THE DEVELOPMENT OF QUASI-OPTICAL TECHNIQUES FOR LONG WAVELENGTH IMAGING

Presented by
Robert K. May, B.Sc. (Hons.)

A thesis submitted for the degree of
Doctor of Philosophy

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

Department of Experimental Physics
NUI Maynooth,
Maynooth, Co. Kildare,
Ireland

October 2008

Head of Department and Research Supervisor
Prof. J. Anthony Murphy

Acknowledgements

First and foremost I wish to thank Prof. J. Anthony Murphy for his supervision and guidance over the last six years. In particular I would like to thank him for sharing with me the great enthusiasm he has for long wavelength optics. On a more practical side I would also like to thank him for ensuring that I didn't stray too far from the goal of completing this thesis.

I would like to thank all the staff of the Department of Experimental Physics, who all contributed in some way to my academic development during both my undergraduate and postgraduate studies at NUI Maynooth. In particular I would like to acknowledge the invaluable assistance provided by members of the THz Optics Group in no particular order Dr. Marcin Gradziel, Dr. Creidhe O'Sullivan, Dr. Neil Trappe and the late Dr. Bill Lanigan.

I must also thank the technical and administrative staff of the Department of Experimental Physics: Mr. David Watson for his skilled work in making the various optical components, phase gratings and other bits and pieces needed to conduct the experimental work; Mr. John Kelly for coming to the rescue whenever the computers decided to play havoc; many thanks as well to the administrative staff: Ms. Grainne Roche and Mr. Derek Gleeson.

My fellow postgraduates over the years, especially those in the THz group who were able to offer their guidance and with whom I enjoyed stimulating conversations concerning various difficulties and challenges encountered during the research over pints in the Roost. Thanks for your friendship over the years. You know who you are.

I wish to acknowledge the financial support provided by Science Foundation Ireland to undertake this research.

Finally, I wish to thank my parents, Breffni and Anita and my sisters Suzanne and Louise for the love and support they have given me throughout my life.

Abstract

This thesis concerns the development of quasi-optical techniques for long-wavelength imaging, which was conducted through a combination of experimentation and computer simulation. This work was conducted as part of a SFI-funded research program undertaken by the THz Optics group of the Department of Experimental Physics at NUI Maynooth, the aim of which was to extend existing quasi-optical techniques through experimental measurements and the development of simulation tools necessary for efficient design and analysis of long-wavelength optical systems.

Description of the upgrading of the 100 GHz test measurement facilities at NUIM and the results obtained from transmission- and reflection-mode active imaging experiments are presented. Numerical simulation of quasi-optical components and systems using scalar wave diffraction techniques was performed. In particular, Gaussian Beam Mode Analysis (GBMA) was applied to the design and analysis of discrete and continuous phase modulating optical multiplexers (phase gratings) for use at 100 GHz. The use of GBMA for iterative phase retrieval was investigated for application to phase grating design. Several phase unwrapping techniques were also investigated in order to simplify manufacture of phase gratings with difficult-to-fabricate profiles. Results of experimental measurements from a number of test gratings are presented and verified using the Maynooth Optical Design and Analysis Laboratory (MODAL) software package. Further improvements to phase grating design are also presented.

Chapter 1.

Introduction

1.1 The drive for the development of THz technology

Terahertz (THz) radiation is defined as the part of the electromagnetic spectrum between 0.1 and 10 THz and technology for this waveband is currently undergoing enormous development [1.1]. Recent identification of a whole range of applications including increased bandwidth communication links, secure networks (possibly because of severe atmospheric attenuation), remote sensing, next generation THz space telescopes, medical imaging (cancers [1.2, 1.3, 1.4]), monitoring complex chemical substances with THz spectroscopy [1.5, 1.6, 1.7, 1.8] (with relevance to contraband detection and medical physics [1.9, 1.10], food sciences and biological sciences [1.11, 1.12]), pharmaceutical process control [1.13], security-screening, weapon systems, etc. have all been responsible for driving advances in THz technology. Substantial basic research and application development, including investigation into optical techniques needed to radiate, guide and collect radiation in a controlled manner, is required if THz technology is to achieve its ultimate potential.

1.2 Background to work described in this thesis

The work described in this thesis was undertaken as part of an SFI-funded principal investigator grant research programme entitled "*The development of an integrated quasi-optical and electromagnetic numerical simulator for the computer aided design (CAD) and analysis of novel terahertz systems*". The objectives of the research programme were to develop quasi-optical techniques, components and advanced software simulation tools necessary to advance THz technology.

The primary goals of the programme were to

1) develop efficient CAD software tools for the unique propagation regime of quasi-optical systems in the THz waveband and
2) investigate components and systems for long-wavelength array imaging

The first aspect involved integrating both methods of theoretical analysis and efficient computational tools into a practical CAD platform to achieve a powerful and efficient environment for optical design and analysis in the THz waveband. The second, with which the author of this thesis was primarily involved, concerned the experimental investigation of novel optical systems and components necessary for the development of THz array imaging systems.

1.2.1 Development of efficient CAD software tools for THz

Although the "THz gap" has been steadily diminishing over the past few years with the development of useful sources and detectors of terahertz radiation, there still exists a gap where reliable, basic optical design and analysis simulation tools in the far-infrared and terahertz wavebands are concerned.

The power that is radiated, guided and collected by quasi-optical components obeys the laws of long wavelength optics in ways different from visible light and which are dominated by diffraction effects. Many basic THz sources and quasi-optical components (corrugated horns, lens antennas, phase gratings, polarising interferometers, off-axis reflectors, etc.) require sophisticated analytical approaches to understand and reliably predict their performance [1.15,1.16]. While good commercial design-analysis and performance-verification tools do exist for optical and millimetre-wave systems (e.g. GRASP[1], CODE V[2], ASAP[3], Zemax[4], etc.) these were never intended for THz applications and so frequently fall short in their capability to accurately predict beam behaviour, propagation and loss in the sub-millimetre region. With no commercially available software that is specifically designed for THz wavelengths many basic THz components cannot be handled in these standard optics packages. To fill the gap that exists many institutes and companies develop their own software tools, but often these tend to be very specific and do not always have a proven track record of achievable accuracy, sensitivity and dynamic range.

The main aim of the research programme undertaken by the THz Optics Group at Maynooth therefore was to develop an optical design-analysis simulator that could cover the broad range of applications of THz quasi-optics. The end result of this aspect of the research programme was the software package called Maynooth Optical Design and Analysis Laboratory (MODAL), a brief description of which is provided in Chapter 2. The author of this thesis was involved with verifications of MODAL for a number of simple specific test cases, the results of which were then compared to results produced using benchmark software (GRASP 8 – a cumbersome, computationally inefficient tool for analysis of THz systems, but which yields accurate results).

---

[1] GRASP: general reflector and antenna farm analysis software from TICRA (www.ticra.com)
[2] CODE V: optical design and analysis software by Optical Research Associates (www.opticalres.com)
[3] ASAP: optical engineering software from Breault Research Corporation, Inc. (www.breault.com)
[4] Zemax: optical design package from Zemax Development Corporation (www.zemax.com)

1.2.2 Quasi-Optical Imaging Systems and Components

Many of the principles and techniques of THz optics are similar to those employed in the millimetre waveband but must be extended to higher frequencies. This implies the development of a whole range of novel optical components and systems. Investigations of novel optical components at Maynooth have concentrated on shaped lens antennas, axicons and multiplexing phase gratings. Lens antennas are required for coupling to free-space because, as devices that are used at lower frequencies (corrugated horn antennas) are pushed towards higher frequencies they become more difficult to realise. Lens antennas were investigated at NUIM by Lavelle [1.17]. An axicon [1.18, 1.19] acts as a kind of lens for an incident propagating beam but produces a focal line rather than a focal spot thus giving significant depth of field, larger than is possible with a lens for the same spot size. With axicons it is even possible to produce pencil beams with radii of the order of a wavelength, promising useful resolution properties.

Currently microwave and millimetre wave systems tend to be single pixel, however array imaging systems will become common in THz applications and multiplexing phase gratings are important for their development [1.20, 1.21]. Sensitive array receivers will ideally rely on heterodyne techniques. Quasi-optical issues that need to be addressed to facilitate the development of feasible large format heterodyne and bolometric array imaging systems include quasi-optically coupling local oscillator schemes for large heterodyne arrays. For sparse arrays this means producing multiple images of an input local oscillator beam so that these can be fed to a detector array through some coupling device (at its simplest a beam-splitter). An elegant and efficient solution to this problem by optical means is to use diffractive phase gratings to produce a set of focused beams which match the spatial distribution of the detector array feeds.

Much work on phase grating development at optical wavelengths should be applicable to the THz waveband. Past work by researchers from the THz Optics group at NUIM has been on the development of simple phase gratings at longer wavelengths for small arrays [1.14]. The development (simulation, design, manufacture and testing) of phase gratings for long wavelength array imaging, which was experimentally investigated in the dedicated THz test laboratory at NUIM, was a major component in the research undertaken by the author of this thesis. In this thesis we begin by investigating the performance and practical limitations of Dammann gratings, which are ideal if a regular square or rectangular array of beams is required. We extend previous

work to include the real limitations introduced in particular by off-axis reflection optics which give rise to aberrations, as well as investigating machine tolerance effects. We then go on to investigate novel Fourier gratings, which provide the best basis for feeding more general sparse detector arrays.

To achieve the goals of concept verification in the development of novel quasi-optical components for array systems and generally investigate the accuracy of quasi-optical models the existing experimental test facilities at NUIM were extended to develop a quasi-optical test laboratory to allow for sensitive near-field and "far-field" measurements at THz wavelengths. The results of which are seen in later chapters, where beam pattern measurements from previously tested phase gratings were obtained with much higher sensitivity than was possible with the pre-existing facilities.

MODAL was used by the author for verification of experimental measurements of phase gratings whose operation is well understood but also for testing novel compact phase grating designs and to determine the best test arrangements to use for experimental testing of these designs. An important aspect for the development of the MODAL software package is verification and the excellent agreement that was found to exist between experimental measurements of quasi-optical components and those predicted by MODAL proved very useful - at least qualitatively.

Besides the development of phase gratings, the other work undertaken by the author (in conjunction with Mr. Ian McCauley and Ms. Leanne Young) was to investigate the use of various transmission and reflection systems for imaging of biological samples, as a means of probing the bio-medical potential of long wavelength terahertz radiation. This is another important step towards the development of array imaging systems, since it is not yet clear what type of geometry will yield the most useful and meaningful images. This is especially true in terms of imaging of biological tissues because of the high absorption by water of THz radiation. To this end various near-field and Fourier-optics type transmission- and reflection-mode imaging systems were investigated using the newly upgraded single-pixel 100 GHz test facility at NUIM. A potentially useful application of terahertz imaging proved to be in wound analysis through layered dressings using a reflection-mode system. Although interference due to standing waves was an issue, a methodology to lessen its impact on imaging results was adopted. It is anticipated that the results obtained from this imaging work will inform the direction that future developments on array imaging at Maynooth should take.

1.3 Summary of Thesis Contents

The long-wavelength components and systems dealt with require appropriate analysis tools. Although a full vector, physical optics (PO) approaches is an accurate means of analysis it often proves computationally intensive. If the field from an optical components propagates in a paraxial manner then scalar approximations are possible. The components considered in this thesis do not contain sub-wavelength features so scalar wave diffraction techniques were used in all instances. The three techniques used for simulations in this thesis were Gaussian Beam Mode Analysis, Fresnel integrals and Fourier transforms. The prevalence of the last two in optical analysis meant that only brief details of implementations of these two methods are provided in Appendix A. *Chapter 2* is devoted to Gaussian Beam Mode Analysis which proves to be an efficient means of analysis for long-wavelength systems because, typically, only a small number of modes is required for accurate representation of the system.

Chapter 3 describes the upgrading of the experimental test facility at NUIM including descriptions of the single-pixel scanning system that was constructed and the design, fabrication and testing of a suite of off-axis reflectors. The rest of Chapter 3 describes the various transmission- and reflection-mode imaging experiments and the results obtained from experiments made on biological and non-biological samples.

Chapters 4 and *5* are concerned with the design, fabrication and testing of phase gratings for sub-mm and THz wavelengths for which they will be of vital importance in large multi-pixel array imaging systems. *Chapter 4* provides an introduction to the concept of the diffractive phase element (DPE), which is characterised by perfect transparency and thus by optimal diffraction efficiencies. Two of the most important applications of DPE's are beam-splitting and beam-shaping. The latter is of great importance in various laser applications such as material processing and pattern projection and is usually interpreted as the transformation of a beam from one shape to another, e.g. Gaussian to top-hat. This thesis concentrates on phase gratings that have multiplexing, or beam-splitting functions. Chapter 4 concentrates on Dammann gratings (binary-level phase gratings). The design, manufacture and operation (including bandwidth characteristics, mechanical tolerances) of these devices are examined through computational simulation (in terms of both Gaussian beam modes and Fourier analysis) and experimental verification.

Chapter 5 examines more efficient types of multiplexing phase gratings, including the multi-level phase grating and phase-only variations of the kinoform: the Fourier phase grating. Fourier gratings have smooth surface profiles and so are more easily fabricated for long-wavelength applications than their discrete-level counterpart – the opposite of the situation at visible wavelengths, for which fabrication techniques favours digitised profiles. Efficient numerical methods for the design of Fourier gratings are discussed and a novel implementation of one described in terms of Gaussian beam modes is used to find the solution to a particular phase grating problem to produce a sparse beam array. Phase unwrapping is then used to produce a smoother equivalent grating design that is easier to manufacture. The design, fabrication and testing of two example phase gratings is also presented. Numerical simulations in MODAL of the sparse beam-array grating was then used to explain how the grating actually produced its output and was also used to redesign the grating to find an alternative solution that would be able to operate in a system including non-ideal optics.

Finally *Chapter 6* concludes with a brief summary of the thesis including discussion of possible future developments that might follow on from the work described here.

Chapter 2.

Scalar wave diffraction with Gaussian Beam Mode Analysis

2.1 Introduction

This chapter describes Gaussian Beam Mode Analysis, one of the propagation techniques used to perform the numerical simulations of beam propagation through quasi-optical systems described in subsequent chapters. The appropriate propagation regime necessary to accurately model a particular optical system is determined by the size of structures (in the optical components) that are encountered by a propagating beam, relative to the wavelength of the electromagnetic radiation.

At visible wavelengths optical elements have dimensions of hundreds or thousands of wavelengths. At these scales the main approach used to calculate beam patterns is referred to as geometrical optics, which involves treating the beam as a bundle of light rays, the straight paths of which are traced through the system from one component to the next. When using geometrical techniques one is concerned only with the intensity of the rays.

At longer wavelengths, such as in the millimetre and sub millimetre range, optical elements tend to be much smaller with dimensions on the scale of several tens of wavelengths. At these scales diffraction effects tend to dominate and geometrical techniques are no longer sufficient to produce accurate results. Instead the expanding wave nature of the beam must be taken into account and one must resort to scalar wave diffraction techniques that are based on simplifications of scalar forms of Maxwell's equations. In general, although the electromagnetic wave is a vector field, if a field propagates in a paraxial manner, scalar approximations are possible. Scalar diffraction theory provides a set of simple equations that govern the propagation of light between two planes. In this regime the propagating beam is treated as a complex-valued wavefront with real and imaginary components which mean that as well as having an intensity profile the beam also has a phase distribution associated with it. The intensity of a complex-valued wavefront, E is proportional to the squared magnitude $|E|^2$, while the phase front is given by the argument, $\mathrm{Arg}\{E\}$.

If the component-to-wavelength ratio becomes even smaller, such that the propagating wavefront encounters components with features at or below a wavelength, the approximations inherent in the scalar formulation become invalid and a more rigorous modelling technique that retains the vector forms of Maxwell's equations is required. A rigorous vector physical optics (PO) approach is required to account for the electromagnetic coupling effects along the boundary of a diffracting profile, i.e. to

account for the interaction between the incident optical field and the surface of the optical component. When working at this scale the polarisation property of beams becomes apparent and a PO model can be used to characterise the polarisation properties of a quasi-optical system. The solution to electromagnetic boundary value problems requires significant computational resources to calculate the electric and magnetic fields induced on the surfaces of optical components. It is therefore usual to limit the application of PO to regions where the effects of electromagnetic coupling are significant. Once the interaction along a boundary has been properly accounted for, the resulting field can be used as a secondary source, the field values of which can then be propagated to a subsequent plane using more efficient scalar wave analysis.

For THz systems the question of whether to use a vector or scalar solution depends on whether the field is paraxial or wide-angle in nature. The types of structures of interest in this thesis (particularly phase gratings) have features with sizes that are at least several times the wavelength so scalar diffraction techniques are appropriate since the wavefronts produced by these components are confined to a relatively narrow angular spread ($\leq 30°$). While initial analysis of phase gratings was undertaken using code written by the author, MODAL was later used to perform verification of experimental results. Three different scalar diffraction techniques were used in this thesis: Fresnel integrals, Fourier transforms and Gaussian beam mode analysis (GBMA). Propagating a field to the next optical component using Fresnel integrals involves calculating diffraction integrals for each observation point and so becomes computationally intensive, especially in two dimensions. Fourier transform theory is an efficient means of propagating through ideal optical systems, especially between two planes one of which is in the Fourier plane of the other. GBMA is an alternative scalar wave diffraction method that is particularly suited to long-wavelength quasi-collimated systems. A modal analysis involves decomposing the field into a set of Gaussian modes and propagating them individually – a straightforward process that consists of slipping the mode phases with respect to each other. Because both Fresnel integrals and Fourier transforms are widely used analysis methods in optical simulation and so are excluded from discussion in this chapter (however details of how they were implemented for performing numerical simulations in this thesis are provided in Appendix A). Instead this chapter provides a description of Gaussian beam mode analysis.

2.2 Overview of GBMA and chapter contents

The THz Optics group in the Department of Experimental Physics at NUI Maynooth has particular expertise in the development and application of GBMA [2.1, 2.2, 2.3, 2.4, 2.5] for the design and analysis of long wavelength quasi-optical systems. One of the goals of the work undertaken by the author of this thesis was to apply GBMA to the simulation and analysis of phase gratings. Previous investigations into the application of GBMA to the description of regular phase gratings have been undertaken at NUIM, where those studies examined how to use GBMA to model the diffraction patterns generated by a number of particular pre-existing binary-level phase gratings. The work described in this thesis builds on that existing body of work, but then extends it by applying GBMA not only to the analysis of more complicated grating designs but also to their design.

One of the main attractions of Gaussian beam mode analysis is that an accurate description of propagating wavefront can usually be achieved using only a small number of modes, which makes it a computationally efficient tool for simulating the propagation of well-behaved beams through (possibly) complicated optical systems. For example, one area where GBMA finds application is in the design and analysis of the quasi-optical systems found in ground- and space-based millimetre wave astronomical telescopes [2.6, 2.7]. The optical systems in such instruments are carefully designed to deliver a beam from the sky to one or more detectors with minimum interference by the optics along the way. Thus an accurate modal description of the beam propagating through such a system can generally be achieved using only a small number of modes. In contrast, some of the imaging experiments that will be described in Chapter 3 use an optical system that is designed to collect and propagate beams that have very different characteristics: very complicated profiles, containing high spatial frequency data. Thus the analysis of these imaging experiments provided a great opportunity to explore the issues involved in using GBMA to describe propagation of these complicated beam patterns.

Our description of GBMA begins in §2.3 with the definition of Gaussian-Hermite beam modes and how they can be used to decompose, reconstruct and propagate beam field that are defined within a Cartesian coordinate system over arbitrary distances. One benefit of GBMA over more traditional scalar wave diffraction techniques (Fresnel integrals and Fourier Transforms) is that it offers greater insight

into the behaviour of a beam as it propagates through a quasi-optical system. The parameters that define a set of Gaussian beam modes can thus be selected to suit the particular system being analysed. In particular a good choice of Gaussian beam modes can improve computational efficiency by choosing a mode-set that best fits the quasioptical system. The behaviour of the Gaussian beam mode parameters: its width W, phase radius of curvature R and phase slippage ϕ are discussed in §2.3.2.

The transverse amplitude variation of a Gaussian beam mode falls off smoothly and rapidly with off-axis distance so a modal analysis can define edges accurately and also truncation effects at the rims of optical components (as described in §2.7). GBMA can become computationally intensive for analysis of beam patterns with complicated profiles because accurate beam description may require a large number of modes. One way to reduce computational overhead is to consider possible symmetry properties of the input field, as described in §2.8.

The software package MODAL that was developed by the THz Optics Group at NUI Maynooth is briefly described in §2.9. As the acronym suggests, principle analysis is based on a modal description of beams in long-wavelength multi-element quasi-optical systems. However, the package also includes efficient PO and scalar diffraction integral options, making it an extremely versatile design and analysis tool.

2.3 Fresnel Transform Gaussian Beam Modes

Gaussian beam modes constitute complete orthonormal sets, each of which are solutions to the paraxial wave equation, thus any arbitrary solution can be expressed as a superposition of a set of such modes. A monochromatic coherent beam represented by a scalar field E, can be written as a linear combination of independently propagating modes. Modal techniques involve expanding a source field as a summation of modes, before individually propagating each mode to another plane some distance z from the original plane. The propagated modes can then be summed with the appropriate coefficients to yield the field due to the propagation of the source field.

The particular choice of mode set depends on the symmetry of the problem. For a system possessing axial symmetry the field can be decomposed into Laguerre-Gaussian modes. However for systems with no specific axial symmetry Hermite-Gaussian modes are more useful, since they are defined as separable one-dimensional

functions in Cartesian coordinates. Hermite-Gaussian modes were chosen for use in our analysis of imaging experiments (see Chapter 3) and phase gratings (see Chapters 4 & 5) because of the Cartesian symmetry involved.

2.3.1 Gaussian-Hermite Modes

A field defined in Cartesian coordinates at a reference plane z_0 (which is taken to be zero at the Gaussian beam waist position) can be expressed as

$$E(x, y, z_0) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} A_{mn} \psi_{mn}(x, y, z_0) \tag{2.1}$$

where the mode coefficients A_{mn} are constants (real- or complex-valued, depending on the field E) that determine the contribution of each Gaussian-Hermite wavefunction ψ_{mn} to the paraxial field $E(x,y,z_0)$. The "standard" independently propagating, two-dimensional Hermite-Gaussian eigenfunction of order $[m,n]$ is of the form

$$\psi_{mn}(x, y, z) = h_m(x;W)\, h_n(y;W) \exp\left[-ik\left(z + \frac{x^2 + y^2}{2R}\right) + i\phi_{mn}(W; R)\right] \tag{2.2}$$

where for convenience the normalised one-dimensional wavefunctions $h_m(x;W)$ and $h_n(y;W)$ are defined (with lower-case h to denote normalisation) in [2.12] as

$$h_m(s;W) = \frac{1}{\sqrt{2^{m-\frac{1}{2}}\, m! \sqrt{\pi W^2}}} H_m\left(\sqrt{2}\frac{s}{W}\right) \exp\left[-\frac{s^2}{W^2}\right] \tag{2.3}$$

where transverse coordinates are denoted by $s \equiv (x, y)$; H_m is a Hermite polynomial of degree, or order m; the last term specifies the transverse Gaussian amplitude variation; the first term is a normalisation factor. Finally the $\sqrt{2}/W$ term in the argument of H_m is a scaling factor that varies with propagation distance z.

The one-dimensional Hermite polynomials are defined by the pure recurrence relation

$$H_m(r) = 2rH_{m-1}(r) - 2(m-1)H_{m-2}(r) \tag{2.4}$$

where the first two polynomials (of order $m = 0$ and $m = 1$) are

$$H_0(r) = 1 \quad , \quad H_1(r) = 2r \tag{2.5}$$

for all values of transverse coordinate r. The higher-order polynomials are then constructed in a recursive manner. For example the second-order polynomial $H_2(r)$ is derived by substitution of $H_1(r)$ and $H_0(r)$ for terms $H_{m-1}(r)$ and $H_{m-2}(r)$, respectively in equation (2.4). When used to define the function $h_m(s;\ W)$, $H_m(r)$ is defined over coordinates $r = \sqrt{2}s/W$.

