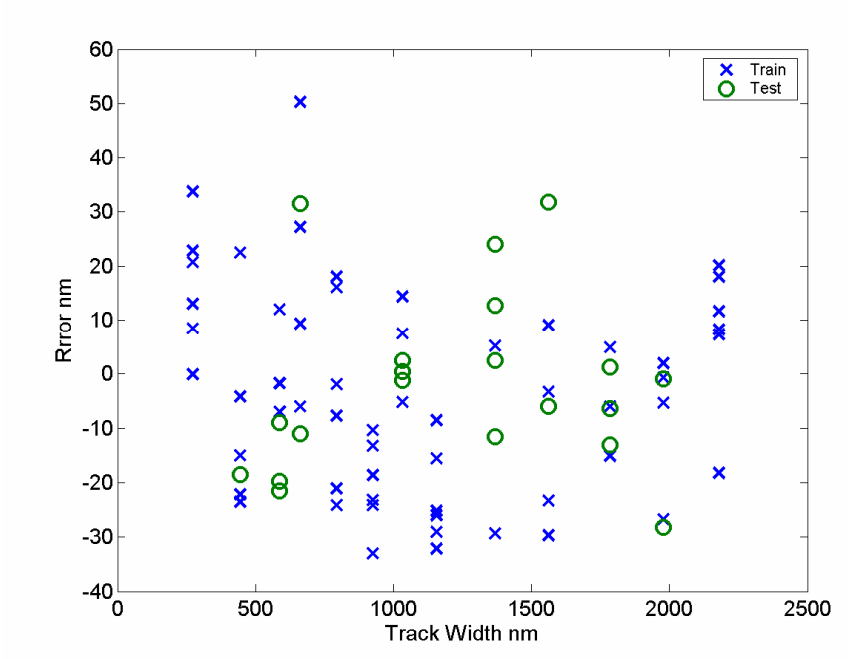
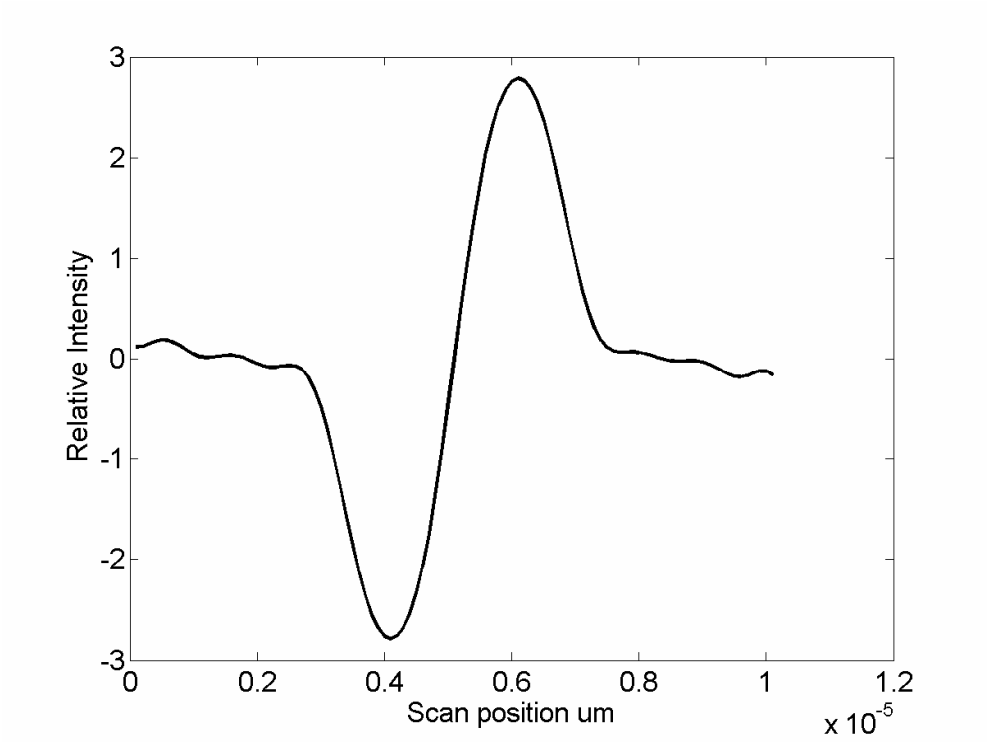


Figure 72 - Training results for BCR sample

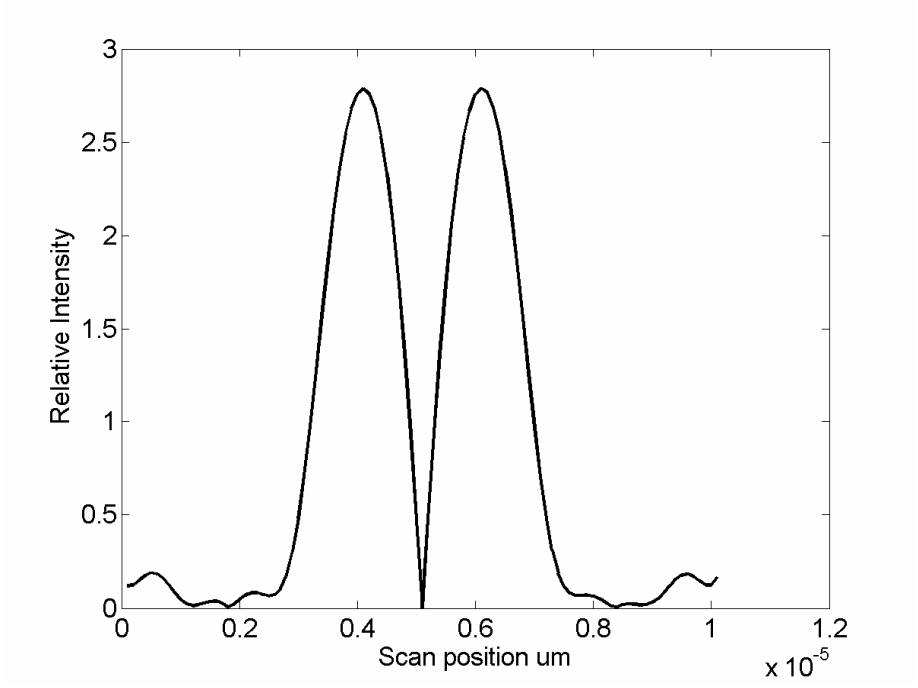
The standard deviation of for the training set and the testing set were 18.8nm and 16.6nm respectively, this is almost 10 times worse than for the DSOM case and this is caused by the sample layout and the size of the PSF in dark field mode. The separation of the two beams is 2.6 microns as calculated in chapter 5 section 5.3, this makes the track response twice as wide as the PSF and as the BCR has reference pads that are very close to the tracks of interest this makes processing afterwards much more difficult as shown in Figure 71. If the beam separation could be reduced this would be less of a problem and performance would then be no worse than for the DSOM case. The beam separation is dependent upon the angle imposed by the Nomarski prism and focal length of the objective. This means that we have little control over the separation of the two beams used in the Nomarski system.

Another factor influencing the training results for the Normarski setup is that we obtain profiles that are the intensity of the differential image (as shown in Figure 73).

The actual desired signal is shown in (a) it has a positive and negative peak, when it is captured by the camera the negative peaks are converted to positive peaks (b). This means that the spectrum is not in the ideal form for the ANN as we still have relatively large low frequency components and a large DC term. This problem is demonstrated and a solution presented in the next section when measuring the silicon sample.



(a) differential signal



(b) signal acquired by camera

Figure 73 Actual and acquired signals

This shows a limitation of this optical system using these particular parameters. The samples used should be well separated or higher NA should be used as this reduces the beam separation / point spread function ratio.

The silicon sample was measured with the Nomarski system in two modes. Firstly, the sample was measured with dark field. This sample was expected to perform better than for the BCR case as the sample tracks are well isolated (60 microns apart and no reference pads)

An example profile of a 180nm track is given in Figure 74 to show how the system performs for these small tracks. Each track was measured 4 times to give 84 measurements for training and testing the network.

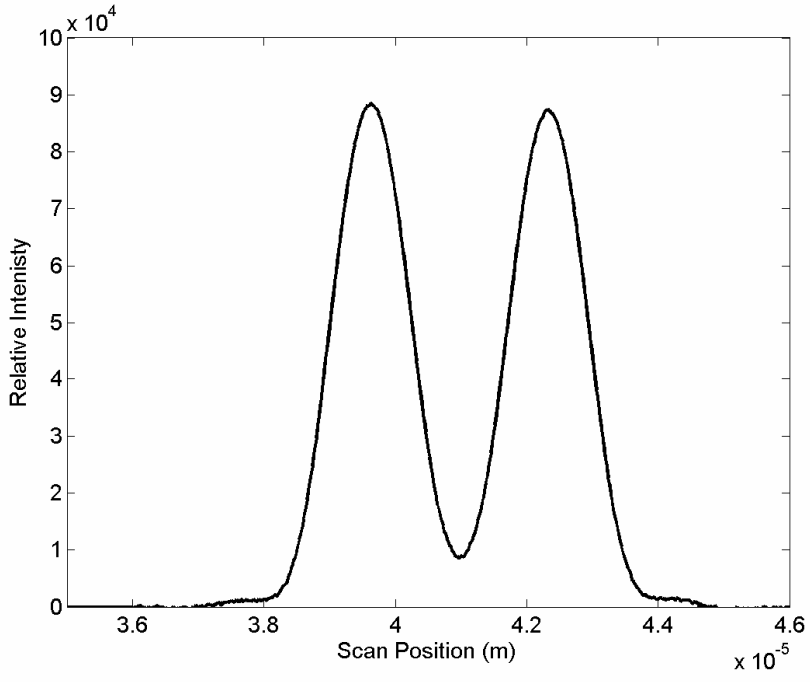


Figure 74 - Dark field profile of 180nm track

The measurements were processed and used to train an ANN and the results of this training are shown in Figure 75.

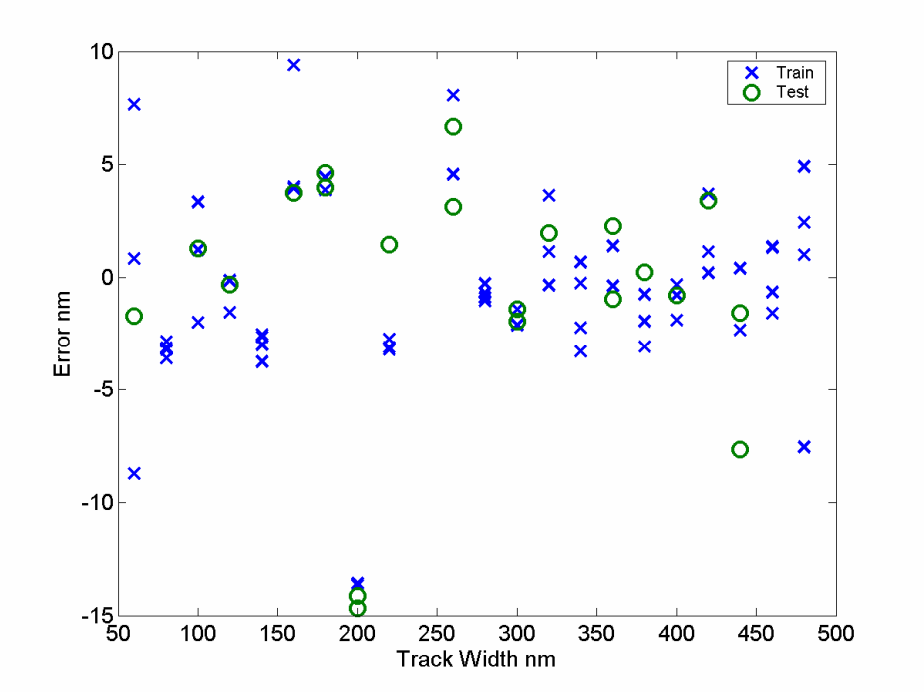


Figure 75 - Training result for silicon sample

Standard deviation of the training and testing sets were 4.1 and 5.5nm respectively. The data processing was the same as before.

The system was then modified so that a modified differential profile could be obtained this was achieved by including a quarter wave plate in the imaging arm. This allowed the possibility of imposing a phase shift between the two beams. This means that the system is no longer operating in dark field mode, instead we obtain the differential signal as shown in Figure 76 for a 200nm track, not the absolute of the differential intensity as for the normal dark field Nomarski case. There are a few blips on the profile caused by vibrations due to changes in the environment during the scan.

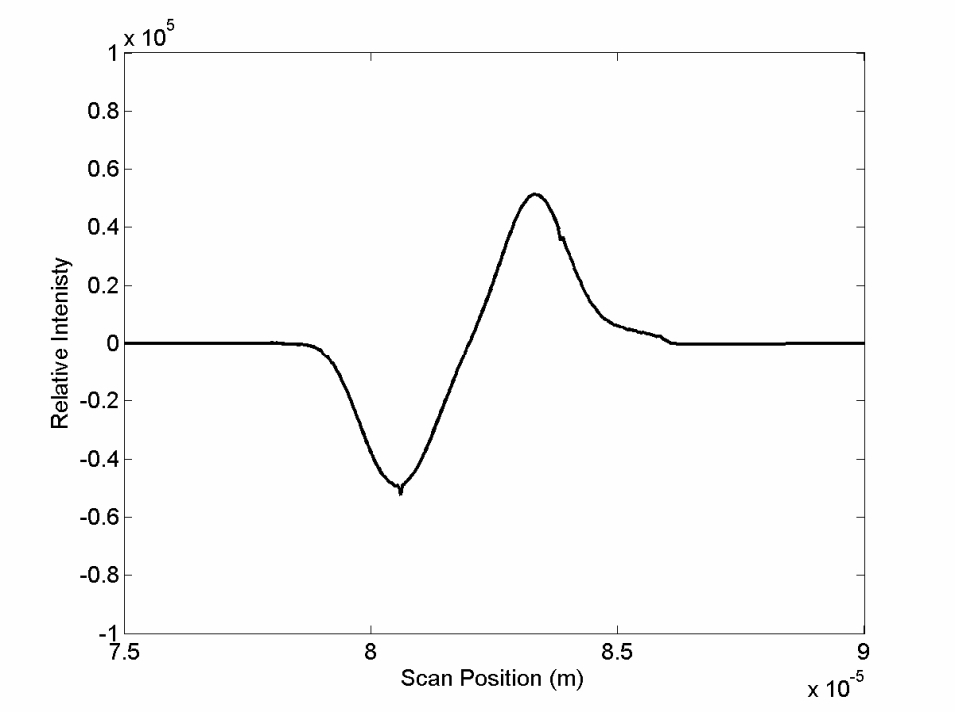


Figure 76 - differential profile obtained by inserting a quarter wave plate

A network was trained and the outputs from the network are shown in Figure 77.

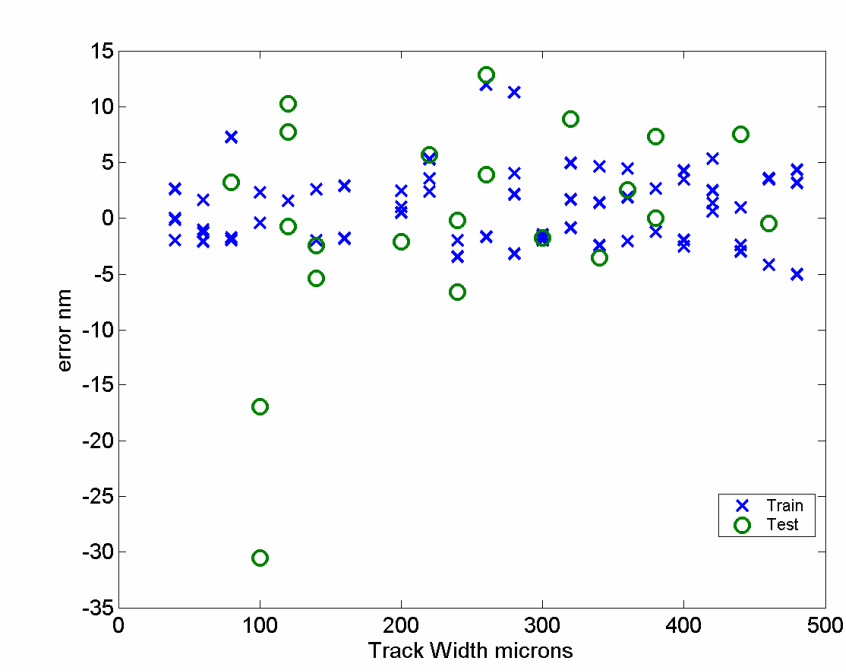


Figure 77 - training result on silicon sample

The standard deviation of the training and testing results are 3.4 and 9.5nm respectively. As can be seen there are two very poor testing results if they are excluded then the testing standard deviation becomes 5.5nm. The testing errors are due to the stability of the system. The repeatability was not as good as previously observed because the measurement environment was changing considerably due to temperature changes as well as pressure waves caused by doors opening and closing.

These training results are better than for the previous case especially considering that in this case the SNR is worse due to the lower light level used. Less light (approximately 10 times smaller) had to be used so that the PSF on the CCD did not saturate the camera where as for the dark field case this was not a problem as the light level could be increased until the differential signal saturated the camera. With better mechanical structure and better choice of Nomarski prism to get the ideal beam separation this system would be very good for the task. The system is relatively easy to use and has different modes of operation, and is suitable for different types of samples.

6.3.3 Hologram

The BCR sample was coated in a thin layer of aluminium, since the large transparent sections of the sample would only reflect a small fraction of the reference beam, thus resulting in very poor contrast in the fringe pattern.

An example of a profile obtained from the microscope is given in Figure 78 for a 2.1 micron track. The optical set up was the same as above with NA 0.3 but this time the wavelength was 633nm.

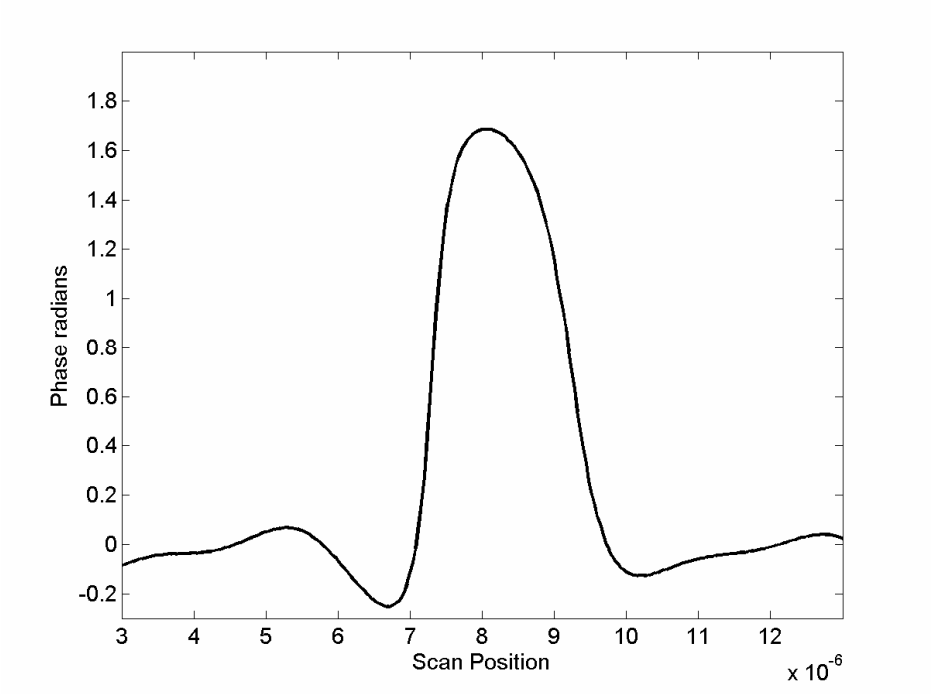


Figure 78 - profile of 2.1 micron track

The track shape is not very symmetrical but this is mainly because the sample is not very uniform, as shown by pictures of the sample taken with a conventional bright field optical microscope. The measured profiles were then processed and used to train a network. The training results are presented in Figure 79.

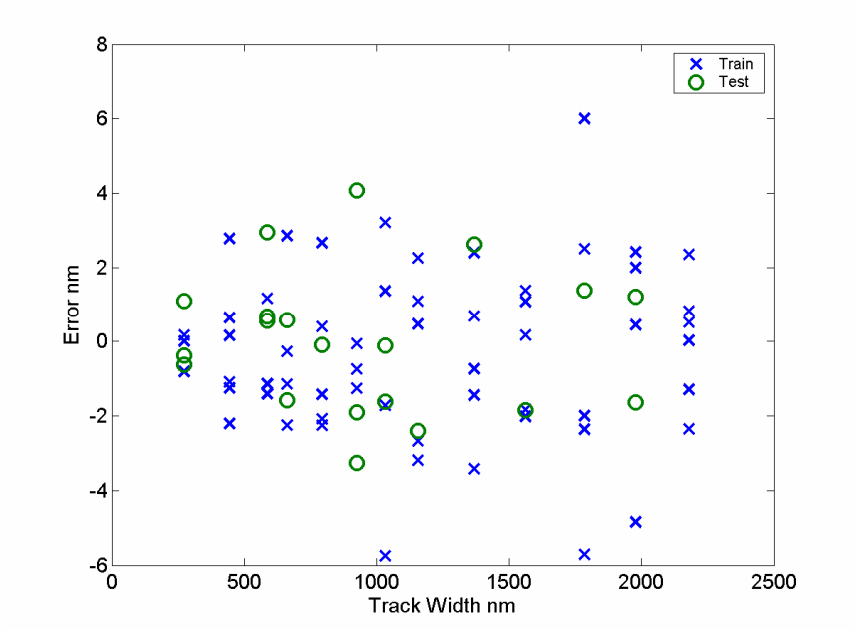


Figure 79 - Training result for BCR sample

The standard deviations of the training and testing set were 2.24 and 1.9nm respectively. Due to the mechanical construction, this system is very stable and produces profiles with good repeatability and high signal to noise ratio.

The silicon sample was also measured with this system and an example profile is shown below in Figure 80.

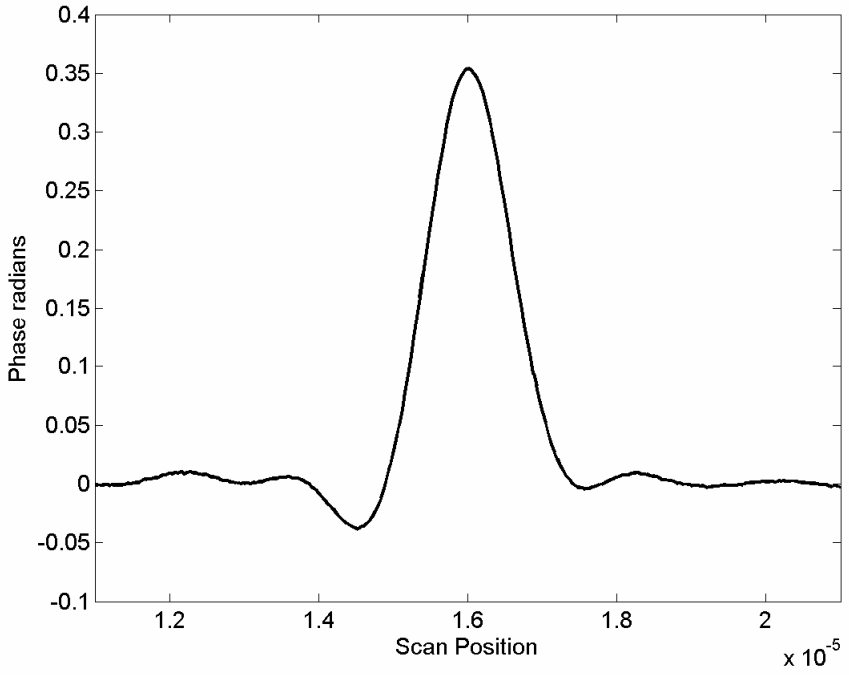


Figure 80 - profile of silicon track

The profiles are much more uniform and symmetrical. This sample is much more uniform in general than for the BCR sample. The training results are excellent and are shown in Figure 81.