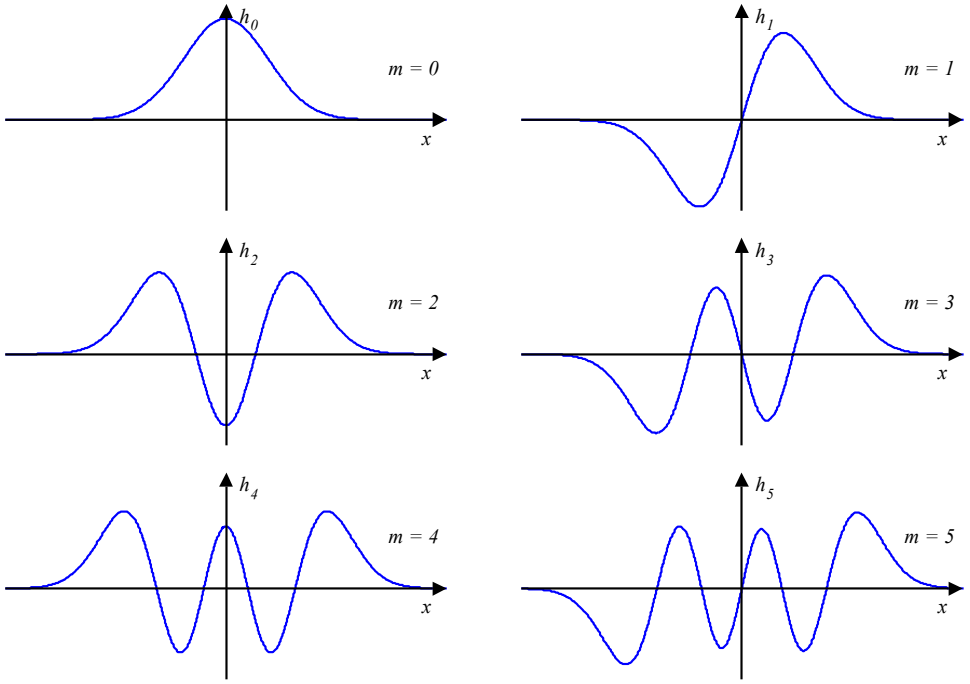


Figure 2-1. The amplitude profile of Gaussian-Hermite modes $h_m(x)$ with indices of $m = 0$ to 5.

The amplitude profiles of the first six one-dimensional Gaussian-Hermite modes $h_m(x)$ are shown in Figure 2-1 to illustrate several features common to Gaussian-Hermite modes. Firstly, even-numbered modes (e.g. $m = 0, 2, 4,\ldots$) are symmetric about the origin, while odd-numbered modes (e.g. $m = 1, 3, 5,\ldots$) are asymmetric about the origin. The mode of order m contains m nulls (points where the amplitude profile crosses the x-axis) and a total of $(m+1)$ extrema (maxima and minima). The quasi-sinusoidal profile of mode m contains $m/2$ full periods (the distance from one peak or trough to the next), the length of period Λ_m being approximately constant over the length of the mode. The period decreases with mode number, i.e. $\Lambda_{m+1} < \Lambda_m$. Another feature, perhaps not so noticeable with the low-order modes shown in Figure 1, is that the two outermost extrema (maxima for even-numbered modes, a minimum and a maximum for odd-numbered modes) of any mode have greater magnitude than the inner peaks and troughs. Figure 2-2 shows the amplitude profile of four two-dimensional Hermite-Gaussian modes $h_{mn}(x,y)$ constructed through different combinations of the one-dimensional functions $h_m(x)$ and $h_n(y)$ shown in Figure 2-1.

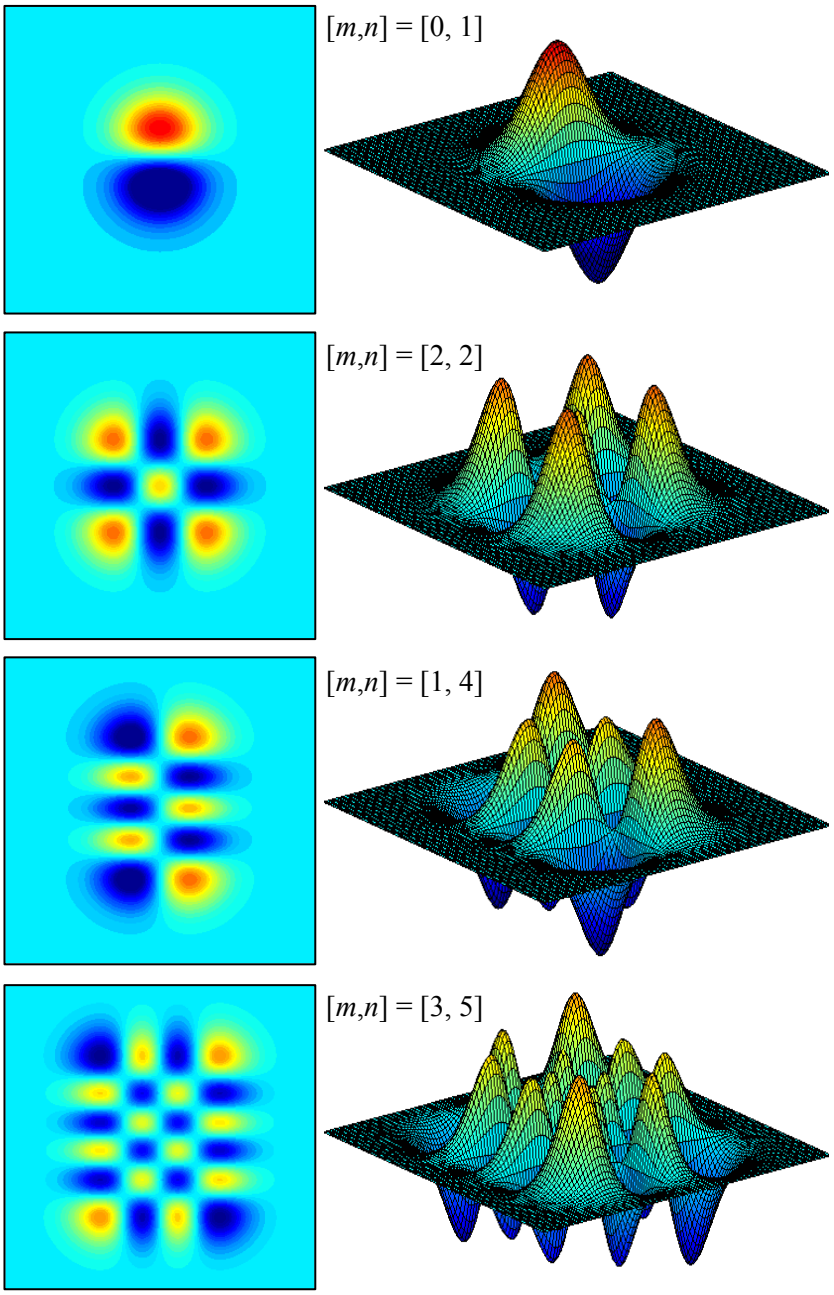


Figure 2-2. False-coloured plots and surface plots of the amplitude profile of four two-dimensional normalised Hermite-Gaussian modes $h_{mn}(x,y)$.

Computational Considerations

Due to its fast execution speed and efficiency in performing matrix calculations numerical simulations of GBMA were performed using the script based language MATLAB[1]. One problem of using MATLAB that became apparent in early GBMA simulations is that it cannot handle as large a range of numbers as, say Mathematica[2].

[1] MATLAB: a numerical computing environment by The Mathworks, Inc. (www.mathworks.com)
[2] Mathematica: a computational software program developed by Wolfram Research (www.wolfram.com)

The minimum and maximum positive floating-point numbers that MATLAB can represent depend on the specific computer used (the range of representable numbers are returned by the built-in functions `realmin` and `realmax`). Calculations that require values outside this range result in an overflow or underflow. The computer used for simulations had minimum and maximum representable numbers of 2.2251e-308 and 1.7977e+308, respectively. These limits result in an inability to compute Hermite polynomials above a certain order using the recurrence relation as defined by equation (2.4) due to the form that the Hermite polynomials take. Apart from the zeroth-order, which is a constant, all higher-order Hermite polynomials increase in absolute value with increasing off-axis distance. The maximum value of $H_m(r)$ is much greater than that of $H_{m-1}(r)$. Eventually at some off-axis position the value of a polynomial of order m will reach the maximum representable floating-point number. Subsequent off-axis values of that polynomial are then represented by $+\infty$. When the next polynomial of order m+1 is evaluated (using the recurrence relation) the values at all off-axis positions corresponding to infinite values in the previous polynomial result in values of Not-a-Number (NaN), since this is the result of any operation with undefined numerical results (such as infinities). To avoid this problem the recurrence relation (2.4) that defines the Hermite polynomials was rephrased to include the normalisation factor from equation (2.3) so that infinite values are not encountered.

The normalisation factor, which we label $N_m(W)$, is given by

$$N_m(W) = \frac{1}{\sqrt{2^{m-\frac{1}{2}}\, m!\sqrt{\pi W^2}}} \tag{2.6}$$

which upon substitution in equation (2.3) leads to the more compact notation

$$h_m(s;\, W) = N_m(W) H_m\left(\sqrt{2}\frac{s}{W}\right)\exp\left[-\frac{s^2}{W^2}\right]$$

for the normalised Gaussian-Hermite function. The normalisation factor can now be transferred into the recurrence relation by multiplying each instance of $H_m(r)$ that occurs in equation (2.4) by the normalisation factor appropriate for that polynomial degree m. The original recurrence relation

$$H_m(r) = 2rH_{m-1}(r) - 2(m-1)H_{m-2}(r)$$

then becomes

$$H_m(r)N_m(W) = 2rH_{m-1}(r)\left(\frac{N_{m-1}(W)}{\sqrt{2m}}\right) - 2(m-1)H_{m-2}(r)\left(\frac{N_{m-2}(W)}{\sqrt{4m(m-1)}}\right)$$

where the denominators in the terms in parenthesis are needed to balance the equation after the normalisation factors have been introduced. Referring to the normalised Hermite polynomials using the notation $H_m(r;\ W) = H_m(r)N_m(W)$, the normalised recurrence relation can now be rewritten as

$$H_m(r;\ W) = \frac{2r\,H_{m-1}(r)}{\sqrt{2m}} - \frac{2(m-1)H_{m-2}(r)}{\sqrt{4m(m-1)}} \tag{2.7}$$

Finally, the normalised zeroth- and first-order Hermite polynomials are given by

$$H_0(r;\ W) = H_0(r)\ N_0(W) = \left(\frac{2}{\pi W^2}\right)^{1/4} \tag{2.8}$$

$$H_1(r;\ W) = H_1(r)\ N_1(W) = 2r\left(\frac{1}{2\pi W^2}\right)^{1/4} \tag{2.9}$$

The normalised Gaussian-Hermite modes are now defined as

$$h_m(s;\ W) = H_m\left(\sqrt{2}\frac{s}{W};\ W\right)\exp\left[-\frac{s^2}{W^2}\right] \tag{2.10}$$

Modal Decomposition and Reconstruction

The contribution that each mode ψ_{mn} makes to the modal expansion of a given field $E(x, y; z)$ is determined by the mode coefficients A_{mn} in equation (2.1). The values of A_{mn} are calculated by first multiplying both sides of equation (2.1) by the complex conjugate of modes ψ_{mn} and then performing an overlap integral in both transverse directions to yield

$$A_{mn} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E(x, y; z)[\psi_{mn}(x, y, z)]^*\,dxdy \tag{2.11}$$

where * indicates complex conjugation. Reconstruction of the output field at a subsequent plane, a further distance z beyond the input plane, is achieved by calculating the modal summation: summing the modes (defined at the z-plane) that are weighted by the mode coefficients A_{mn} as in equation (2.1).

The power of a modal description is that the mode coefficients need only be calculated once (assuming there is no further scattering of power between modes). The evolution of the input beam can then be followed by simply recalculating the modal sum at any given z plane. Although calculating the mode coefficients may be computationally intensive, thereafter the subsequent re-summing of modes at various planes is a relatively trivial task compared to say performing a series of numerical integrations to compute the Fresnel integrals.

Cleary for a given field E, the values of A_{mn} depend on the choice of mode-set ψ_{mn}, as defined by the values of W_0 at the waist plane z_0. Since there is no unique way of choosing the beam waist radius W_0, this scaling parameter can be chosen so as to produce an expansion that best fits the physical constraints of the problem. To maximise computational efficiency the goal is to define a mode set that can reproduce the source field with reasonable accuracy using the least number of modes possible. How to choose an appropriate mode-set scaling factor for efficient analysis of phase gratings will be discussed in Chapter 4.

2.3.2 Gaussian Beam Mode Parameters

An understanding of the physical parameters of the Gaussian beam mode formulation can simplify how the mode set at each plane is represented. The parameters that affect the wavefunction ψ_{mn} as it moves through free-space are the beam radius $W(z)$, the wavefront phase radius of curvature $R(z)$ and the phase slippage $\phi_{mn}(z)$. These three beam parameters can be described in terms of the *Rayleigh range*

$$z_R = \frac{\pi W_0^2}{\lambda} \tag{2.12}$$

where W_0 is the beam waist radius and λ the wavelength of radiation in the medium through which the beam propagates. The Rayleigh range refers to the distance that a collimated beam travels before it begins to diverge significantly and serves as a good indicator of the approximate boundary between the near-field (or Fresenel) and far-field (or Fraunhofer) zones. At the distance z_R, the beam radius $W = \sqrt{2}W_0$. Beyond this distance the beam is said to be in the far-field region. Since the beam profile does not change significantly as z increases in the far-field range, any arbitrary distance that is much greater than z_R can be taken to be the far-field range. For example when calculating the far-field diffraction pattern from a phase grating the output plane was (arbitrarily) set to lie at a distance of $z = 100z_R$ from the grating plane, which easily satisfies the far-field criterion that $z >> z_R$.

In some literature Gaussian beams are characterised by the confocal distance z_c, which can be defined in terms of a Gaussian beam that is sent through a focus. It is the distance between the planes on opposing sides of the focal plane at which the beam

half-width is equal to $\sqrt{2}W_0$. In other words it is the distance between the Rayleigh ranges on either side of the focus,

$$z_c = 2z_R$$

The reason why Hermite-Gaussian and Laguerre-Gaussian polynomials are chosen as the mode basis set over others (for example Zernicke polynomials) is that these modes have the physically desirable property of maintaining their profile at all transverse planes in z. That is, although the width of each mode increases with increasing propagation distance z, as $W(z)$ does, and acquires spherical curvature $R(z)$, the shape of each mode remains unchanged. The shape of the propagating beam as a whole changes only due to the evolution of the phase slippage, $\phi_{mn}(z)$ associated with each mode. This term is a measure of the degree to which individual modes slip in and out of phase with each other and depends on both the mode number and distance from the waist position, but not on transverse position (x, y).

Gaussian Beam Radius, $W(z)$

The beam width parameter $W(z)$ is defined as the off axis distance where the field magnitude of the fundamental mode – a single Gaussian – drops to $1/e$ of its on-axis value, or equivalently where its intensity has dropped to $1/e^2$ of its on-axis value. As the beam propagates from its waist position (the plane where W has its minimum value W_0 and which is taken to be at $z_0 = 0$ for convenience) diffraction causes the beam width to increase with distance. The rate at which beam expansion occurs is given by the non-linear equation (derived in Appendix A.3)

$$W(z) = W_0\sqrt{1+\left(\frac{z}{z_R}\right)^2} \tag{2.13}$$

Figure 2-3 shows the $1/e$ width of an expanding Gaussian beam (solid blue curve) that is propagating from left to right, from its waist position (at $z = 0$) to a plane in the far-field (at $z = 10z_R$). Upon leaving the waist position, the beam expands slowly at first before reaching a value of $W = \sqrt{2}W_0$ at a distance of z_R from the waist position z_0. As we move further from the waist position the beam expansion increases until it becomes approximately linear in the far-field range (for $z >> z_R$), as indicated by the dashed green line in Figure 2-3. Since $z >> z_R$, the 1 inside the square-root of equation (2.13) becomes negligible and the radius can be expressed as

$$W(z) = W_0\left(\frac{z}{z_R}\right) \tag{2.14}$$

Thus we can say that in the near-field the beam radius is approximately constant $W(z) \approx W_0$, compared to the far-field where $W(z) \propto z$.

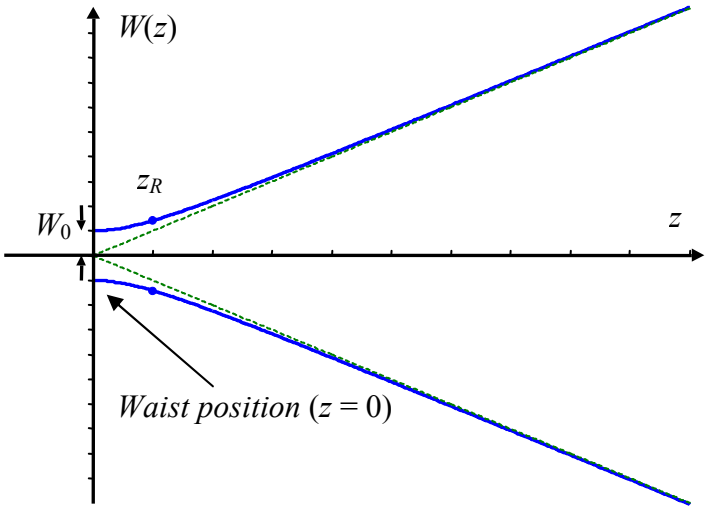


Figure 2-3. Variation in beam radius $W(z)$ of an expanding Gaussian from the beams waist position ($z = 0$, where $W = W_0$, the beam waist radius), to the far-field ($z >> z_R$) where beam expansion is approximately linear. The distance z = zR is indicated by the circular blue markers superimposed on the line plot of beam radius $W(z)$.

As Figure 2-4 shows, diffraction effects mean that the rate of beam expansion is dependent on initial beam waist radius W_0. Here two beams with waist radii $W_{0,1}$ and $W_{0,2}$, such that $W_{0,2} > W_{0,1}$, are shown propagating from left to right through a common waist position (at $z = 0$). The beam with the smaller initial radius expands more rapidly and results in a larger beam size than the other due to diffraction.

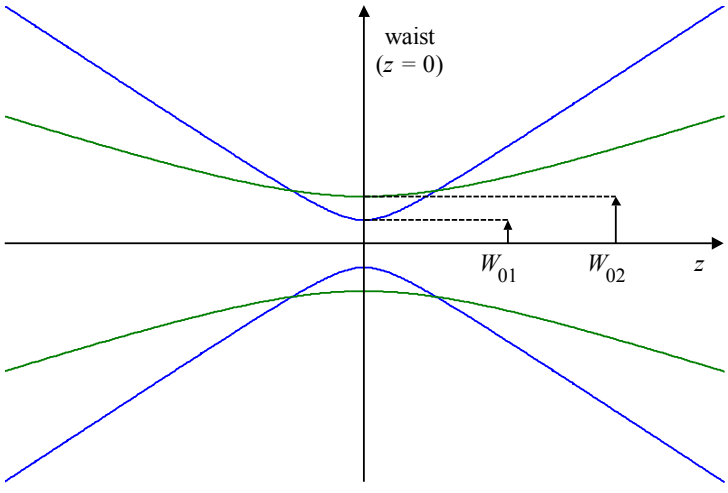


Figure 2-4. Diffractive expansion of two Gaussian beams with waist radii $W_{0,1}$ and $W_{0,2}$ ($W_{0,1} < W_{0,2}$) that share a common waist position at $z = 0$.

Phase Radius of Curvature, *R*(*z*)

The phase radius of curvature $R(z)$ describes the constant phase surface of a beam with respect to a plane transverse to the direction of propagation, which translates to the phase delay of

$$\phi(r; z) = \frac{\pi r^2}{2R(z)} \tag{2.15}$$

seen in equation (2.2). The beam waist position is the point along the axis of propagation where the beam is planar. In other words if the beam waist position is taken to be at $z = 0$ the radius of curvature at that point is infinite: $R(0) = \infty$. As the beam spreads outwards from the waist position the propagating wavefront acquires curvature, resulting in a rapidly changing finite radius of curvature. When the wavefront has reached the Rayleigh range, z_R the radius of curvature acquires its minimum value of $R(z_R) = 2z_R$. At this point therefore the centre of curvature of the wavefront is located at the same distance behind the waist position at $z = -z_R$. Beyond the Rayleigh range the radius of curvature increases once again. When z is much greater than z_R, the Gaussian beam begins to behave like a spherical wavefront centred at the waist position and consequently $R(z)$ increases approximately linearly with increasing distance.

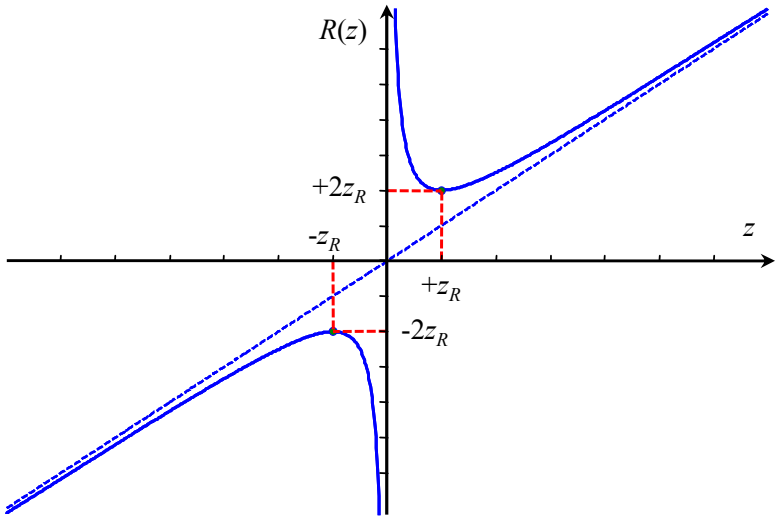


Figure 2-5. Radius of curvature $R(z)$ as a function of propagation distance z through a beams waist ($z = 0$). For $z > 0$ the minimum value of $R(z) = 2z_R$ occurs at $z = +z_R$, which means to an observer at this point the beam appears to be a spherical wavefront centred on $z = –zR$.

The behaviour of the radius of curvature is illustrated in Figure 2-5 and summarised by (as derived in Appendix A.3)

$$R(z) = z + \frac{z_R^{\,2}}{z} \quad \approx \quad \begin{cases} \infty & \text{for} \quad z << z_R \\ 2z_R & \text{for} \quad z = z_R \\ z & \text{for} \quad z >> z_R \end{cases} \tag{2.16}$$

Phase Slippage, $\phi_{mn}(z)$

As well as changing beam width $W(z)$ and phase radius of curvature $R(z)$, the propagating wavefunction ψ_{mn} is also affected by a phase shift or slippage, in addition to the plane-wave phase shift given by exp[-*ikz*], which appears in the exponential term of equation (2.2). This additional mode-dependent axial phase slippage term takes the form of

$$\phi_{mn}(z;\, W_0) = [m + n + 1]\tan^{-1}\left(\frac{z}{z_R}\right) \tag{2.17}$$

for two-dimensional modes where the individual transverse contributions are

$$\phi_m(z;\, W_0) = [m + ½]\tan^{-1}\left(\frac{z}{z_R}\right) \tag{2.18}$$

Thus for all modes m and n with common W_0, (i.e. of the same mode set) the evolution of the phase shift (the shape of which is given by the arctangent term) is the same for each mode. Thus the phase slippage associated with any given mode is exactly the same as for the fundamental mode (m = 0) except for a scaling factor determined by its mode index m. This point is illustrated in Figure 2-6, which shows the evolution of the phase slippage $\phi_m(z;\, W_0)$ for the first four (m = 0…3) one-dimensional modes on passing through a waist position at z = 0, i.e. where the beam is brought to a focus and it expands thereafter. The effect of the term ϕ_m on each mode m is an accumulation of phase as the beam crosses the waist position, for which at finite distances z the magnitude is determined by equation (2.18). In the far-field, where z tends towards infinity, the phase slippage associated with each 1-D mode is given by

$$\Delta\phi_m = \pm(2m + 1)\ \pi/2 \tag{2.19}$$

The question of whether the phase shift should be added or subtracted depends on the direction of propagation (+z or –z). To calculate the phase in the direction of propagation of the beam (from left to right in Figure 2-6) the phase shift $\Delta\phi_m$ is added. While for calculation of the phase at a previous plane, i.e. opposite to the direction of propagation $\Delta\phi_m$ is subtracted.

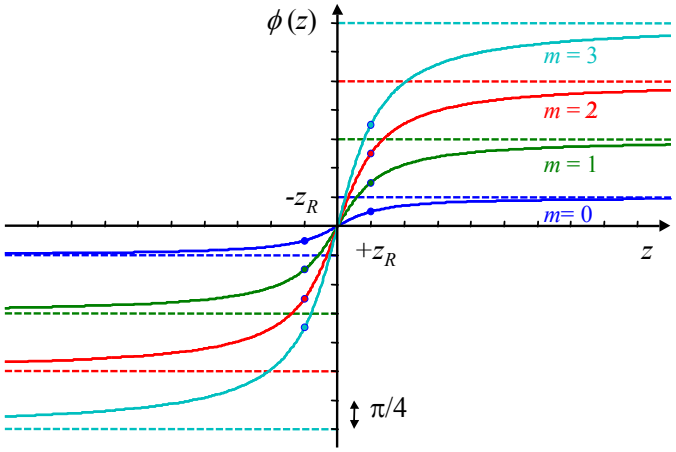


Figure 2-6. Evolution of the phase slippage term $\phi_m(z)$ for modes m = [0…3] through a beam waist position at $z = 0$.

The physical interpretation of the phase shift gained on passing through a waist is that the phase velocity of each mode increases. This is a Gaussian beam version of the *Guoy* effect (valid for any optical beam passing through a focus), which says that any reasonably simple cross section will acquire an extra half-cycle of phase-shift (i.e. $+\pi/2$) in passing through a focal region. We see from equation (2.19) that this statement is exactly true for the fundamental mode ($m = 0$) and for higher-order modes requires only the addition of a $(2m+1)$ scaling factor.

Since each mode gains a different amount of added phase each mode travels with a different phase velocity through a quasioptical system. Therefore, although all modes may be in phase at the waist plane and so can be made to combine so as to represent a certain field profile at that position, as the modes propagate they do so at different speeds and therefore at subsequent planes will not combine in the same manner as they did at the previous plane, which gives rise to interference effects.

Far-field approximation

Some simplifications can be made to the exponential term in equation (2.2) when calculating the wavefunction ψ_{mn} at the two extreme planes that we are interested in propagating beams between when analysing the operation of for example a phase grating. If the grating plane is defined at the beam waist position (at $z = 0$) then we are interested in calculating its far-field diffraction pattern (at $z \approx \infty$). Since the plane-wave phase term e^{-ikz} is independent of mode number and therefore the same for all modes

and oscillates much more rapidly with distance than any other term it can be factored out of the calculation of the wavefunction at all planes. Thus equation (2.2) can be written as

$$\psi_{mn}(x, y, z) = h_m(x;W(z))\, h_n(y;W(z)) \exp\left[-ik\left(\frac{x^2+y^2}{2R(z)}\right) + i\phi_{mn}(z;\, W_0)\right]$$

In terms of Gaussian beam mode analysis of a phase grating if the grating is situated so that it coincides with the beam waist position, at $z = 0$ the wavefront illuminating the grating has a flat wavefront. Therefore all the extra phase structure imprinted on the beam performs the phase modulation associated with the grating profile that is necessary to produce the required far field intensity distribution. With this assumption of grating position, from equation (2.16) the wavefront radius of curvature $R(z)$ at the grating plane is taken to be infinite and therefore the input phase front is uniform. Similarly, at $z = 0$ the phase slippage term $\phi_{mn}(z;\, W_0) = 0$ since all modes are in phase at the waist position. Consequently the exponential term is reduced to unity and the resulting input wavefunctions at the grating plane consist solely of real two-dimensional normalised Gaussian-Hermite functions

$$\psi_{mn}(x, y, z) = h_m(x;W(z))\, h_n(y;W(z)) \tag{2.20}$$

Similarly, since the image formed by a phase grating is produced at its far-field, the wavefunctions at this plane can also be simplified somewhat. Again the plane-wave can be removed, as can the spherical phase term (provided the paraxial approximation is valid). As z tends towards infinity so the phase slippage $\phi_m \rightarrow (m + ½)\pi/2$ so the remaining exponential term $\exp[i\phi_m]$ can be expressed as $\exp[i\pi/2]^{\,m+\,½} = i^{(m+\,½)}$. Therefore the far-field wavefunction on a spherical wavefront as $z \rightarrow \infty$ has the form of

$$\psi_{mn}(x_0, y_0, z) = h_m(x_0;W(z))\, h_n(y_0;W(z))\, i^{(m+\,½)} \tag{2.21}$$

where transverse coordinates (x_0, y_0) are those in the far-field image plane.

Visualising the phase for near-field diffraction

Consider the case where some complex field, E_0 produced by a radiating source S at the GBM waist position z_0 is to be propagated a finite distance $z < z_{FF}$, i.e. so that the far-field approximation is not appropriate. In this situation the propagated field, E_z is calculated using Gaussian-Hermite modes of the form given by equation (2.2). If we are interested in examining the phase distribution at z, which is given by $\phi_z = \mathrm{Arg}\{E_z(x, y)\}$, we will see that it is dominated by a spherical phase component which obscures the

underlying structure of interest. In order to be able to visualise the latter more clearly it is necessary to remove the spherical phase component that is introduced by the term

$$\exp\left[-ik\left(\frac{x^2+y^2}{2R(z)}\right)\right]$$

in equation (2.2) that describes the expanding spherical-wave nature of the Gaussian beam modes. This component is removed by simply multiplying $E_z(x, y)$ by the complex conjugate of the above term before extracting the phase distribution.

2.3.3 Aperture Size and Beam Width

Consider the case of a propagating Gaussian beam, which, for example, could be used to illuminate a phase grating of a certain size. When the grating aperture is placed in the path of the beam, the beam suffers truncation as its flanks are blocked by the edges of the obstructing aperture. Since the intensity of a Gaussian beam falls off rapidly with increasing off-axis distance, for a sufficiently large aperture the amount of clipping of the beam is negligible. For an aperture with radius a equal to the input beam radius W approximately 86% of beam power is transmitted through the aperture [2.10]. If the aperture diameter is increased to $2a = \pi W$, then approximately 99% of incident beam power passes the aperture.

However as well as simply causing truncation a sharp-edged aperture also produces significant diffraction effects in the sense of changing the form of the transmitted beam. This is especially true in quasioptical systems where component size is on the scale of tens of wavelengths. Therefore large apertures are needed not only to reduce truncation effects, but also to minimise edge diffraction effects. It can be shown that even for large apertures with diameters measuring $d = \pi W$ near-field diffraction effects occur with peak intensity ripples measuring ~17% of maximum beam intensity. Only when aperture size is increased to $d = \sim 4.6W$ is the peak intensity of such diffraction ripples reduced to ~1% of maximum beam intensity. If a component, such as a grating, is designed for illumination with a particular predetermined beam size, the aperture size can be chosen so as to minimise both truncation and diffraction effects. Conversely if the components aperture size component is predetermined then one must match the illuminating beam size so as to minimise these unwanted effects.

Effective Beam Mode Width, L_m

The transverse Gaussian amplitude variation $\exp[-r^2/W(z)^2]$ effectively sets the mode size by tapering off power the further one moves from the axis of propagation. In this context radius W refers to the beam radius associated with the Hermite-Gaussian mode set $h_m(x)$ and should not be confused with the radius of an illuminating Gaussian beam whose phase-modulated field we may wish to expand in terms of that mode set. It is important to note that as well as being dependent on radius W, the effective size of a given mode depends also on the particular mode-order m. Figure 2-7 shows the amplitude profile of three one-dimensional Gaussian-Hermite modes of index m = 6, 10 and 20. Clearly the effective mode width, L_m – defined here as the separation between the two outermost peaks (maxima or minima) – increases with mode-order m.

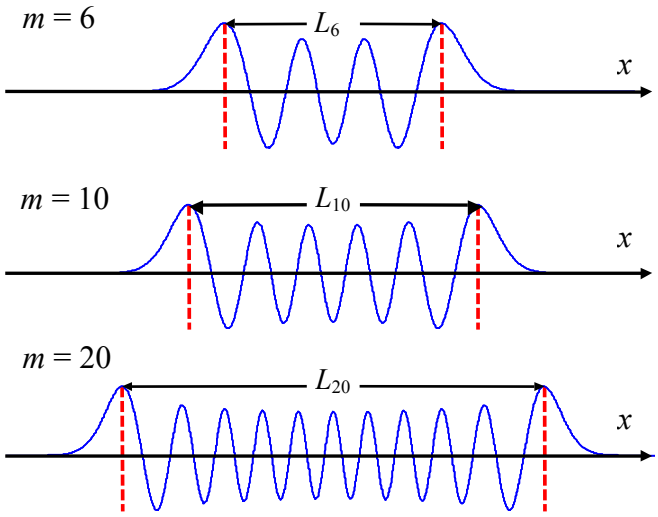


Figure 2-7. Amplitude profile of three one-dimensional Gaussian-Hermite modes $h_m(x)$ with mode-order m = 6, 10, 20. The effective mode width L_m refers to the distance between the two outer maxima.

If a given beam is to be expanded in terms of a set of Gaussian beam modes (GBM's) it is clearly useful to match the width of the modes with respect to the phase-modulated input beam. Consider for example a grating or aperture of radius, or half-width a, the field across which is to be expanded in terms of a set of GBM's. Clearly any mode whose effective width exceeds the aperture diameter will not be permitted to pass untruncated through the aperture. Furthermore, for efficient representation of the beam at the aperture/grating plane all modes in the chosen mode set should contribute above some level to the field. It is therefore necessary to choose a mode set in which the majority of modes can just pass through the aperture without incurring significant levels

of truncation. For example the mode set could be scaled such that the highest-order mode half-width is approximately equal to the aperture radius.

Figure 2-8 shows a plot of L_m, as defined as above (distance between two outermost maxima), for the first 30 Gaussian-Hermite modes. An approximately linear relationship is observed between mode width L_m and $\sqrt{m}$ (represented by the dashed green line in Figure 2-8). Hence the mode-width L_m of a Hermite-Gaussian mode of order m is related to mode-order m and mode radius W_x as follows

$$\frac{L_m}{2W_x} \approx \sqrt{m} \tag{2.22}$$

which yields an expression for the effective mode width as $L_m \approx 2W_x\sqrt{m}$. For a given aperture or beam width of $2a$ whose field is to be expanded in terms of a Gaussian-Hermite mode set with highest-order mode m, a suitable Gaussian width W_x for the mode set is obtained by substituting $2a$ for L_m to yield a value of

$$W_x \approx \frac{a}{\sqrt{m_{\max}}} \tag{2.23}$$

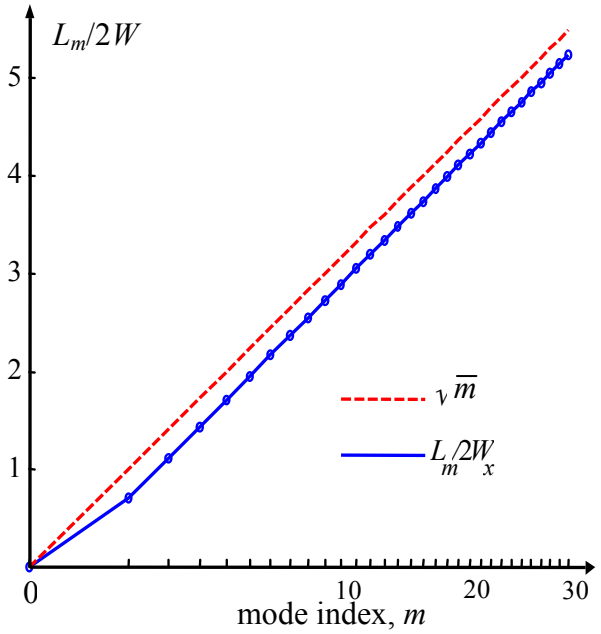


Figure 2-8. Plot of effective mode width L_m divided by twice the mode scaling parameter Wx for Gaussian-Hermite mode orders m = [0…30]. The square-root of mode indices m are plotted along the x-axis, revealing an approximately linear relationship between $L_m/2W_x$ and $\sqrt{m}$.

Of course this definition of L_m – the separation between the outermost maxima (or maximum and minimum for odd-numbered modes) in a modes intensity pattern – is rather arbitrary. Any point on the mode profile where mode amplitude falls to some predefined value could serve equally well for defining the edge of a mode, provided the

mode profile can be accurately represented between the two chosen edges, i.e. provided that all features (peaks and troughs) of the highest-order mode are included between the two endpoints. For example, to be consistent with the definition of the width of a Gaussian beam, we could alternatively define the edge of a mode as being the off-axis distance where mode amplitude is $1/e$ of its maximum value. Figure 2-9 shows mode $h_6(x)$ with vertical dashed lines at the $1/e$ amplitude points. Figure 2-10 shows the plot of $L_m/2W_x$ against $\sqrt{m}$ using this alternative definition for the mode width.

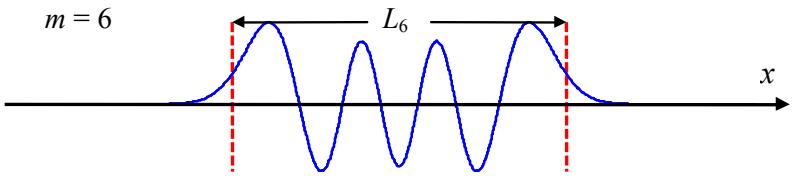


Figure 2-9. One-dimensional Hermite-Gaussian modes $h_6(x)$ with mode widths L_m defined as the distance between the points where mode amplitude is $1/e$ of its maximum value.

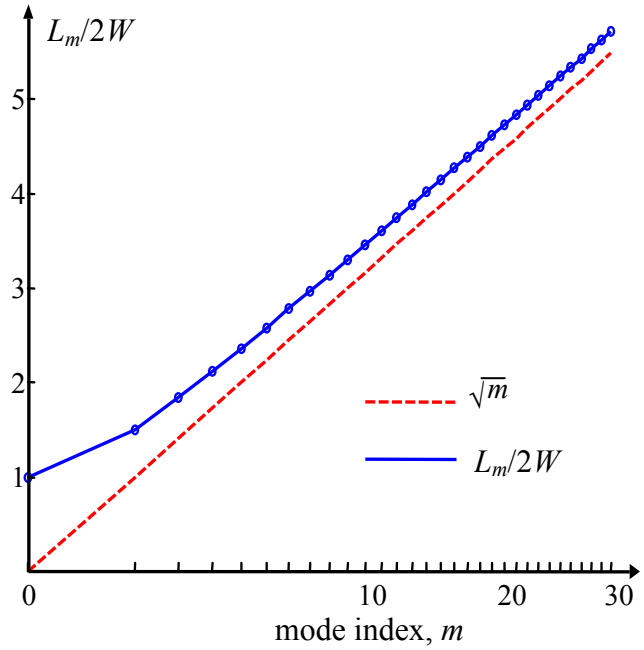


Figure 2-10. Plot of mode width L_m versus mode index m for the first 30 Gaussian-Hermite modes, using the definition of L_m as being the distance between the endpoints where the amplitude is $1/e$ of its maximum value.

One useful application of the relationship between mode scaling factor W_x and mode width L_m as given by equation (2.23) is that it can be used to determine the highest order mode for given values of mode width L_m and W_x. If an aperture or grating field of width

$2a$ is to be expanded using a mode set whose Gaussian parameter is W_x then rearranging equation (2.23) gives

$$m \approx \left(\frac{L_m}{2W_x}\right)^2$$

Then since we require that the width of the highest-order mode $m_{\max}$ match the aperture size as closely as possible, the index of the highest-order mode that will fit the aperture is found by setting its width L_m equal to $2a$, which yields

$$m_{\max} \approx \left(\frac{a}{W_x}\right)^2 \tag{2.24}$$

Therefore a modal expansion of the given field need only those modes up to mode index $m_{\max}$, since all modes of a higher index will expand beyond the dimensions within which the field is defined and so will contribute little to the reconstructed field.

It is worth noting that if significant truncation occurs at the edges of an aperture/grating then the high spatial frequency components introduced as a result may not be accurately reproduced in the propagating beam. It would then be necessary to increase the highest order mode index $m_{\max}$ appropriately. For example for top-hat like fields $m_{\max}$ may need to be increased by a factor of two to reproduce truncation effects.

Spatial Period of Gaussian-Hermite Beam Modes

A Gaussian-Hermite mode of order m contains $m/2$ full quasi-sinusoidal periods of approximately equal width Λ_m across its effective mode width L_m. Thus effective mode width L_m can be expressed in terms of mode period as

$$L_m = \frac{m\Lambda_m}{2} \tag{2.25}$$

From equation (2.23) the approximate spatial period of mode m is thus related to the Gaussian parameter W_x as follows

$$\Lambda_m \approx \frac{4W_x}{\sqrt{m}} \tag{2.26}$$

Equation (2.24) states that the maximum mode index needed to describe an arbitrary field of width $2a$ is $m_{\max} = (a/W_x)^2$. However determination of the number of terms to include in a modal expansion by this method requires prior knowledge of the mode-set Gaussian parameter W_x. It is therefore necessary to choose a value of W_x that defines a

mode-set suitable for efficient reconstruction of the input beam. In order to do so one must consider the properties of the input beam itself.

If the size of the smallest fluctuations in the input beam is δ and we require that a modal analysis be capable of reproducing features of this size then the mode-set used to expand the input beam must contain modes whose spatial period satisfies $\Lambda_m \leq \delta$. At the same time the mode-set must be able to describe the beam over its entire length. The highest-order mode must therefore simultaneously satisfy the following criteria:

a) its spatial-frequency must be sufficient to reconstruct the smallest feature in the input beam: $\Lambda_{m,\max} \approx 4W_x/\sqrt{m_{\max}} \leq \delta$

b) it must have a mode-width at least as large as the beam diameter: $m_{\max} \approx (a/Wx)^2$

Solving for W_x in conditions *a*) and *b*) gives

$$W_x \approx \frac{a}{\sqrt{m_{\max}}} \leq \frac{\delta\sqrt{m_{\max}}}{4}$$

which imposes the following criterion on the choice of maximum mode index

$$m_{\max} \geq \frac{4a}{\delta} \tag{2.27}$$

needed to reconstruct a beam of radius a whose minimum feature size is δ. Now equation (2.24) implies that

$$m_{\max} \approx \left(\frac{a}{Wx}\right)^2 \geq \frac{4a}{\delta}$$

which yields an expression for the maximum permissible value of beam mode parameter W_x in terms of beam/aperture radius a and minimum feature size δ, as follows

$$W_x \leq \sqrt{\frac{\delta a}{4}} \tag{2.28}$$

Together equations (2.27) and (2.28) allow one to define a mode-set capable of reproducing an arbitrary wavefront of radius a and minimum feature size δ. The Sampling Theorem says that in order to describe a function $f(x)$ the function must be sampled at a rate of at least two samples per spatial period. Equation (2.27) implies that a mode set with a minimum highest-order mode $m_{\max} = (4a/\delta)$ is required to accurately reproduce a field or function (of radius a and minimum feature size δ). The highest-order mode contains $m_{\max}/2$ quasi-sinusoidal periods and thus requires $(4a/\delta)$ sample points. Thus equation (2.27) could be viewed as a Gaussian beam mode version of the Sampling Theorem.

2.4 Decomposition of 1- and 2-D top-hat fields

In this section we present an example to demonstrate the applicability of GBMA to one- and two-dimensional beam reconstruction, while at the same time illustrating the importance of choosing a suitable mode-set with which to expand a given field.

First we consider the reconstruction of a one-dimensional top-hat field of width $2a$, representing a one-dimensional cut through the field produced by a uniformly illuminated slit of the same width. The field at the aperture consists of an amplitude distribution with a top-hat profile and constant phase distribution, corresponding to a plane wave incident on the narrow slit. Since the phase distribution is flat, the radius of curvature of the beam R is infinite. In other words the aperture plane coincides with the beam waist position. Because the phase distribution does not vary across the aperture plane, the E-field is not complex but can be considered to be real. The aperture field can thus be decomposed using the set of one-dimensional Gaussian-Hermite modes

$$\psi_m(x, z_0) = h_m(x; W_0)$$

After the input field has been expanded in terms of these modes, the reconstructed field is propagated to a plane some finite distance z away. We will examine four different methods of scaling the set of Gaussian-Hermite modes used to reconstruct the top-hat field by varying the beam width parameter W_0, or W_x for the one-dimensional case.

The particular choice of mode-set (determined by the value of the mode width parameter W_0) is crucial to a computationally efficient expansion of the top-hat field. The approach often taken (to modelling a top-hat field) is to choose a mode set that maximises power in the fundamental mode. However this approach results in the remaining beam power being distributed between many higher-order modes and in fact no power is coupled to the next highest-order symmetric mode. Although all of the higher-order modes contain very little power, because of the sharp edges a large number of modes are needed to produce an accurate reconstruction of the field.

Figure 2-11 shows the mode coefficients A_m for the first thirteen Hermite-Gaussian modes (m = 0…12) over a range of values of the beam width parameter W_x. For each value of W_x a mode-set $h_m(W_x)$ is created and the overlap integral (Equation 2.11) performed to yield the mode coefficients A_m corresponding to that particular set of modes. Since a top-hat function is symmetric about the origin, only those modes that are also symmetric about their centre can contribute to the modal expansion. Therefore only

the even-numbered (symmetric) mode coefficients are shown in Figure 2-11 because the odd-numbered modes do not contain any power.

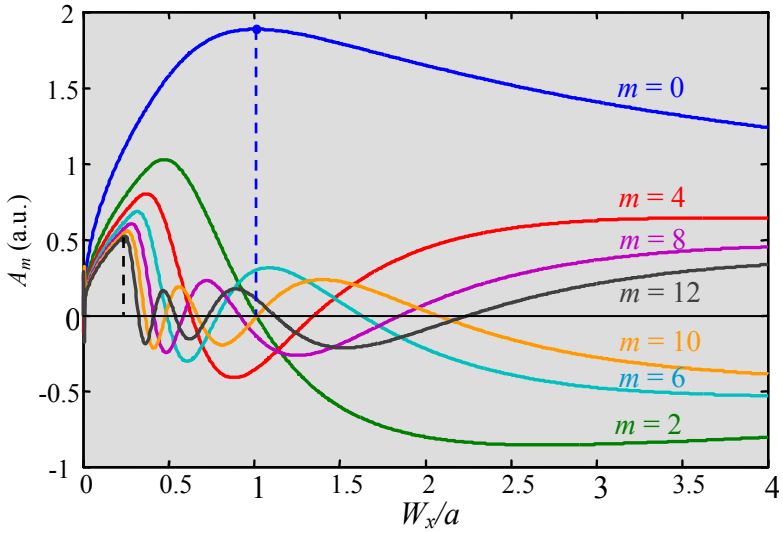


Figure 2-11. Real-valued mode coefficients A_m used in the modal decomposition of a top-hat field of width $2a$. The values of the even-numbered mode coefficients m = [0, 2, 4…12] are displayed for various values of the beam width parameter W_x, resulting from integration of each mode with the top-hat function.

Figure 2-11 shows that the mode set in which the fundamental mode $|A_0|$ has maximum power occurs for a value of $W_x \approx a$. Note that this particular mode-set results in $|A_2| = 0$ and $|A_{10}| = 0$ and that the remaining higher-order modes have relatively low power levels compared to the fundamental mode, thus necessitating the inclusion of many higher-order modes to accurately reproduce the top-hat field.

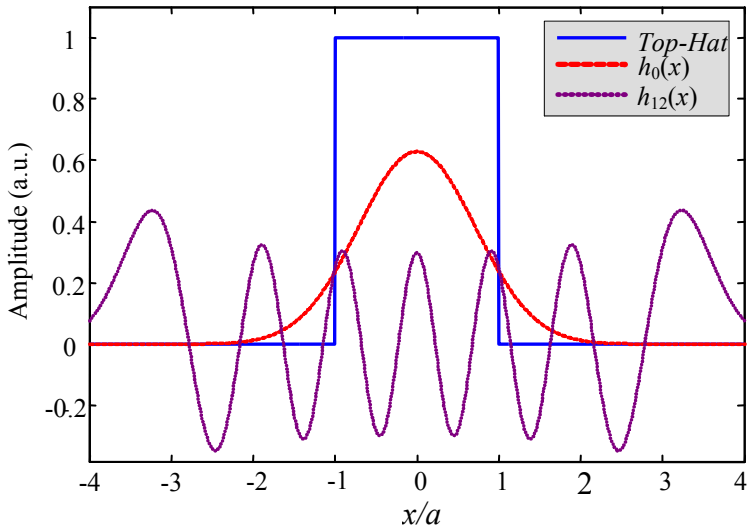


Figure 2-12. Amplitude profiles of top-hat function and Gaussian-Hermite modes m = 0 (fundamental) and m = 12 for the mode set in which power in the fundamental mode is maximised.

Figure 2-12 shows the extent of the amplitudes of modes $h_0(x)$ (the fundamental mode) and $h_{12}(x)$ in relation to the top-hat function. Clearly mode $h_{12}(x)$ stretches well beyond the limits of the top-hat function ($x = \pm a$) and as such contribute only a small amount of power to the beam expansion.

A modal expansion of the top-hat field was performed using a mode-set in which the fundamental mode was optimised and in which the highest-order mode was set to $m = 200$. The magnitudes of the even-numbered mode coefficients are shown in Figure 2-13. The first thing to notice is that the value of A_0 is considerably greater than any other mode coefficient: the values of A_m for $m > 0$ fall off rapidly over the first few modes, after which the rate of decline slows. Notice also the quasi-periodic distribution that the mode coefficients values take, which resembles the far-field diffraction pattern from a top-hat function.