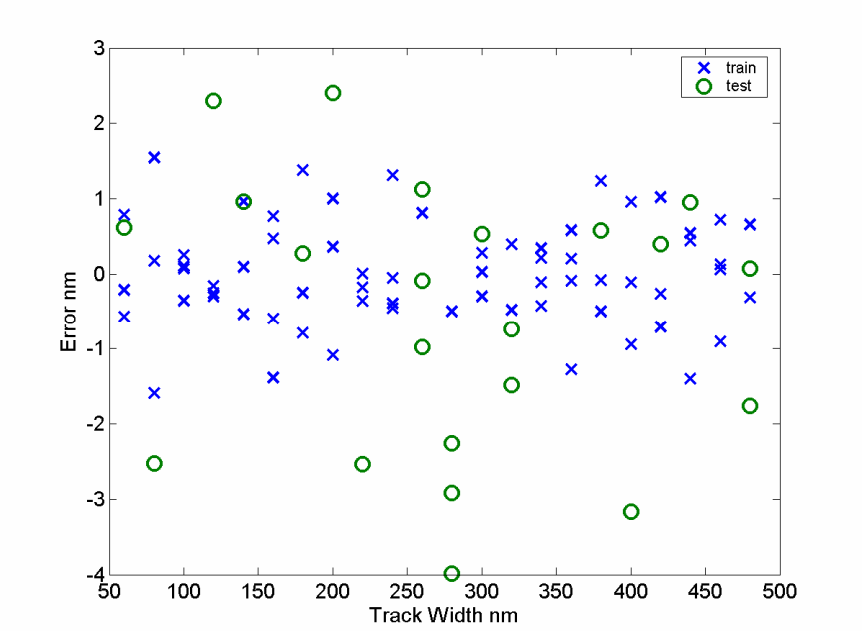


Figure 81 hologram silicon sample result

The standard deviation for the training and testing sets are 0.69 and 1.78nm respectively. This demonstrates the power of this technique, as an optical system with 0.3NA and using a wavelength of 633nm was able to successfully measure 60nm track widths with standard deviation of less than 2nm, this track width corresponds to **1/43rd** of the optical spot size. The experiments were repeated several times, under different conditions, and similar results were obtained.

6.3.4 Comparison of training results

This section summarises the main training results that have been present thus far. In all cases the objective has an NA of 0.3. The DSOM and Nomarski used a wavelength of 688nm and so the PSF was 2.798 microns wide. The hologram system used a wavelength of 633 nm and the spot size was therefore 2.574 microns.

A Summary of the samples measured and the training results for each system is presented in Table 27.

Table 27 - comparison of testing results (nm) for different samples and systems

Sample Name	1-3u	BCR	Silicon
Range	1-3 microns	0.273-2.1 microns	0.06-0.48 microns
DSOM error (nm)	10.86	2.55	-
Nomarski error (nm)	-	18.8	5.5
Hologram error (nm)	-	1.9	1.7

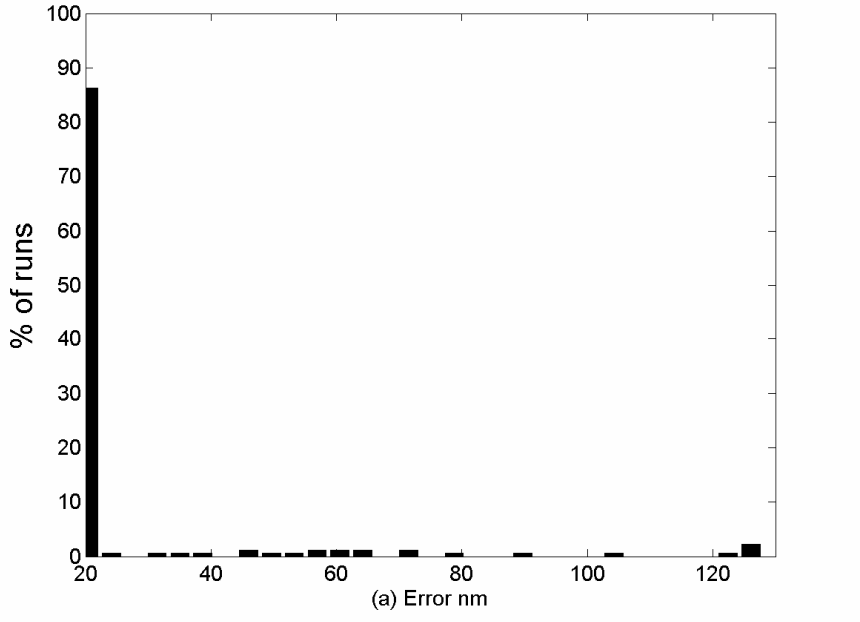
The system that has the best performance is the hologram system. This is not surprising as this system was the most mechanically stable system and had the highest signal to noise ratio, which are two of the most important factors to ensure proper operation of the technique. However, the differentiation was not done optically as for the other systems.

The scanning Nomarski system suffered from poor environmental conditions as well as an un-optimal setup imposed by the beam separation of the prism. Using a more suitable prism the training results would improve greatly

The next section of this chapter looks at various aspects of the ANN training and how choice of parameters, such as number of training patterns, choice of inputs etc. can influence the training results. The general nature of the network response is shown and a method for correcting target errors is demonstrated.

6.4 Reduced training sets

The amount of training data is important for the performance of the network. This is demonstrated by Figure 82a-c. During the experiment each of the 20 different width tracks were measured 4 times each. During training only 1,2 or 3 of the profiles from each track width were used. In the first case (a) only one copy from the four profiles taken for each track was used to train the network. This was repeated where the number of tracks picked at random was increased to 2 then 3. This was performed 500 times and the results were tabulated and shown in a bar chart in Figure 82.



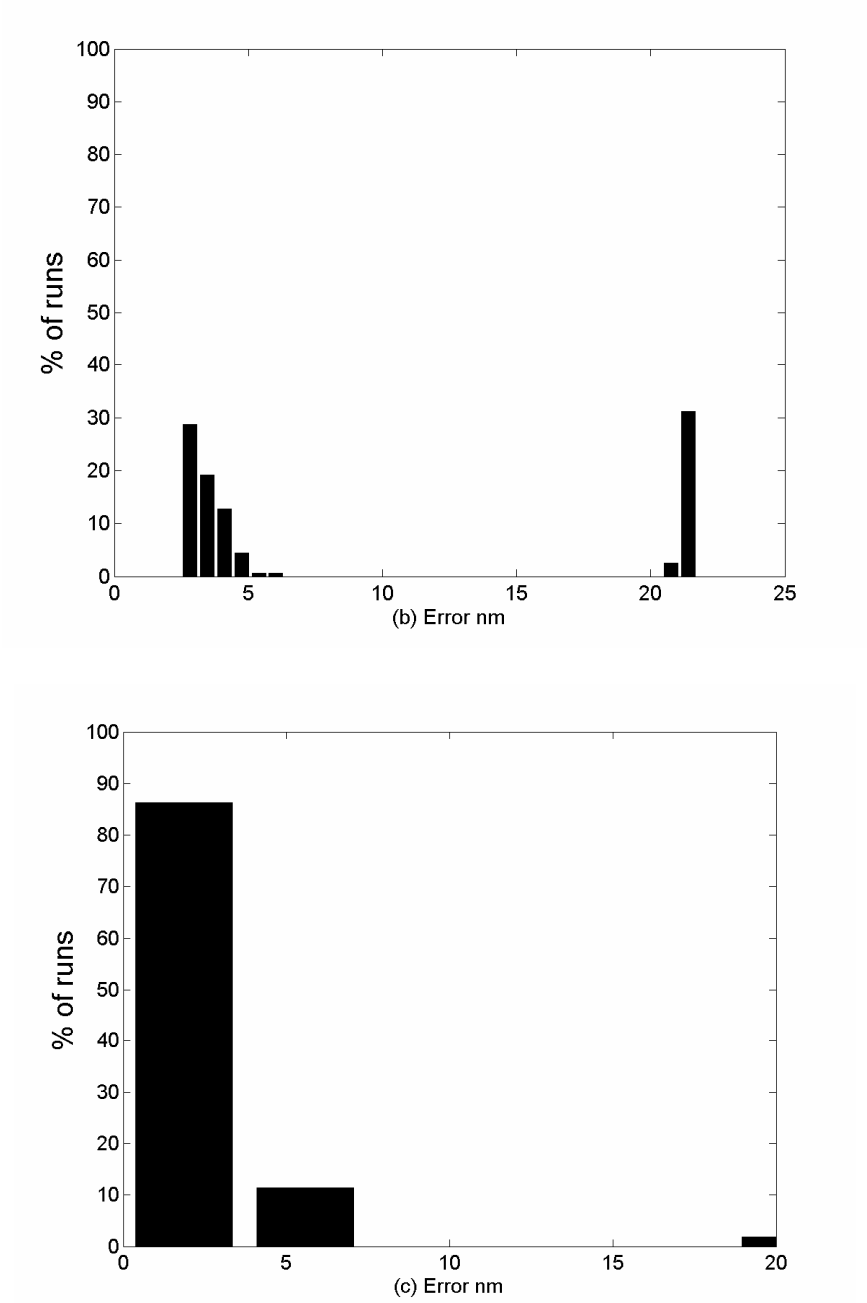


Figure 82 - Amount of training data

In the first case none of the 500 networks trained had a testing set error below 20nm. In graph b where 2 of the tracks were picked at random this improved the training dramatically most of the networks had errors below 5nm although there is still some

that stopped training with errors around 20nm. The final case had 3 out of 4 examples used to train and this produces the most reliable training with 85% below 3nm although there is still the occasional network that stopped training at 20nm.

As expected more training examples produce more reliable and better training performance. This shows that wherever possible, as many training examples as possible should be used. If time allowed many more scans of each track width would be used to train the network. Having many examples of each track improves training because one very noisy input pattern will not influence the network as much if there are lots of other patterns that agree more closely for that target.

The reason that some of the training stops at larger values even when a lot of training examples are used is discussed in some detail at the end of this chapter in the section on repeatability errors.

6.5 Missing tracks left out at random

The network produces a general model of the input output relationship. This means that if tracks are removed completely from the training set then the training results should still be valid for those track widths if the profiles are presented to the finished network. This is presented in Figure 83 where the 100nm and the 240nm tracks have been left out of the training process completely.

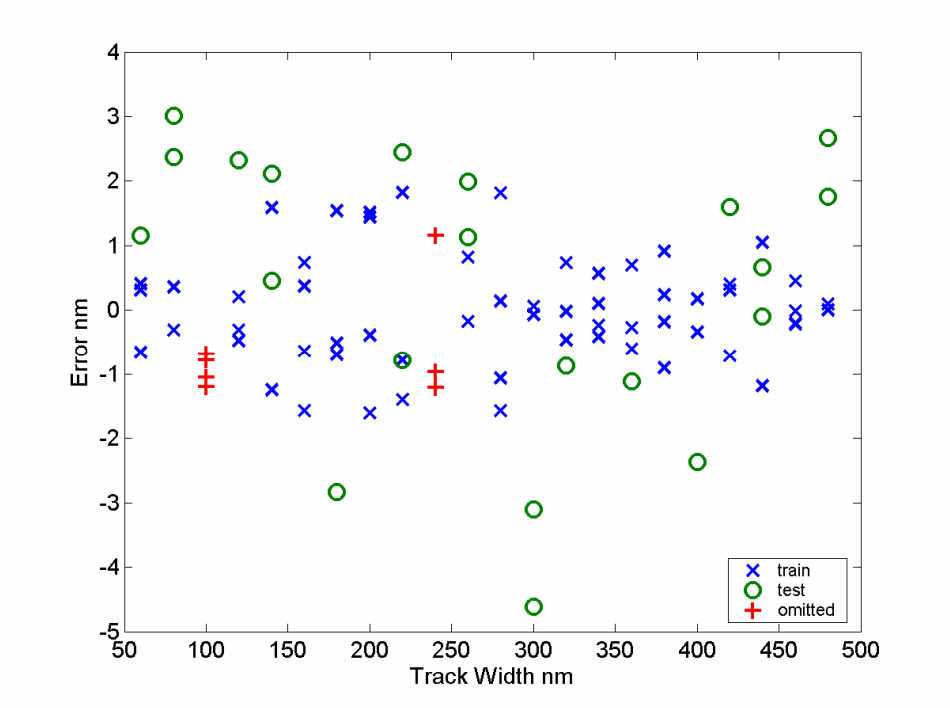


Figure 83 - 2 widths missing from training data

After training the input data for the 100nm and 240nm tracks are then presented to the finished network and as can be seen from Figure 83 the track width errors are comparable to those for the training and testing sets. This shows that the network has produced a general result and that the training set is not required to cover all possible test values. This was presented as simulations in chapter 4 page 16 where a test only set was presented to the network and the errors for this set were comparable to the training and testing sets. This shows that the experimental results and simulations are consistent.

6.6 Out of range

The network that has been trained will only be valid for the range of tracks in the training set. This is demonstrated in Figure 84 where the network was trained on the

range of 60nm-460nm and after the network had finished training the 480nm tracks were presented to the network and the errors plotted.

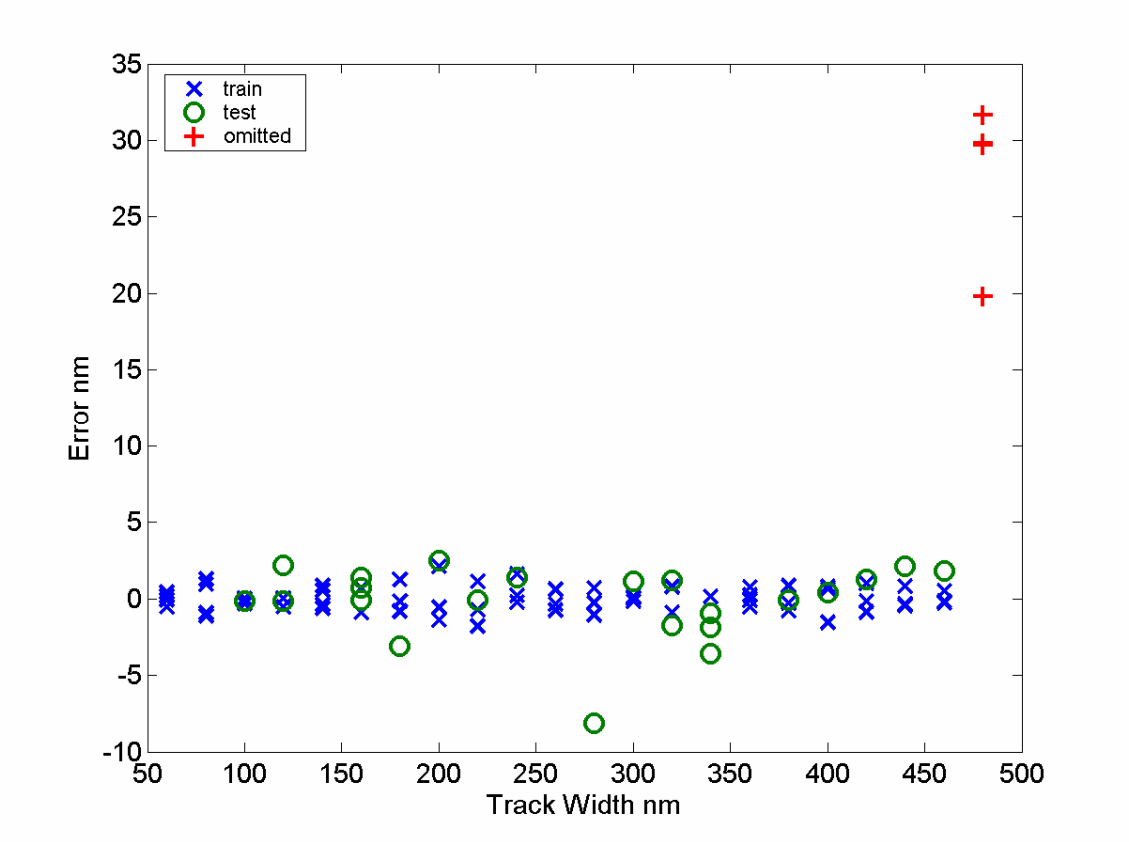


Figure 84 hologram silicon sample out of range

The error for the out of range tracks is considerably higher, the mean is approximately 30nm and the standard deviation is several times greater than for the training set. This demonstrates the importance of knowing the working range for the network used and will be an important consideration to the final system design in practice. If line widths are required for a specific range then the training sample must cover at least this range otherwise it will not perform correctly for the out of range tracks. This demonstrates that these networks are very good at interpolation across the training range but are poor at extrapolation outside of the training range.

6.7 Input points

A simple experiment to investigate the impact of the choice of input points used for training has been performed. A variety of networks were trained where the input patterns were varied. Some networks only used low frequency components to train others only high frequency some a combination of the two. An example of the points available for a specific track is shown in Figure 85.

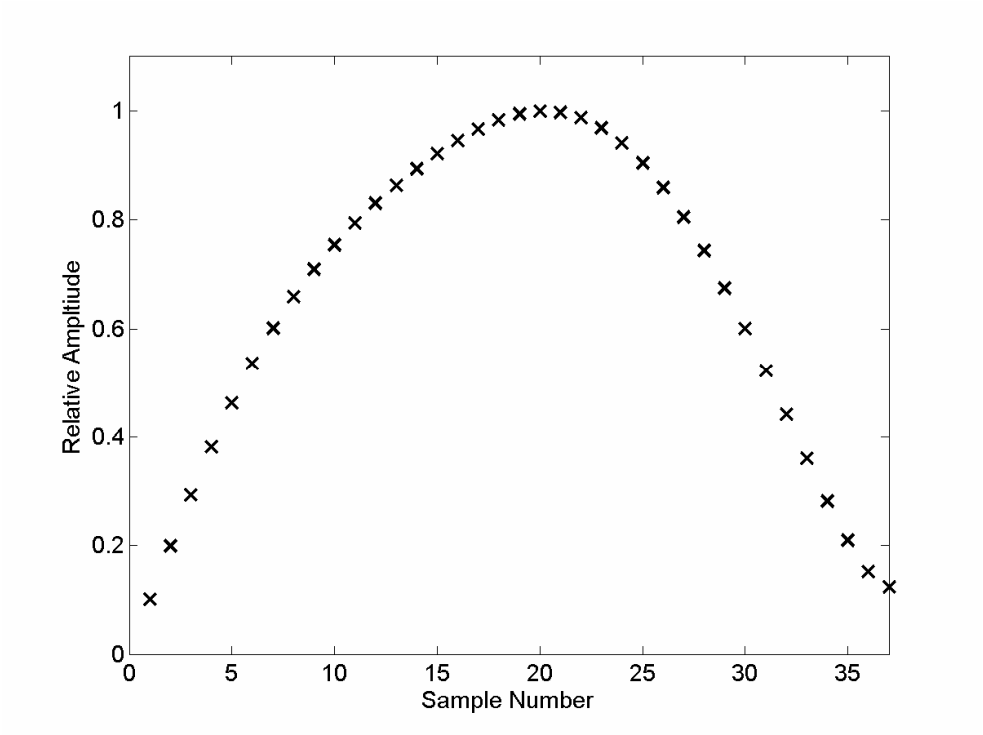


Figure 85 - Sample numbers used for training

Figure 86 and Table 28 below shows the sample numbers used for each network and its corresponding training error. The points with the circle around are the input points used for that training type.

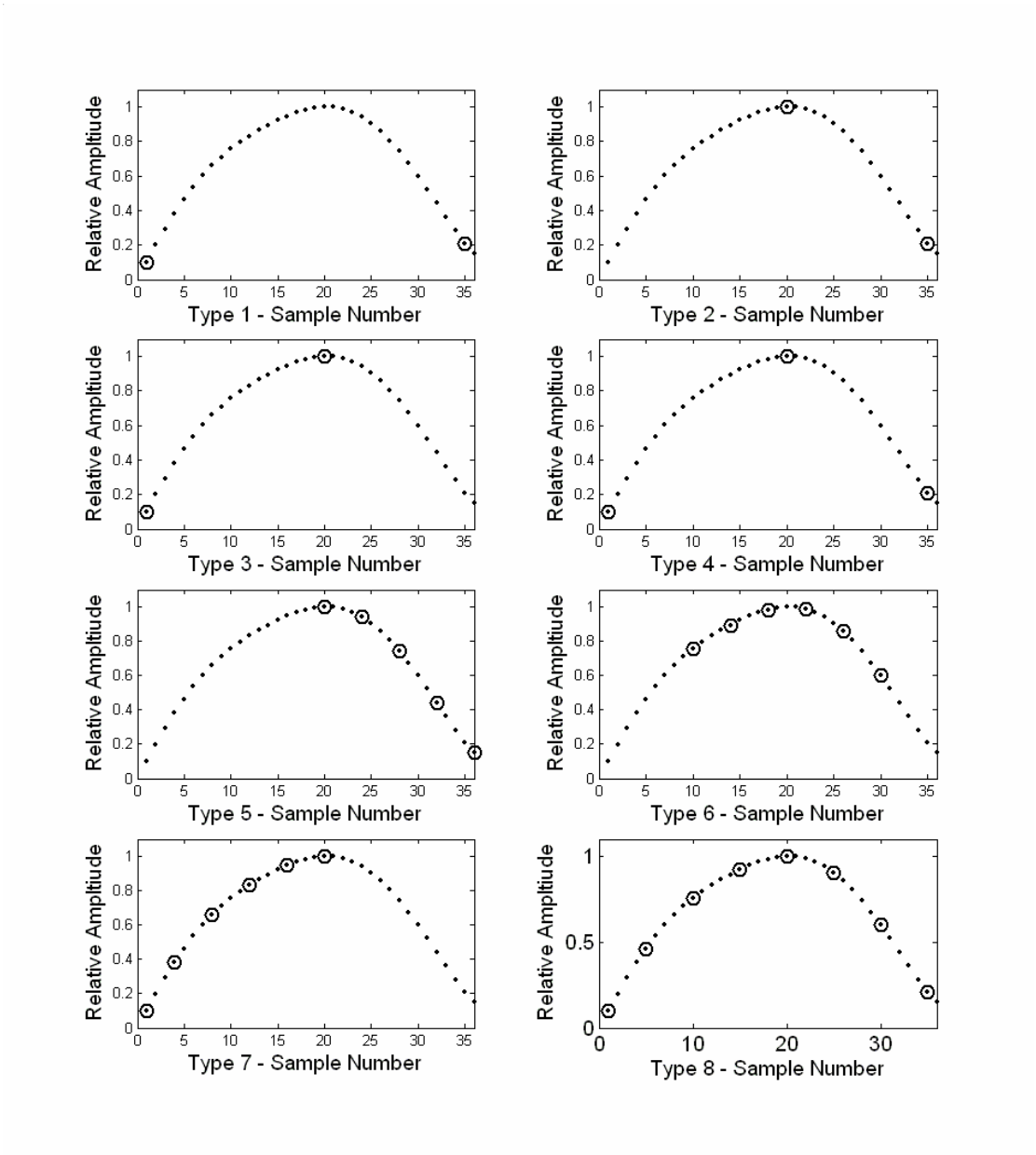


Figure 86 - Input points used

The input points were chosen so that groups of low, middle and high frequencies were used as well as the usual equally spaced points.