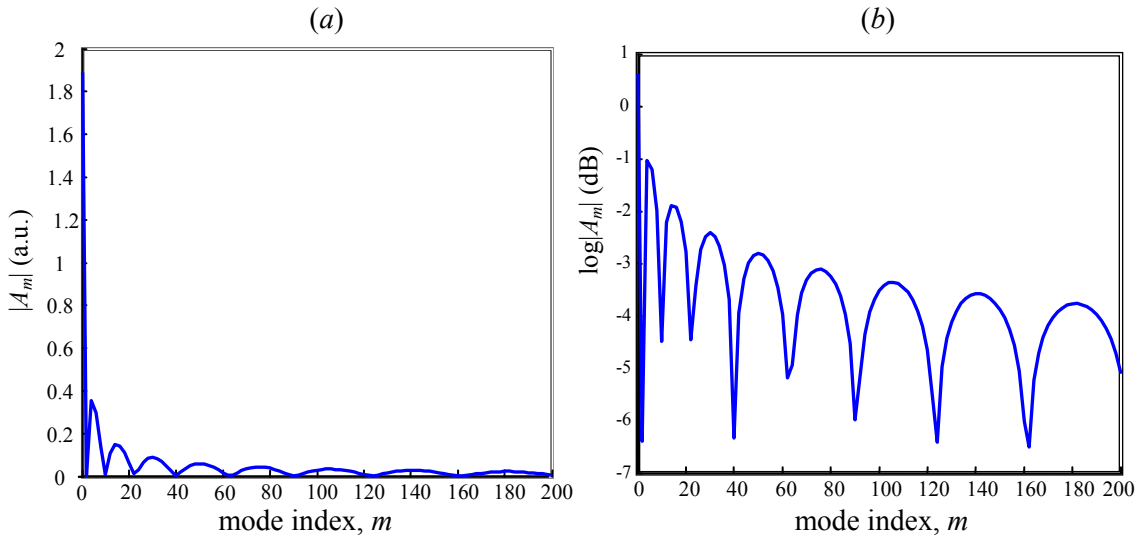


Figure 2-13. (*a*) Linear- and (*b*) log-scale plots of mode coefficients $|A_m|$ of all even-numbered Gaussian-Hermite modes (with mode indices $m = 0 \ldots 200$) used in the expansion of a top-hat field. The mode-set used is one in which the fundamental mode power is maximised.

Figure 2-14 compares the input top-hat amplitude distribution with the reconstructed field amplitude distribution $|E_{rec}|$ using the mode-set in which the fundamental mode is optimised and with $m_{max} = 50$ and $m_{max} = 200$. Clearly the more modes are included, the more accurate the reconstructed field. The reconstruction in which $m_{max} = 200$ manages to suppress power in unwanted side-lobes that appear past the real physical extent of the top-hat function, $|x| > a$. Also by using more higher-order modes the closer the reconstructed field approximates the sharp vertical edges of the top-hat function at $x = \pm a$.

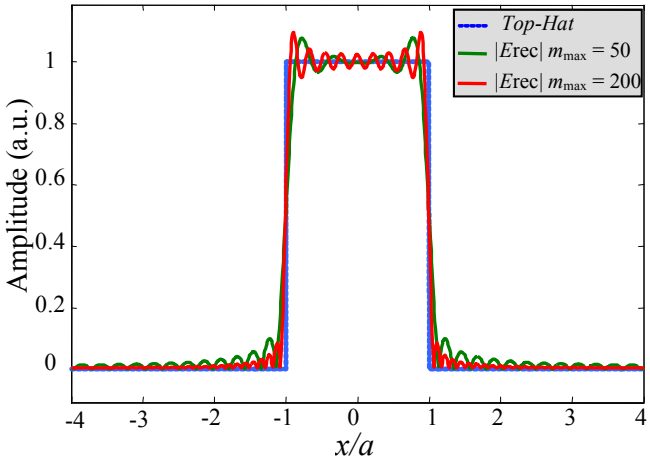


Figure 2-14. Top-hat function (blue) and amplitude of modal reconstructions $|E_{rec}|$ produced using a mode-set in which the fundamental mode has been optimised and in which the highest-order mode index is m_{max} = 50 (green) and 200 (red).

This example shows that using this particular choice of mode-set (in which the fundamental mode is optimised), the accuracy of the reconstruction is increased by increasing the number of modes. However to reduce computational overhead it would be desirable to be able to efficiently reconstruct an input field with as few modes as possible. Therefore an alternative choice of mode-set might be to try other values of W_x than that which optimises power in the fundamental and instead find a value that maximises power in some higher-order mode. For the case of the top-hat field, Figure 2-11 shows that a value of $W_x = \cong 0.1305a$ will produce a mode-set that is scaled such that power in mode $h_{12}(x)$ is maximised.

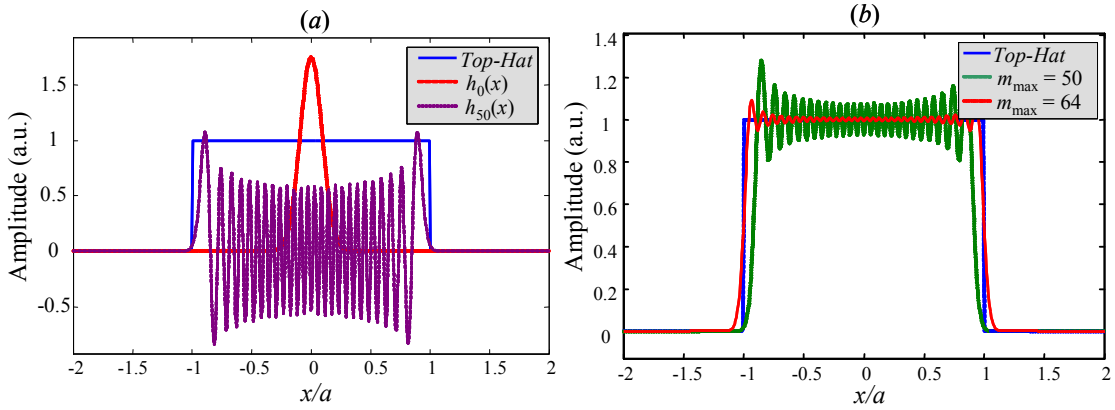


Figure 2-15. Reconstruction of top-hat function with a mode-set scaled so that power in mode $h_{50}(x)$ is maximised. (*a*) Top-hat (blue) with modes $h_0(x)$ and $h_{50}(x)$ superimposed. (*b*) Reconstructions of top-hat using mode-set in which highest-order mode index is m_{max} = 50 (green) and m_{max} = 64 (red).

A modal expansion of the top-hat function was performed using a mode-set scaled such that power in mode $h_{50}(x)$ was maximised. Figure 2-15(a) shows the top-hat function with the modes $h_0(x)$ and $h_{50}(x)$ superimposed. All modes $m = 0\ldots50$ fit inside the extent of the top-hat so no side-lobes are present in the reconstructed field shown by the green curve in Figure 2-15(b). However two problems arise with this choice of mode-set. Firstly the edges of the reconstructed field do not extend as far as those in the original top-hat function because the highest-order mode is smaller than the top-hat full width $2a$. Secondly the magnitude of peak-to-peak ripples in the reconstructed field is much greater than those produced using a mode-set in which the fundamental mode was optimised. Examination of the amplitude coefficients (Figure 2-16) for this mode-set (in which mode $m = 50$ is optimised) reveals that while mode coefficient power falls off smoothly, it does not drop to zero. This implies that the inclusion of some higher-order modes with indices $m > 50$ is needed to produce a more accurate reconstruction. If the number of modes is increased until the magnitude of the last mode coefficient reaches zero a better description of the input field would be achieved. A second reconstruction with a mode-set including all modes up to $m = 64$ was calculated. The resulting amplitude distribution $|E_{rec}|$ shown in by the red curve in Figure 2-15(b) is a much better approximation of the original top hat beam: the peak-to-peak ripples are reduced in magnitude and the extent of the reconstructed field matches closely that of the top hat full-width, $2a$.

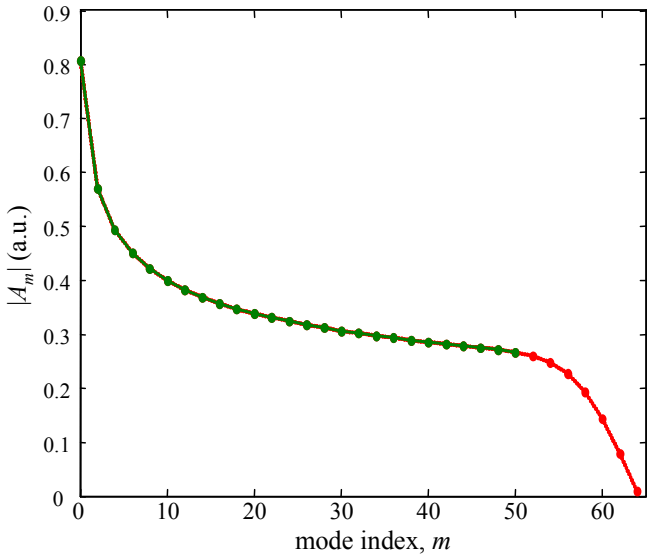


Figure 2-16. Mode coefficient amplitude $|A_m|$ for a mode-set in which the power in mode $m = 50$ is maximised. The accuracy of reconstruction is improved by including modes with indices up to $m = 64$.

It was previously reported in [2.11] that to successfully model the edges of a top hat function high-order modes should be included in the mode-set because their high-spatial frequencies are ideal for describing the sharp discontinuity present at an edge. It was stated that by scaling the mode-set such that the positions of the outer zero-crossings of the highest-order mode match the positions of the edges of the top hat, the field reconstruction is optimised in the sense that the highest-order mode might be expected to recreate the edge without extending too far beyond it. Figure 2-17 shows mode m = 12 whose beam width parameter W_x has been chosen so as to satisfy the above condition, which required a value of $W_x = 0.3625a$.

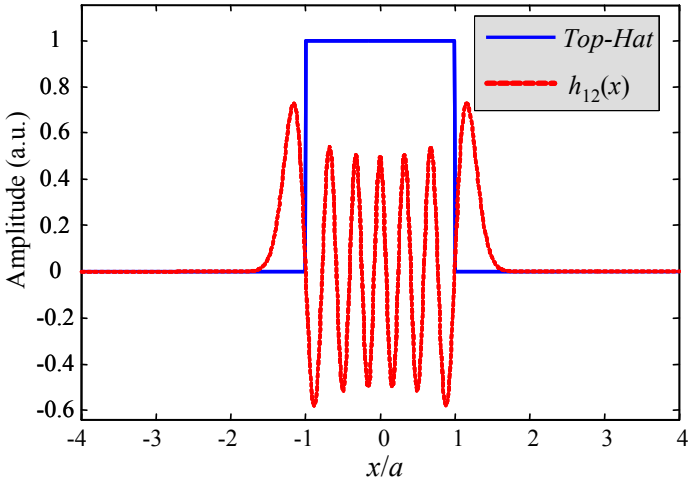


Figure 2-17. Gaussian-Hermite mode of order m = 12 scaled (by choosing an appropriate value of W_x) such that its first and last zero-crossings coincide with the vertical edges of the top hat amplitude.

Figure 2-18 shows a plot of beam width W_x values (blue curve) that correspond to mode-sets in which the first and last zero-crossings of the mode m match the position of the vertical edges of a top hat of full width $2a$. The values of W_x that produces a mode-set in which the zero-crossings of mode m correspond to positions of vertical edges is given approximately by the expression

$$W_x = \frac{a}{0.18m} \tag{2.29}$$

which is in close agreement with the calculated mode widths L_m of the higher-order modes (blue curve in Figure 2-18).

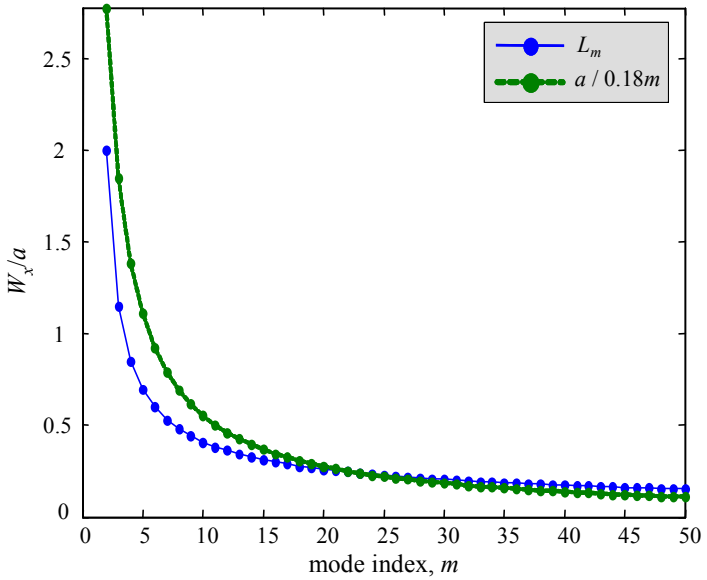


Figure 2-18. Mode width parameter W_x for various mode sets of maximum mode-order m = [2…50], that results in the first and last zero-crossings coinciding with the top-hat edges (at $x = \pm a$).

The top hat field was expanded in terms of a mode set with maximum mode-order $m_{\max}$ = 50, in which the beam width parameter W_x was chosen so as to satisfy the above zero-crossing criterion. The reconstructed top-hat amplitude distribution (Figure 2-19) shows only a single side-lobe beyond the vertical edges of the top-hat. The amplitudes of the mode coefficients (Figure 2-20) are seen to fall off smoothly and rapidly with increasing mode index m and approach a value of zero for the highest-order mode, $m_{\max}$ = 50.

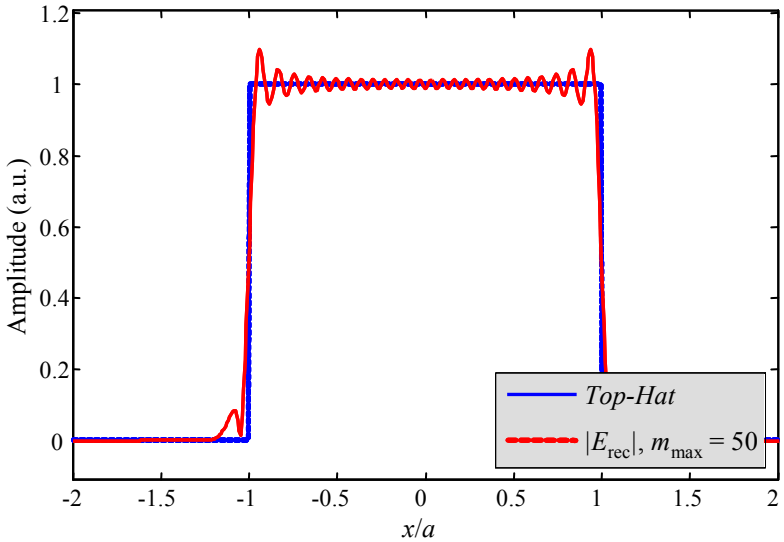


Figure 2-19. Modal reconstruction of top-hat function with a mode-set whose highest-order mode index is $m_{\max}$ = 50. The mode-set was scaled by choosing W_x such that the first and last zero-crossings of mode $m_{\max}$ coincide with the top-hat edges (at $x = \pm a$).

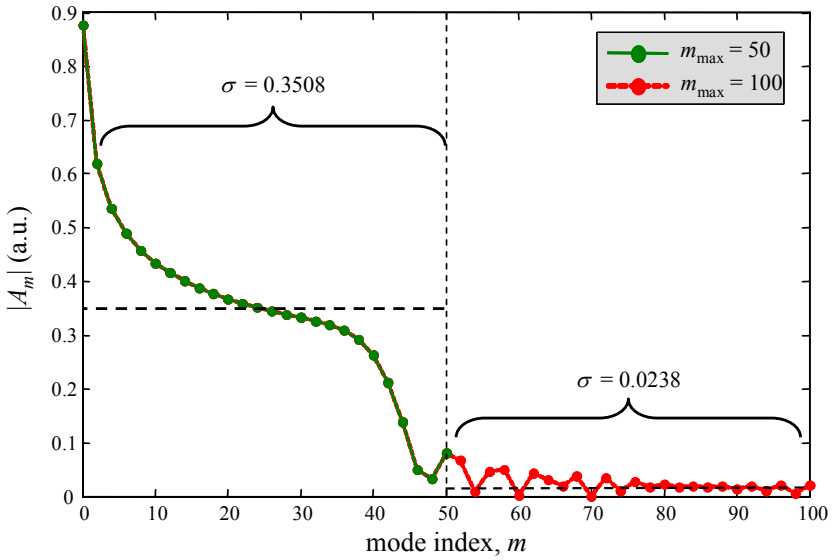


Figure 2-20. Mode coefficient amplitudes $|A_m|$ for a modal expansion of the top-hat function in which the mode-set is scaled so that the first and last zero-crossings of mode $h_{50}(x)$ coincide with the edges of the top-hat (at $x = \pm a$). The mean value σ of $|A_m|$ is 0.3508 for modes $m = 0\ldots50$ but only 0.0238 for modes $m = 50\ldots100$.

In order to evaluate the accuracy of such a finite mode expansion (i.e. using only 50 modes) we now increase the number of higher-order modes to include modes with indices up to $m = 100$ but with the mode-set scaling unchanged, i.e. using the same beam width parameter W_x. Since the first fifty modes already do a good job of reconstructing the top-hat function, we expect that the new higher-order modes ($m = 51\ldots100$) are not as important to the expansion and therefore should not contribute much to the reconstructed beam profile. This is verified by examining the values of amplitude coefficients for modes with $m = 51\ldots100$ (red curve in Figure 2-20). The mode coefficient amplitudes $|A_m|$ for all modes with $m > 50$ is less than 10% of that from the highest contributing mode (the fundamental mode). Furthermore, the average σ of amplitude coefficients $|A_m|$ for modes with $m \leq 50$ is 0.3508, whereas that for modes $m = 50\ldots100$ is only 0.0238. The amplitude distribution of the reconstructed field using the extended mode set is shown in Figure 2-21. The modes added cause some reduction in the magnitude of peak-to-peak ripples across the width of the top hat and produce sharper edges, but they also produce multiple side lobes beyond the edges of the top hat due to the fact that the modes added extend beyond the limits of the top hat function.

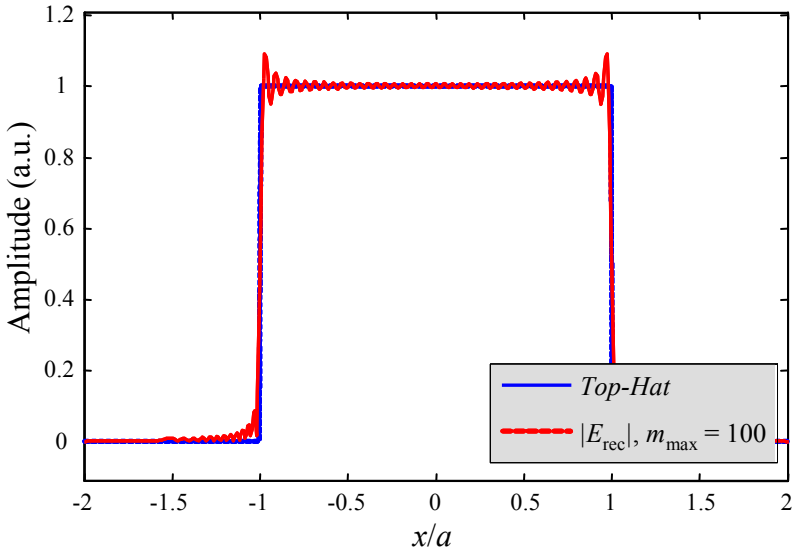


Figure 2-21. Modal reconstruction of top-hat function using a mode-set which has been scaled such that the "zero-crossings" criterion applies to the Gaussian-Hermite mode of index $m = 50$, but which also has a highest-order mode index of $m_{max} = 100$ (unlike that shown in Figure 2-19 where $m_{max} = 50$).

The method just described produces an unintended maximum just beyond the edge of the top-hat (see Figure 2-19) and indicates that if we wish to suppress this artefact we should slightly reduce the extent of the highest-order mode with respect to the top-hat. Therefore the final method of choosing a mode-set to reconstruct a top-hat field is to use a value of the beam width parameter W_x such that the outer extrema (maxima or a minimum and a maximum depending on mode symmetry) of the highest-order mode coincide with the edges of the top hat (of full-width $2a$). In other words we require that

$$D_m(W_x) = 2a \tag{2.22}$$

for the highest-order mode $m = m_{max}$, where the outer extrema separation, D_m of mode m is dependent on the beam width parameter W_x.

Figure 2-23 shows a Gaussian-Hermite mode of order $m = 10$ which has been scaled so that its outer maxima coincide with the edges of the top hat function. A value of $W_x = 0.3438a$ was needed to fit mode $h_{10}(x)$ to the top hat function in this way.

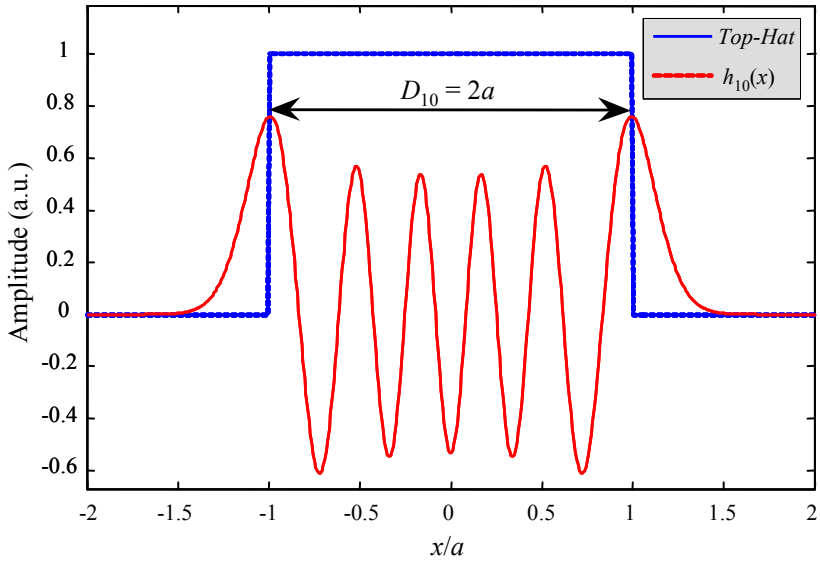


Figure 2-23. Gaussian-Hermite mode of order m = 10 that has been scaled such that the distance between its outer maxima, D_m coincide with the vertical edges of the top-hat function of full-width $2a$.

A reconstruction of the top-hat function was performed using a mode-set that was scaled so as to satisfy equation (2.22) with highest-order mode $m_{\max}$ = 50. The reconstructed amplitude distribution (Figure 2-24) is comparable to that produced using the previous scaling method (Figure 2-19) except that the extra side lobes have been removed. Unfortunately, the vertical edges now appear smoother than before.

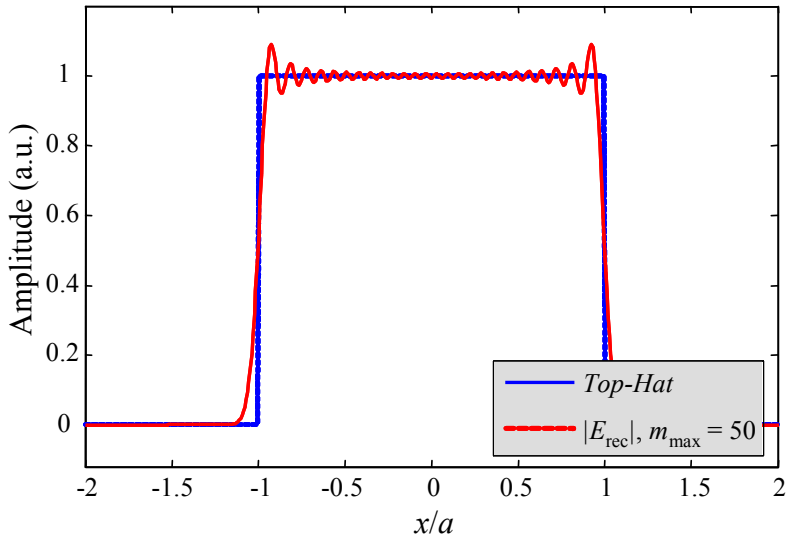


Figure 2-24. Modal reconstruction of top-hat function using a mode-set whose highest-order mode $m_{\max}$ = 50 has been scaled such that the distance between its outer maxima coincide with the vertical edges of the top-hat function (at $x = \pm a$).

The amplitude of mode coefficients $|A_m|$ shown in Figure 2-25 have a distribution similar to that of Figure 2-20 except that now the fall-off in mode amplitude occurs

slightly later, i.e. more modes contribute to the reconstructed top-hat function. As before the contribution from modes with indices above $m = 50$ is extremely small.