Table 28 - Training results for different input types

Sample type	Mean of 20/30 runs of standard deviation of test (nm)	Samples number used	Description
1	15.6	1 35	End

2	6.7	20 35	Mid and end
3	20.6	1 20	Mid 1st
4	5.6	1 20 35	End 2 and mid
5	2.8	20 24 28 32 36	High
6	8.0	10 14 18 22 26 30	Mid
7	12.6	1 4 8 12 16 20	Low
8	4.3	1 5 10 15 20 25 30 35	Equal spaced

Table 28 shows some very interesting results. In general we can see that the high frequency samples perform better than using the low frequency samples (Type 2 is much better than type 1 and 3 and type 5 is the best performing network). Type 8, using equally spaced samples is also a good network but not as good if just the high frequencies were used. The value for type 8 is worse than the previously presented results in section 6.3 page 19. This is because in this case it is an average of many runs and so if any networks do not train well then the poor results affect the average. It appears that more reliable training is obtained if just the high frequencies are used as they produce consistent good training results and have a low average training error. The equally spaced network can produce well-trained networks (2nm errors) but the training is more likely to stop early then for the high frequency case.

This is very interesting as the high frequency components are the ones that should contain the most significant information regarding the changing of track width, as the high components are related to higher resolution.

In general, for single track objects the high frequency components should be used for training, as this will produce the best results. For other object types it is best to start

with equal spaced samples as this should always perform well and then the optimal points can be investigated.

6.8 Auto correction for target errors

The auto correction idea makes use of the fact that a general rule is produced for the input/target relationship and if any target is incorrect then for that specific track/target pair this relationship no longer holds. This means that during training the error for this track increases and training stops. The auto correction idea looks for a training sample that has abnormally high error and adjusts the target and retrains the network until the error is comparable to that of the training set. This idea was applied to experimental data. In this case the target belonging to the 320nm tracks was altered and the network was trained and the error corrected after several iterations.

The actual track width is 320 nm wide but has been increased by 8% to 345.6nm. After each iteration the target was updated by subtracting the mean error for the 4 examples of tracks of this width, so for example after the first iteration the target is adjusted from 345.6nm to331.29nm. The target was successfully corrected to 321.76 nm after 4 iterations as shown in Table 29. After 4 iterations the errors associated with this track were comparable to the rest of the training/testing sets.

Table 29 - Auto correct updates

	Iteration 1	**Iteration 2**	**Iteration 3**	**Iteration 4**
Target (nm)	345.6	331.29	328.15	322.84
Update value (nm)	14.31	3.14	5.30	1.08
New target (nm)	331.29	328.15	322.84	321.76

It is interesting to look at the errors for the whole set as well. As can be seen after each iteration the mean errors for all tracks tend to reduce as the target/pattern error for the 320nm track being incorrect will stop training early for the first few iterations. As shown in Figure 87.

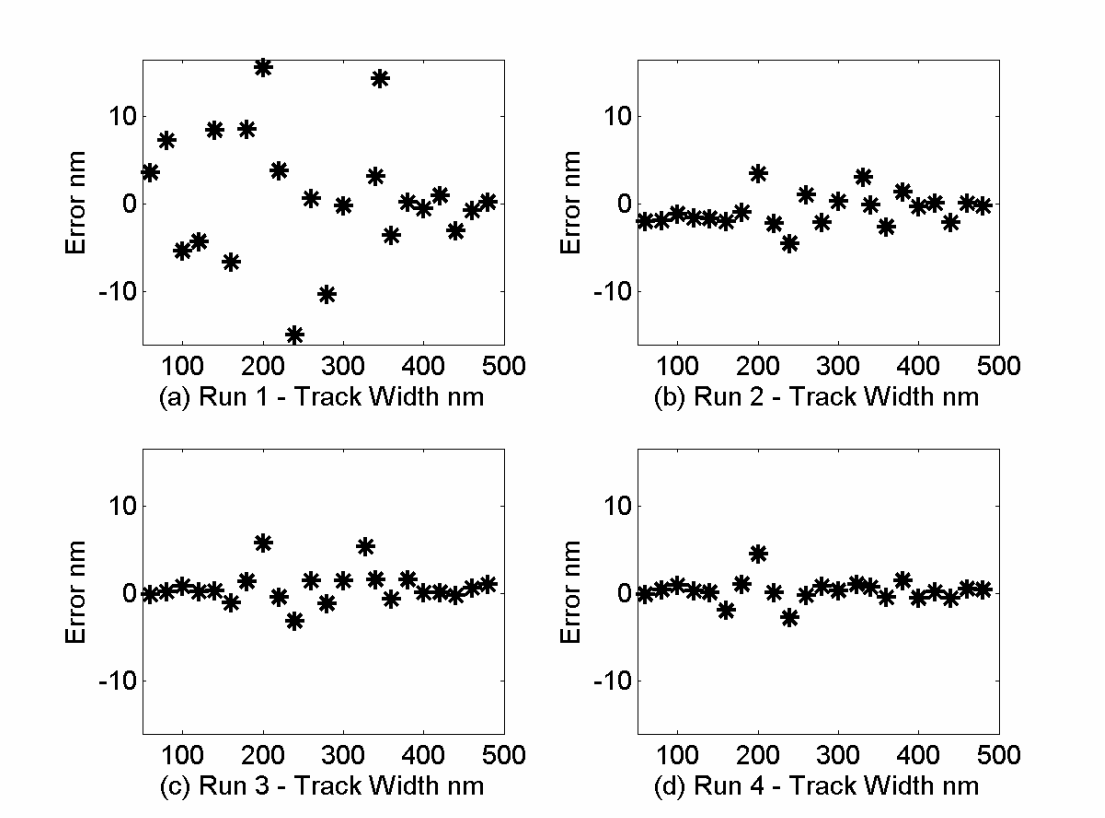


Figure 87 - Mean values for each track with after each iteration

As can be seen the error for the 320nm track reduces after each iteration and this is also the case for most other tracks.

Table 30 - Training results for auto correct

	Iteration 1 Std Error (nm)	**Iteration 2 Std Error (nm)**	**Iteration 3 Std Error (nm)**	**Iteration 4 Std Error (nm)**
Mean (320nm)	14.31	3.14	5.30	1.08
Mean(mean all)	0.80	-0.69	0.67	0.30
Std (320nm)	0.33	0.45	0.65	0.66

Mean (std all)	1.95	1.65	1.83	1.65

Table 30 shows the means and standard deviation for the 320nm track all of the other tracks after each iteration. The mean is reducing after each iteration. The standard deviation also reduces although it did increase after the 3rd iteration before reducing again.

This demonstrates that the network can be used to correct for target errors although the extent to which the network can cope with these errors is still to be investigated fully.

An interesting question is raised by this technique. Firstly is there any benefit to including the incorrect target point or data to the solution would it just be best to train a network and leave the data out? Also if it is included and corrected where did this extra information come from?

The answer to the first question would be yes, there is an important reason to include the data point if the correct target can be established as by including the point the solution over that region of track width is constrained by the inclusion of the data point. It will therefore help to improve the local solution between those nearby track width values and should therefore be included if the target can be established.

The second question is more interesting. No additional information is being added to the system and yet additional information (namely the correct target) is being obtained. This means that the network performance as a whole must be degraded to

obey the information content laws discussed in chapter 3 page 39, but as the degradation is spread out across the whole training range then the degradation is lost among the noise and it appears that the additional information is obtained for free. Put another way across the training range the error increases by a tiny amount to give an increase in performance in the local area around the incorrect target.

The extent to which this technique work and the underlying limitations need to be understood more and should be looked at in more detail in the future.

The technique will now be applied to double track structures to demonstrate that multiple parameters can be extracted.

6.9 Double Track Experiment

A sample that contained double tracks was measure with the DSOM system. The sample was the 1-3 micron sample. The double tracks ranged from 1-3 microns with separations up to 4.8 microns. The double tracks were measured 3 times each and in total there were 162 tracks to use for training. The training targets were derived from the 0.7NA scans of the tracks during the main experiment as for the single track case.

The NA was 0.18 for the main scans and 0.7NA for the high resolution scans from which the width and separation would be obtained. Each scan comprised of 500 points with 40nm increments between points giving a total scan length of 20 microns. A flat region to the side of the track was used to normalise the signals so that any intensity variation between different track scans would be removed. The tracks were then processed in the usual manner.

Initially just one network was trained to obtain the width and separation. The training results for this network are presented in Figure 88 and Figure 89.

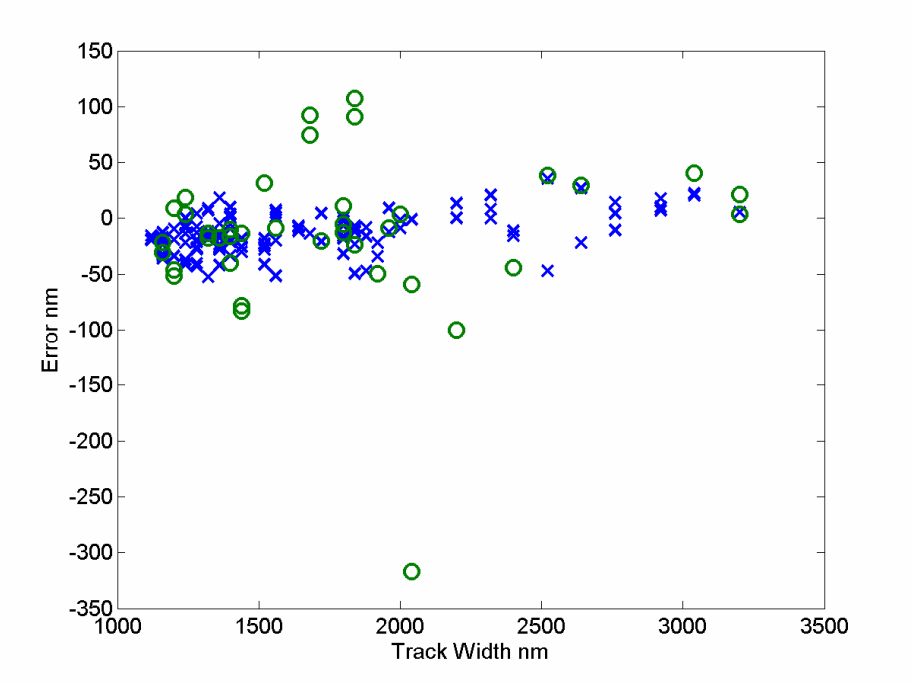


Figure 88 - Double track width results

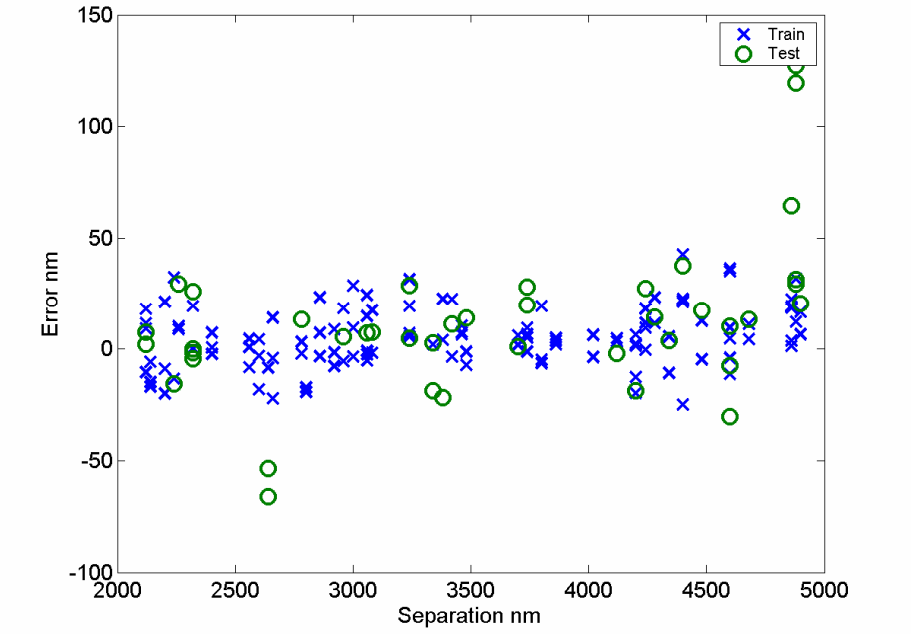


Figure 89 - Double track separation results

The standard deviations for the training and testing sets are given in table 7. For the testing set the 5 worst performing tracks were removed. There were a number of tracks in the testing set where the error was higher. This was because there were very few examples of each separation and width value available on this sample.

Two other networks were then trained to increase the training performance. The other networks produced only the width or the separation value and the training results are also presented in Table 31.

Table 31- Double track training results

	Two Output Network		**One Output Network**	**One Output Network**
	Width nm(%)	**Sep nm(%)**	**Width (%)**	**Sep (%)**
Std Train	17.3 (0.78)	13.4 (0.37)	5.20 (0.17)	3.42 (0.058)
Std Test	25.4 (1.33)	14.5 (0.31)	9.52 (0.63)	9.29 (0.28)

The testing set results translate to an error of 2.7% for the width parameter and 0.65% for the separation parameter using one network for each parameter individually. If the 5 worst performing tracks are removed the error for the testing set is 0.63% for the width and 0.28% for the separation. These training results would be greatly improved by increasing the number of examples of width and separation included in the training set. Using a better quality sample with a more appropriate range of sizes would also improve the results.

6.10 Single Track / Double Track classifier

If a sample contained a mixture of single and double track structures it would be useful to be able to sort the tracks out into the two types so that they could be sent to the correct networks before the parameters were calculated. This can be achieved by using a classifier. In this case all of the single tracks and all of the double tracks are assigned a specific target for their type. A classifier was trained based on the data from the 1-3 micron single tracks and the double tracks measured in the experiment above. The results from this classifier are shown in Figure 90 and the two targets (T1 & T2) used were given in Table 32. The network type used was similar to those used for previous networks and as such they may not be the best networks for this task. A self-organising map [62] maybe more suitable, possible improvements to the ANN used is discussed later in chapter 7 section 7.6.

Table 32 - targets for classifier

	T1	**T2**
Double Track	0.8	0
Single Track	0	0.8

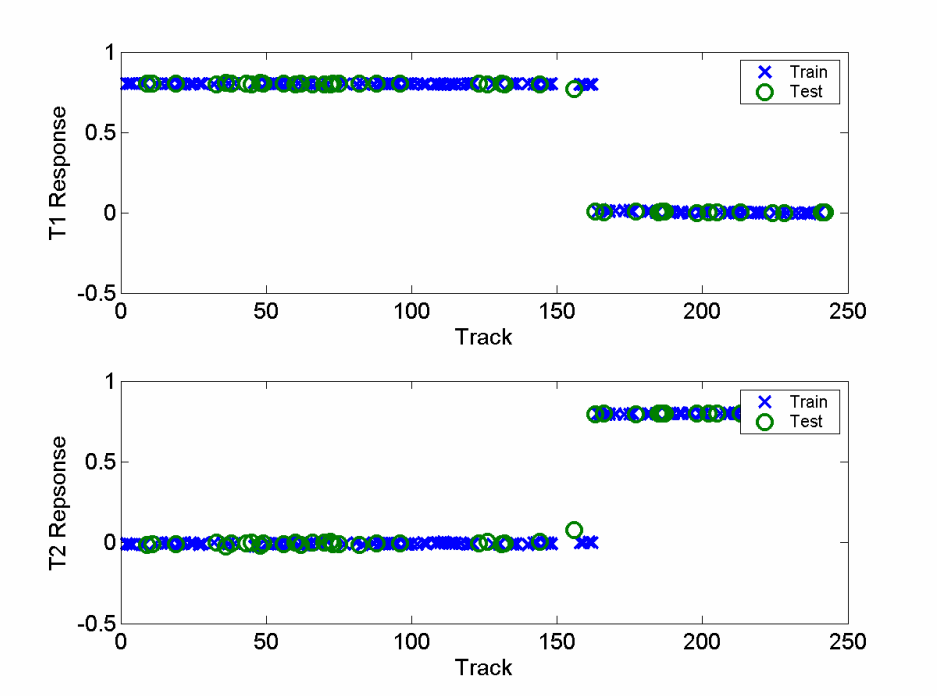


Figure 90 - Double track or single track classifier results

The double-track/single-track classifier worked well, all tracks were classified correctly. The first 160 tracks are the double tracks and the last 80 tracks are the single tracks, the targets can be seen to be distinct for the two sets and none of the tracks were misclassified.

This worked well because the 1-3 micron single tracks and the double tracks are quite different. However as simulations have shown previously this should still work well even if the double track and single-track profiles are very similar, for example, if a double track of 500nm width and separation of 50 nm is measured the classifier should be able to distinguish it from a single track of 1050nm.

6.11 Repeatability of Training

This section attempts to address the reasons behind the variation in the training results for successive training. For the simulation cases the repeatability of training was excellent. For the experimental data this is not always the case. There are several important differences between the simulations and the experimental data. The noise associated on each track is not necessarily the same as was for the experimental case where it was simple white noise. In the experiment there are other factors influencing the noise, for example, how much the temperature changed during each scan, were there any large vibrations etc.

The experimental training sets are smaller than for the simulation cases and so the noise problems make the results worse as there is less data to help with the training. Also if there is any abnormal profiles in validation set training will stop early. It is therefore important to remove any suspect data from the training process (or correct it with the autocorrect procedure discussed earlier). The way in which the data is assigned to the training and testing sets can also cause training to stop early in some instances. The testing set data is picked at random from the total number of tracks available. Sometimes all of the examples of the smallest or largest tracks happen to be in the testing set and this causes larger errors as they are outside the training range of the network, training stops early and the results are poor. By modifying the way in which the data is allocated to the training and testing sets this problem could be overcome.

Another reason for the repeatability errors is due to the initial network state. The initial state of the network is important; the network is initialised by setting all of the

weight and bias values to small random values. These small values corresponded to a starting location in the error space. Due to the relatively small number of patterns the problem is ill conditioned and so the starting location of the network in the error space is important as the network can have problems traversing large distances thorough the error space and this means that it can be possible to start so far away from a minimum that training stops almost immediately as all updates possible cause the error to increase considerably. For example there could be a local minimum near to the start location that is very far away from the global minimum, the final training errors of this local minimum could be many times worse than for the global case but as the network cannot get out of the minimum due to the distance and topography of the error space training stops. This is helped by having a good choice of initial weights to use, many programs distribute the weights in a more suitable manner than just small random values. Also using positive and negative input points and targets can also help depending on the form of the input data.

The impact of all of these various factors influencing the repeatability of training needs to be investigated. As there is usually a large difference between a well trained network and a poor one, however, it is easy to spot by using a testing set of data. This means that in a practical situation the network can be retrained until a good run has taken place.

6.12 Overall Uncertainty

For this system to be used to provide calibrated linewidth standards the measurement needs to be traceable. All sources of error need to be considered and taken into account when deciding what the overall uncertainty is.

For example although we state that the standard deviation of the training error is *x* nm, this is not the uncertainty of the measurement, as this will need to take into account numerous factors such as:

- Uncertainty in the profile measurement due to errors in the optical microscope
- Uncertainty in the profile measurement due to errors in known stage position
- Uncertainty of the training targets
- Any error introduced by the network
- Any uncertainty due to the signal processes/pre-processing of tracks
- Effect of signal digitisation
- Effect of shot noise / vibration and other noise sources
- Drift in the system
- Laser stability and wavelength

All of these sources of uncertainty need to be considered and combined appropriately to form an uncertainty budget.

It is therefore important to note that this technique cannot have a lower uncertainty than the AFM or SEM used to calibrate the gold standard sample. This system does however provide a way to do rapid measurements of various sample parameters on a many types of samples for relatively low cost.

7 Future Work

This chapter describes other parameters that can be measured using this technique, these have been investigated through simulations due to lack of suitable experimental samples. A simulation investigating the measurement of sidewall slopes on tracks is presented. Similar work extracting height information is also given. Then follows a discussion of future work tasks regarding improvement to the network design and overall architecture for multiple parameter extraction as well as other topics such as choice of input parameters etc.

7.1 Slope simulation

In theory the ANN method can also be applied to extract other parameters, as long as the signal to noise ratio is high enough for there to be a measurable effect on the chosen input data. The first parameter of interest was to obtain a measure of the sidewall slopes of a track as illustrated by Figure 91.

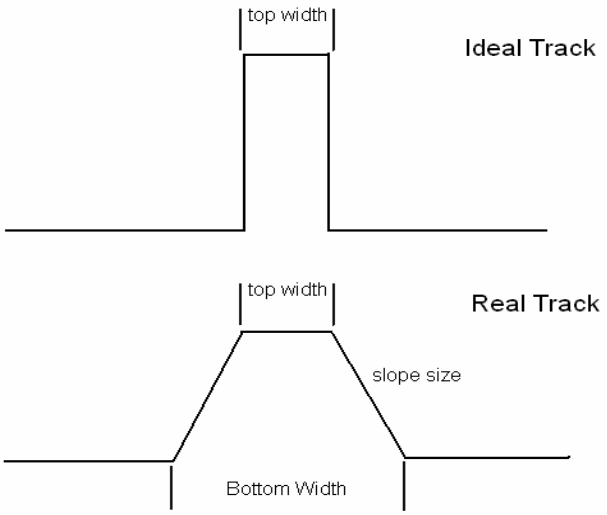


Figure 91 - Ideal Track and Track with sloped sides

The real track is convolution of two rectangular functions of non equal widths (w_1 & w_2 & $w_1 > w_2$), so the spectrum is the multiplication of two sinc functions.

$$T_{(f_x)} = \int_{-\frac{w}{2}}^{+\frac{w}{2}} \exp(j2\pi f_x x)dx \qquad \text{Equation 7-1}$$

$$T_{(f_x)} = w\,sinc(\pi f_x w) \qquad \text{Equation 7-2}$$

$$\text{where } sinc(\pi x) = \frac{sin(\pi x)}{(\pi x)} \qquad \text{Equation 7-3}$$

The resultant for the two widths is therefore the multiplication of the two sinc functions for the two widths w_1 and w_2:

$$R_{(f_x)} = w_1 sinc(\pi f_x w_1) \times w_2 sinc(\pi f_x w_2) \qquad \text{Equation 7-4}$$

$$r_{(x)} = \Im^{-1}\{R_{(f_x)}\} \qquad \text{Equation 7-5}$$

The relationship between the two rectangular functions and the width of the sloped section and the top width of the track is a follows:

w_1 = top width + slope width

w_2 = slope width

Using the above equations it is possible to generate a set of tracks with constant width and variable slopes as shown in Figure 92.

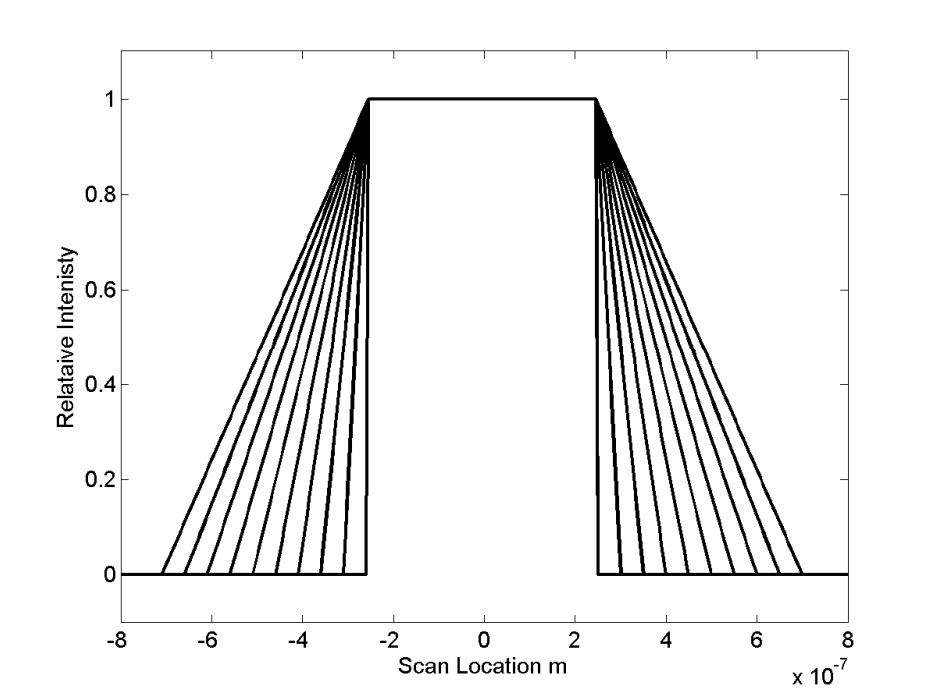


Figure 92 - 500nm top width track with various different slope sizes

This a very simple model of the phase response of a track structure and in a real situation the phase response may be different as this model does not take into account surface scattering effects. This simple model is used to demonstrate the possibility of extracting other parameters.

An ANN was trained where the top width of the tracks was kept constant at 1 micron and the slope sizes were varied from 2 to 248nm. The wavelength used was 633nm and the NA of the system was 0.3. The training results from this network are shown in Figure 93.

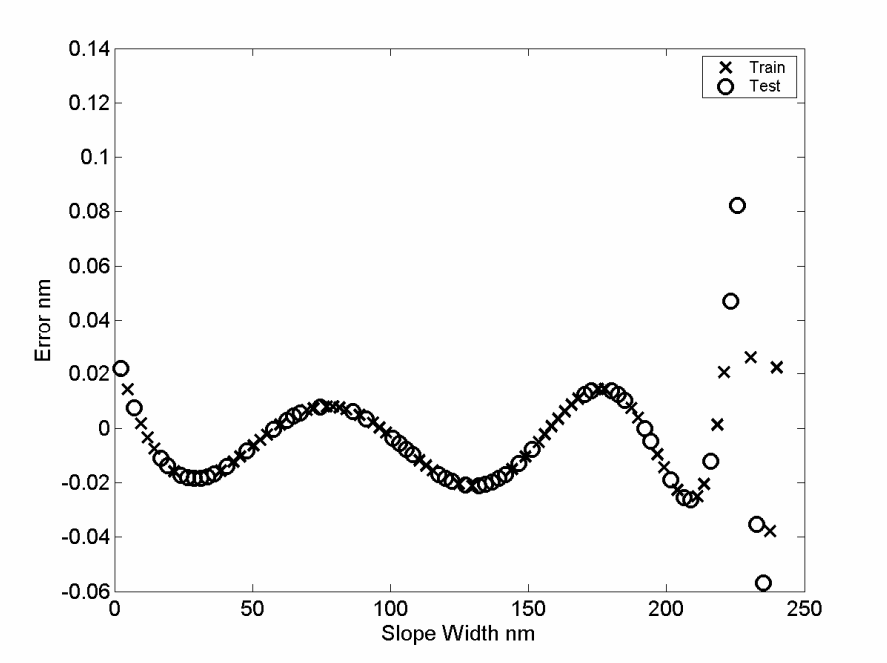


Figure 93 - constant width variable slope size results

The network was easily able to calculate the slope size for this very simple noiseless example. The standard deviation of the error for the testing samples was 0.028nm.

7.1.1 Impact of varying sloped tracks on network trained on tracks with fixed slope value

The effect on the errors produced when tracks of varying slope sizes were applied to a network train on a set of tracks with a fixed slope was investigated. For all of the tracks with different slope values the error for the track width and slope value was considerably higher than for the data that was used to train the network. The error was, however, fairly linear across the range and is roughly the same as the slope difference between the set used to train the network and the actual slope value. e.g. tracks with a 40nm slope size applied to a network trained with 50nm slopes have approximately 10nm error on the width.

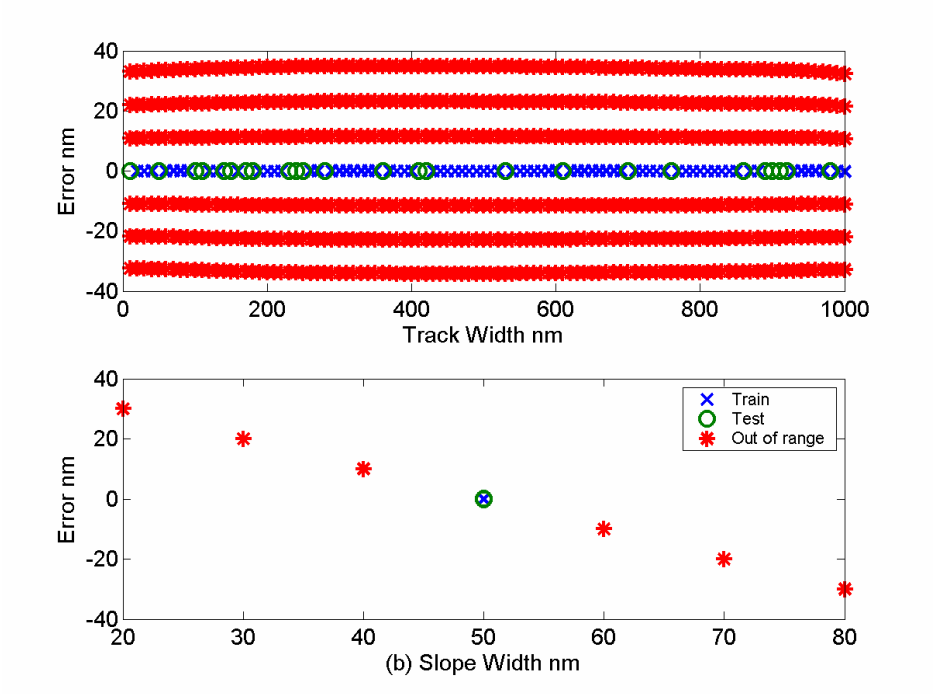


Figure 94 - Effect of slope size on training error

A second network was trained where several slope values (30-70nm) are used in the training set see Figure 95 with the aim of producing a more general network capable of producing accurate track width values for several different slope values. In the second case the training set is 5-6 times larger than the first.

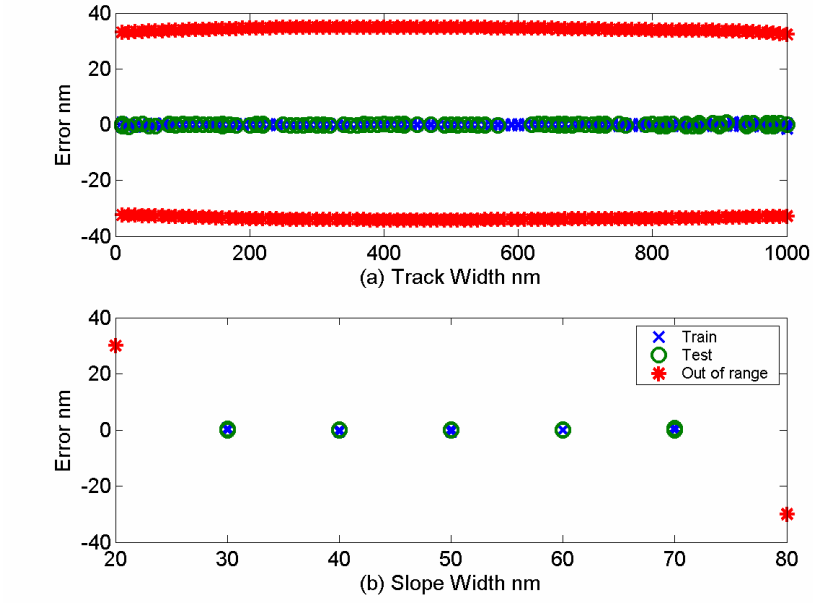


Figure 95 - results for random slope sized network

As can be seen for the tracks with slopes 30-70nm the errors are very small. But for the 20 and 80nm slopes that were not contained in the network training the error is larger – but still fairly linear and related to the difference between to the central slope value of the training set and the current value. e.g. error is approximately 30nm for the 20 nm slope value (50nm is central value). The standard deviations for the training, testing and out of range sets is given in Table 33.

Table 33 - Edge slope and width results

Standard deviation of error (nm)	**Trained on 1 slope value**	**Trained on 5 slope values**
Train-width	0.044	0.31
Train-slope	0.000019	0.13
Test-width	0.032	0.29
Test-slope	0.000022	0.13
Out of range - width	24.48	34.05
Out of range - slope	21.62	30.08

The errors for the network trained with 5 slope values is approximately ten times higher than for the single slope value case. However the error is still sub nanometre and the network is far more robust – if there are small variations in the slope values of the sample then the network is capable of still measuring the width and slope values correctly.

From a practical view point the work above could lead to another source of error. If the golden sample used to train the ANN has a fixed specific slope value then any tracks to be measured that differ from this will have an error in the width measurement due to the difference in the golden standard slope value and the their own slope value. If however the golden standard had tracks with random slopes then

this would not happen as the network would become more general as shown above. Also it is not necessary to actually know the slope values to get this general track width network, as it would become general just by using random sloped tracks in the training set. If the slope value is of interest then a range of slopes can be used in the training to obtain this parameter this is demonstrated later.

7.2 Height simulation

Another important parameter of interest is the height of a track structure. As the height varies it will modify the spectrum of the measured profile due to increase of scattering and the change in peak phase. More high frequency components are created as the height of the features increase. In addition, the apparent height of the features at the image plane will decrease as the width decreases.

A simple situation was simulated. The height of a track was varied (45-100nm) while the width was kept constant at 500nm and a network was trained. The results from this training are presented in Figure 96.

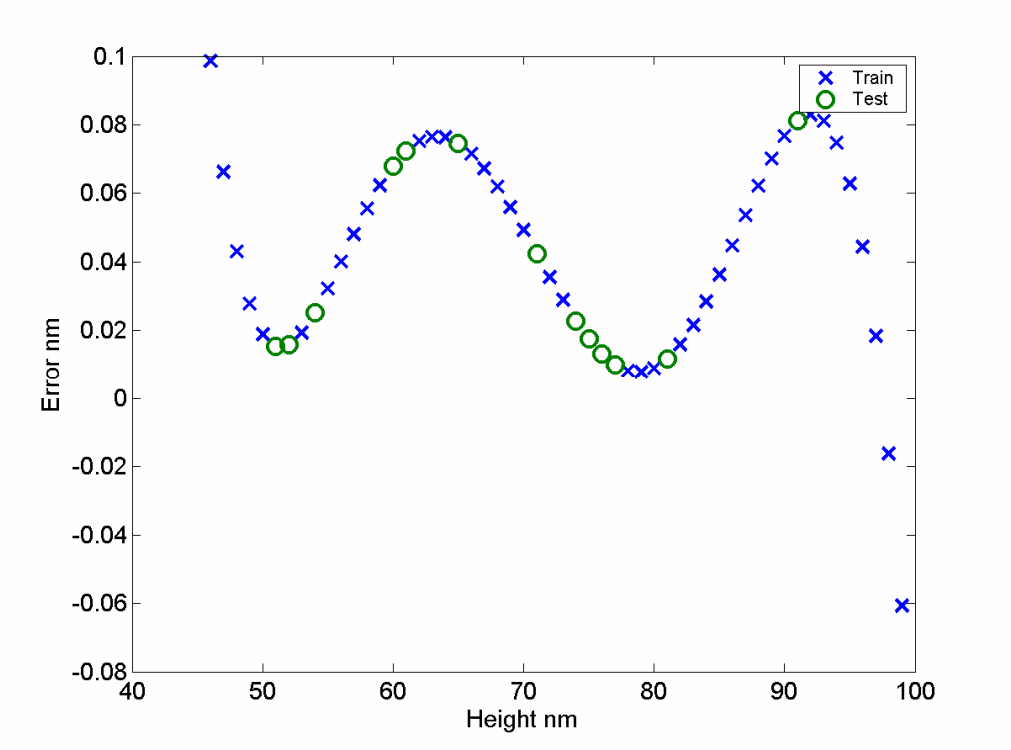


Figure 96 - Varying height constant width network

The standard deviation of the test set was 0.028 nm. This shows the network is more than capable of extracting the height information for this simple case.

7.2.1 Height / range of heights

An investigation into the effect on the errors produced when tracks of varying height are applied to a network trained on only one height value has been performed. A network was trained on tracks with widths from 20nm-1 microns where the height of the tracks was 45nm. Tracks with the same range of widths but with heights ranging from 10-52nm were applied to this network and the errors plotted in Figure 97. The curved lines correspond to the tracks with different heights. The lower have smaller heights and the upper curves larger heights than the training set which is in the middle.

As can be seen when the height changes the error in the width value increases considerably. For example a 5nm increase in height leads to a track width error of around 30nm for a 600nm track.

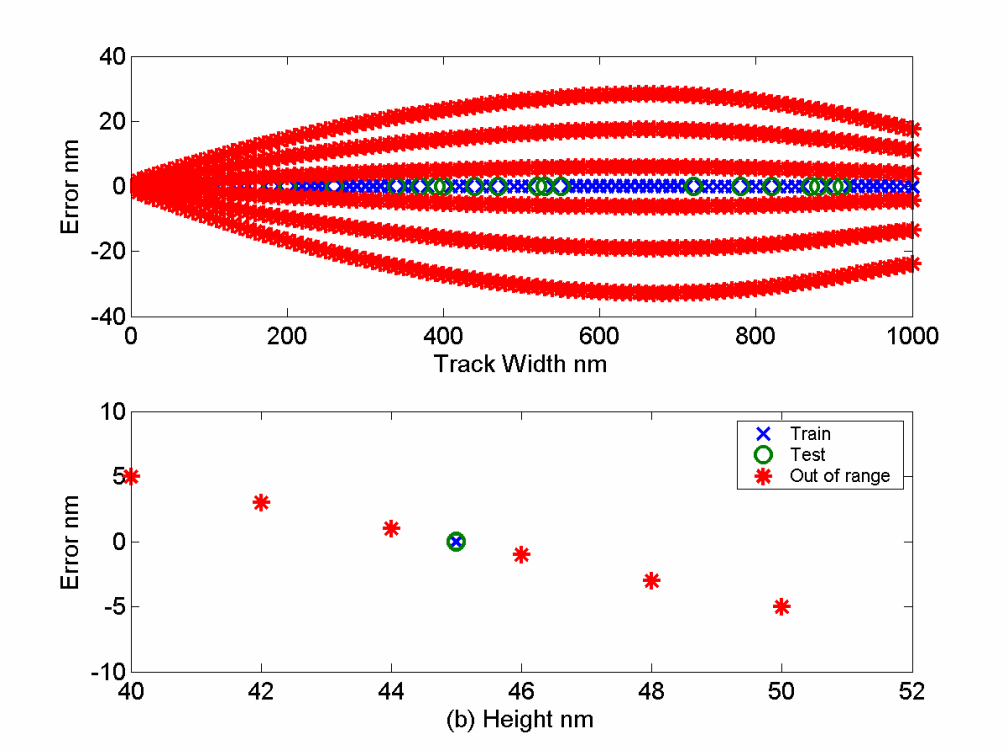


Figure 97 - Set 1 (T1 Top, T2 Bottom)

To produce a more robust, general network, another network was trained by including several height values (42-48nm) and the same range of widths in the training set. The results of this training are shown in Figure 98. By including the extra tracks covering a range of both height and width values the network now produces accurate results for both the height and width of tracks in the range of 42-48nm high and 20-1000nm wide. The two poorly performing lines on Figure 98 are due to 40 & 50nm high tracks which were not included in the training range and are therefore expected to be poor.

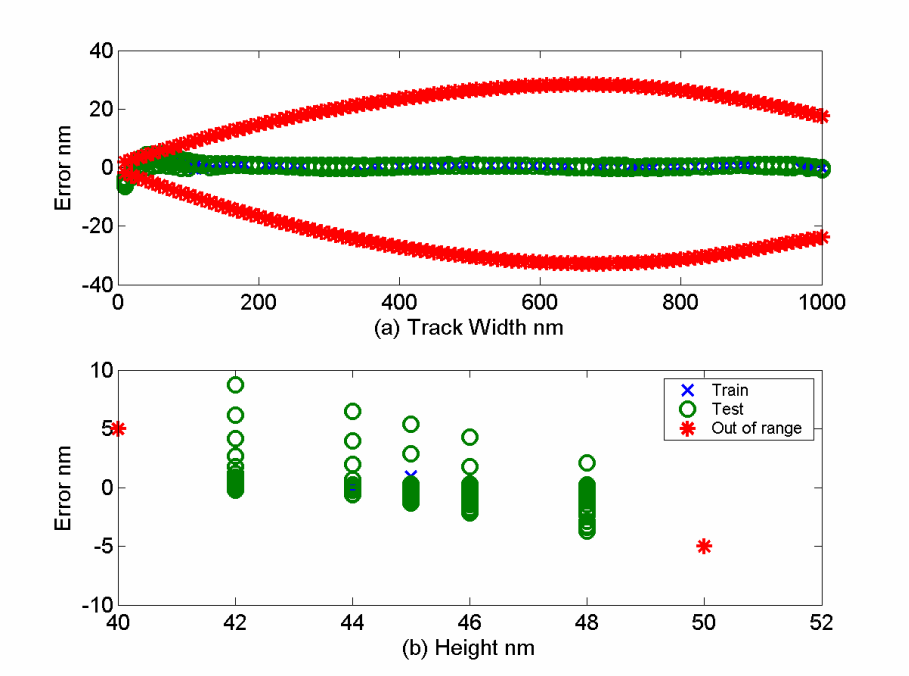


Figure 98 - Set 2(T1 Top, T2 Bottom)

The standard deviations for the training set, testing set and the out of range tracks are given in Table 34 below.

Table 34 - Height and width results

Standard deviation of error	Trained on 1 height value	Trained on 5 height values
Train-width (nm)	0.037	0.32
Train-height (nm)	0.000005	0.067
Test-width (nm)	0.028	0.30
Test- height (nm)	0.000004	0.080
Out of range – width (nm)	16.45	24.12
Out of range – height (nm)	3.42	5.013

The errors for the width values have increased 10 fold, the height errors have increased but are still sub nanometre. This final network is much more robust as it still

correctly gets the width value even if there is variation in the heights of the tracks, but this comes at the price of higher errors. It is still important to know the range that the network is applicable to as the out of range errors get large very quickly once outside the training range.

7.3 Architectures

This technique could be extended to get many parameters for different types of structures, by employing a tree of networks, which classify the structures into different types, calculate which network to send the profile before obtaining the track parameters. An example diagram for this approach is given in Figure 99 for double and single-track objects. Using multiple networks like this is required as in general the networks used perform extremely well when calculating one object parameter over a relatively small range. If more complicated measurements are required then combining many smaller networks will produce superior results as opposed to one very large network.

Most of the networks in the system will be classifiers in that they will decide which branch to send the data down. As long as there are no classification errors (some form of error checking will be required) then the errors in the parameter measurements will be due to the final network used to obtain the parameter value. This means that the final error will not be dependent on the preceding number of layers. It also means that other trees and networks can be added to the system easily as they will not have an impact on the other networks. The total number of networks involved in this type of architecture could grow very quickly especially if large ranges and many parameters need to be measured. This is not a large problem however as the structure of the

network is hidden from the end user as when in use the data will flow through the tree and produce the measurement values.

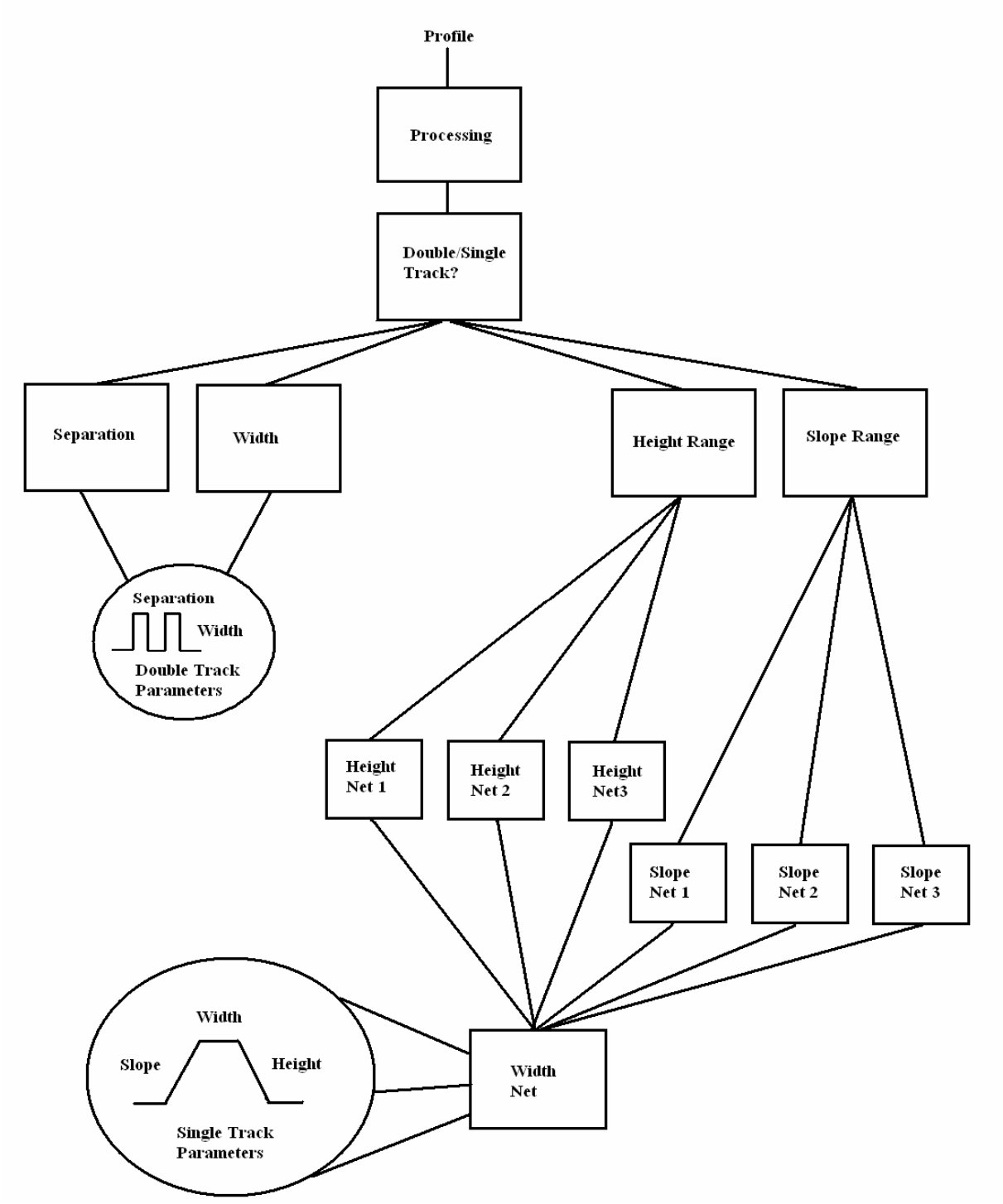


Figure 99 Schematic of multiple networks to obtain object parameters for different objects

The individual elements have been simulated and will be presented. It should be noted that this is just one approach and is by no means optimal. More work is required in this area to see if this is the most suitable method.