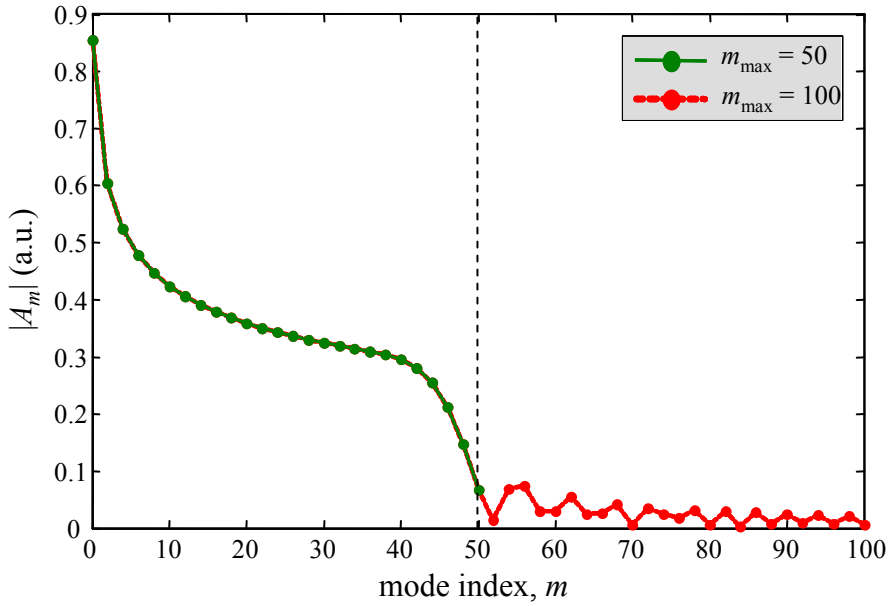


Figure 2-25. Mode coefficient amplitudes $|A_m|$ for Gaussian-Hermite modes $h_m(x)$ used to reconstruct a top-hat function for the case where the modes have been scaled such that the distance between the outer maxima of the mode of order $m = 50$ is equal to the top-hat full-width $2a$. With the mode-set scaled thus, modes of indices $m > 50$ contribute little to the reconstructed top-hat.

To illustrate the difference between all four methods of mode-set scaling discussed in this section Figure 2-26 shows the mode coefficient amplitudes $|A_m|$ for each method (for a highest-order mode $m_{max} = 50$) superimposed. For clarity the scale on the y-axis has been truncated to a value of 1 since the value of $|A_0|$ for the mode-set produced by maximising the power in the fundamental mode is approximately twice the maximum amplitude of any mode scaled using the other three methods. Maximising power in the fundamental mode is clearly the most computationally inefficient of the four methods. Maximising power in the highest-order mode is much more efficient but it too requires the use of several more modes (above the intended highest-order mode, $m_{max} = 50$) to achieve good beam reconstruction. The results produced from the remaining two methods that involve matching the dimensions (either the distance between outermost zero-crossings or between outermost maxima) of the highest-order mode to the top-hat full-width are quite similar. With both of these methods, because the value of $|A_m|$ for the intended user-defined highest-order mode m_{max} approaches zero, one can be confident of achieving good beam reconstruction using a mode-set consisting of a

maximum of (m_{max}+1) modes without needing to include any higher-order modes (as required by the two previous scaling methods).

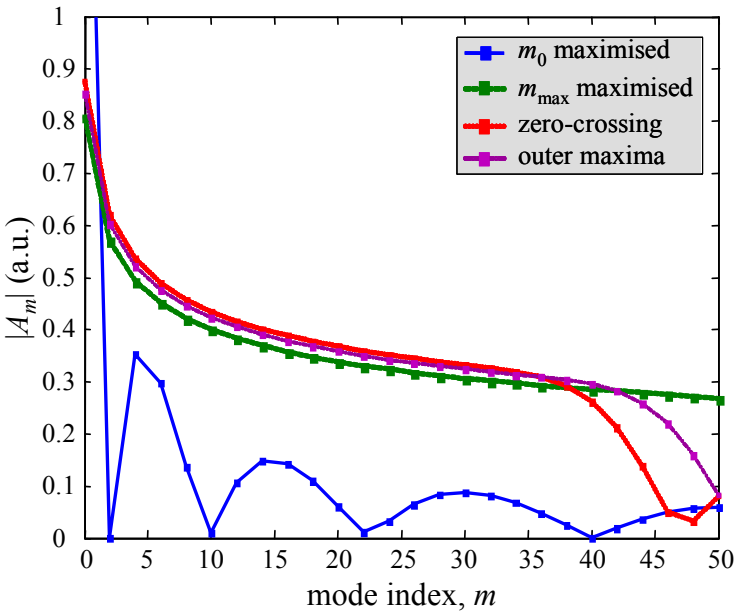


Figure 2-26. Mode coefficient amplitudes $|A_m|$ for mode-set with highest-order mode m_{max} = 50 for each of the mode-set scaling methods described above: 1) maximum power in fundamental mode, m_0; 2) maximum power in the highest-order mode, m_{max}; 3) coincidence of highest-order mode zero-crossings with top hat edges; 4) coincidence of highest-order mode outer maxima with top hat edges.

In conclusion, how one chooses to scale a mode-set (through manipulation of the mode-set beam width parameter W_x) is critical to the efficient decomposition of a given field. The goal is to reconstruct an input field with as few modes as possible to reduce computational overhead (in terms of memory and execution time), which is achieved by choosing a value of W_x appropriate to the problem at hand. For example maximising power in the fundamental mode is an inappropriate scaling option for an input field that has a high-spatial frequency content (such as a top-hat function) because a good reconstruction will only be possible by using a large number of higher-order modes. On the other hand the difficulty with scaling a mode-set to maximise power in an intended highest-order mode is that such a mode-set would not couple well to a field with low spatial frequency content (such as a pure Gaussian). An alternative method of scaling the mode-set to fit a top-hat function would be to select a value of W_x that results in the maximum power in all modes up to the intended highest-order mode m_{max}.

The direct two-dimensional analogue of a one-dimensional top hat function due to a slit is a square top hat function of width and height $2a$ due to a square aperture. In this case the decomposition can be performed separately in each transverse dimension (x and y) and the one-dimensional results cross-multiplied to give the fully two-dimensional result. Since we already know how to scale a mode-set to reconstruct a one-dimensional top hat function, by extension the two-dimensional case is trivial. For example using the outer-maxima criteria developed in the last section we can construct a one-dimensional mode-set $h_m(x;W_x)$ with highest-order mode index, $m_{\max} = 40$. Again due to symmetry only even-numbered modes are considered, i.e. $h_m(x)$ includes only 21 modes with indices m = [0, 2, 4…40]. The two-dimensional modes required for the expansion of the square top-hat function are a product of the two one-dimensional modes

$$h_{mn}(x, y; W_0) = h_m(x; W_x) \times h_n(y; W_y)$$

where $W_0 = W_x = W_y$. Since for the one-dimensional top-hat function

$$E_{\mathrm{TH}}(x) = \sum_{m=0}^{m_{\max}} A_m h_m(x; W_x)$$

therefore for the separable two-dimensional case we have

$$\begin{aligned} E_{\mathrm{TH}}(x, y) &= E_{\mathrm{TH}}(x) \cdot E_{\mathrm{TH}}(y) \\ &= \sum_{m=0}^{m_{\max}} A_m h_m(x; W_x) \sum_{n=0}^{n_{\max}} A_n h_n(y; W_y) \\ &= \sum_{m=0}^{m_{\max}} \sum_{n=0}^{n_{\max}} A_{mn} h_{mn}(x, y; W_0) \end{aligned}$$

Figure 2-27 shows the mode coefficients amplitudes $|A_{mn}|$ as well as the amplitude distribution of the reconstructed two-dimensional top-hat function. Note that only even-numbered mode coefficients are plotted in Figure 2-27(a) since the odd-numbered modes do not contribute. Plots of this type showing mode coefficient magnitudes will appear numerous times in this thesis and contain the fundamental mode located at the lower-left corner and the highest-order mode of order ($m_{\max}$, $n_{\max}$) in the upper-right corner. A horizontal (or vertical) cut through the plot of $|A_{mn}|$ would reveal a distribution resembling the one-dimensional mode coefficients $|A_m|$ shown in Figure 2-25 expect that in the previous case the highest-order mode index was $m_{\max} = 50$.

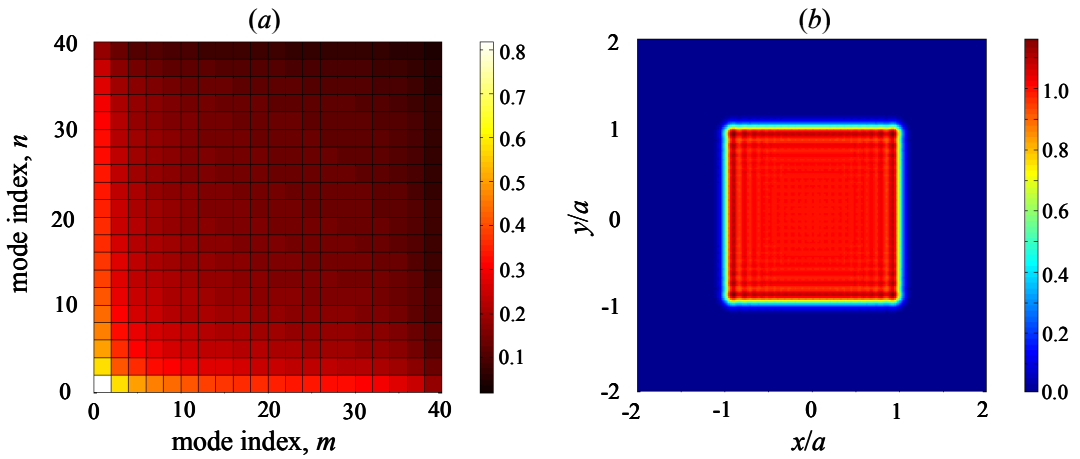


Figure 2-27. Modal expansion of a two-dimensional top-hat function. (*a*) Mode coefficient amplitudes, $|A_{mn}|$ of a 21×21 element mode-set (highest-order mode index, m_{max} = 40 but only even-numbered modes contribute) and (*b*) the amplitude distribution of the reconstructed square top hat function $E_{rec}(x,y)$. The mode-set was scaled by matching the distance between the outer maxima of the highest-order 1-D modes to the positions of the top-hat edges in both *x* and *y* directions.

For two-dimensional problems in which the input field is not separable into two one-dimensional fields, the values of mode coefficients must be determined by operating in two dimensions. Since numerical integration must be performed with sufficiently high resolution two-dimensional numerical integration becomes quite computationally intensive and execution times increases dramatically. An alternative, more efficient method for solving the overlap integral is required and will be discussed in §2.6. For now however the means by which mode coefficients A_{mn} are calculated is not important.

The solution to the two-dimensional square top hat function can be found by simply solving the one-dimensional top hat function. However a more interesting problem is that of a uniformly illuminated circular aperture, which is represented by a circular top-hat function of radius *r*. Although this problem possesses radial symmetry and so decomposition can be achieved using Laguerre modes, it is interesting to consider decomposition in terms of Gaussian-Hermite modes. Clearly since this problem is inseparable in Cartesian coordinates a full two-dimensional approach must be taken. Again a mode-set with highest-order mode indices $m_{max} = n_{max} = 40$ was used. Each two-dimensional mode $h_{mn}(x, y)$ was constructed from the one-dimensional modes $h_m(x)$ and $h_n(y)$. Because of the even symmetry of the target field only even-numbered modes were used.

For the first trial reconstruction of the circular top-hat function the mode-set scaling factor W_0 was chosen so as to maximise power in the fundamental mode $h_{0,0}(x,$

y), i.e. to produce a maximum value of $|A_{0,0}|$. The results of this reconstruction, using a mode-scaling factor of $W_0 = 0.89r$ and which results in a value of 96.06% for the intensity correlation between the ideal and reconstructed field intensities, are shown in Figure 2-28. The reconstructed intensity distribution in Figure 2-28(b) is slowly undulating within the circular aperture and across the aperture edges the intensity drops off smoothly, rather than sharply. Furthermore, the maximum side-lobe level (outside the aperture) is ~20dB. The low-quality reconstruction occurs because an insufficient number of higher-order modes were used. The mode-map in Figure 2-28(a) exhibits an oscillatory pattern similar to what was observed when the same approach was used to reconstruct a one-dimensional top-hat function. As in that example, an improvement in reconstruction quality using the same value of W_0 is only possible by including additional higher-order modes.

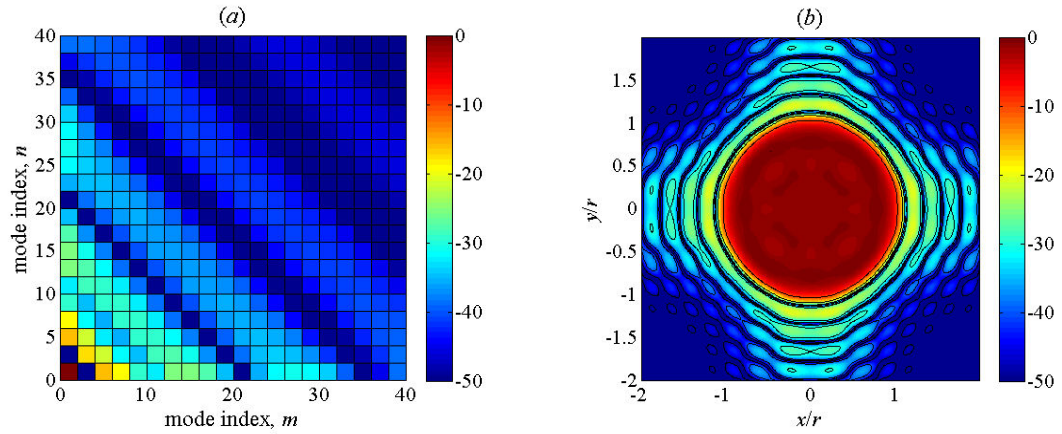


Figure 2-28. (a) Log-scale plot of GBM coefficients $|A_{mn}|$ to decompose a circular top-hat function, whose reconstructed intensity profile is shown in (b), for the case where mode-set scaling factor W_0 was chosen to maximise $|A_{0,0}|$ - the fundamental mode coefficient. The value for W_0 (= $0.89r$) results in 96.06% correlation between target and reconstructed field intensities. Maximum side-lobe level is ~20 dB.

A search was undertaken to find the mode-set scaling factor that yields optimum reconstruction (maximum intensity correlation between target and reconstructed field intensities). A maximum intensity correlation of 98.74% was found to occur for a value of $W_0 \approx 0.21r$. Figure 2-29 shows the mode-map and reconstructed field intensity for optimum mode-set scaling. The reconstructed field intensity shown in Figure 2-29(b) exhibits much sharper edges than before as well as an almost uniform intensity level across the circular aperture. The side-lobes are at a much lower level (maximum of ~28 dB) than those in Figure 2-28(b). Also, many of the mode coefficients in Figure 2-29(a) contribute relatively little to the reconstruction. In fact by removing all modes with

mode coefficient intensities below 30 dB an intensity correlation of 98.14% can be achieved using only one third of the total number of modes.

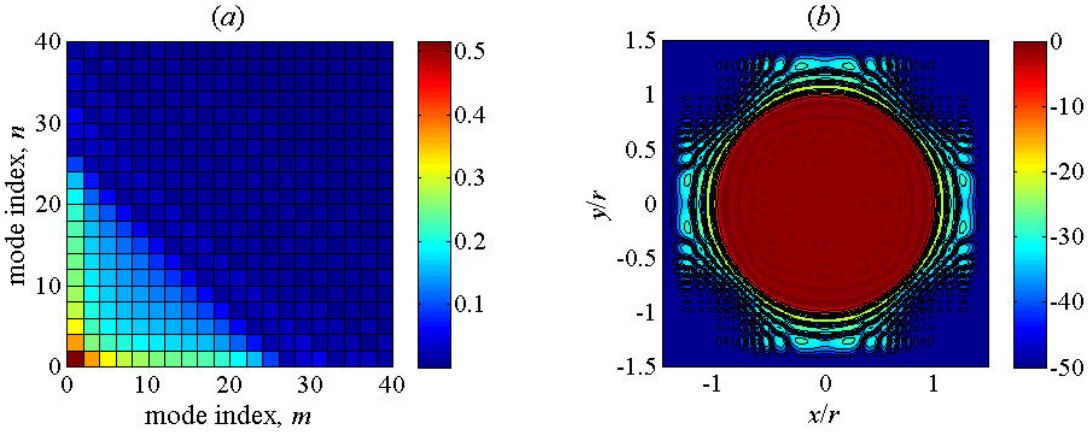


Figure 2-29. (*a*) Linear-scale plot of the GBM coefficients $|A_{mn}|$ and (*b*) the corresponding log-scale plot of the optimum reconstructed amplitude profile, i.e. for a mode-set with a scaling factor W_0 (= 0.21*r*) that results in the maximum intensity correlation (of 98.74%) between target and reconstructed field intensities. Again contour levels in (*b*) are set at intervals of 10 dB. The maximum side-lobe level now occurs at approximately 28 dB.

2.5 The ABCD Matrix Method

While a quasi-optical beam does not have a focus, as exists for rays in geometrical optics where radiation is focused to a point, it does have a waist position associated with it which can be transformed to another waist using lenses and mirrors. These components are used in quasioptical systems to change the radius of curvature of the beam to produce a diverging beam so as to confine the beam of radiation and so avoid the monotonic growth of the Gaussian beam that would otherwise occur. When analysing an optical system it is important to be able to keep track of the Gaussian beam parameters: its radius *W* and radius of curvature *R* at various planes in the system. It is also important to be able to know the associated phase slippages between modes in order to analyse diffraction effects.

The path of paraxial rays through a linear geometrical system can be determined by analysing the effect that various elements have on the radius of curvature of the geometrical optical beam. A useful tool for analysing beam coupling between sources, detectors, lenses and mirrors is the ray transfer or ABCD matrix. Although developed for use with linear geometrical optics systems this technique can be applied to quasioptical systems in which a beam is approximated as a finite sum of Gaussian beam

modes by incorporating the complex beam parameter, q in place of the radius of curvature of a geometrical optics beam. Here we emphasize the fundamental Gaussian beam mode. However, since the behaviour of the beam radius W and radius of curvature R is the same for all Gaussian beam modes, the formulas derived here are applicable to all higher order modes. Apart from having a more complex transverse variation, the only significant differences between the higher-order modes and the fundamental mode is in the axial phase shift, which is mode-dependent and varies as a function of z. A practical difference is that the effective mode size increases with mode number, so truncation effects will be greater for higher-order modes.

In geometrical optics an ideal spherical wavefront with radius of curvature R can be viewed as a collection of rays emanating from a point source. Each ray is characterised by its off-axis position r and slope r' (the angle that the ray makes with the axis of propagation) and within the paraxial approximation is related to the radius of curvature R by

$$R = \frac{r}{\tan r'} \approx r/r' \tag{2.30}$$

If a spherical wavefront (with radius of curvature R_{in}) passes through some paraxial system, then the wavefront that emerges at the output of the system is also spherical with radius of curvature R_{out}. The effect that propagation through an optical system has on a single ray is defined by a pair of linear equations that relates input position and slope to those at the output to the system

$$r_{\text{out}} = A \cdot r_{\text{in}} + B \cdot r'_{\text{in}} \tag{2.31}$$

$$r'_{\text{out}} = C \cdot r_{\text{in}} + D \cdot r'_{\text{in}} \tag{2.32}$$

The coefficients A, B, C and D characterise the paraxial beam transforming properties of the particular optical system. Using equations (2.31) and (2.32) we can derive an expression relating the input and output radii of curvature as

$$R_{\text{out}} = \frac{A \cdot R_{\text{in}} + B}{C \cdot R_{\text{in}} + D} \tag{2.33}$$

Alternatively this system of linear equations can be represented in matrix form as

$$\begin{bmatrix} r_{out} \\ r'_{out} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} r_{in} \\ r'_{in} \end{bmatrix} \tag{2.34}$$

Thus the transformation performed by the optical system from the input beam (with radius of curvature R_{in}) to the output beam (with radius of curvature R_{out}) is characterised by a 2×2 ray transfer matrix **M** with entries A, B, C and D.

The ABCD matrix method is applicable to many optical systems from simple systems, such as the interface between two media, to those containing multiple components. Each element in the system is represented by a corresponding 2×2 matrix that describes the beam transforming effect it has on an input beam. If the system consists of several cascaded elements with corresponding matrices $\mathbf{M}_1$, $\mathbf{M}_{2,\ldots,}$ $\mathbf{M}_n$ the transformation due to the system as a whole is then given by the system matrix

$$\mathbf{M}_{\mathrm{TOT}} = \mathbf{M}_n \cdot \mathbf{M}_{n-1\ldots}\mathbf{M}_2 \cdot \mathbf{M}_1 \tag{2.35}$$

In order to be able to extend the ABCD matrix technique to represent beam transformations by a quasioptical system we now replace the geometrical radius of curvature R in equation (2.33) with a complex beam parameter q (see Appendix A-3)

$$\frac{1}{q(z)} = \frac{1}{R(z)} - i\frac{\lambda}{\pi W^2(z)} \tag{2.36}$$

to give the equivalent Gaussian beam formalism

$$q_{\mathrm{out}} = \frac{A \cdot q_{\mathrm{in}} + B}{C \cdot q_{\mathrm{in}} + D} \tag{2.37}$$

Given an input Gaussian beam with radius W_{in} and radius of curvature R_{in}, the input complex parameter q_{in} is defined as

$$\frac{1}{q_{\mathrm{in}}} = \frac{1}{R_{\mathrm{in}}} - i\frac{\lambda}{\pi {W_{\mathrm{in}}}^2} \tag{2.38}$$

From equation (2.36) the output Gaussian beam radius W_{out} and radius of curvature R_{out} are then given by

$$W_{\mathrm{out}} = \left[\frac{\lambda}{\pi\ \mathrm{Im}(-1/q_{\mathrm{out}})}\right]^{0.5} \tag{2.39}$$

$$R_{\mathrm{out}} = \left[\mathrm{Re}\left\{\frac{1}{q_{\mathrm{out}}}\right\}\right]^{-1} \tag{2.40}$$

or, in terms of input Gaussian beam parameters (W_{in} and R_{in}) and the system matrix coefficients (A, B, C and D), as

$$W_{\mathrm{out}} = \left[\frac{-\lambda}{\pi\ \mathrm{Im}\left\{\dfrac{C + D/R_{\mathrm{in}} - iD\lambda/\pi {W_{\mathrm{in}}}^2}{A + B/R_{\mathrm{in}} - iB\lambda/\pi\ {W_{\mathrm{in}}}^2}\right\}}\right]^{0.5} \tag{2.41}$$

$$R_{\mathrm{out}} = \mathrm{Re}\left\{\frac{C + D/R_{\mathrm{in}} - iD\lambda/\pi {W_{\mathrm{in}}}^2}{A + B/R_{\mathrm{in}} - iB\lambda/\pi {W_{\mathrm{in}}}^2}\right\} \tag{2.42}$$

The concepts developed so far describe the beam transformation that is imparted on a fundamental Gaussian beam mode. This representation is quite adequate for describing the radiation patterns from a variety of antenna types and feeds. The behaviour of a quasioptical system can thus be satisfactorily described by examining the transformations induced on the fundamental Gaussian beam mode representing the radiation pattern from a particular antenna. However the accurate description of more complicated systems requires a multi-mode Gaussian beam approach. Extending the ABCD matrix technique to high order modes is easily accommodated since although effective the effective mode size increases with mode number, the mathematical expressions for the higher-order modes share the same waist radius $W(z)$ and radius of curvature $R(z)$ as the fundamental mode and vary as a function of propagation distance z in exactly the same manner. The only significant difference between the fundamental and higher order modes appears in the phase slippage term $\phi_{mn}(z)$, whose variation with z is mode number dependent – see Eq. (2.17). For the fundamental mode the transformation induced on the phase slippage by an ideal phase transformer is given in [2.14] by

$$\Delta\phi_0 = (\phi_{0,\text{out}} - \phi_{0,\text{in}}) = -\text{Arg}\{A + B(1/q_{\text{in}})\} \tag{2.43}$$

where Arg{} denotes the argument of the bracketed quantity and $\phi_{0,\text{in}}$ and $\phi_{0,\text{out}}$ are the phase slippage terms associated with the fundamental Gaussian beam mode at the input and output planes, respectively, in other words the $\tan^{-1}(z/z_R)$ term in equations (2.17) and (2.18). Expanding equation (2.43) in terms of the real and imaginary parts of q_{in} yields

$$\Delta\phi_0 = -\tan^{-1}\left(\frac{\text{Im}\{A + B(1/q_{\text{in}})\}}{\text{Re}\{A + B(1/q_{\text{in}})\}}\right) = \tan^{-1}\left(\frac{B(\lambda/\pi W_{in}^2)}{A + B(1/R_{\text{in}})}\right) \tag{2.44}$$

The phase slippage for higher-order modes is then calculated by scaling ϕ_{out} with the relevant mode indices to yield

$$\phi_{mn,\text{out}} = [m + n + 1]\phi_{\text{out}} \tag{2.45}$$

for the two-dimensional Gauss-Hermite modes[†] h_{mn} defined in §2.3.