7.3.1 Single Track Tree Simulation

The simulation of the single track tree requires a 3 dimensional data set in width, height and slopes. Unfortunately due to the time involved in generating that many tracks only two 2-dimensional sets were simulated, these were a set of tracks with fixed height and varying width and slopes and a set with fixed slopes and varying height and widths. These will be used to demonstrate the tree structure approach to multiple parameter measurements.

Using one network to obtain both the slope and width parameters was not very successful, as the network had far too much to do, as shown in Figure 100. This is very similar to the situation in section 7.1 but in this case the range of slopes are much higher 0-280nm. By improving the design and altering the structure of the network and increasing the number of inputs it may be possible to achieve better training performance. This will be a trade off with training time as with very large networks the training time increases dramatically, although training time is of no consequence to the end user.

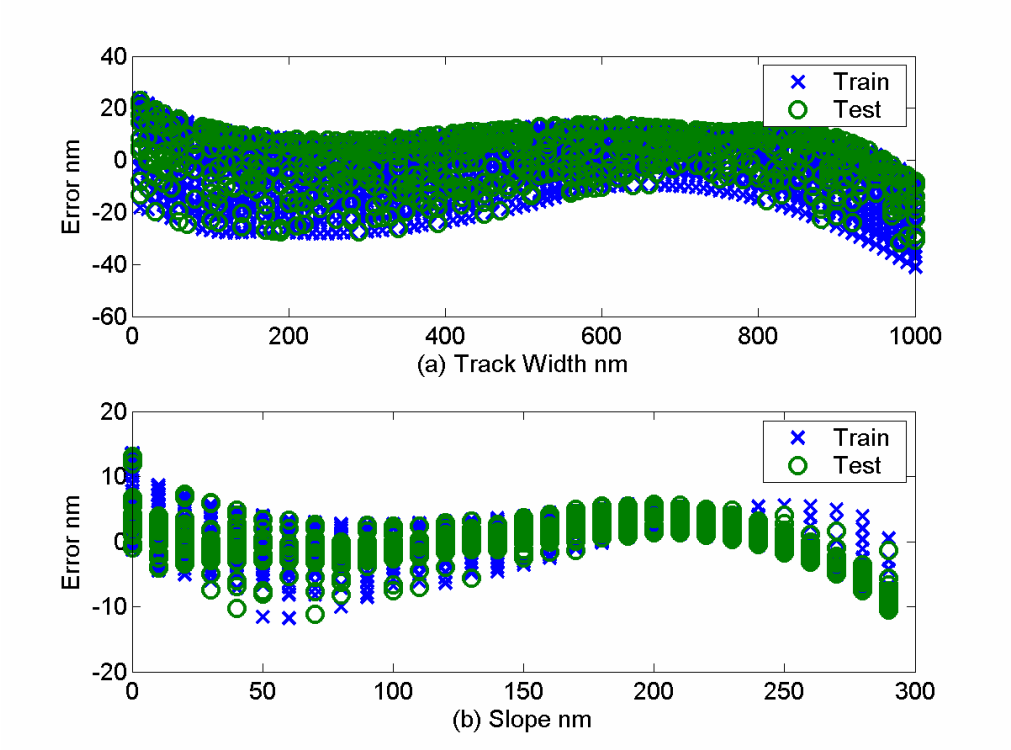


Figure 100 - One network for both width and slope parameters

When two networks were used, one to obtain the width and the other to obtain the slope the training was much more successful, as shown in Figure 101. This allows simpler networks, which are easier to train to be used. In this case there are 3000 different tracks in total with varying width (20-1000nm) and slopes (0-280nm). 75% of them form the training set the other 25% are the testing set.

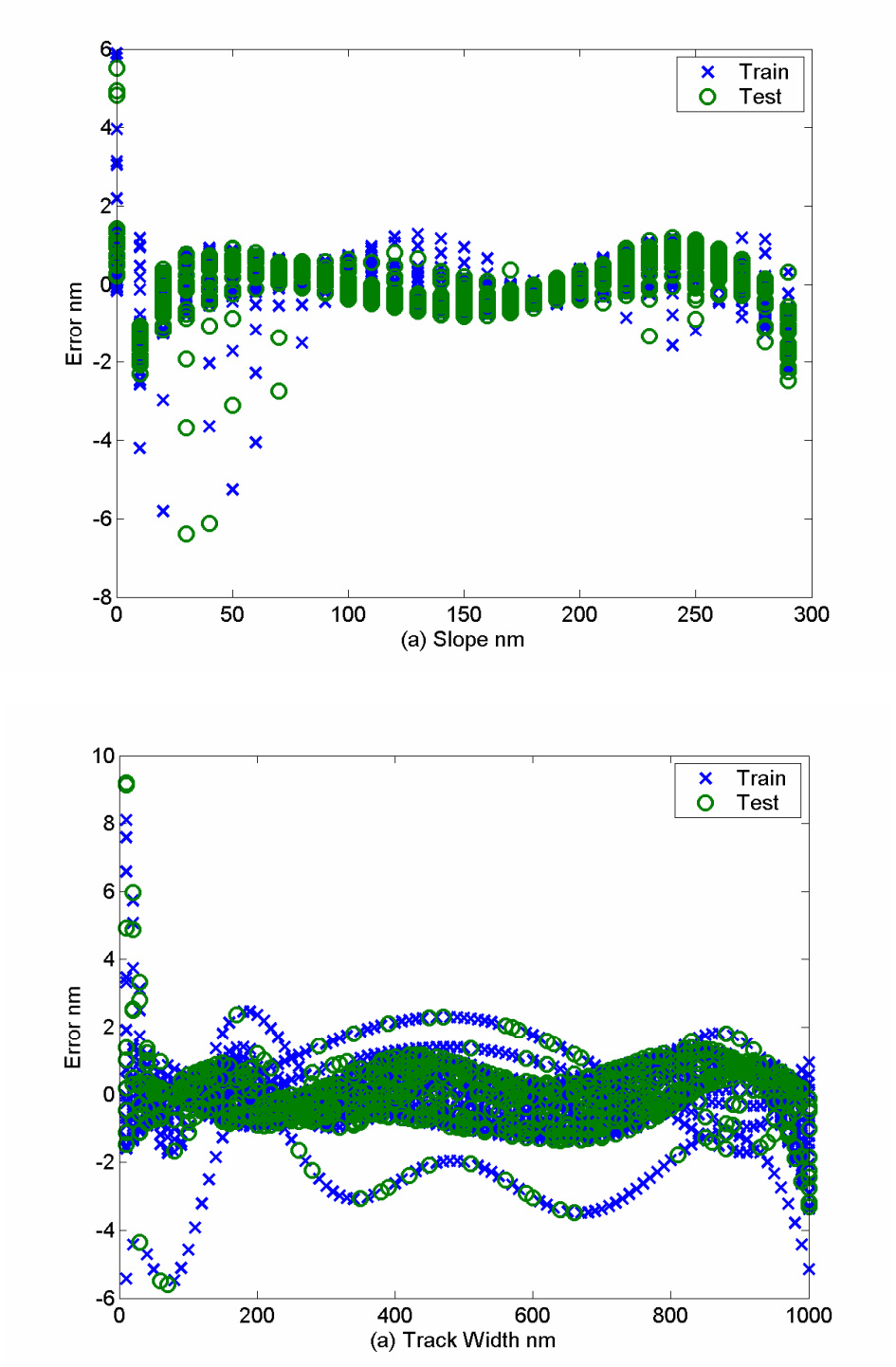


Figure 101 - Using one network for each parameter

This same process was repeated for the height. Again a single network performed poorly (Figure 102), but using one network to get each parameter improved training greatly (Figure 103).

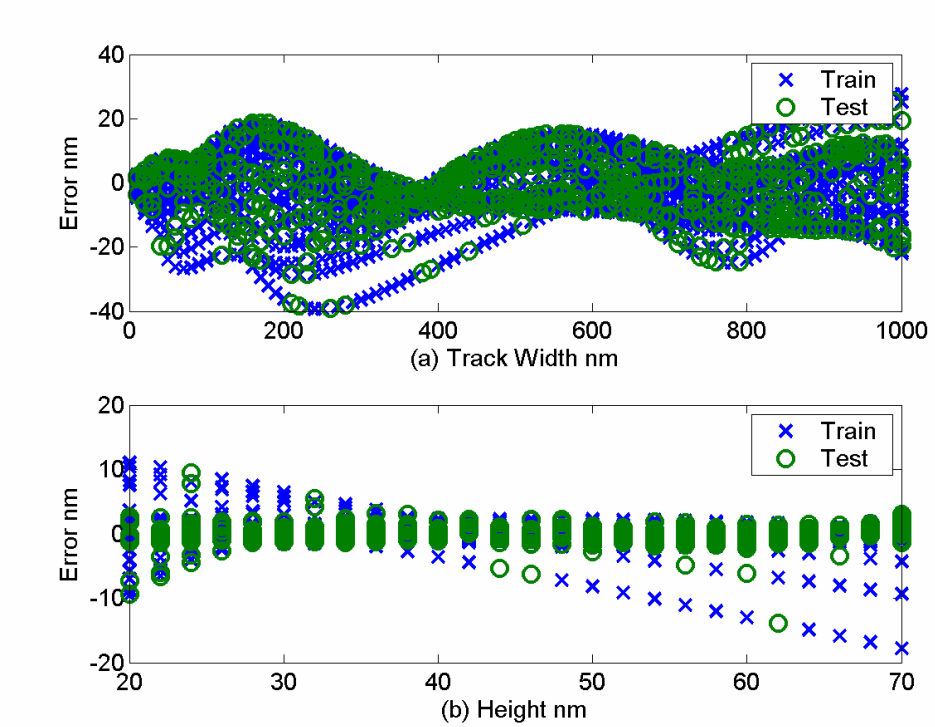
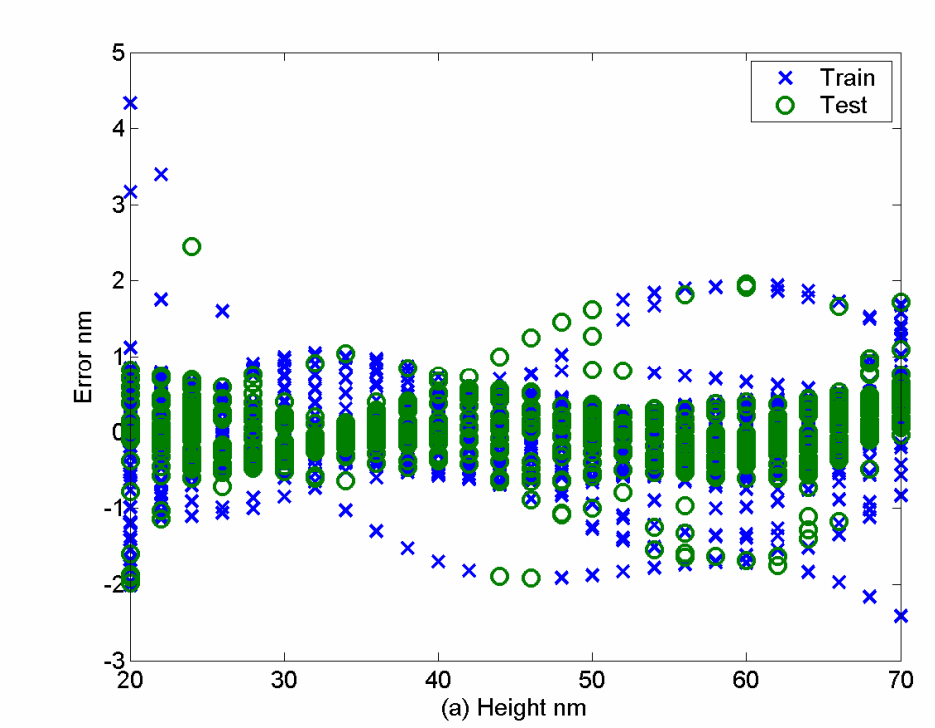


Figure 102 - One network for height and width values



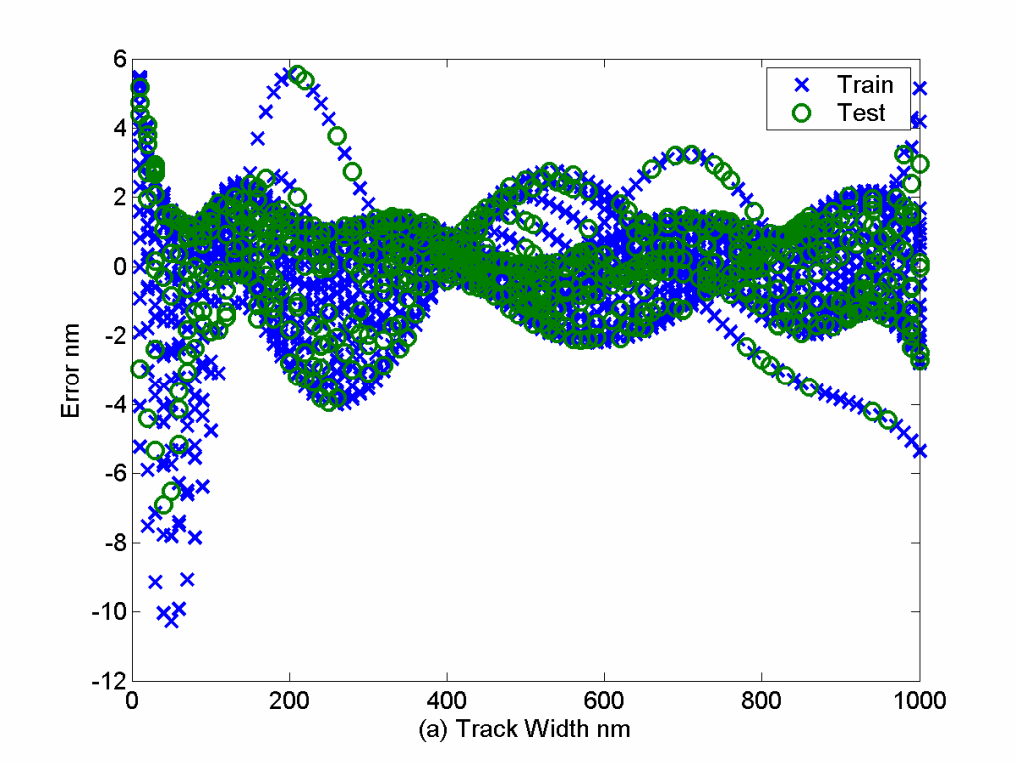


Figure 103 - One network for each parameter

A summary of the training results for the above examples is presented in Table 35.

Table 35 - Training results for slope and height values

	Slope Varying Tracks			**Height Varying Tracks**		
	Dual Net	Width Net	Slope Net	Dual Net	Width Net	Height Net
Std Train w (nm)	11.17	0.98		10.25	1.71	-
Std Train s (nm)	3.35	-	0.71	-	-	-
Std Train h (nm)	-	-		1.72		0.50
Std Test w (nm)	10.13	1.09		10.40	1.52	-
Std Test s (nm)	3.36		0.80	-	-	-
Std Test h (nm)	-	-		1.50		0.49

As can be seen the separate networks for each parameter produce much better results. It should be possible to improve these training results by reducing the range that the network has to train over. This will make each situation simpler and so the training easier. This has been carried out where the height and slope ranges have been split into 4 sets. This means that we now require a range classifier to send the tracks to the correct networks and we then require four networks for both the slope and height values to cover the ranges. These networks will calculate the width value as well as the slope or height for that range.

By having a set of networks to concentrate in a specific range the training errors can be reduced. This requires two things, firstly a pre-classifier to establish the range of the current track to be measured. And secondly a network trained for that specific range to establish the parameter of interest for the track to be measured. This is performed for height and slope and the results are discussed in the following sections.

7.3.2 Single Track Height Classifier

The range of heights have been broken into four subsets so that they can be processed by four more specialise networks. Each subset is assigned a target value. The output targets are chosen to be a 2-bit number so that fewer outputs are required. The bit order is chosen to keep the range a smooth a possible i.e. only one bit changes at a time. The Range and targets are shown in Table 36.

Table 36 - Height range classifier outputs

Range	T1	T2
1	-1	-1

2	-1	1
3	1	1
4	1	-1

This network is similar to the others previously used and as such they do not cope well with discontinuities, so there are transition targets at the edges to reduce the discontinuities. With a better choice of network and network design this would be less of a problem.

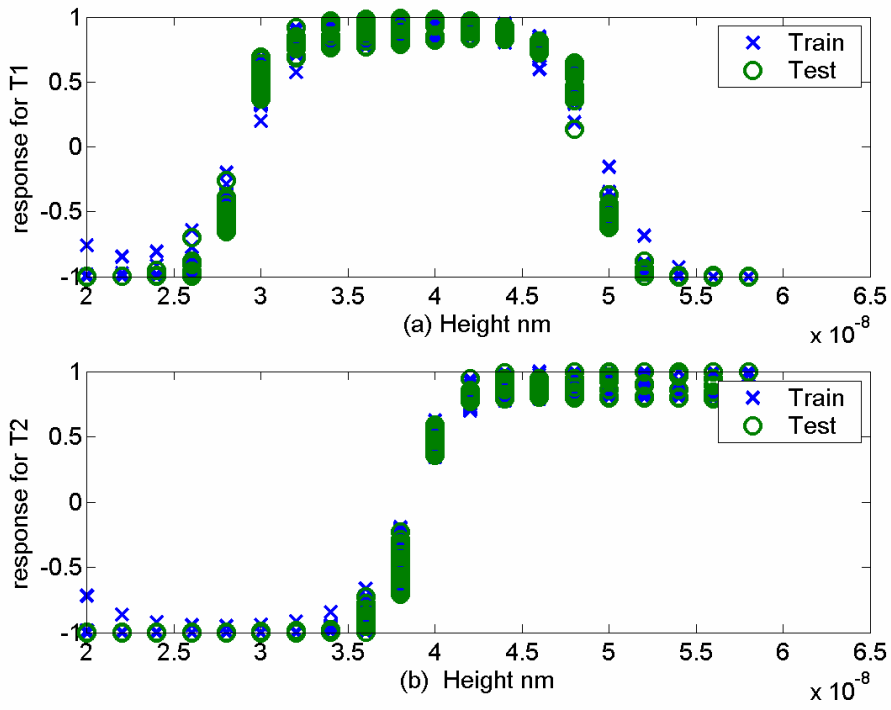


Figure 104 - Height range classifier results

After the network was trained (Figure 104) the results were thresholded and any thing positive became +1 anything negative became –1. This shows that the results are as desired; everything was classified correctly and would be sent to the correct specific height range network (Figure 105).

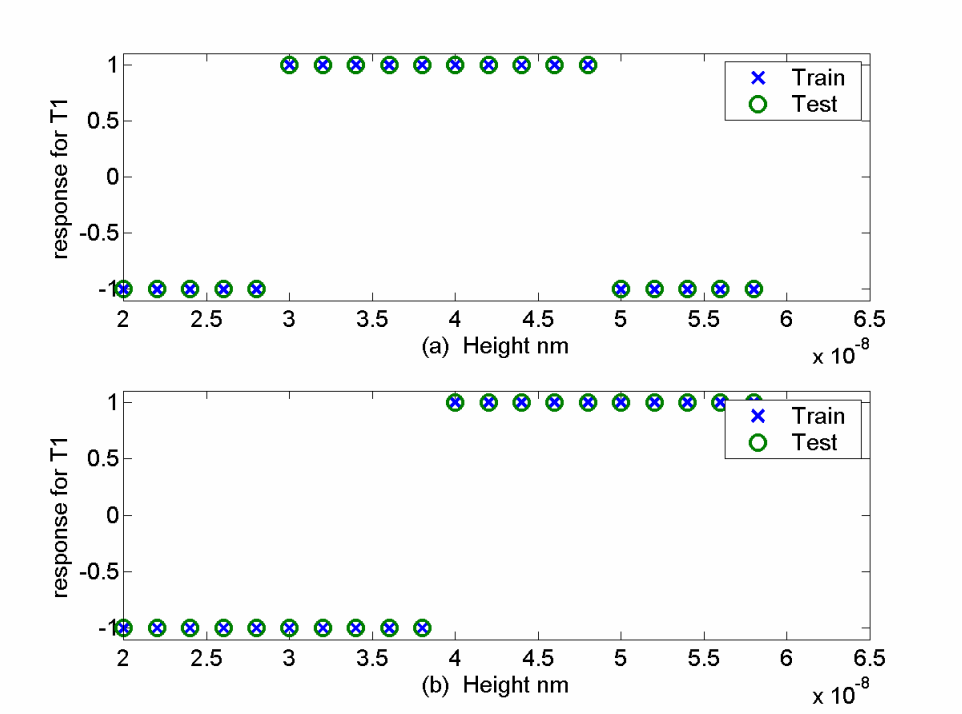


Figure 105 - Results after threshold

The four ranges covered by the new networks will be:

- Net1 (H 20-28nm W 50-1000nm)
- Net2 (H 30-38nm W 50-1000nm)
- Net3 (H 40-48nm W 50-1000nm)
- Net4 (H 50-58nm W 50-1000nm)

An example of the trained network is given for network 4 in Figure 106. As can be seen the errors are several times smaller than for the case where all of the height values were trained on at once.

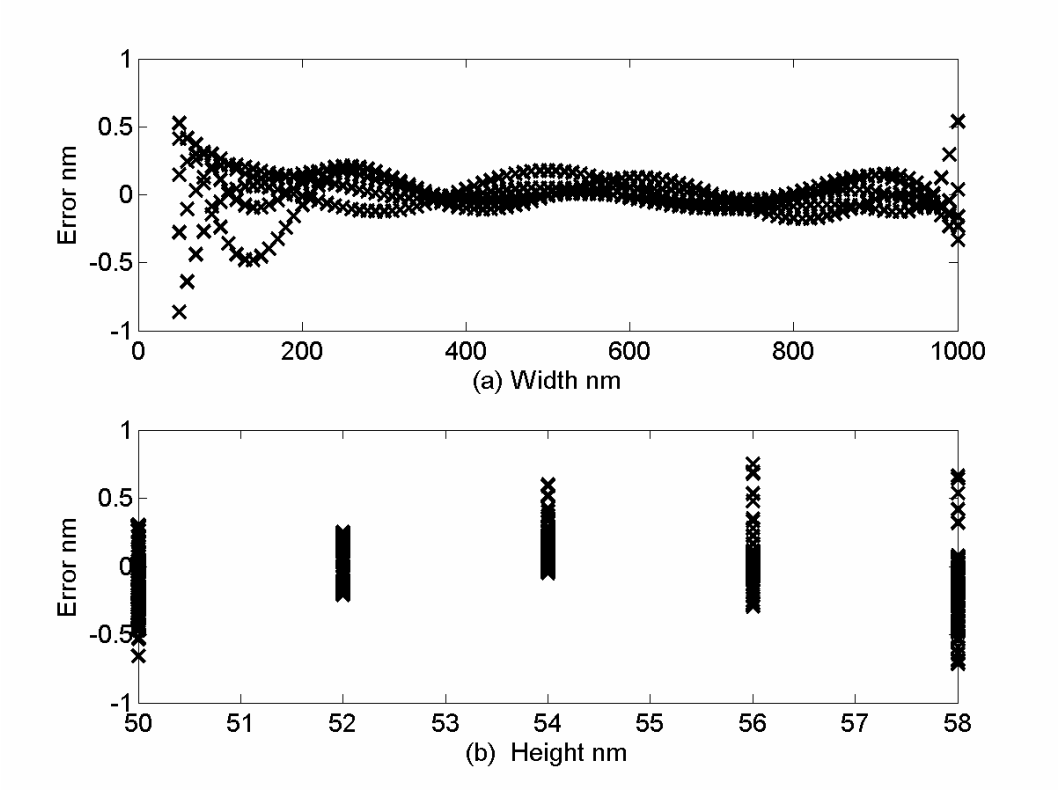


Figure 106 - Reduced range training results for Net 4

The training results for the 4 networks are given in the Table 37.

Table 37 - Training results for the four height networks

	Net 1	Net 2	Net 3	Net 4
Std width (nm)	0.186	0.184	0.186	0.165
Std height (nm)	0.06	0.075	0.097	0.10

The trained results are 7-9 times better for the width value and 5 times better for the height value.

The network could be improved by having overlapping zones at the edges as it is the tracks closest to the range transitions that are most likely to be misclassified. In this

case if the networks overlap slightly then the misclassified tracks will still be sent to a network designed for them as they will be in the overlap zone and so the correct parameter values will be calculated. Obviously there could be another layer of networks, to get the width range and then have several networks to just look at the width range for that specific height range. 1 class – 4 height nets – 4 width range nets (1 for each height) then 16 widths nets (4 for each height) – this should improve the training but increases complexity and amount of required training data.

7.3.3 Single Track Edge Slope Classifier

The same process as described above is performed for the edge slope data. Again the range of slopes has been divided into four ranges and the targets are the same as for the height case. Training the ANN produces the results shown in Figure 107. As can be seen the transitions are much less well defined but all tracks is still classified correctly (Figure 108). The slopes range from 0-280nm and as the height of the tracks is 45nm the angles that the sides make with the track substrate range from 90-9.1 degrees.

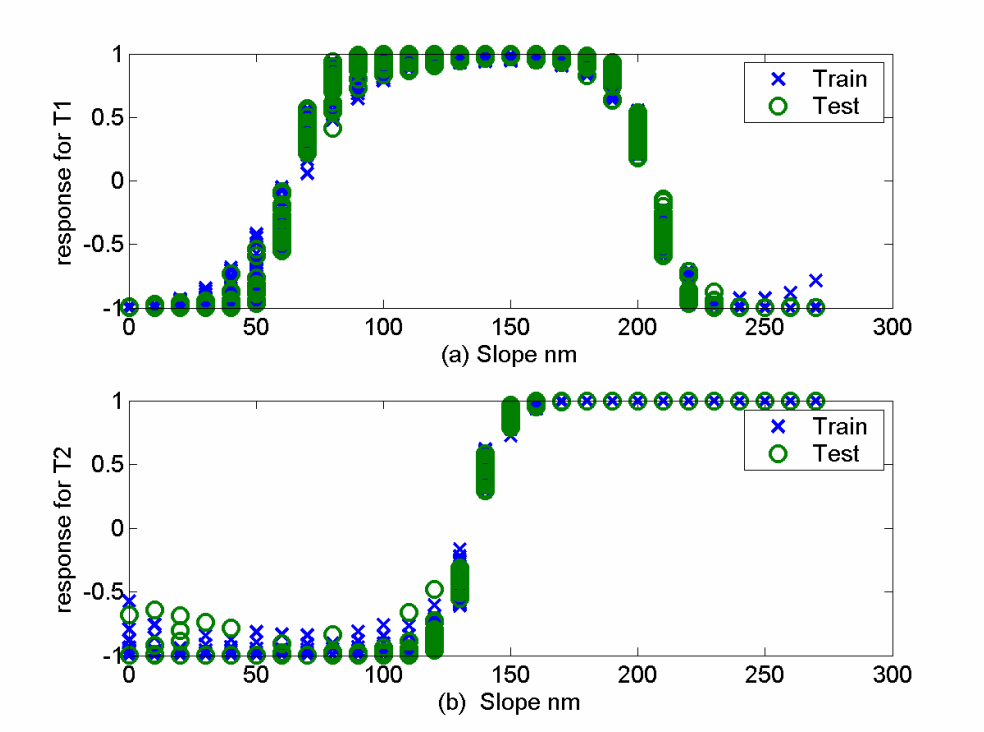


Figure 107 - Slope range classifier results

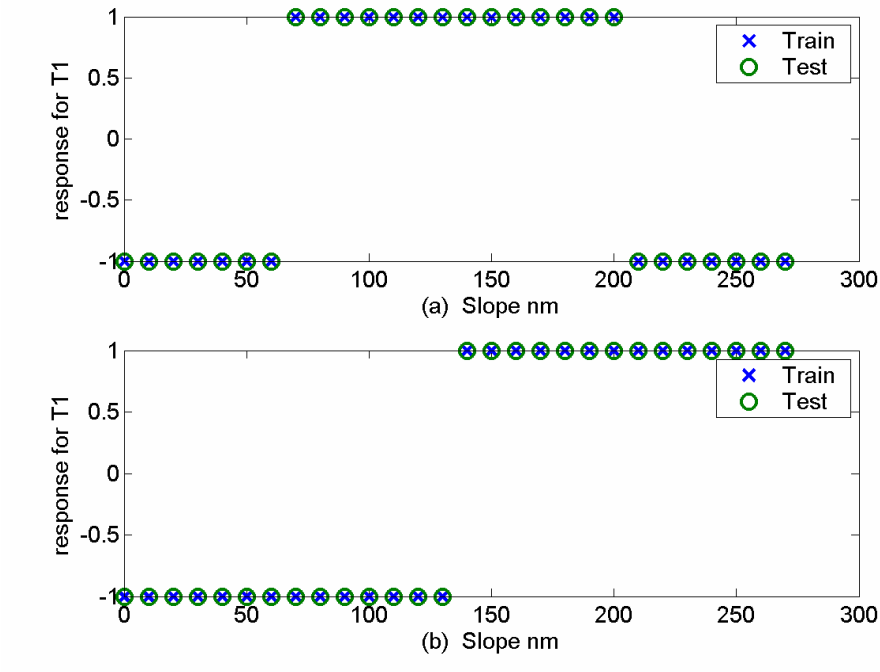


Figure 108 - Training results after threshold

The four ranges covered by the new networks will be:

- Net1 (S 0-60nm W 50-1000nm)

- Net2 (S 70-130nm W 50-1000nm)
- Net3 (S 140-200nm W 50-1000nm)
- Net4 (S 210-280nm W 50-1000nm)

We now require four networks to calculate the width and slope angle for the 4 ranges that the data was split into above. An example of the training for one of these four networks is presented in Figure 109. It is clear to see the 7 different slope values produce slightly different width values but the errors are sub nanometre where as the slope range is 60nm in this case.

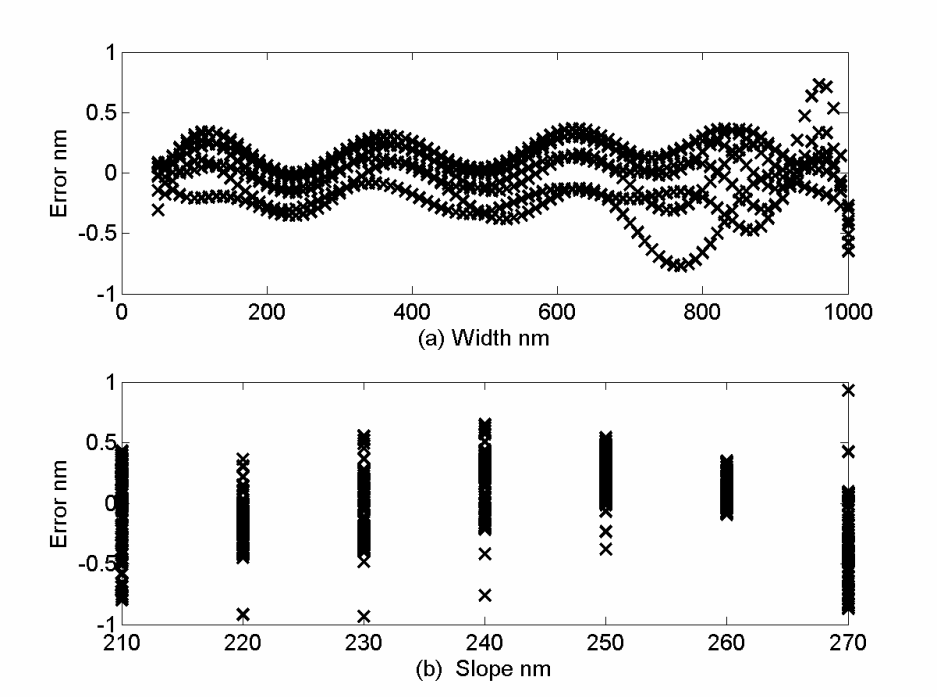


Figure 109 - Training results for net 4

Training results for all four of the networks are presented in Table 38. The standard deviations are relatively small.

Table 38 - Training results for the 4 slope range nets

Heights 20-1000 nm Slopes :	Net 1 0-60nm	Net 2 70-130nm	Net 3 140-200nm	Net 4 210-280
Std width (nm)	0.76	0.173	0.175	0.172
Std Slope (nm)	1.22	0.43	0.252	0.19

The training results are in general better than for the single range case. The width is 5 times better apart from in range 1 where it is similar. The slope value is 2-3 times better again except for Net 1. It would seem that the network 1 did not train as well. This is not unexpected as this network contains the smallest slopes and therefore the smallest tracks.

The network could be improved by reducing the range of widths that each height range looks at as described for the height case. Improvements to the network design for the range classifier will reduce the chances of misclassifications.

7.4 Profiles

The preceding section demonstrates that it should be possible to build up a good picture of the structure of an object. Having accurate values for all of the main parameters such as height, width and sidewall slopes it may be possible to extend this work to obtain a profile of the object.

The idea is simple and illustrated in Figure 110, firstly a track with certain parameters is generated, this then passes through an optical system model, the resulting image is then compared to the actual image obtained from the microscope, how well the two images match is stored and the track parameters are adjusted and the process iterates

through all possible permutations of the track parameters. Then the actual track parameters are found by finding the best global fit for all the parameters tried. There are however a number of problems with this idea. Firstly the scale of the problem is huge, especially if you do not know what the object is. Secondly the optical model has to be extremely accurate and thirdly the matching process must produce a quantitative measure of good fit. Overcoming these obstacles is vital to making this technique a success.

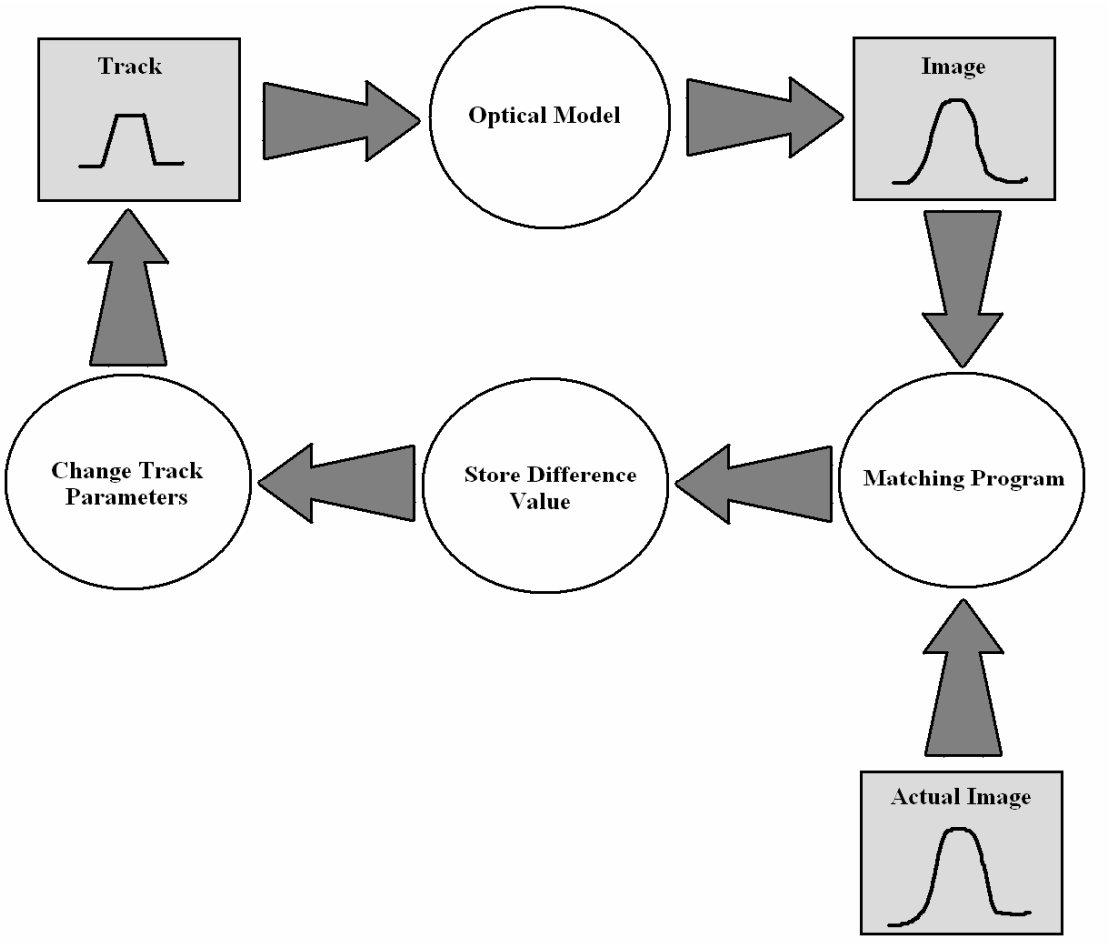


Figure 110 - profile generation

The scale of the problem is reduced considerably by the use of the ANN as this provides values for the height, slope, width and separation for the track. This means

that the starting point for the track structure is well known therefore reducing the dimensionality of the problem, this means that the parameters of the principal components are well defined. The scale of the search problem can be reduced by using more suitable parameters for the next iteration of the program instead of just trying every possible combination of parameter value and seeing which are best, search algorithms could be used to home in rapidly on the best solution. Genetic algorithms are particularly well suited to these types of problems and would reduce the calculation time considerably.

The optical model of the system could be developed from vector diffraction theory of the system, however this may not be accurate enough for this task as no matter how good the model is there will always be differences between the actual real world microscope and the model, as the effects of aberrations and impact of environmental conditions can not always be known. It would therefore be much better to obtain the optical model from the actual optical system itself, this could be done by measuring some known sample or examining the point spread function of the system.

The matching process could be achieved in a number of ways. A conventional match filter can be used, with the spatially reversed image acting as the filter impulse response, and the modelled image as the filter input. This approach, however, does not readily provide suitable criteria to test the matching. A modified match filter may work better for this application. With this approach, instead of taking the integral of the product of the two functions at each shift location, the integral of the difference will be taken instead. A null signal will therefore indicate a perfect match. After each iteration

the difference value is stored, and the problem is to find the global minimum of the error value.

This technique relies heavily on computing power as the process will require many iterations and the optical model and matching process can also be computationally demanding. However, with the ever-increasing power available of modern computers this sort of problem is becoming viable. This will enable super resolved profiles to be obtained.

7.5 Input points

The choice of the inputs to use is crucial to the success of this approach. This concerns the format of the input points i.e. using the differential spectrum but also which points from the processed data to use. A simple investigation of the importance of the different spectral components was presented in chapter 6. This showed that for the case of the single track the high frequency components were more significant. This work needs to be extended for multiple parameter objects to discover which spectral components are best for identifying various parameters. For example in case of a double track object certain spectral components may be dominated by a zero in the cosine term due to the track separation and other spectral components that are more influenced by the *sinc* term of the track widths. By only presenting inputs that are the most relevant to the parameter of interest better training could be achieved. Obviously this will become more complex as additional parameters are included and their effects on the spectrum are investigated it may not be possible to separate the effects of various parameters in the spectrum in which case other processing methods may become more suitable.

One promising technique is to use the most significant singular values of the profiles or differential profiles. This has the advantage of condensing the useful information to only a few parameters. For example for a double track structure the spectrum is quite complicated over many points but there are only two or three significant singular values. This processing method needs to be looked at in more detail.

Another approach to improve the inputs could be achieved by using a High NA objective to measure the small track or groove structures as this can yield additional information if polarised light is used. A paper by Morgan et al [63] showed that the response of small groove differed greatly depending on the choice of polarisation used this was demonstrated experimentally and was in good agreement with theory. This difference in signal may be utilised as inputs for the ANN. One part of inputs could come from profiles measured with one polarisation. A second part could come the orthogonal polarisation. The way in which the inputs vary as the tracks change will be different for the two parts of the inputs and this may make training easier for the network as there is more contrast in the inputs.

7.6 Network Development

The ANNs used have been rather simple and ‘off the shelf’ networks, specific work to develop networks optimally suited to this work could improve the effectiveness of the networks. This is especially the case for the classifier networks that have not dealt well with the discontinuities in the target outputs.

Networks that have different types of inputs may benefit by having additional layers that are not fully connected (see Figure 111) so that some pre-processing of the input

information is carried out by the network on the different kinds of inputs – for example from using different polarisations to scan the a track it could have two sets of inputs each derived from a scan with either *p* or *s* polarisations. After the first partially connected layer the network proceeds as usual and calculates the parameter of interest.

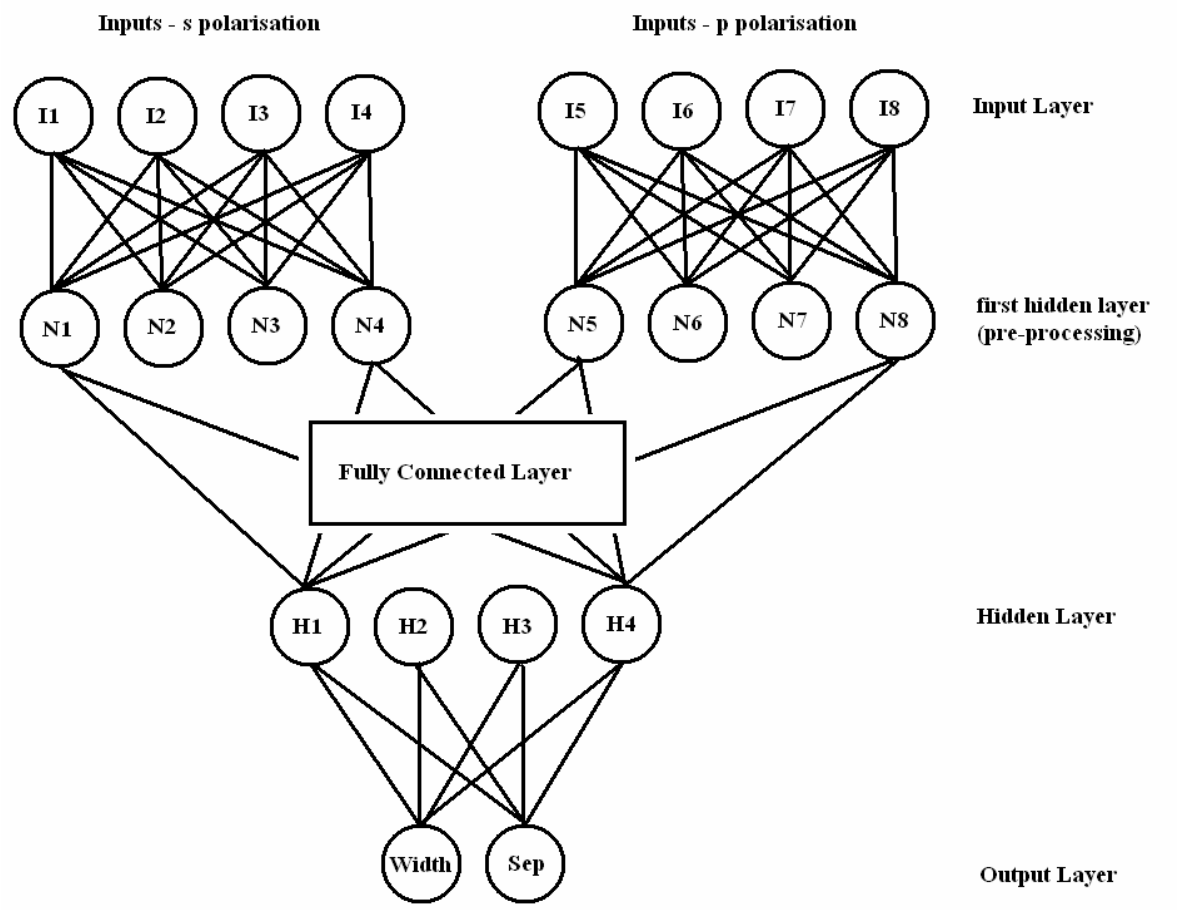


Figure 111 - example of a different network design

7.7 Future of Optical Microscopes

Ultimately, the optical microscope is going to depend more and more on novel signal processing techniques and computational power to keep pace with the demand of industry. Eventually the optical microscope will not be distinct from the computer, the merging of the systems will produce smart, adaptive systems where the response of the optical system is modified on the fly to best suit the object being measured. For

example the illumination conditions can be modified to change the systems response to aid the measurement of specific object parameters such as sidewall angle. Signal processing will extract many object parameters such as the height, width, side wall angle, edge quality etc. and optical models based on the actual experimental response of the current optical setup will be used to produce super resolved profiles, these may be based on analysis of the point spread function of the current optical setup or analysis from measurements of test known samples. This may seem a long way off but computing power is increasing at a tremendous pace opening up the possibility of this approach in the coming years.

7.8 Providing a user friendly system

The system needs to be modified before it can be used in a general-purpose lab at NPL for routine measurements. One way to do this is to construct the main microscope section around a traditional upright microscope body. This will provide robust mechanical form and usability. Coarse and fine motion control stages will be required to provide sample navigation and fine scanning control.

A wide field imaging arm also needs to be incorporated into the system to aid with sample alignment and navigation, which should be relatively simple to accomplish if the above approach is adopted. A wave front sensor at the Fourier plane should be incorporated to help with alignment of the sample. This will ensure that the sample is normal to the optical axis, and is in focus, to within a certain tolerance. In an ideal world, a set of samples designed for diagnostic purposes would be available. The sample is measured by the system and the profiles are processed. A program will tell

the user if the system is in perfect alignment or whether something needs adjusting with suggestions of what to adjust.

Master samples will be required for several different objects of various parameters. These samples will be calibrated by an SEM or AFM. This will produce a system that is appropriate for lab-trained technicians to use as part of the measurement service for the NPL providing calibrated track width samples for industry. A full uncertainty budget needs to be completed for these results to be traceable.

7.9 Summary

In conclusion the ANN approach is very powerful as many different parameters can be measured to a high degree of accuracy. By splitting large ranges of interest into smaller groups better training results can be obtained. A summary of the main training results presented in this chapter is given in Table 39.

Table 39- summary of training results for different parameters presented

	Slope Range (nm)	**Width Range (nm)**	**Errors (nm)**
Slope	0-60,70-130,140-200,210-280	50-1000	Slope <1.22 (mean 0.52) Width <0.8 (mean 0.32)
Height	20-28, 30-38, 40-48, 50-58	50-1000	Height <0.1 (mean 0.083) Width <0.19 (mean 0.18)
Classifiers	4 ranges	50-1000	All correct

Further improvements can be made by tailoring the inputs used to the parameter of interest. This could be achieved in a number of ways by, for example, picking specific spectral components or using other processing methods such as singular values and

optimising the network design. The overall architecture (such as the tree approach presented) enables good measurements to take place over a large range for many different parameters.