Hence the beam transformation imparted on a multi-moded field $E_{\text{in}} = \Sigma A_{mn}\Psi_{mn}(x,y;W_{\text{in}};R_{\text{in}};\phi_{\text{in}})$ by a quasioptical system is determined by simply re-calculating the modal summation but using modes Ψ_{mn} that are now defined in terms of

[†] For one-dimensional Hermite modes h_m the phase slippage is $\phi_{m,\text{out}} = [\text{m} + ½]\phi_{\text{out}}$

the output Gaussian beam mode parameters W_{out}, R_{out} and $\phi_{mn,out}$ as given by equations (2.41), (2.42) and (2.45).

The phase slippage tells us how the higher-order modes interfere to change the form of a beam. In particular several extreme cases are worthy of note. Firstly, if $B = 0$ from equation (2.43) the phase slippage, $\Delta\phi_0 = 0$ or $-\pi$. In other words the output image of the input beam is real and inverted. The other important case occurs when $A = 0$ and $1/R_{in} = 0$, in which case $1/q_{in} = -i(\lambda/\pi Win^2)$ and so $\Delta\phi_0 = -\text{Arg}\{B/q_{in}\} = \pi/2$. Thus the output image produced by such a system is equivalent to the image of the far-field of the input waist, i.e. the Fourier Transform of the input field. The same result occurs when $A = -B(1/R_{in})$.

2.5.1 ABCD Matrices of commonly encountered components

The following is a description of the ray transfer matrices (RTM) or ABCD matrices associated with a number of commonly encountered quasioptical components.

Propagation through a Uniform Medium

The simplest and most fundamental ABCD matrix is that representing propagation through a medium of uniform refractive index, such as free-space. A ray with initial off-axis position r_{in} and slope r'_{in} (= dr/dz) will have, after propagating a distance L, the same slope but an off-axis position proportional to r'_{in}. Thus in the paraxial limit

$$r_{out} = r_{in} + Lr'_{in}$$
$$r'_{out} = 0 + r'_{in}$$

and the corresponding ray transfer matrix is

$$\mathbf{M}_{dist} = \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \tag{2.46}$$

irrespective of refractive index.

Curved Interface

Another matrix of fundamental importance is that describing a curved interface between two media with different refractive indices n_1 and n_2

$$\mathbf{M}_{\text{curved interface}} = \begin{bmatrix} 1 & 0 \\ \frac{n_2 - n_1}{n_2 R} & \frac{n_1}{n_2} \end{bmatrix} \tag{2.47}$$

with $R < 0$ for a surface concave to the left and $R \to \infty$ for a flat surface.

Thin Lens

From the above two basic matrices more complex and useful optical systems can be constructed. For example a thin lens can be treated as two curved interfaces side-by-side, with the thickness of the lens – the separation between the curved surfaces – neglected. If the lens material has refractive index n_2 and is embedded in material of refractive index n_1, the thin lens matrix is computed by multiplying a matrix for the first curved surface (on the left side of the lens, with curvature R_1) by one for the second surface (on the right side of the lens, with curvature R_2), i.e.

$$\mathbf{M}_{\text{thin lens}} = \begin{bmatrix} 1 & 0 \\ \frac{n_1 - n_2}{n_1 R_2} & \frac{n_2}{n_1} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ \frac{n_2 - n_1}{n_2 R_1} & \frac{n_1}{n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{n_1 - n_1}{n_1}\left(\frac{1}{R_2} - \frac{1}{R_1}\right) & 1 \end{bmatrix} \tag{2.48}$$

Since the focal length f of a thin lens is described by the *Lens Makers Formula* [2.9] as

$$\frac{1}{f} = \left(\frac{n_2 - n_1}{n_1}\right)\left(\frac{1}{R_2} - \frac{1}{R_1}\right) \tag{2.49}$$

equation (2.48) can be rewritten as

$$\mathbf{M}_{\text{thin lens}} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \tag{2.50}$$

where the sign convention used is that a positive value for f indicates a converging, or biconvex, lens with $R_1 < 0$ and $R_2 > 0$. A plano-convex lens having one convex surface (with radius of curvature R) and a flat surface (with infinite radius of curvature) is also described by matrix (2.50) since the focal length of such a lens is given by $1/f = (n_2 - n_1)/n_1 R$.

Thick Lens

The only difference between a thin lens and a thick lens is that in the latter the axial thickness of the lens is taken into account, i.e. after encountering the first curved surface of the lens, the beam propagates a distance d to the second surface. Therefore the ray

transfer matrix must include an extra matrix representing propagation between the two surfaces between the matrices representing the curved surfaces as follows

$$\mathbf{M}_{\text{thick lens}} = \begin{bmatrix} 1 & 0 \\ \frac{n_1-n_2}{n_1 R_2} & \frac{n_2}{n_1} \end{bmatrix} \cdot \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ \frac{n_2-n_1}{n_2 R_1} & \frac{n_1}{n_2} \end{bmatrix} = \begin{bmatrix} 1+\frac{(n_2-n_1)d}{n_2 R_1} & \frac{n_1 d}{n_2} \\ -\frac{1}{f}-\frac{(n_2-n_1)^2 d}{n_1 n_2 R_1 R_2} & 1+\frac{(n_1-n_2)d}{n_2 R_2} \end{bmatrix}$$

where again the refractive indices of the lens material and the surrounding material are n_2 and n_1, respectively and the lens makers formula holds for f.

Spherical and Ellipsoidal Mirror

While mirrors may introduce beam distortion, which is especially true of fast mirrors and those with off-axis configuration, ideal mirrors can be approximated reasonably well with thin lenses, since both transform the beam's radius of curvature R without affecting its radius W. The beam transformation due to a spherical or ellipsoidal mirror can thus be estimated using the ABCD matrix for a thin lens using matrix (2.50). A spherical mirror employed in normal incidence is equivalent to a thin lens of focal length $f = R/2$, so its matrix is

$$\mathbf{M}_{\text{spherical mirror}} = \begin{bmatrix} 1 & 0 \\ -2/R & 1 \end{bmatrix} \tag{2.51}$$

An ellipsoidal mirror can be treated as a pair of thin lenses located at the centre of an ellipsoidal surface section that act in series to bring a beam from a point source at one focus of the ellipse to a quasi point-like image of the source at the other. The first lens (of focal length R_1) collimates the beam from one focal point to produce a parallel beam at the mirrors centre. The second lens (of focal length R_2) then focuses this beam to the second ellipse focus. Focal lengths R_1 and R_2 are the distances from the point of reflection on the ellipse centre to its two foci. The ABCD matrix is then formed from the product of two thin lens matrices of appropriate focal lengths

$$\mathbf{M}_{\text{ellipsoidal mirror}} = \begin{bmatrix} 1 & 0 \\ -1/R_2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -1/R_1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\left(\frac{1}{R_1}+\frac{1}{R_2}\right) & 1 \end{bmatrix} \tag{2.52}$$

However, since the focal length of an ellipsoidal mirror is given by

$$\frac{1}{f} = \frac{1}{R_1} + \frac{1}{R_2} \tag{2.53}$$

matrix (2.52) reduces to that for a single thin lens of focal length f.

2.5.2 The Gaussian Beam Telescope

An important quasi-optical system is the Gaussian beam telescope (GBT), as illustrated in Figure 2-30. It consists of a pair of focusing elements (mirrors or lenses with focal lengths f_1 and f_2) that are separated by the sum of their focal lengths ($f_1 + f_2$). The input beam waist is located at z_{in}, (a distance d_{in} in front of lens L_1) and the output beam waist at z_{out} (a distance d_{out} beyond lens L_2).

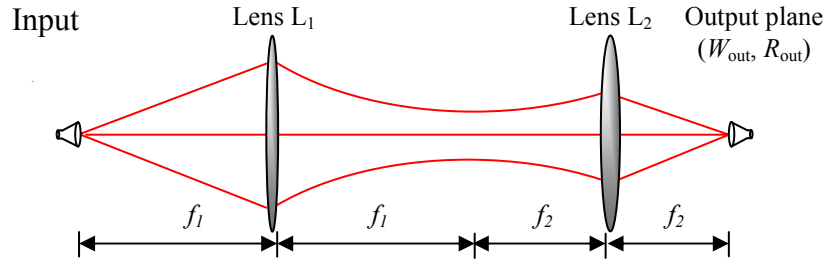


Figure 2-30. A Gaussian beam telescope consisting of two thin-lenses L_1 and L_2 of focal lengths f_1 and f_2 (where in this case $f_1 \neq f_2$) that are separated by ($f_1 + f_2$). The two horn antennas (left and right) represent the positions of the input and output Gaussian beam waist positions which are located at distance of $d_{in} = f_1$ in front of lens L_1 and a distance $d_{out} = f_2$ beyond lens L_2, respectively.

In terms of ABCD matrices the system can be described by combining two thin lens matrices

$$\mathbf{L}_1 = \begin{bmatrix} 1 & 0 \\ -1/f_1 & 1 \end{bmatrix} \quad , \quad \mathbf{L}_2 = \begin{bmatrix} 1 & 0 \\ -1/f_2 & 1 \end{bmatrix}$$

to represent lenses L_1 and L_2 (of focal lengths f_1 and f_2) with three free-space propagation matrices

$$\mathbf{P}_1 = \begin{bmatrix} 1 & d_{in} \\ 0 & 1 \end{bmatrix} \quad , \quad \mathbf{P}_2 = \begin{bmatrix} 1 & (f_1 + f_2) \\ 0 & 1 \end{bmatrix} \quad , \quad \mathbf{P}_3 = \begin{bmatrix} 1 & d_{out} \\ 0 & 1 \end{bmatrix}$$

to account for propagation from input to output planes via the two lenses. By cascading these five matrices the system matrix of the GBT is then given by

$$\mathbf{M}_{GBT} = \mathbf{P}_3 \cdot \mathbf{L}_2 \cdot \mathbf{P}_2 \cdot \mathbf{L}_1 \cdot \mathbf{P}_1$$

When the input and output planes are situated so as to coincide with the focal planes of lenses L_1 and L_2 so that $d_{in} = f_1$ and $d_{out} = f_2$, this results in the matrix

$$\mathbf{M}_{GBT} = \begin{bmatrix} -f_2/f_1 & 0 \\ 0 & -f_1/f_2 \end{bmatrix} \tag{2.54}$$

The significance of the GBT is that the output Gaussian beam parameters W_{out} and R_{out} are wavelength independent and depend solely on the ratio of focal lengths f_1 and f_2, being given by

$$W_{out} = W_{in}\left(\frac{f_2}{f_1}\right) \tag{2.55}$$

$$R_{out} = R_{in}\left(\frac{f_2}{f_1}\right)^2 \tag{2.56}$$

From equation (2.55) we see that magnification (the ratio of output to input beam radii) of the GBT is simply equal to

$$\square = f_2/f_1 \tag{2.57}$$

Furthermore the output beam waist position d_{out} is also wavelength independent and occurs at a distance of

$$d_{out} = \frac{f_2}{f_1}\left(f_1 + f_2 - \frac{f_2}{f_1}d_{in}\right) \tag{2.58}$$

So for $d_{in} = f_1$, $d_{out} = f_2$. These wavelength-independent properties of the GBT make it particularly useful for broadband applications. For a GBT with $f_1 = f_2 = f$, the system matrix reduces to the identity matrix and system yields unit magnification with $d_{out} = d_{in} = f$. In other words the beam produced at the output plane is an image of the input beam. Such a system features in the transmission imaging experiments reported in Chapter 3.

Figure 2-31 shows how the parameters W, R and ϕ_m of a multi-mode Gaussian beam vary through a GBT. In this example the two thin lenses were chosen to have focal lengths f_1 = 350mm and f_2 = 500mm. At the intermediate focal plane between the two lenses (a propagation distance of $z = 2f_1$ = 700mm from the input plane) the beam has a waist position – the radius of curvature R going from $-\infty$ to $+\infty$. The ray transfer matrix at this plane is given by

$$\begin{bmatrix} 1 & f_1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -1/f_1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & d_{in} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & f_1 \\ -1/f_1 & 0 \end{bmatrix}$$

since $d_{in} = f_1$. The input plane is defined to be at the input beam waist position with $R_{in} = +\infty$ and because $A = 0$ in the system matrix above the phase slippage of the fundamental Gaussian beam mode at the intermediate focal plane is $\Delta\phi_0 = \pi/2$. In other words the intermediate focal plane corresponds to the Fourier plane of the input beam. Similarly since the output plane (located at $z = (2f_1 + 2f_2)$ = 1700 mm in Figure 2-31) is defined at the output beam waist position, it acts as the Fourier plane of the intermediate focal plane. Thus, besides magnification, the image produced at the output plane has the same

form as that at the input plane, which is verified since because $B = 0$ in the GBT matrix the fundamental beam mode phase slippage at the output plane is $\Delta\phi_0 = \pi$. At the output plane the higher-order modes have the same relative phase shifts as at the input plane so the output and input plane images are the same.

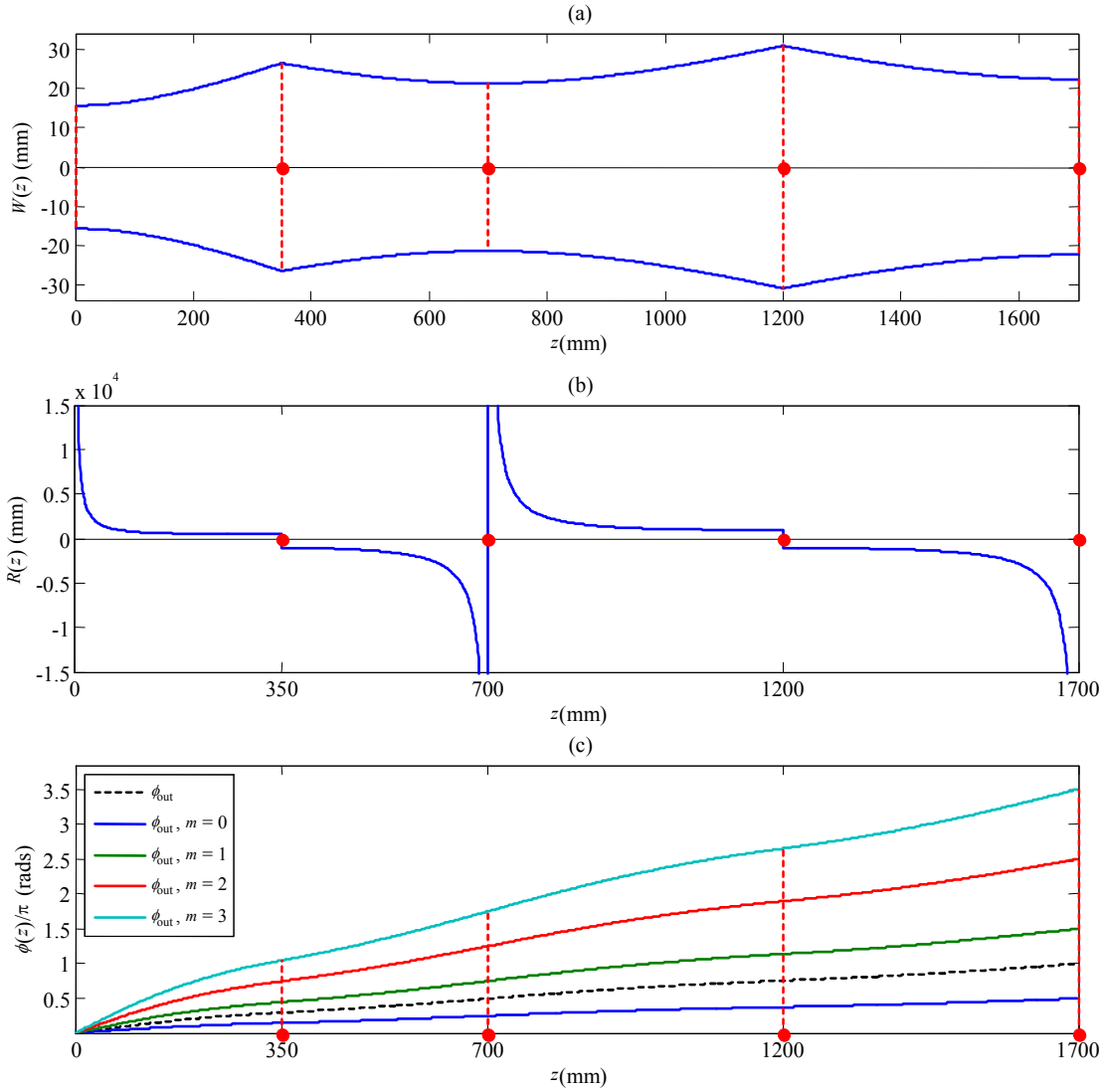


Figure 2-31. Variation of the Gaussian beam mode parameters radius $W(z)$, radius of curvature $R(z)$ and phase slippage $\phi(z)$ of four Gauss-Hermite modes ($m = 0 \ldots 3$) when propagated through a Gaussian Beam Telescope consisting of two thin lenses with focal lengths $f_1 = 350$mm and $f_2 = 500$mm. The values of W and R are the same for all four modes.

2.6 Singular Value Decomposition in GBMA

As we have seen a known field $E(x,y,z_0)$, defined at a plane z_0, can be approximated in terms of a set of Gaussian Beam modes $\psi_{mn}(x, y, z_0)$ as

$$E(x, y, z_0) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} A_{mn} \psi_{mn}(x, y, z_0)$$

where the relative contribution from each mode is specified by the values of the mode coefficients A_{mn}. One method of calculating values for A_{mn} is to numerically evaluate the overlap integral

$$A_{mn} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} E(x, y; z)[\psi_{mn}(x, y, z)]^* \, dxdy$$

However since both the field and beam modes must be densely sampled (below the Nyquist rate of $\lambda/2$) to avoid aliasing and obtain convergence this calculation is computationally intensive. An alternative and faster approach based on least-squares curve fitting that uses Singular Value Decomposition (SVD) is now discussed.

The GBM-approximated field value at the point (x_i, y_j) is given by

$$E(x_i, y_j, z_0) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} A_{mn} \psi_{mn}(x_i, y_i, z_0)$$

which, in expanded form is

$$E(x_i,y_j) = \psi_{0,0}(x_i,y_j)A_{0,0} + \psi_{1,0}(x_i,y_j)A_{1,0} + \ldots + \psi_{m,0}(x_i,y_j)A_{m,0} + \ldots + \psi_{m,n}(x_i,y_j)A_{m,n}$$

The values of field $E(x, y)$ at all discretely sampled points in the (x, y) plane are then

$$\begin{aligned}
E(x_1, y_1) &= \psi_{0,0}(x_1,y_1)A_{0,0} + \psi_{1,0}(x_1,y_1)A_{1,0} + \ldots + \psi_{m,0}(x_1,y_1)A_{m,0} + \ldots + \psi_{m,n}(x_1,y_1)A_{m,n} \\
E(x_2, y_1) &= \psi_{0,0}(x_2,y_1)A_{0,0} + \psi_{1,0}(x_2,y_1)A_{1,0} + \ldots + \psi_{m,0}(x_2,y_1)A_{m,0} + \ldots + \psi_{m,n}(x_2,y_1)A_{m,n} \\
&\vdots \qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\
E(x_p,y_1) &= \psi_{0,0}(x_p,y_1)A_{0,0} + \psi_{1,0}(x_p,y_1)A_{1,0} + \ldots + \psi_{m,0}(x_p,y_1)A_{m,0} + \ldots + \psi_{m,n}(x_p,y_1)A_{m,n} \\
&\vdots \qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\
E(x_p,y_q) &= \psi_{0,0}(x_p,y_q)A_{0,0} + \psi_{1,0}(x_p,y_q)A_{1,0} + \ldots + \psi_{m,0}(x_p,y_q)A_{m,0} + \ldots + \psi_{m,n}(x_p,y_q)A_{m,n}
\end{aligned}$$

where p and q are the number of discrete samples in x and y directions and the mode-set contains $m \times n$ modes. The above is thus a linear system with pq equations in mn unknowns and so can be written in matrix form as

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_{pq} \end{bmatrix} = \begin{bmatrix} \psi_{1,1} & \psi_{1,2} & \cdots & \psi_{1,mn} \\ \psi_{2,1} & \psi_{2,2} & \cdots & \psi_{2,mn} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{pq,1} & \psi_{pq,2} & \cdots & \psi_{pq,mn} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_{mn} \end{bmatrix} = \mathbf{\Psi A} \tag{2.59}$$

where $\mathbf{\Psi}$ is a *pq*-by-*mn* matrix representing the set of modes $\psi_{mn}(x, y, z_0)$, $\mathbf{A}$ is a *mn*-by-1 column vector whose entries correspond to the unknown mode coefficients A_{mn} and $\mathbf{E}$ is a *pq*-by-1 column vector whose entries are the discretely sampled values of the reconstructed field – an approximation to the known field $E(x, y, z_0)$. Ordinarily a system of linear equations (assuming necessary linear independence and that $mn = pq$) is solved by multiplying on both sides by the inverse of the matrix containing the system coefficients (in this case $\mathbf{\Psi}$) as

$$\mathbf{\Psi}^{-1}\mathbf{E} = (\mathbf{\Psi}^{-1}\mathbf{\Psi})\mathbf{A} = \mathbf{I}\mathbf{A} = \mathbf{A}$$

where $\mathbf{I}$ is the *mn*-by-*mn* identity matrix. However because here $pq > mn$, $\mathbf{\Psi}$ is rectangular and therefore its inverse, $\mathbf{\Psi}^{-1}$ does not exist. A system with more equations (pq) than variables (mn) is said to be overdetermined and as such has no exact solution. However an approximate solution for $\mathbf{A}$ can be found by using the optimisation technique of linear least squares fitting, which attempts to minimise the Euclidian norm squared of the residual $\mathbf{\Psi A}-\mathbf{E}$, which is given by

$$\|\mathbf{\Psi A}-\mathbf{E}\|^2 = ([\mathbf{\Psi A}]_1-\mathbf{E}_1)^2 + ([\mathbf{\Psi A}]_2-\mathbf{E}_2)^2 + \ldots + ([\mathbf{\Psi A}]_{mn}-\mathbf{E}_{mn})^2 \tag{2.60}$$

where $[\mathbf{\Psi A}]_i$ is the i^{th} entry of the column vector $\mathbf{\Psi A}$. Since the dot product of two real-valued vector columns $\mathbf{a}$ and $\mathbf{b}$ can be written as $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^{\mathrm{T}}\mathbf{b}$, where $\mathbf{T}$ indicates the transpose, equation (2.60) can be written as

$$\begin{aligned} \|\mathbf{\Psi A}-\mathbf{E}\|^2 &= (\mathbf{\Psi A}-\mathbf{E})^{\mathrm{T}}(\mathbf{\Psi A}-\mathbf{E}) = (\mathbf{\Psi A})^{\mathrm{T}}\mathbf{\Psi A} - (\mathbf{\Psi A})^{\mathrm{T}}\mathbf{E} - (\mathbf{E})^{\mathrm{T}}(\mathbf{\Psi A}) + (\mathbf{E})^{\mathrm{T}}\mathbf{E} \\ &= \mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}\mathbf{A}^2 - 2\mathbf{\Psi}^{\mathrm{T}}\mathbf{E}\mathbf{A} \end{aligned}$$

the minimum of which is found when the derivative with respect to $\mathbf{A}$ equals zero, ie

$$\mathbf{\Psi}^{\mathrm{T}}(\mathbf{\Psi A}) - \mathbf{\Psi}^{T}\mathbf{E} = 0$$

or as a system of linear equations

$$(\mathbf{\Psi}^{T}\mathbf{\Psi})\mathbf{A} = \mathbf{\Psi}^{T}\mathbf{E}$$

which is the normal system associated with the system $\mathbf{\Psi A} = \mathbf{E}$, the normal equations of which have the unique solution given by

$$\mathbf{A} = (\mathbf{\Psi}^{T}\mathbf{\Psi})^{-1}\mathbf{\Psi}^{T}\mathbf{E} = \mathbf{\Psi}^{+}\mathbf{E} \tag{2.61}$$

where the pseudo-inverse of $\mathbf{\Psi}$ is

$$\mathbf{\Psi}^{+} = (\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi})^{-1}\mathbf{\Psi}^{\mathrm{T}}$$

When $\boldsymbol{\Psi}$ is complex-valued the conjugate-transpose replaces the transpose, i.e. $\boldsymbol{\Psi}^* = \boldsymbol{\Psi}^{\mathrm{T}}$ and the pseudoinverse is written as

$$\boldsymbol{\Psi}^{+} = (\boldsymbol{\Psi}^{*}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^{*} \tag{2.62}$$

Note that the square matrix $(\boldsymbol{\Psi}^{*}\boldsymbol{\Psi})$ is only invertible if $\boldsymbol{\Psi}$ has full column rank, i.e. if rank$\boldsymbol{\Psi} = m$. This criterion is satisfied for the matrix $\boldsymbol{\Psi}$ since all columns correspond to different Gaussian-Hermite modes and these are all linearly independent.

The pseudo-inverse $\mathbf{S}^{+}$ (in particular we consider the Moore-Penrose generalised inverse [2.15]) of an m-by-n matrix $\mathbf{S}$ is a generalisation of the inverse matrix that satisfies the conditions

$\mathbf{S}(\mathbf{S}^{+}\mathbf{S}) = \mathbf{SI} = \mathbf{S}$ (i.e. $\mathbf{S}^{+}$ is a left inverse of $\mathbf{S}$: $\mathbf{S}^{+}\mathbf{S} = \mathbf{I}$)

$(\mathbf{S}^{+}\mathbf{S})\mathbf{S}^{+} = \mathbf{IS}^{+} = \mathbf{S}^{+}$

$(\mathbf{SS}^{+})^{*} = \mathbf{SS}^{+}$ (i.e. $\mathbf{SS}^{+}$ is Hermitian)

$(\mathbf{SS}^{+})^{*} = \mathbf{SS}^{+}$ ($\mathbf{S}^{+}\mathbf{S}$ is also Hermitian)

and has the property that it is its own inverse, i.e. $(\mathbf{S}^{+})^{+} = \mathbf{S}$.

2.6.1 Singular-Value Decomposition (SVD)

A computationally simpler way of calculating the pseudoinverse than by equation (2.62) is to use singular-value decomposition (SVD), which is based on the theorem that any m-by-n matrix $\mathbf{S}$ can be decomposed into the product of three matrices of the form

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{*} \tag{2.63}$$

where $\mathbf{U}$ is an m-by-m unitary matrix, $\boldsymbol{\Sigma}$ is an m-by-n diagonal matrix whose non-negative real entries are the singular values ($\sigma_{i\,=\,1:n}$) of $\mathbf{S}$, arranged in descending order and $\mathbf{V}$ is an n-by-n unitary matrix as shown below

$$\begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ s_{3,1} & s_{3,2} & \cdots & s_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,n} \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,m} \\ u_{2,1} & u_{2,2} & u_{2,3} & \cdots & u_{2,m} \\ u_{3,1} & u_{3,2} & u_{3,3} & \cdots & u_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{m,1} & u_{m,2} & u_{m,3} & \cdots & u_{m,m} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & 0 & \sigma_n \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \cdots & v_{n,n} \end{bmatrix}$$

A singular value σ (a real non-negative number) and its corresponding left-singular and right-singular vectors $\mathbf{u}$ and $\mathbf{v}$ for a rectangular matrix $\mathbf{S}$ satisfy the conditions

$$\mathbf{Sv} = \sigma\mathbf{u}$$

$$\mathbf{S}^{\mathrm{T}}\mathbf{u} = \sigma\mathbf{v}$$

The diagonal entries of the diagonal matrix $\mathbf{\Sigma}$ are the singular values of $\mathbf{S}$ whose corresponding left- and right-singular vectors form the columns of the unitary matrices $\mathbf{U}$ and $\mathbf{V}$ such that

$$\mathbf{SV} = \mathbf{U\Sigma} \tag{2.64}$$

$$\mathbf{S}^{\mathrm{T}}\mathbf{U} = \mathbf{V\Sigma} \tag{2.65}$$

Since $\mathbf{U}$ and $\mathbf{V}$ are unitary, i.e. $\mathbf{U}^{-1} = \mathbf{U}^{*}$ and $\mathbf{V}^{-1} = \mathbf{V}^{*}$ equation (2.64) becomes

$$\mathbf{S} = \mathbf{U\Sigma V}^{*} \tag{2.66}$$

which is the singular value decomposition of the m-by-n matrix $\mathbf{S}$. An m-by-n matrix has at least 1 and at most $\min(m,n)$ singular values. Incidentally, the singular values calculated by SVD can be used to calculate the rank of $\mathbf{S}$ and is equal to the number of singular values above a suitable threshold.

If $\mathbf{S}$ contains many more rows than columns, i.e. if $m \gg n$, then $\mathbf{U}$ becomes extremely large resulting in slow computation and the need for large storage. However since $\mathbf{\Sigma}$ contains only n non-zero diagonal entries only the first n columns of $\mathbf{U}$ are required (since all further columns are multiplied by zero) so a more compact SVD can be used. The so-called reduced SVD has matrices with the following dimensions: $\mathbf{U}$ is m-by-n, $\mathbf{\Sigma}$ is n-by-n and $\mathbf{V}$ is n-by-n (as before) as illustrated below

$$\begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ s_{3,1} & s_{3,2} & \cdots & s_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,n} \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,n} \\ u_{3,1} & u_{3,2} & \cdots & u_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m,1} & u_{m,2} & \cdots & u_{m,n} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \cdots & v_{n,n} \end{bmatrix}$$

This version of SVD is referred to as thin SVD, or "economy size" decomposition as it is referred to in MATLAB. The first step in calculating the thin SVD (as is done with the LAPACK routines implemented by the MATLAB function `svd.m`) is usually a QR factorisation of $\mathbf{S}$, after which matrix $\mathbf{R}$ is reduced to a bidiagonal matrix. The singular values and vectors are then found by performing a bidiagonal QR iteration [2.16].

The pseudoinverse $\mathbf{S}^{+}$ is calculated by reversing the order of the component matrices, transposing the singular vector matrices and taking the reciprocal of the diagonal entries of the central matrix as follows

$$\mathbf{S}^{+} = \mathbf{V\Sigma}^{+}\mathbf{U}^{*} \tag{2.67}$$

where $\mathbf{\Sigma}^+$ is the transpose of $\mathbf{\Sigma}$ with every non-zero entry replaced by its reciprocal

$$\mathbf{\Sigma}^+ = \mathrm{diag}\left(\frac{1}{\sigma_1},\frac{1}{\sigma_2},\ldots,\frac{1}{\sigma_n}\right) = \begin{bmatrix} 1/\sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & 1/\sigma_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/\sigma_n \end{bmatrix} \tag{2.68}$$

While this method is computationally intensive it is useful if $\mathbf{S}$ is ill-conditioned (if the condition number of $\mathbf{S}$ multiplied by the round-off error is large), in which case very small singular values when inverted increase numerical noise in the solution. To obtain a more stable result all singular values below a threshold are rounded to zero before calculating $\mathbf{\Sigma}^+$. For example the MATLAB function `pinv.m`, calculates the Moore-Penrose pseudoinverse using SVD and accepts an optional input argument `tol` below which any singular values are treated as zero.

Incremental algorithms exist for calculating the pseudoinverse of a matrix $\mathbf{R}$ that is related to a matrix $\mathbf{S}$ for which the pseudoinverse $\mathbf{S}^+$ is already known. In particular, if $\mathbf{R}$ differs from $\mathbf{S}$ by only a changed, added or deleted row or column, then such algorithms may require less work than calculating $\mathbf{R}^+$ with SVD. This may be of use with the present application, for example if the pseudoinverse of one mode matrix $\mathbf{\Psi}_1$ is known then the pseudoinverse of a second, closely related matrix $\mathbf{\Psi}_2$ could be determined with less computational effort.

2.7 Truncation Analysis with Gaussian Beam Modes

A distinct advantage of GBMA is that it can account for truncation effects that occur when a propagating wavefront passes the finite aperture of a lens, mirror, etc. The method developed here to analyse truncation effects due to finite-sized elements in terms of Gaussian beam modes will later be applied to explain truncation effects observed in beam pattern measurements presented in Chapters 3, 4 and 5.

Truncation occurs because of the finite radius of components in a quasi-optical system. If a propagating wavefront is significantly truncated unforeseen diffraction effects will occur, which affects the subsequent behaviour of the field [2.12, 2.13]. Consider an arbitrary wavefront that is defined by a finite sum of Gaussian beam modes at the origin, $z = 0$. Then wavefront incident on the plane of a lens L (of focal length f) is

$$E_{in}(x, y) = \sum_{n=0}^{n_{max}} \sum_{m=0}^{m_{max}} A_{mn} \cdot \psi_{mn}(x, y; z) \tag{2.69}$$

where the modes $\psi_m(x, y; z)$ are defined in the plane of the lens aperture (a distance $z = +f$). The ABCD matrix technique [2.14] is used to keep track of the mode set parameters (beam radius W, radius of curvature R and phase slippage ϕ_n between the modes) as the wavefront propagates from, for example, a phase grating (at $z = 0$), through the lens L and onto the output (Fourier) focal plane of L. To model truncation one assumes that any radiation incident on the surface of the truncating component (lens or mirror) is transmitted but that any radiation that arrives outside the perimeter of the truncating aperture is essentially lost. The finite aperture of a lens is treated as a circularly symmetric stop of radius a so the output field $E_{out}(x, y)$ transmitted from the lens aperture is related to the input field $E_{in}(x, y)$ as follows

$$E_{out}(x, y) = \begin{cases} E_{in}(x, y) & r \le a \\ 0 & r > a \end{cases} \tag{2.70}$$

where r is the distance from the optical axis (z-axis), as illustrated in Figure 2-32. After truncation is performed a new set of mode coefficients B_m that describe the truncated field $E_{out}(x, y)$ is calculated by performing the following overlap integral

$$B_{mn} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \psi^*_{mn}(x, y; z) \cdot E_{out}(x, y) dx dy \tag{2.71}$$

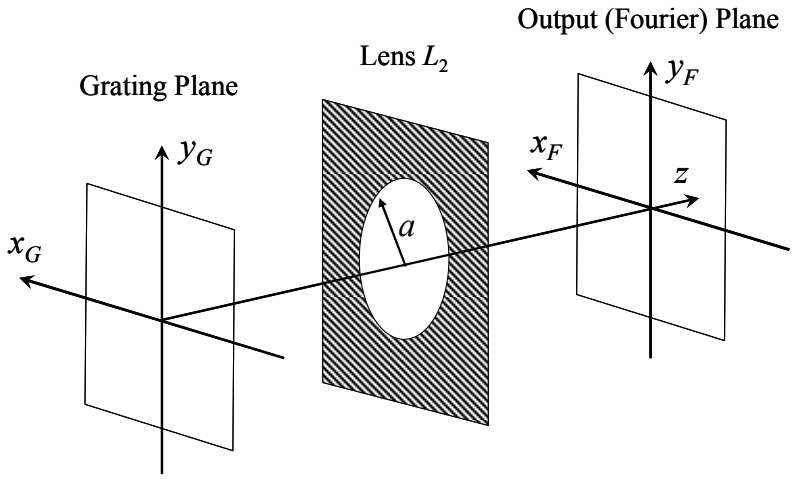


Figure 2-32. Schematic of the model used for truncation of the diffraction pattern produced by a diffraction grating (left) by a lens L_2 of radius a (centre). Any radiation incident on the shaded region ($r > a$) at the lens plane is truncated, i.e. set equal to zero.

Finally the truncated field is propagated to the output plane (a distance f_2 beyond L_2). The mode parameters W and R at the output plane are calculated using ABCD matrices, from which the focal plane mode set is derived and the output plane field is then given by

$$E_F(x_F, y_F) = \sum_{n=0}^{n_{max}} \sum_{m=0}^{m_{max}} B_{mn} \cdot \psi_{mn}(x_F, y_F; z) \tag{2.72}$$

where the modes $\psi_m(x_F, x_F; z)$ are now defined at the output focal plane. In order to compare the truncated field $E_F(x_F, y_F)$ with the field that would result from an ideal non-truncating lens we simply replace mode coefficients B_{mn} with A_{mn} in Eq. (2.72). In GBMA mirrors are treated as in-line phase transforming devices, so truncation by a mirror is treated in exactly the same way as for a lens.

Alternatively, because within the aperture S of the truncating element the output field is just equal to the incident field we can write

$$B_{mn} = \int_S \psi'^{*}_{mn}(x, y; z) \cdot E_{in}(x, y) dxdy \tag{2.73}$$

where the truncated modes $\psi'(x, y; z)$ are

$$\psi'(x, y; z) = \begin{cases} \psi(x, y; z) & r \le a \\ 0 & r > a \end{cases} \tag{2.74}$$

Substituting for $E_{in}(x,y)$ then gives

$$B_{mn} = \sum_{n=0}^{n_{max}} \sum_{m=0}^{m_{max}} A_{mn} \int_S \psi'^{*}_{mn}(x, y; z) \cdot \psi_{mn}(x, y; z) dxdy \tag{2.75}$$

Now we define a scattering matrix

$$S_{mn} = \int \psi'^{*}_{mn}(x,y;z) \cdot \psi_{mn}(x,y;z) dxdy \tag{2.76}$$

which determines how the power in the input mode coefficients A_{mn} is redistributed amongst the output mode coefficients B_{mn} as follows

$$B_{mn} = S_{mn} \cdot A_{mn} \tag{2.77}$$

In the case where truncation effects are not included $B_{mn} = A_{mn}$ and S_{mn} simplifies to an identity matrix.

Note that the calculation of the 2-D matrix S_{mn} requires M^2 or $(M \cdot N)^2$ numerical integrations (M and N being the number of Gaussian Hermite modes in x and y), depending on the dimensionality of the propagating field. Thus for two-dimensional problems it is computationally more efficient to calculate the output mode coefficients B_{mn} directly with Eq. (2.73) rather than by calculating S_{mn} first.

One advantage of using a scattering matrix approach to determine truncation at a finite aperture is that it can be extended it to a system comprising a number of elements with truncating apertures. The scattering matrix for the entire system is given by

$$S_{tot} = S_{\mathrm{N}} \cdot S_{\mathrm{N-1}} \cdot \ldots \cdot S_2 \cdot S_1 \tag{2.78}$$

the product of the scattering matrices S_i for each of the N elements of the system. Any input field to the system, which is described by mode coefficients A_{mn}, will then produce an output field, the mode coefficients of which are

$$B_{mn} = S_{tot} \cdot A_{mn} \tag{2.79}$$

Of course all input fields to the system must be described in terms of the same set of Gaussian beam modes. Similarly another scatter matrix can be calculated to allow for backward propagation through the system from output to input planes.

2.8 Symmetry Considerations in GBMA

Further reductions in computational overhead required to implement GBMA can be achieved by taking into account possible symmetries of the input field. Consider for example the amplitude distribution shown in Figure 2-33(*a*) of a complex-valued field $E(x, y)$[†]. A modal decomposition of this field was performed using a Gaussian-Hermite mode-set, $h_{mn}(x, y)$ in which the highest-order modes indices in x and y are $m = n = 40$. The mode coefficient amplitudes $|A_{mn}|$ for the reconstructed field are shown in a negative greyscale in Figure 2-33(*b*). The particular choice of mode-set resulted in very good agreement (root-mean squared error of 0.14% and intensity correlation of 99.91%) between the input and reconstructed fields.

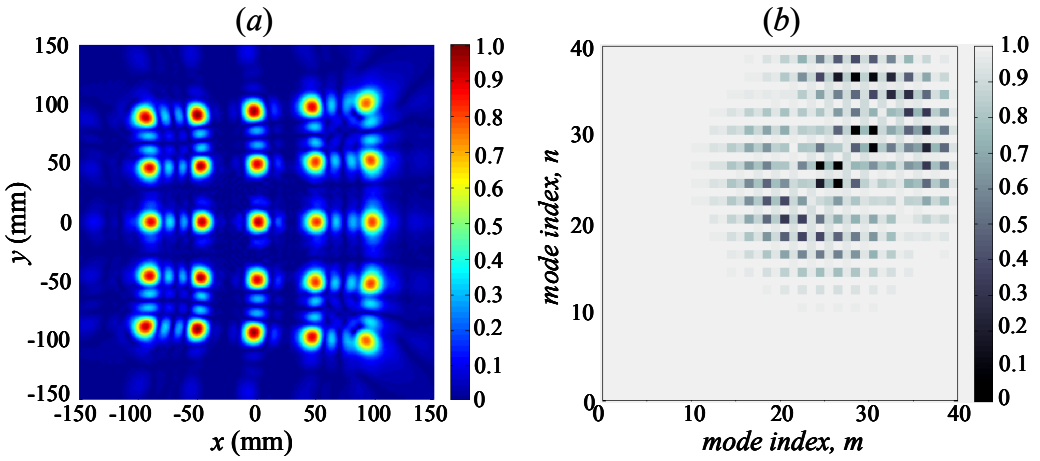


Figure 2-33. A modal decomposition (with a set of Gaussian-Hermite modes $h_{mn}(x, y)$) of the field whose amplitude distribution is shown in (*a*) produces a set of mode coefficients whose amplitudes, $|A_{mn}|$ are plotted in (*b*). Power exists in both even-numbered and odd-numbered modes.

Power exists in both even- and odd-numbered modes, but clearly not all of the available modes contribute to the decomposition. If we could know in advance (before calculating A_{mn}) which modes were necessary and which were redundant, then a reduced mod-set containing only relevant modes could be used, which would lead to increased computational efficiency by reducing the time needed to calculate mode coefficients. Alternatively the gain in execution time could be traded for increased spatial resolution (by using a more densely sampled field) or modal frequency resolution (by including more higher-order modes in the decomposition). While the exact contribution from each individual mode cannot be known priori in mode coefficient calculations, the

[†] The field $E(x, y)$ was simulated using MODAL and is the output field produced by one of the Dammann gratings discussed in Chapter 4 (tested using a 4-*f* arrangement consisting of two 500mm focal length ellipsoidal mirrors).

contribution due to particular subsets of modes (grouped according to the axial symmetry exhibited by individual modes within a particular subset) can be determined quite easily. Each mode subset is a group of modes in which each mode exhibits similar axial symmetry. In one dimension modes fall into only one of three such categories:

a) even-numbered modes (with even symmetry)

b) odd-numbered-modes (with odd symmetry)

c) both even- and odd-numbered modes (with both even and odd symmetry)

Each mode subset corresponds to a portion of a function $f(x)$ with specific axial symmetry as follows. Any one-dimensional function, $f(x)$ can be expressed as

$$f(x) = f_e(x) + f_o(x) \tag{2.80}$$

the sum of its even part $f_e(x)$ and its odd part $f_o(x)$ [2.17], which are given by

$$f_e(x) = \tfrac{1}{2}\big[f(x) + f(-x)\big] \tag{2.81}$$

and

$$f_o(x) = \tfrac{1}{2}\big[f(x) - f(-x)\big] \tag{2.82}$$

If $f_e(x) = 0$ then $f(x)$ is described exactly by $f_o(x)$ and the function $f(x)$ is itself an odd function, and vice versa. Computationally $f(x)$ is represented by a one-dimensional array and $f(-x)$ is obtained by simply reversing the ordering of elements in $f(x)$.

A full modes-set (containing both even and odd modes) with highest-order mode index m_{max} contains $M = (m_{max}+1)$ modes. A function $f(x)$ that has both even and odd parts, i.e. given by Eq. (2.80), requires both even and odd modes to accurately describe it. If, however $f(x)$ is either even or odd, i.e. if $f(x) = f_e(x)$ or $f_o(x)$ then accurate modal decomposition can be achieved using a mode-set comprising only those modes with the same symmetry and thus only $\sim M/2$ modes. The relationship between the even/odd portions of a one-dimensional function and corresponding mode subset is summarised in Table 2-1.

Function $f(x) =$	Symmetry	mode indices, m		number of modes ($M = m_{max}+1$)
		even	odd	
$f_e(x) + f_o(x)$	–	✓	✓	M
$f_e(x)$	even	✓	×	$M/2$
$f_o(x)$	odd	×	✓	$M/2$

Table 2-1. Relationship between even and odd portions of one-dimensional function $f(x)$ and the mode subsets needed for their reconstruction. The function $E_x + O_x$ is nether even or odd.

Extending this concept to two-dimensions, any two-dimensional function $f(x, y)$ can be expressed in terms of its even and odd parts in one of three ways

$$f(x, y) = E_x(x, y) + O_x(x, y) \tag{2.83}$$

$$f(x, y) = E_y(x, y) + O_y(x, y) \tag{2.84}$$

$$f(x, y) = EE_{xy}(x, y) + EO_{xy}(x, y) + OO_{xy}(x, y) + OE_{xy}(x, y) \tag{2.85}$$

where $E_x(x, y)$ and $O_x(x, y)$ in Eq. (2.83) refer to the parts of $f(x, y)$ that are even and odd about the x-axis and therefore describe even and odd symmetry of $f(x, y)$ in the y-direction. Similarly terms $E_y(x, y) + O_y(x, y)$ in Eq. (2.84) are the parts of $f(x, y)$ that are even and odd about the y-axis. Note that each of the four terms in Eq. (2.85) are just that: individual terms. For example $EO_{xy}(x, y)$ denotes the part of $f(x, y)$ that is both even about the x-axis and odd about the y-axis, and should not be read as being the product of terms $E_x(x, y)$ and $O_y(x, y)$ that appear in the previous two equations.

We define two operators, $\mathbf{E}_s\{\ \}$ and $\mathbf{O}_s\{\ \}$ (in bold, un-italicised font to distinguish them from functions of the same name, with subscript s denoting transverse coordinate x or y) for calculating the various even/odd parts of $f(x, y)$. Both operate on a 2-D function and return another function of the same size, which is the even or odd part of the input function, the axis about which the symmetry is tested being specified by the subscript. Thus we have the following expressions for the terms in Eq. (2.83) and (2.84)

$$E_x(x, y) = \mathbf{E}_x\{f(x, y)\} = \tfrac{1}{2}[f(x,y) + f(x,-y)]$$

$$O_x(x, y) = \mathbf{O}_x\{f(x, y)\} = \tfrac{1}{2}[f(x,y) - f(x,-y)]$$

$$E_y(x, y) = \mathbf{E}_y\{f(x, y)\} = \tfrac{1}{2}[f(x,y) + f(-x,y)]$$

$$O_y(x, y) = \mathbf{O}_y\{f(x, y)\} = \tfrac{1}{2}[f(x,y) - f(-x,y)]$$

The four terms in Eq. (2.85) are calculated by using appropriate combinations of operators $\mathbf{E}_s\{\ \}$ and $\mathbf{O}_s\{\ \}$ in succession as follows

$$EE_{xy}(x, y) = \mathbf{E}_x\{E_y(x, y)\} = \mathbf{E}_x\{\mathbf{E}_y\{f(x, y)\}\}$$

$$OO_{xy}(x, y) = \mathbf{O}_x\{O_y(x, y)\} = \mathbf{O}_x\{\mathbf{O}_y\{f(x, y)\}\}$$

$$EO_{xy}(x, y) = \mathbf{E}_x\{O_y(x, y)\} = \mathbf{E}_x\{\mathbf{O}_y\{f(x, y)\}\}$$

$$OE_{xy}(x, y) = \mathbf{O}_x\{E_y(x, y)\} = \mathbf{O}_x\{\mathbf{E}_y\{f(x, y)\}\}$$

For example

$$EE_{xy}(x, y) = \tfrac{1}{4}[f(x, y) + f(x,-y) + f(-x, y) + f(-x,-y)]$$

Note that the order in which operators $\mathbf{E}_s\{\ \}$ and $\mathbf{O}_s\{\ \}$ are applied is irrelevant.

In two dimensions the different permutations of even- and odd-numbered modes results in eight distinct mode subsets, each one corresponding to one of the eight terms in equations (2.83, 2.84 and 2.85). The modes needed to represent each even/odd term are listed in Table 2-2, where symmetry in the transverse directions x and y is indicated for each term – where such a statement can be made. For example the term $E_x(x, y)$, which is even about the x-axis (in the y-direction) may or may not be even about the y-axis. Also listed in Table 2-2 is the total number of modes required to represent each function, where the number of modes in the x- and y-directions are $M = m_{\max}+1$ and $N = n_{\max}+1$, respectively, where the total number of modes in the full two-dimensional mode-set is $M \times N$.

Function	symmetric in…		M		N		# of modes
	x	y	even	odd	even	odd	
$f(x, y)$	–	–	✓	✓	✓	✓	MN
$E_x(x, y)$	–	✓	✓	✓	✓	×	$\frac{1}{2}MN$
$O_x(x, y)$	–	×	✓	✓	×	✓	$\frac{1}{2}MN$
$E_y(x, y)$	✓	–	✓	×	✓	✓	$\frac{1}{2}MN$
$O_y(x, y)$	×	–	×	✓	✓	✓	$\frac{1}{2}MN$
$EE_{xy}(x, y)$	✓	✓	✓	×	✓	×	$\frac{1}{4}MN$
$OO_{xy}(x, y)$	×	×	×	✓	×	✓	$\frac{1}{4}MN$
$EO_{xy}(x, y)$	×	✓	×	✓	✓	×	$\frac{1}{4}MN$
$OE_{xy}(x, y)$	✓	×	✓	×	×	✓	$\frac{1}{4}MN$

Table 2-2. The relationship between even and odd-portioned functions of a two-dimensional function $f(x,y)$ and the mode sets (of order m and n) that they can be decomposed into. The final column contains the number of modes required to reconstruct the given function, where $M = m_{\max}+1$ and $N = n_{\max}+1$. Columns 2 and 3 (indicating the symmetry of a specific function in x and y directions) are only specified for cases where the symmetry is exactly determined. Functions in which the symmetry in a given direction is ambiguous are left blank.