8 Conclusions

The main purpose of this research was to enable submicron object parameters to be measured with an optical system by harnessing the power of signal processing techniques – namely artificially neural networks.

The limited resolution of an optical microscope arises from the finite pass band of the aperture of the optical system and by the wavelength of the radiation used. The finite resolution means that objects smaller than approximately 200 nm cannot be resolved optically. Due to the many advantages of optical systems, such as, non destructive, non contact, ease of use and the types of samples that can be imaged etc. they are very desirable in many measurement situations, and therefore being able to extend the measurement range of the optical system would be very useful for many applications.

Our approach has moved away from the work previously attempted by others where a super-resolved profile was the ultimate goal, for many applications it is the measurement of the object parameters that is ultimately required and so our focus was to provide this ability.

The technique is primarily optical system independent, this means that the most suitable system for the sample under investigation can be utilised. There are, however, a few key requirements for the technique to be successful, such as, high signal to noise ratio, high repeatability/stability. This technique does depend upon having access to a much higher resolution system to calibrate a sample that can be used to train the neural networks. The choice of system (e.g. AFM or SEM) is not crucial as long as it provides

traceable measurements. However it must be noted that that system will have its own measurement uncertainty and the optical system with the ANN will be effectively modelling the system that calibrated the 'master' sample.

The optical systems used for this work were the DSOM, the scanning Nomarski microscope in two modes of operation, and the hologram system. These systems all had good signal to noise ratios and were very repeatable. A summary of the optical systems used is given in Table 40.

Table 40 - Comparison of optical systems

	Hologram	**DSOM**	**Nomarski**
Sample types	All types some restrictions	Reflectivity or large phase structures	All types, well spaced
Amplitude profiles	✓		+
Phase profiles	✓		+
Intensity	✓	✓*	✓*
Practical SNR achieved	Amplitude 1 in 3000+ Phase 0.5mrad	Intensity 1 in 1500 Differential intensity 1 in 3-400 Differentiation should not affect the SNR	Bright field 1 in 1100 Dark field better
Complexity	Medium	Low	Medium

*and/or differential intensity
+Phase stepping can be used to obtain amplitude and phase information

Several different samples were measured to demonstrate the technique experimentally. These ranged in size from a 1-3 micron chrome on glass sample to a 60-480nm silicon track sample. Double tracks structures from 1-3 micron width with 1-4.5 micron separations were also measured. A simple classifier for double or single tracks was trained for the chrome on glass sample.

The main results from the networks are presented in Table 41.

Table 41 - comparison of training results for different samples and systems

Sample Name	1-3u	BCR	Silicon
Range	1-3μm	0.273-2.1 μm	0.06-0.48 μm
DSOM	10.86 nm	2.55 nm	-
Scanning Nomarski	-	18.8 nm	5.5 nm
Hologram	-	1.9 nm	1.7 nm

The hologram system produced the best training results. This system is the most stable due to its mechanical design and common path nature.

This approach is also capable of measuring double track structures and providing accurate width and separation values. The errors are better than 10nm, which corresponds to smaller than 1% error across the whole range of widths and separations.

The measurement of other parameters such as sidewall slope and track height has been presented through simulations; due to lack of suitable samples these have not been confirmed experimentally. The fact that many parameters can be measured is a testament to the power of this technique.

Where multiple parameters and large ranges need to be covered a tree structure of different networks can be utilised to keep the training results accurate. This is a very suitable structure as the overall number of networks is hidden from the end user and new networks are relatively easy to add to the structure to provide new measurements, as they do not impact on existing networks.

Further improvements could be made by designing artificial neural networks that are optimised for this type of work. Choosing more suitable input data or performing other types of data transformation or signal processing to extract the most pertinent data for the parameter of interest will also increase training performance.

This technique has demonstrated great improvements in the measurement capability of optical systems. By increasing the NA and reducing the wavelength, providing a more stable environment (vibration isolation and temperature stability) it should be possible to measure track structures considerably smaller than those presented.

Table 42 - ratio of smallest track to optical spot size

Spot size	2574.2 nm
Smallest track	60 nm
Ratio	42.9

The ratio of the smallest track measured to optical spot size is given in Table 42. This ratio should still apply if the wavelength is fixed and the NA is increased. This allows a prediction of the smallest tracks that it should be possible to measure for different setups. This is shown in Table 43.

Table 43 - predicted smallest track sizes for different wavelength and NA in nm

NA	0.3		0.5		0.7		0.95	
Wavelength	**Spot d**	**Track w**	**Spot d**	**Track w**	**Spot d**	**Track w**	**Spot d**	**Track w**
633	2574.2	60.0	1544.5	36.0	1103.2	25.7	812.9	18.9
470	1911.3	44.5	1146.8	26.7	819.1	19.1	603.6	14.1

Using high NA and a short wavelength it should be possible to obtain track width measurements down to 14nm. The ratio used to generate this table is limited not by the approach but by size of the smallest feature on the samples available to us. There is a strong possibility that this technique could be used to measure track widths as small as 10nm.

Table 44 summarises the key work tasks and achievements of this project.

Table 44 - Work tasks and achievements

	Optical Systems	**ANN**
Aims	Have a number of systems for different sample types for obtaining profiles for ANNs	Robust networks for feature extraction
Initial Requirements	• Be able to measure phase and intensity samples. • Have high repeatability and stability and high SNR. • Mechanically sound systems.	• Extract track width for 100nm track • Extract width and separation for double tracks • Other parameters
Achievements	• Improve mechanical stability and optimise setup for hologram system • Combined system for DSOM and scanning Nomarski. • Investigate sources of noise and vibration • Measure variety of samples with system to	• Train on simulated data and experimental data for single and double tracks • Test robustness of networks • Investigate effect of node, input points, target errors • Investigate other parameters for extraction • Effect of noise on networks

	determines suitability	
Capabilities	• Good repeatability of all systems. • High SNRs • Relatively compact and mechanically robust designs for systems	• Successful networks for single tracks down to 60nm - 2nm errors • Successful network for double tracks 1-3 microns approx. 10nm errors • Working classifier for single or double tracks • Auto correction for target errors • Other parameter extraction possible • Networks relatively insensitive to noise. 1nm error approx SNR 70dB

This thesis has shown that the useful measurement range of optical systems can be extended beyond conventional limits to provide specific parameter measurements for several object types by utilising the power of artificial neural networks. Tracks as small as 60nm have been correctly measured with optical systems with 0.3 NA. By increasing the NA and reducing the wavelength tracks as small as 10nm should be measurable.

The future of the optical microscope is intertwined with the rise of computing power and the application of novel signal processing techniques. This will allow more and more information to be extracted from the data recorded by the optical system. As the years progress the optical microscope will depend more and more on computing systems and signal processing to provide the measurements demanded by industry. This thesis has shown the huge improvement to the measurement range of optical microscopes that is possible using the power of computers and signal processing.

9 Appendix 1 - Generalised Delta Rule and Back propagation

The generalised delta rule [64] was first used for training Adalines. The derivation of the weight update process is relatively straightforward and for batch mode training the training error after all training patterns have been presented once to the network (after one epoch or iteration) is defined as:

$$E = \frac{1}{2}\sum_{i=1}^{n_N}\left[d(i) - a_N(i)\right]^2 \qquad \text{Equation A1-1}$$

Where:

i the current training pattern presented

N = layers

w_L = weights between layers L = 1,2….N

d = desired outputs

a = output of the layer

The weight update rule is based on the gradient of the error with respect to the weights. So that by updating the weight value we can move in the opposite direction to the gradient and the overall error will decrease.

$$\Delta w_N(i,j) = -\alpha \frac{\partial E}{\partial w_N(i,j)} \qquad \text{Equation A1-2}$$

E is a function of network output a_N, which in turn is a function of the output of the previous layer.

$$a_N(i) = f\left(y_N(i)\right) \qquad \text{Equation A1-3}$$

where

$$y_N(i) = \sum_j w_N(i,j) a_{N-1}(j)$$ Equation A1-4

Equation (2) expanded by chain rule

$$\frac{\partial E}{\partial w_N(i,j)} = \frac{\partial E}{\partial a_N(i)} \frac{\partial a_N(i)}{\partial y_N(i)} \frac{\partial y_N(i)}{\partial w_N(i,j)}$$ Equation A1-5

$$\frac{\partial E}{\partial a_N(i)} = -[d(i) - a_N(i)]$$ Equation A1-6

from equation (3)

$$\frac{\partial a_N(i)}{\partial y_N(i)} = f'(y_N(i))$$ Equation A1-7

from equation (4)

$$\frac{\partial y_N(i)}{\partial w_N(i,j)} = a_{N-1}(j)$$ Equation A1-8

$$\Delta w_N(i,j) = \alpha[d(i) - a_N(i)] f'(y_N(i)) a_{N-1}(j)$$ Equation A1-9

define:

$$\delta_N(i) = \frac{\partial E}{\partial y_N(i)} = \frac{\partial E}{\partial a_N(i)} \frac{\partial a_N(i)}{\partial y_N(i)} = [d(i) - a_N(i)] f'(y_N(i))$$ Equation A1-10

Equation (9) becomes

$$\Delta w_N(i,j) = \alpha \delta_N(i) a_{N-1}(j)$$ Equation A1-11

This is the Generalised Delta Rule.

Training the output layer is a direct application of the delta rule, with the input replaced by the outputs from the previous layer. However it only works for the output layer due to the fact that training is dependant on the 'error' *d(i)-a(i)*

Training on the hidden layer is performed by back propagation. Back propagation is an extension of Generalised Delta Rule used to train intermediate layers. For intermediate layer *m*, weight updates are given by:

$$\Delta w_m(i,j) = -\alpha \frac{\partial E}{\partial w_m(i,j)}$$ Equation A1-12

Following same procedure for GDR:

$$\Delta w_m(i,j) = \alpha \left[\frac{-\partial E}{\partial a_m(i)} f'(y_m(i)) \right] a_{m-1}(j)$$ Equation A1-13

The problem now is that E is not a direct function of the output $a_m(i)$as was for the output layer. We have to use the chain rule again.

$$\frac{\partial E}{\partial a_m(i)} = \sum_k \frac{\partial E}{\partial y_{m+1}(k)} \frac{\partial y_{m+1}(k)}{\partial a_m(i)}$$ Equation A1-14

The first term in the summation is $\delta_{m+1}(k)$, the second term is just $w_{m+1}(k)$, substituting:

$$\Delta w_m(i,j) = -\alpha \left[\sum_k \delta_{m+1}(k) w_{m+1}(k,i) f'(y_m(i)) \right] a_{m-1}(j)$$ Equation A1-15

Equation (15) is similar to Equation (11) where

$$\delta_m(i) = \left[\sum_k \delta_{m+1}(k) w_{m+1}(k,i) f'(y_m(i)) \right]$$ Equation A1-16

Therefore

$$\Delta w_m(i,j) = \alpha \delta_m(i) a_{m-1}(j)$$ Equation A1-17

δ_m is the back propagated error of the network. Training an intermediate layer of a network is the same as training the output layer but instead of using the errors *d(i) – a(i)*, which are valid only for the output layer, we use another version of the error, which is the weighted sum of the errors from the following layer.

10 Appendix 2 - Converting Phase noise to photon noise

This was achieved by simulating the data acquisition of the hologram system. In this case a fringe pattern of 128x50 pixels is used. There were 8 pixels per fringe and the fringe contrast was 0.8. An example fringe pattern is shown in Figure 112

The value for the maximum number of photons per pixel was then varied and shot noise added to the interferogram by adding to each pixel the square root of its value multiplied by a random number. This is shown in equation A2-1

$$I_n = I + sqrt(I \times randn) \qquad \text{Equation A2-1}$$

The random number had a normal distribution with zero mean and standard deviation of one.

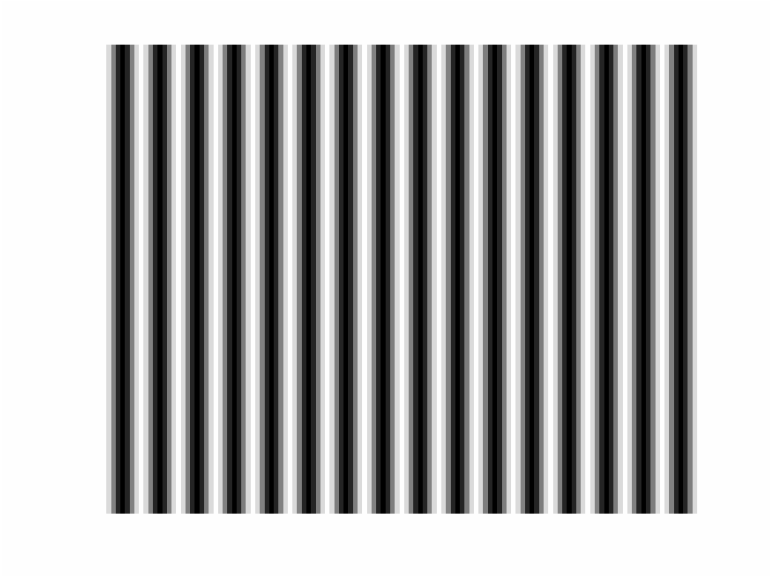


Figure 112 - Example fringe pattern

This produces a noisy interferogram as shown in Figure 113

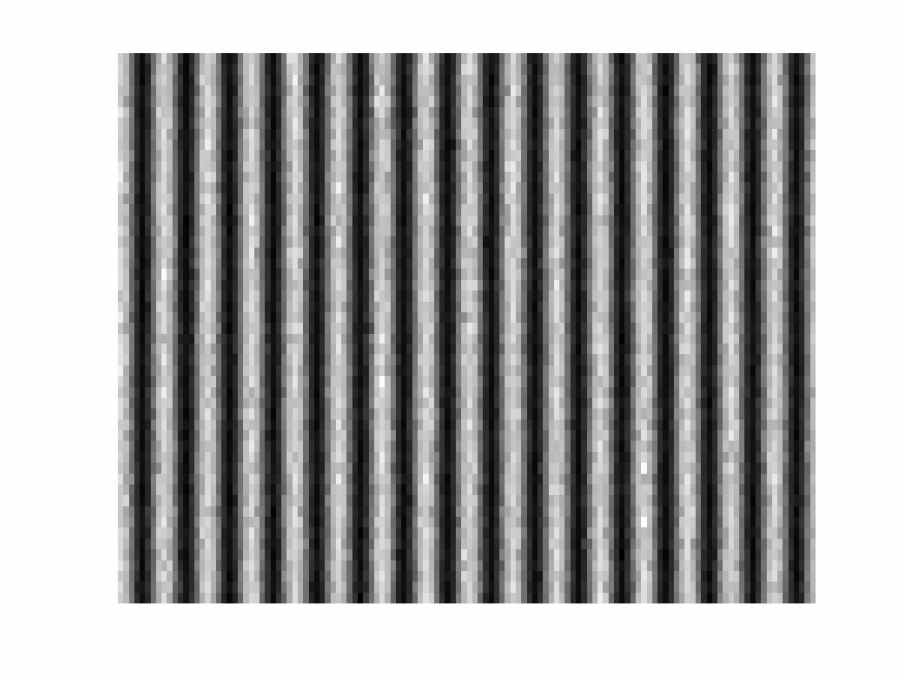


Figure 113 - Interferogram with shot noise

The fringes are then summed down to produce a line profile as shown in Figure 114.

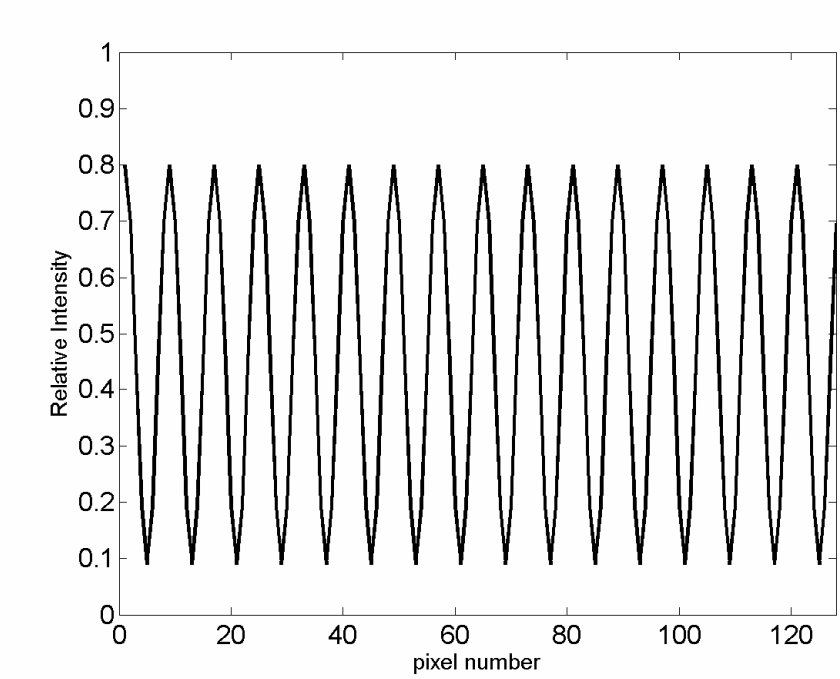


Figure 114 - averaged window

The next stage is to take the Fourier transform and store the phase and amplitude values at the point corresponding to the fringe frequency, as shown in Figure 115.

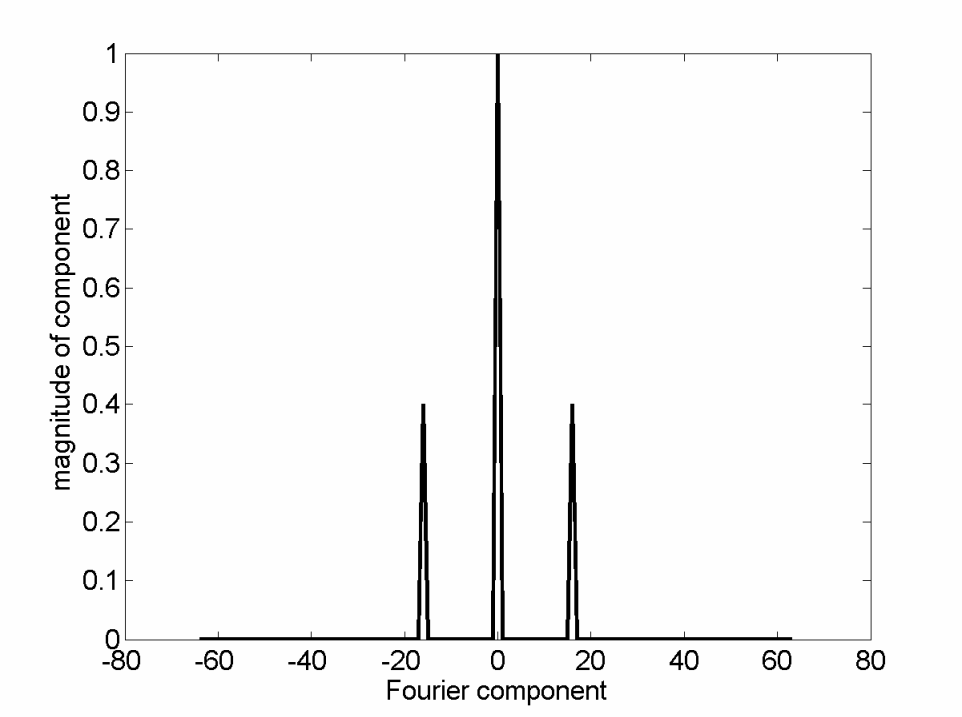


Figure 115 - FFT of averaged signal

This process is iterated 500 times to obtain a set of phase and amplitudes values. If there was no noise present the phase values would be identical but because of the shot noise the phases will vary. Taking the standard deviation gives a measure of the phase noise. The total number of photons used in the measurement can also be calculated by summing the photons in the nosy fringe pattern, allowing the signal to noise ratio to be calculated.

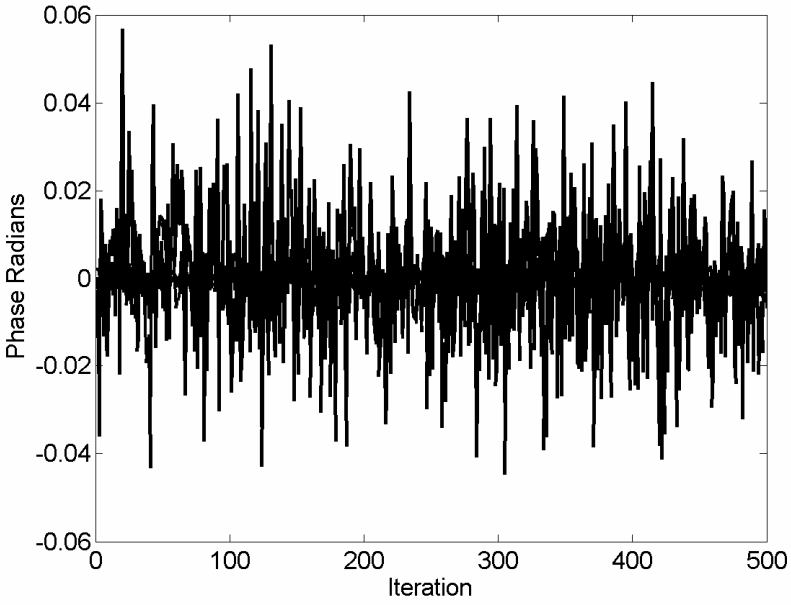


Figure 116 - phase values for 500 iterations

For example if the total number of photons in the interferogram was 1 million, after 500 iterations the mean of all of the amplitude values was 0.4001 and the standard deviation of the amplitude was 0.0005797, giving a SNR of 691. The corresponding phase noise due to the photon noise was 0.0018 radians.

The full table of values is given in chapter 4, table 3.

11 Appendix 3 – Effect of vibration for the DSOM system

This simulation was achieved by generating a 2D PSF and a shifted version of the same PSF. The PSF was calculated as a $Jinc^2$ function [2] and the size was chosen so that the full width half maximum occupied a similar number of pixels as for the practical case on the CCD camera. The PSF was scaled to have a maximum value of 50000 photons, which is approximately the pixel saturation level on the CCD.

Keeping the detection location fixed and shifting the PSF simulates vibrations. The change in the number of detected photons for a specific shift and detector size can be calculated. The SNR is then calculated as the mean value of photons detected for the original PSF divided by the difference in the shifted value and the original value.

There were two regions examined to simulate the practical experiment. One detector was placed offset from the maximum of the PSF by approximately one quarter of the optical spot size and ranged in size from 5x5 to 11x11 pixels. A larger window was used to simulate the effect of vibration on the detector used to monitor intensity fluctuations and this window was varied from 50x50 to 200x200 pixels.

---

[2] Where the Jinc function is define as: $jinc = \frac{J_1(kNAr)}{kNAr}$, where J_1 is the first order Bessel function of the first kind, $k = 2\pi/\lambda$, r is the radial coordinates of the image plane and NA is the numerical aperture of the system.

11.1 Right hand side window

The effect of vibration for the right hand detector is shown below. The top figure shows the SNR for the different window sizes for different amounts of shift. The shift values are given as a percentage of a pixel.

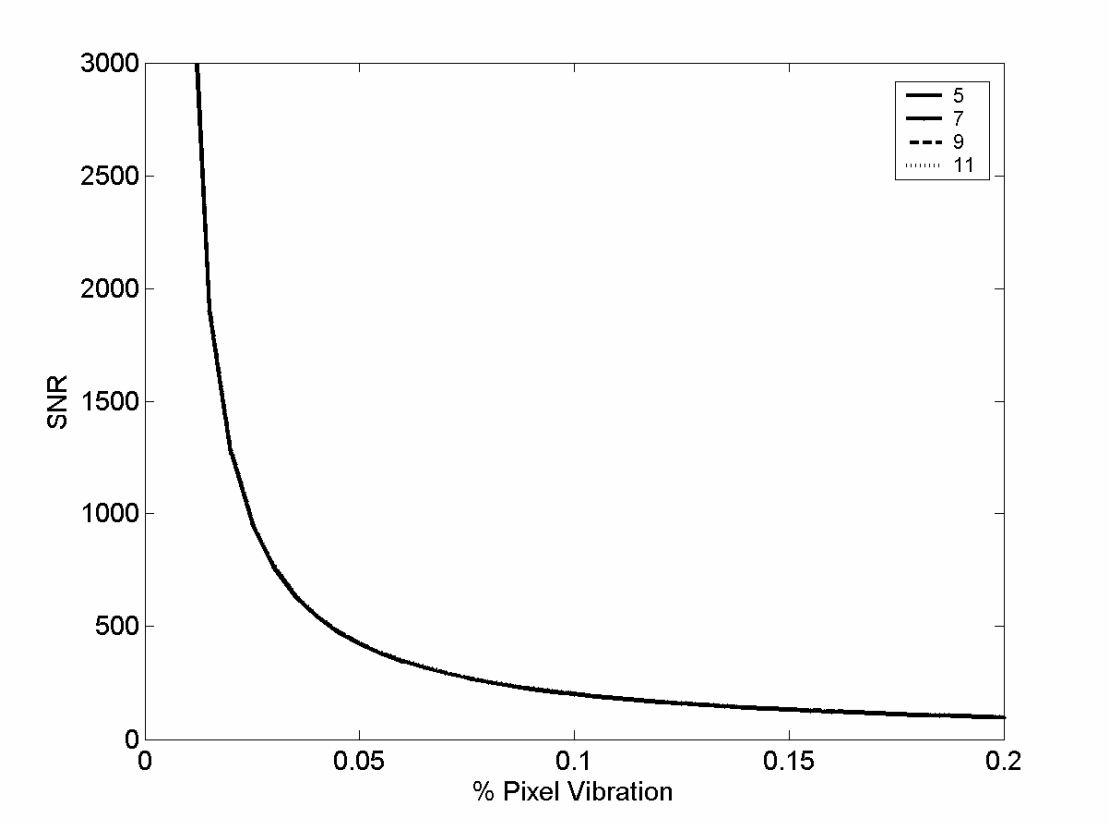


Figure 117 - SNR for right hand window

As can be seen from the graph above the SNR quickly drops off due to the vibration. Increasing the size of the window used has little effect on the SNR. It shows that the small detectors used for the differential intensity signals are very sensitive to vibration. A vibration level of 0.05% of a pixel gives a SNR of 1 in 500.

The vibration could come from several locations. The camera itself could be moving in this case the amount of movement corresponding to 0.05% of a pixel is 0.05%x11μm = 0.55um. (Where the camera pixel size was 11 μm)

If however there were vibration in any of the optics before the spot on the sample was magnified then the effect would be much greater. In this case the magnification is x444 this would mean a vibration of 1.2nm would lead to the same SNR as above.

11.2 Full Window Size

The effect of vibration for the intensity level monitoring detector window is shown below. The figure shows the SNR for the different window sizes for different amounts of shift. The shift values are given as a percentage of a pixel.

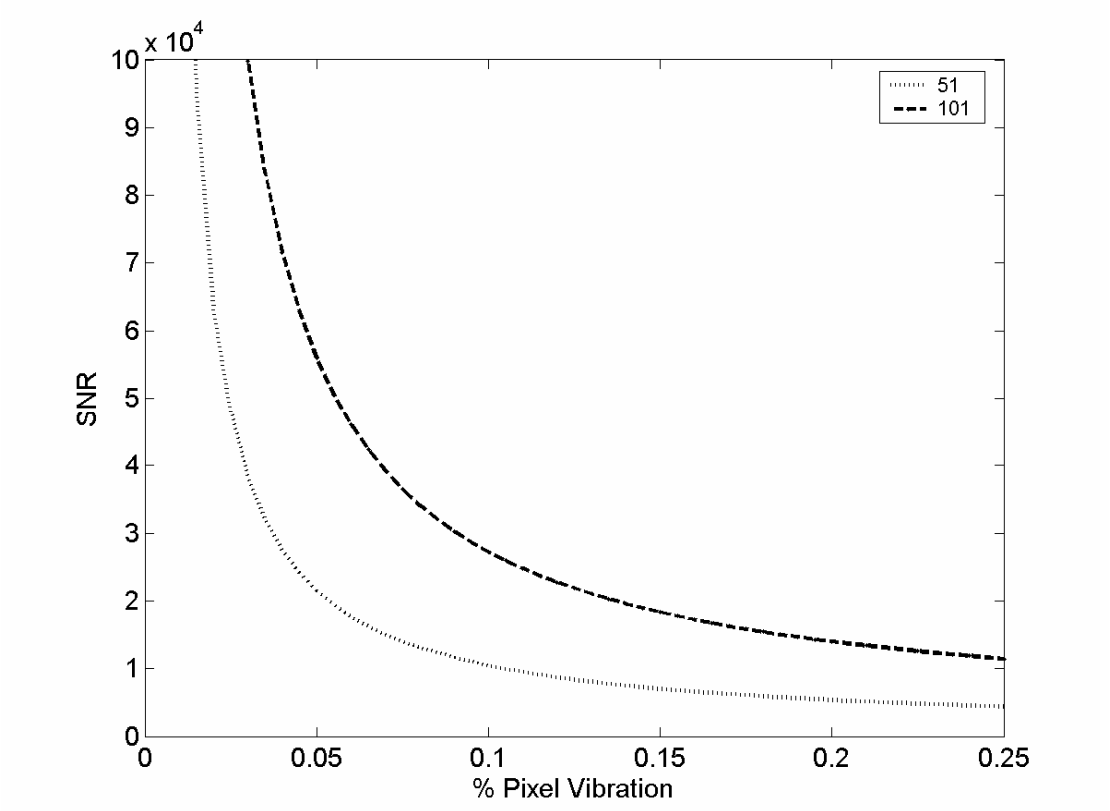


Figure 118 - SNR for large window

The graph is plotted for two smaller windows the 200x200 window has a considerably higher SNR and does not fit on this graph. Even using the 101x101 window size for a vibration of 0.1% pixel the SNR is 30000.

This is very similar to the graph for the right hand detector, but the size of the window has more of an effect. When the window is larger than 200x200 the SNR goes up very quickly as the window is quite a bit larger than the FWHM of PSF so only very small changes in side lobes are measured as PSF moves and so SNR is good.

For the smaller windows this is not the case. Practically though a smaller window has to be used as the camera pixels noise has an increasing effect as the window size is increased. This is because no additional signal is obtained by increasing the window size only pixel noise is added, thus reducing the SNR.

12 Appendix 4 – Derivation of system mathematics for linear input polarisation for Nomarski system

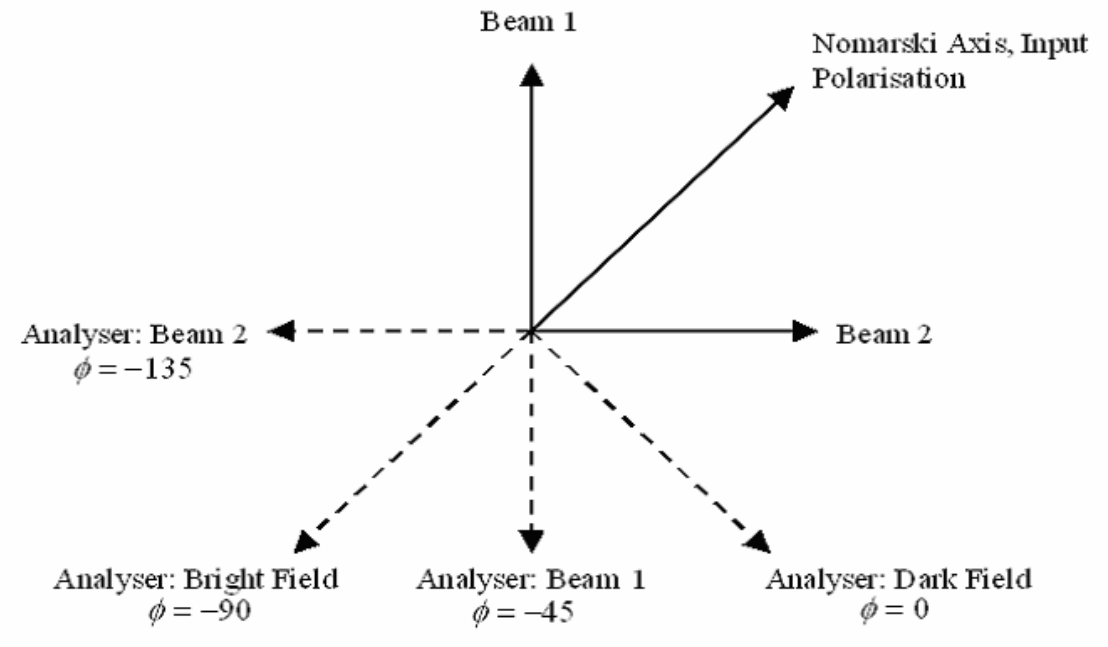


Figure 119 - Nomarski system alignment

The two beams coming from the Nomarski prism incident with linearly polarised light are:

$$s_1^{'} = E\exp(j\varphi)$$

$$s_2^{'} = E\exp(j\varphi)$$

E is the initial amplitude of the beams and is the same for both beam and is equal to 1 to simplify things. φ is the initial phase of the both beams.

The beams then interact with the object and return with the object information.

$$s_1 = E_1 \times \exp(j\vartheta_1 + \varphi)$$

$$s_2 = E_2 \times \exp(j\vartheta_2 + \varphi)$$

The beams are then resolved in the direction of the analyser; analyser angle measured 90 degrees from the Nomarski prism axis in this case.

$$s_{1a} = s_1 \cos(135 - \phi)$$

$$s_{2a} = s_2 \cos(45 - \phi)$$

$$i = s_{1a} + s_{2a}$$

$$I = i \times i^*$$

Take out some factors so can look at signal of interest

$$E_{1x} = E_1 \times \cos(135 - \phi)$$

$$E_{2x} = E_2 \times \cos(45 - \phi)$$

$$i = E_{1x} \exp(j\vartheta_1 + \varphi) + E_{2x} \exp(j\vartheta_2 + \varphi)$$

$$I = \left(E_{1x} \exp(j\vartheta_1 + \varphi) + E_{2x} \exp(j\vartheta_2 + \varphi)\right) \times \left(E_{1x} \exp(j\vartheta_1 + \varphi) + E_{2x} \exp(j\vartheta_2 + \varphi)\right)$$
$$I = E_{1x}^2 + E_{2x}^2 + 2E_{1x}E_{2x} \cos(\vartheta_1 - \vartheta_2)$$

Now include E_{1x} and E_{2x}

$$E_{1x}^2 = \left[E_1 \times \cos(135 - \phi)\right]^2$$

$$E_{2x}^2 = \left[E_2 \times \cos(45 - \phi)\right]^2$$

$$2E_{1x}E_{2x} = 2 \times \left[E_1 \times \cos(135 - \phi)\right] \times \left[E_2 \times \cos(45 - \phi)\right]$$

$$2E_{1x}E_{2x} = 2 \times \left[E_1\right] \times \left[E_2\right] \cos(135 - \phi) \cos(45 - \phi)$$

combining cos (135) and cos (45) bits from above equation

A = ϕ B = 135 C = 45

$$\cos(135 - \phi)\cos(45 - \phi) = \frac{1}{2}\left[\exp(j(B - A)) + \exp(-j(B - A))\right] \times \frac{1}{2}\left[\exp(j(B - A)) + \exp(-j(B - A))\right]$$
$$= \frac{1}{4}\left[\exp(j(B - A + C - A)) + \exp(-j(B - A - C + A)) + \exp(j(B - A - C + A)) + \exp(-j(B - A - C + A))\right]$$

$$= \frac{1}{4}\left[2\cos(B + C - 2A) + 2\cos(B - C)\right]$$

$$= \frac{1}{4}\left[2\cos(135 + 45 - 2\phi) + 2\cos(135 - 45)\right]$$

$$= \frac{1}{4}\left[2\cos(180 - 2\phi) + 2\cos(90)\right]$$

$$= -\frac{1}{2}\left[\cos(-2\phi)\right]$$

The squared terms become:

$$E_{1x}^2 = \left[E_1 \times \cos(135 - \phi)\right]^2$$

$$E_{1x}^2 = {E_1}^2 \times \cos^2(135 - \phi)$$

$$E_{1x}^2 = {E_1}^2 \times \frac{1}{2}(1 + \cos(135 - \phi))$$

$$E_{2x}^2 = \left[E_2 \times \cos(45 - \phi)\right]^2$$

$$E_{2x}^2 = {E_2}^2 \times \cos^2(45 - \phi)$$

$$E_{2x}^2 = {E_2}^2 \times \frac{1}{2}(1 + \cos(90 - \phi))$$

Therefore combining the above equations the complete system equation is:

$$I = {E_1}^2 \times \frac{1}{2}(1 + \cos(90 - \phi)) + {E_2}^2 \times \frac{1}{2}(1 + \cos(270 - \phi)) - E_2 E_1 \left[\cos(-2\phi)\right]\cos(\vartheta_1 - \vartheta_2)$$

Depending on the conditions set the microscope will operate in bright field and dark field differential or as a normal scanning microscope with either beam 1 or 2.

for ϕ =0

$$I = \frac{1}{2}E_1^2 + \frac{1}{2}E_2^2 - E_1 E_2 \cos(\vartheta_1 - \vartheta_2)$$

if $\vartheta_1 = \vartheta_2$ and $E_1 = E_2$ then I = 0 Hence dark field system

for ϕ = 90

$$I = \frac{1}{2}E_1^2 + \frac{1}{2}E_2^2 + E_1E_2\cos(\vartheta_1 - \vartheta_2)$$

if $\vartheta_1 = \vartheta_2$ and $E_1 = E_2$ then $I = 2E_1^2$ Hence bright field system

for ϕ = +45

$$I = 0 + E_2^2 + 0$$

for ϕ = -45

$$I = E_1^2 + 0 + 0$$

13 References

[1] Forest preserve district council, *history of microscope*, 10/07/2006, www.newton.dep.anl.gov/nat61tn/500-599/nb506.htm,

[2] About, *timeline – history of microscopes*, 03/08/2005 http://inventors.about.com/od/mstartinventions/a/microscopes.htm

[3] H Smith, *history*, 10/08/2006, http://bama.ua.edu/~hsmithso/class/bsc_656/websites/history.html

[4] Molecular expressions, science optics and you, "*pioneers in optics*", 8/15/2006, www.micro.magnet.fsu.edu/optics/timeline/people/lister.html

[5] E. Hecht, *Optics, 3th ed.*, Addison Wesley Longman, New York, 2002.Chapter 13, p603.

[6] E. Hecht, *Optics, 3th ed.*, Addison Wesley Longman, New York, 2002.Chapter 10, p463.

[7] Mic-UK, *the history of the microscope*, 08/08/2005 www.microscopy-uk.org.uk/intro/histo.html,

[8] Wikipedia, *phase contrast microscope*, 14/08/2006, http://en.wikipedia.org/wiki/microscopy

[9] Fracon, *Progress in microscopy*, Pergamon Press, London, 1961, page 149

[10] J. W. Goodman, *Introduction to Fourier optics, 2nd ed.*, McGraw-Hill, New York, 1996.

[11] E. Hecht, *Optics, 4th ed.*, Addison Wesley Longman, New York, 2002. Chapter 10, p461.

[12] E. Hecht, *Optics, 4th ed.*, Addison Wesley Longman, New York, 2002.

[13] R. Shannon J Wyant, *Applied Optics and optical engineering*, Academic Press, New York, 1983, p140

[14] National Conference of Standard, *papyrus story*, 08/11/2005 Laboratories,http://www.ncsli.org/misc/cubit.cfm,

[15] ISO, *how it all started* 08/08/2005, http://www.iso.org/en/aboutiso/introduction/index.html#four

[16] ISO, *What international standardisation means* 08/08/2005, http://www.iso.org/en/aboutiso/introduction/index.html

[17] J. L. Harris, *Diffraction and resolving power*, J. Opt. Soc. Am. **54** (7) (1964) 931-936.

[18] C. W. Barnes, *Object restoration in a diffraction limited imaging system*, J. Opt. Soc. Am. **56** (5) (1966) 575-578.

[19] R. W. Gerchberg, *Super-resolution through error energy reduction*, Opt. Acta **21** (9) (1974) 709-720.

[20] S.J. Howard, *Method for continuing Fourier spectra given by the fast Fourier transform*. J. Opt. Soc. Am. **71**(1) (1981) 95-98

[21] S M Rowe, *Light Distribution in the Defocused Image of a Coherently Illuminated Edge*., J. Opt. Soc. Am., **59**, (1969), 711-714.

[22] M E Barnett and N P Turner, *Symmetry in the Coherent Spread Function for a Semi-transparent edge*, Optik, **75**(2), (1987),85-87

[23] Nunn J, Mirande W, Jacobsen H and Talene N 1997 *Challenges in the calibration of a photomask linewidth standard developed for the European Commission,* VDE-VDI Conf. Proc.: Mask Technology for Integrated Circuits and Micro-components pp 53–68

[24] Olympus, *LEXT*, 16/08/06, http://www.olympus-europa.com/medical/26_LEXT.htm

[25] Zygo, *Newview 6000 optical profiler*, 16/08/06, http://www.zygo.com/?/products/nv6000/.

[26] C.W.See, M. Vaez Iravani, and H.K. Wickramasinghe, *Scanning Differential Phase Contrast Optical Microscope: Application to Surface Studies*, Appl. Opt., Vol. 24 (15), 2373 – 2379, 1985.

[27] G.E.Sommargren, and B.J.Thompson, *Linear Phase Microscopy*, Appl. Opt., Vol. 12 (9), 2130 – 2138, 1973.

[28] T. Sawatari, *Optical Heterodyne Scanning Microscope*, Appl. Opt., Vol. 12 (11), 2768 – 2775, 1974

[29] R.L. Jungerman, P.C.D. Hobbs, and G.S.Kino, *Phase Sensititive Scanning Optical Microscope*, Appl. Phys. Lett., Vol. 45 (8), 846 – 848, 1984.

[30] C.C. Huang, *Optical Heterodyne Profilometry*, Opt. Eng., Vol. 23 (4), 365 – 370, 1984.

[31] G. Makosch, and B. Drollinger, *Surface Profile Measurement with a Scanning Differential ac Interferometer*, Appl. Opt., Vol. 23 (24), 4544 – 4553, 1984.

[32] C See et al, *Scanning differential optical profilometer for simultaneous measurement of amplitude and phase variation*, Appl. Phys. Lett, 1988, Vol 53, No 1.

[33] R. Pike and S. H. Jiang, *Ultrahigh-resolution optical imaging of colloidal particles*, J. Phys.: Condens. Matter **14** (2002) 7749-7756

[34] J Stewart et al, *Experimental demonstration of polarisation-assisted transverse and axial optical superresolution*, Optics Communications, 2004,vol 241, pp315-319

[35]S. Sherif and P Torok, *Pupil plane masks for super-resolution in high-numerical-aperture focusing journal of modern optics*, 10 September 2004 vol. 51, no. 13, 2007–2019

[36] G Toraldo di Francia, *Resolving power and information*, J. Opt. Soc. Am., 1955, Vol. 45, No. 7,

[37] E A Guillemin, *the mathematics of circuit analysis*, J Wiley, New York, 1951, p288,290

[38] E Whittaker G Watson, *A Course of Modern Analysis*, Cambridge university press, Cambridge,1973.

[39] C.K. Rushforth, and R. W. Harris, *Restoration, Resolution, and Noise*, JOSA, Vol 58 (4), 539 – 545, 1968.

[40] W. Lukosz, *Optical system with resolving powers exceeding the classical limit*, J. Opt. Soc. Am. **56** (11) (1966) 1463-1472

[41] W. Lukosz, *Optical Systems with Resolving Powers Exceeding the Classical Limit. II*, JOSA, Vol 57 (7), 932 – 941, 1967.

[42] G. Toraldo di Francia, *Degrees of Freedom of an Image*, JOSA, Vol 59 (7), 799 – 804, 1969.

[43] I Cox and C Sheppard, *information capacity and resolution in an optical system*, J. Opt. Soc. Am., 1986, Vol 3, No 8pp1152-1158

[44] C.J.R. Sheppard, and K.G.Larkin, *Information Capacity and Resolution in Three-Dimensional Imaging*, Optik, Vol 113 (12), 548 – 550, 2003.

[45] Minami et al, *Superresolution of Fourier transform spectra by autoregressive model fitting with singular value decomposition*, Applied Optics,1985, Vol 24, No 2.

[46] J Walker, *Optical imaging with resolution exceeding the Rayleigh criterion*, Optica Acta, 1983, Vol 30, No. 9, pp1197-1202.

[47] B Widrow, *30 years of adaptive neural networks: Perceptron, madaline and backpropagation*, Proceeding of IEEE, 1990, Vol 78, No9

[48] Nelson, *a practical guide to neural nets,* Addison Wesley publishing, Wokingham England,1990

[49] Gurney, *an introduction to neural networks*, UCL press limited, London, 1997.

[50] M Chester, *Neural Networks a tutorial*, Prentice hall inc, New jersey 1993

[51] P Lisboa, *Neural Networks – Current Applications*, Chapman & Hall, London,1992

[52] *How many kinds of ANN exist?*, 25/02/2005, www.faqs.org/faqs/ai-faq/neural-nets/part1/section-10.html

[53] Schalkoff, *Artificial neural networks*, McGraw-Hill,London,1997

[54] Simon Haykin, *Neural Networks: A Comprehensive Foundation*, Second Edition, Prentice Hall New Jersey 1999

[55] M Hayes, *Statistical digital signal processing and modelling*, John Wiley and sons inc, New York, 1996 p51

[56] K. Levenberg, *A method for the resolution of certain problems in least square*, Quart. Appl. Math. **2** (1944) 164-168.

[57] D. Marquendt, *An algorithm for least squares estimation of non-linear parameters*, SIAM J. Appl. Math **11** (1963) 431-441.

[58] Bishop, *Neural networks for pattern recognition*, Oxford University Press, Oxford, 1995, p130

[59] Sarle, *how many hidden layers should I use?*, 30/03/2005 , ftp://ftp.sas.com/pub/neural/FAQ3.html.

[60] N Sawyer et al, *Ultrastable absolute phase common path optical profiler based on computer generated holography*, Applied Optics, 1998, Vol. 37, No. 28.

[61] M Suddendorf et al, *Single probe beam differential amplitude and phase scanning interferometer*, Applied Optics, 1997, Vol. 36, No. 25,

[62] J Dayhoff, *Neural network architectures – an introduction*, Van Nostrand Reinhold, New York, 1990.

[63] S Morgan et al, *Interferometric optical microscopy of subwavelength grooves*, Optics Communications,2001, vol 187, pp29-38.

[64] K. Swingler, *Applying neural networks: a practical guide*, Morgan Kaufmann Publishers, San Francisco, 1996.