Figure 2-34 illustrates the above concepts whereby individual modes with indices (m,n) are identified by their position within a square grid representing (in this example) a 36-element mode-set. The expression in Eq. (2.83), $f(x,y) = E_x(x, y) + O_x(x, y)$, is shown in Figure 2-34(a) where individual modes that contribute to $E_x(x, y)$ and $O_x(x, y)$ are shown with different shades. Half of the modes contribute to $E_x(x, y)$ and the rest to $O_x(x, y)$, as seen in Table 2-2.

(*a*)

5	O_x	O_x	O_x	O_x	O_x	O_x
4	E_x	E_x	E_x	E_x	E_x	E_x
3	O_x	O_x	O_x	O_x	O_x	O_x
2	E_x	E_x	E_x	E_x	E_x	E_x
1	O_x	O_x	O_x	O_x	O_x	O_x
0	E_x	E_x	E_x	E_x	E_x	E_x
n/*m*	0	1	2	3	4	5

(*b*)

5	E_y	O_y	E_y	O_y	E_y	O_y
4	E_y	O_y	E_y	O_y	E_y	O_y
3	E_y	O_y	E_y	O_y	E_y	O_y
2	E_y	O_y	E_y	O_y	E_y	O_y
1	E_y	O_y	E_y	O_y	E_y	O_y
0	E_y	O_y	E_y	O_y	E_y	O_y
n/*m*	0	1	2	3	4	5

(*c*)

5	OE_{xy}	OO_{xy}	OE_{xy}	OO_{xy}	OE_{xy}	OO_{xy}
4	EE_{xy}	EO_{xy}	EE_{xy}	EO_{xy}	EE_{xy}	EO_{xy}
3	OE_{xy}	OO_{xy}	OE_{xy}	OO_{xy}	OE_{xy}	OO_{xy}
2	EE_{xy}	EO_{xy}	EE_{xy}	EO_{xy}	EE_{xy}	EO_{xy}
1	OE_{xy}	OO_{xy}	OE_{xy}	OO_{xy}	OE_{xy}	OO_{xy}
0	EE_{xy}	EO_{xy}	EE_{xy}	EO_{xy}	EE_{xy}	EO_{xy}
n/*m*	0	1	2	3	4	5

Figure 2-34. Grids showing a mode set (with modes whose indices span the range m, n = 0…5) and the relationship between individual modes of index (m, n) and the particular even/odd portion of a 2-*D* field $f(x, y)$ that they contribute to for the three methods of expressing $f(x,y)$ as a sum of its even/odd parts. $f(x,y)$ = (*a*) $E_x(x,y) + O_x(x,y)$, (*b*) $E_y(x,y) + O_y(x,y)$ and (*c*) $EE_{xy}(x,y) + OO_{xy}(x,y) + EO_{xy}(x,y) + OE_{xy}(x,y)$.

If the eight even/odd parts of a 2-D function $f(x, y)$ are known, each can then be compared to $f(x, y)$ to test for any possible symmetries. If one of the terms matches closely the original function then a modal analysis can be achieved using the mode subset corresponding to that even/odd term since all other modes will be redundant. If more than one term is found to be an equally good match then the one corresponding to a smaller mode-set (with fewer modes – see Table 2-2) is chosen. If no term resembles the original function closely enough then we conclude that $f(x, y)$ possesses no axial symmetry and accurate modal analysis will require use of all modes in the mode-set.

The 'goodness of fit' between input function $f(x, y)$ and the various trial even/odd parts was quantified in terms of correlation, $h(x,y)$. The correlation between two functions $f(x,y)$ and $g(x,y)$, whose Fourier transforms are $F(u, v)$ and $G(u, v)$, is given by

$$\Im\{h(x,y)\} = H(u,v) = \Im\{f(x,y) \cdot g(x,y)\} = F(u,v) \otimes G(u,v)$$

and is a method used for pattern recognition [2.20]. If $g(x,y)$ matches exactly $f(x,y)$ then maximum throughput occurs and a bright spot is observed at $H(0,0)$, corresponding to the zeroth-order spectral point or DC component of the spectrum $H(u,v)$. If $g(x, y)$ does not match $f(x, y)$ then the intensity value of $H(0,0)$ is reduced accordingly.

The sample field shown in Figure 2-33(*a*) was analysed using the method described. The field was first decomposed into its various even/odd symmetry components and each of these compared with the original field. Each of the symmetry component fields is plotted alongside a plot of the corresponding mode coefficient amplitudes, $|A_{mn}|$.

Figure 2-35 shows the constituent parts of the input field when described in terms of even and odd symmetry about the x-axis: $E_x(x, y)$ and $O_x(x, y)$. The former is almost identical to the input field, whereas the latter does not reproduce any of the Gaussian spots to significant intensity levels (note the intensity levels on the colour bar) and so is not a good candidate for representing the input field. Thus we can say that to a good approximation $f(x, y) = E_x(x, y)$.

Figure 2-36 shows the parts of $f(x, y)$ with even and odd symmetry about the y-axis: $E_y(x, y)$ and $O_y(x, y)$. Although the first term, $E_y(x, y)$ does succeed in reproducing all 25 Gaussian beams to reasonable intensity levels, distortion effects in the input field are not symmetric about the y-axis so they cannot be correctly described using only the part of the input field that is. Thus $f(x, y)$ cannot be accurately described in terms of just this term but must include the odd term, $O_y(x, y)$ as well.

None of the four mixed-symmetry parts of $f(x, y)$ shown in Figure 2-37 match the input field. Note that terms $EE_{xy}(x, y)$ and $EO_{xy}(x, y)$ have features with the highest intensity and are in fact seen to be almost identical to terms $E_y(x, y)$ and $O_y(x, y)$ shown in Figure 2-36. Thus the input field can be adequately described without the inclusion of the other two terms shown in Figure 2-37.

Figure 2-38 shows a bar-chart of correlation coefficients values – the value of $H(0,0)$ – calculated between the input field and each even/odd part. Clearly the correlation coefficients for the individual terms in each of the equations (2.83), (2.84) and (2.85) will sum to unity. We see that the correlation coefficient corresponding to the term $E_x(x, y)$ in Eq. (2.83) is itself equal to one, therefore a modal reconstruction of the input field requires a mode-set containing just those corresponding modes, i.e. those with indices $m = [0, 1, 2, \ldots m_{\max}]$ and $n = [0, 2, 4, \ldots n_{\max}]$.

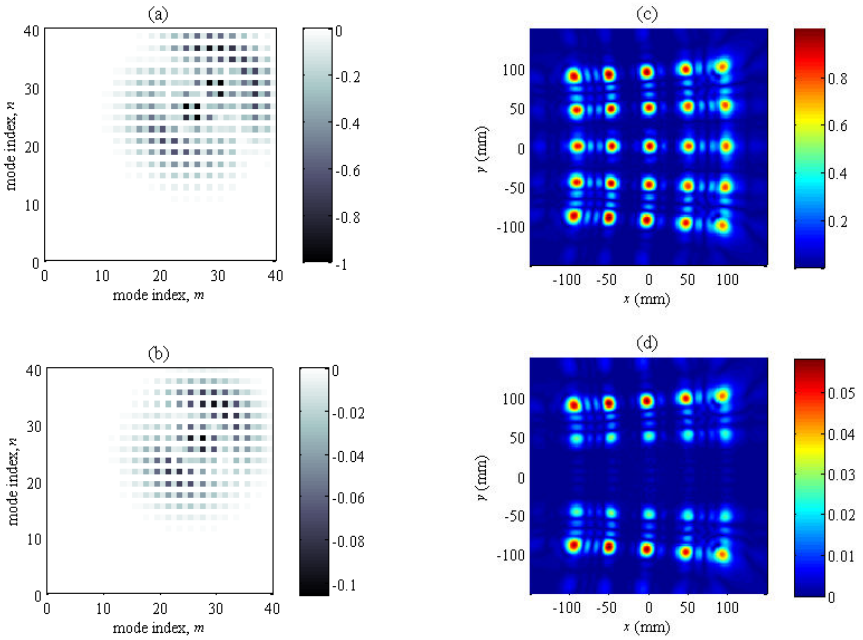


Figure 2-35. (*a*-*b*) Mode coefficient amplitudes, $|A_{mn}|$ and (*c*-*d*) corresponding reconstructed amplitude distributions $E_x(x, y)$ and $O_x(x, y)$.

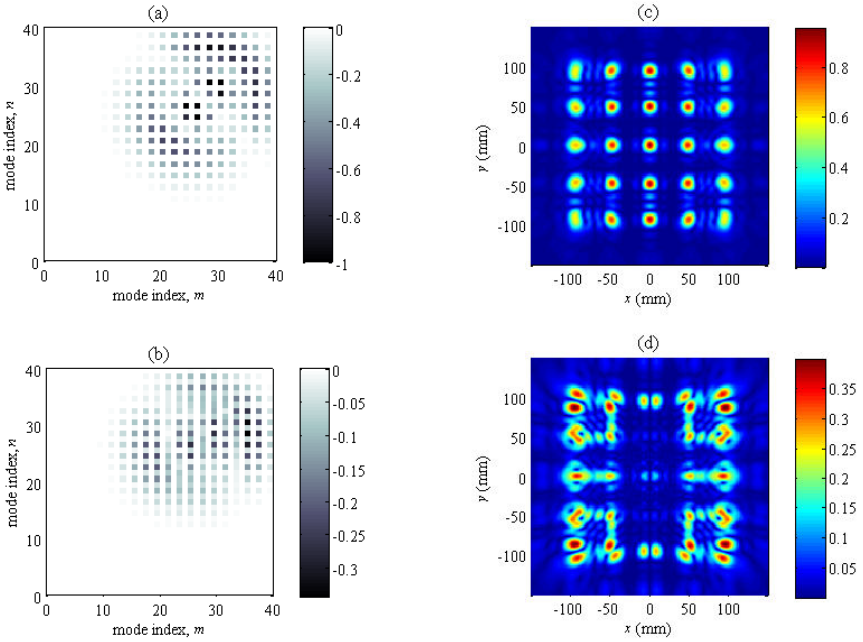


Figure 2-36. (*a*–*b*) Mode coefficient amplitudes and (*c* – *d*) their corresponding reconstructed amplitude distributions $E_y(x, y)$ and $O_y(x, y)$.

The technique described here for identifying symmetries of a given field can be routinely applied as a quick pre-processing step to any 2-D field before attempting modal reconstruction so as to identify redundant modes and thereby increase the computational efficiency of two-dimensional Gaussian Beam Mode Analysis.

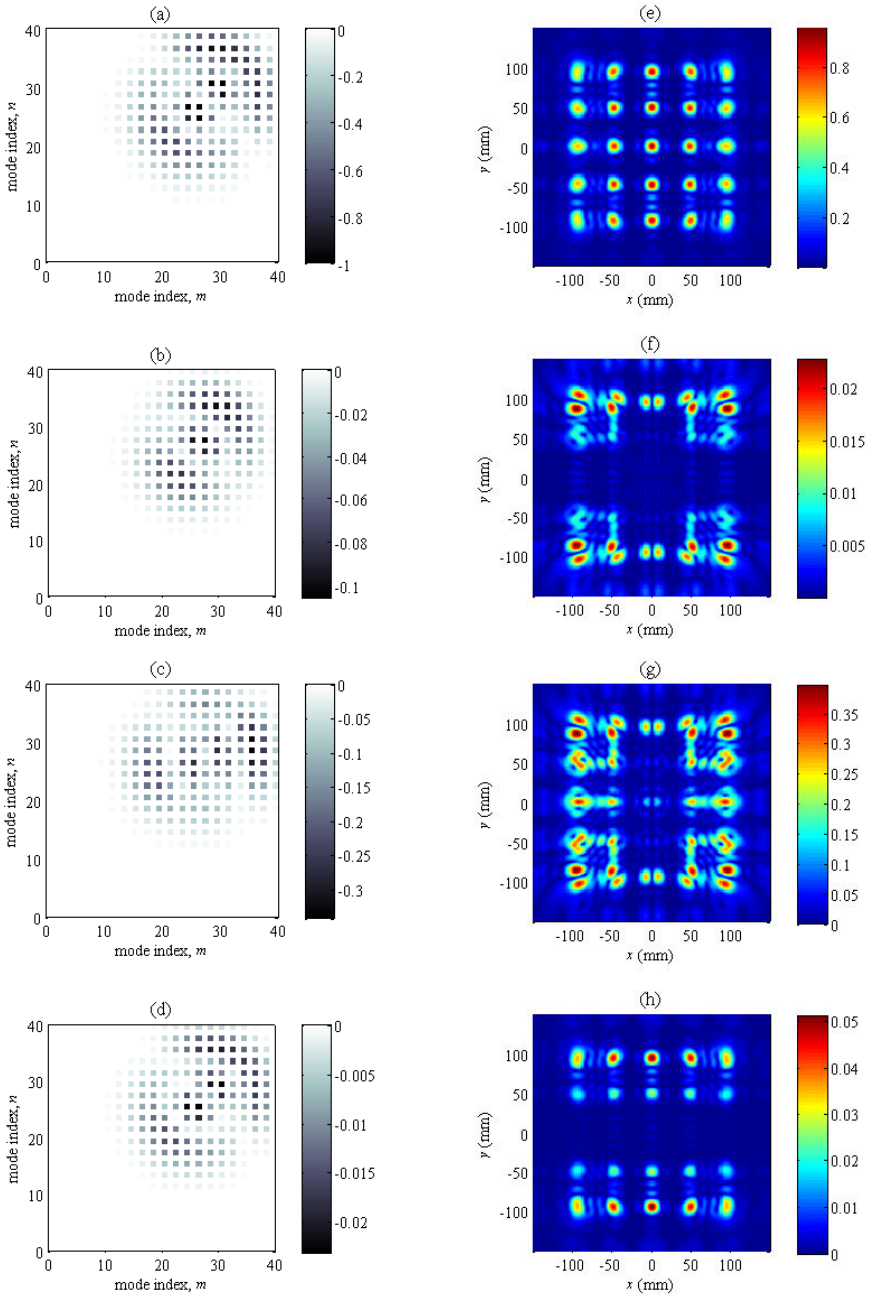


Figure 2-37. (*a–d*) Mode coefficients amplitudes and (*e–h*) their corresponding reconstructed amplitude distributions (from top to bottom) $EE_{xy}(x, y)$, $OO_{xy}(x, y)$, $EO_{xy}(x, y)$ and $OE_{xy}(x, y)$.

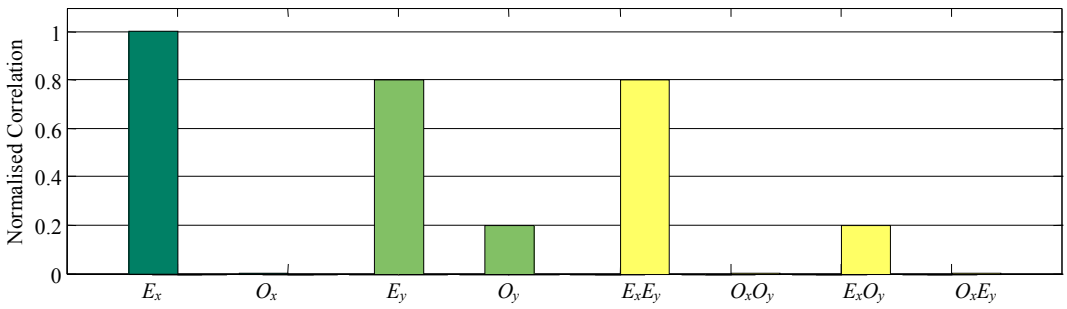


Figure 2-38. Correlation coefficient values between input $f(x,y)$ and its constituent even/odd parts. Bars are colour-coded according to terms in equations (2.83), (2.84) and (2.85). Dark green correspond to terms in Eq. (2.83), bright green to terms in Eq. (2.84) and yellow to terms in Eq. (2.85).

2.9 Maynooth Optical Design and Analysis Laboratory (MODAL)

Optical design in the THz waveband suffers from a lack of dedicated software tools for modelling the range of electromagnetic and quasi-optical propagation conditions encountered in typical systems [2.21]. One is forced to use commercial optical design software packages (GLAD, ASAP, CODE V, Zemax) written for very different wavelength systems and there is often a lack of confidence in results because of possible inappropriate underlying physical principles. Thus a major component of the SFI research program, referred to in Chapter 1, was to develop a physical optics design and analysis computer-aided design (CAD) software package. The result of work by Dr. Marcin Gradziel and following on from previous work by Dr. David White was the Maynooth Optical Design and Analysis Laboratory (MODAL). A brief description of MODAL and how it was used in this thesis follows.

MODAL incorporates analytical techniques that have been developed for long-wavelength design and analysis in the THz waveband. The basic approach used by MODAL to model long-wavelength propagation is the application of Gaussian beam mode analysis, which has been extended to include the efficient description of off-axis (tilted) components such as simple curved reflectors [2.11]. As a rigorous model of electromagnetic wave propagation, physical optics (PO) can be used to accurately characterise complete systems. However, for the initial design or preliminary analysis of large multi-element optical systems, the straightforward PO approach proves to be unsuitable, being as it is computationally intensive. MODAL incorporates different propagation models that can be used within the same framework from approximate methods (ray tracing and paraxial beam modes) that prove extremely efficient and accurate in certain situations as well as fast PO software developed at NUI Maynooth [2.27] to improve the computational efficiency of the usual PO approach, when a more rigorous approach is required [2.22, 2.23].

Although MODAL is aimed primarily at the increasingly important THz range of the electromagnetic spectrum, it should also prove a useful tool for wavelengths ranging from X-rays to radio waves [2.24]. MODAL combines an OpenGL-based user interface, for easy definition and manipulation of optical systems, with a powerful and flexible analysis engine that implements multiple propagation methods, ranging from plane wave decomposition to full physical optics approach. The package provides built-

in presentation facilities, as well as export filters for analysis using external software. Calculations can be accelerated by running the code on a parallel computer composed of a heterogenous collection of machines (using PVM) and has been developed for use in Windows and Linux.

MODAL keeps track of the best-fit Gaussian width, the phase radius of curvature and waist positions to aid in the design process. Off-axis mirrors can be designed by specifying the angle of throw, and the input and output beam waist positions and sizes. Lenses are also facilitated, propagation through which is done with multiple reflections between the curved input and output surfaces (the radii of which are specified by the user).

MODAL provides a range of different analysis techniques in one package, making it an extremely flexible tool for both the design and analysis phases of instrumentation at THz frequencies. A combination of techniques is typically used to model a complete system. Design and analysis methods based on SVD Gaussian Beam Mode decomposition is a potentially very useful tool for quasi-optics. In particular it can be used to quickly generate useful results during system design, whereas rigorous PO calculations can be prohibitively slow. In future MODAL will also incorporate other code written in NUI Maynooth for related work, such as SCATTER [2.25] 2.26] (to predict beam patterns from shaped corrugated horn antennas) and FIRPOS [2.27] (fast physical optics code).

MODAL was utilised in this thesis as a design verification tool, in particular for the testing the diffractive phase gratings that are described in Chapters 4 and 5. Experimentally obtained measurements were compared with results computed using MODAL simulations of the same optical systems. In all cases where results computed using MODAL are shown, analysis was performed using a combination of the modal option (with SVD) and scalar diffraction (Fresnel integrals) option. The PO propagation option was not used because the size of the features that are of interest here meant that sufficiently accurate results could be produced through analysis with the alternative, more computationally efficient methods provided by MODAL.

Chapter 3.

Active Imaging at 100 GHz

3.1 Introduction

This chapter describes the experiments undertaken as part of the SFI funded Principal Investigator Research programme in terahertz optics (led by Prof. J.A. Murphy of NUI Maynooth), one of the objectives of which was to investigate the potential of terahertz imaging techniques for bio-medical applications.

The imaging work undertaken was intended to act as part of the strategy to combine imaging modalities from the two ends of the infrared spectrum: imaging at millimetre wavelengths (conducted by the THz Optics Group in the Department of Experimental Physics) would complement imaging at near-infrared wavelengths (as pursued by colleagues in the Department of Electronic Engineering at NUIM). The original goal was investigate the possibility of combining measurements from different wavelength bands to extract more useful, higher quality images with higher information content. Initially it was envisaged that THz imaging techniques could prove useful for deep tissue imaging in the human body, e.g. tumours, deep wounds and even brain imaging, which is possible at near-IR wavelengths [3.3,3.4]. However from initial experiments conducted at Maynooth [3.1,3.2] it was clear that the aspirations to perform deep tissue imaging could not be realised due to the inherent limitations of THz radiation, in particular the strong absorption properties of THz radiation by water. However, wound healing particularly beneath bandages was seen as a valuable potential application and was pursued at near-IR wavelengths also [3.5].

Pre-existing experimental test facilities at Maynooth were inadequate for performing two-dimensional imaging experiments. A brief description of a new raster-scanned system, as well as details of the optical components needed to conduct various experiments and which were designed by the author of this thesis, are presented first. The rest of the chapter describes several transmission and reflection mode imaging arrangements and the results obtained from each. As well as experimental measurements, the results of numerical simulations of some of the experiments, which were undertaken by the author, are presented for comparison.

Imaging is achieved using passive or active imaging systems, as illustrated in Figure 3-1. At millimetre wavelengths passive imaging involves detecting electromagnetic radiation naturally emitted by an object due to its temperature. In active imaging involves illuminating the object(s) with a transmitter, or source, and collecting the light that is transmitted, scattered and reflected through/from the object. These

definitions of passive and active imaging systems are of course generalisations since all but the simplest imaging system also requires relay optics, such as lenses and reflectors, to guide radiation in a controlled manner through the detection system in order to form a clear image at the detector.

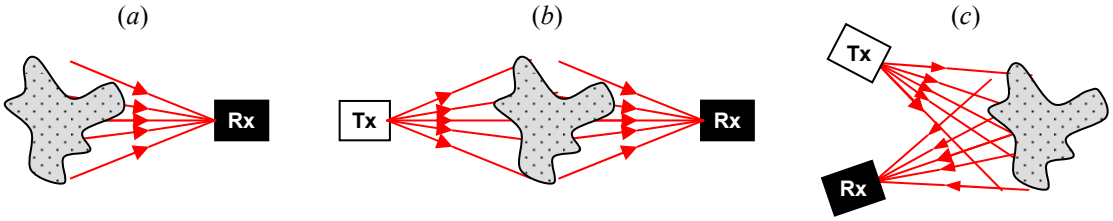


Figure 3-1. Passive imaging (*a*) requires only a detector (or receiver Rx) for detecting naturally emitted light. Active imaging (*b*-*c*) involves actively illuminating the object(s) with a source (or transmitter Tx). Arrangements (*b*) and (*c*) illustrate transmission- and reflection-type active imaging systems, respectively.

At millimetre wavelengths the low power levels of radiation naturally emitted by an object close to room temperature, and small differences in temperature between these objects and their surroundings, means that passive imaging requires extremely sensitive detection techniques. The experiments described in this thesis involved making intensity measurements using room-temperature bolometers. In order to obtain higher sensitivity than would be possible using such devices in a passive arrangement, measurements were instead made using various active imaging arrangements. In the experiments described here incoherent detection techniques were used. Active imaging can also be undertaken using heterodyne techniques, where clearly the detected signal is down-converted to a lower frequency through subsequent electronics. In fact to investigate these approaches a vector network analyser (VNA) was recently acquired by the THz Group at NUIM. However, such systems are currently too expensive for practical applications. Previous work at Maynooth has concentrated on design and testing of components at and around 100 GHz, i.e. within the W-band, which spans the frequency range 75-110 GHz. Although THz sources have become more readily available in recent years, the cost and complexity of operating in the THz range is still relatively high compared to sources operating up to a few hundred GHz, while the increase in resolving power is not necessarily required. Since working in the W-band is relatively inexpensive and therefore more practically applicable and also because of the ease of access to sources, detectors and accessories operating in this frequency range, all the imaging experiments described here were performed in the W-band, particularly at 100 GHz.

The various active imaging arrangements that were experimented with can be classed according to transmission and reflection geometries.

In transmission mode the object under test was illuminated from behind and the radiation transmitted through the object measured. Two transmission mode arrangements were investigated. In near-field transmission imaging the transmitted radiation intensity in the plane immediately behind the illuminated sample was measured. In "re-imaged" transmission imaging experiments relay optics were used to produce an image of the transmitted radiation onto an image plane where the intensity pattern was then measured. The latter Fourier optics arrangement included the facility to perform spatial frequency filtering.

In the reflection mode imaging experiments, a near-field arrangement was employed, several variations of which were experimented with so as to obtain highest spatial resolution with the components at our disposal.

3.2 THz quasi-optical test facility at Maynooth

The pre-existing scanning systems available to the THz Optics group prior to the experiments described here included a 2-D raster scanner and an azimuthal scanner. The azimuthally-scanned arrangement is used primarily in the measurement of horn antenna beam patterns [3.6]. Another variation includes a motorised z-axis translation stage which allows for precise control of the source-detector separation and has been used in the experimental analysis of standing wave effects established in horn-to-horn systems. The 2-D image acquisition system, referred to as GHOST (GHz Optical Scanning Tool), was constructed specifically for the purpose of measuring intensity beam patterns from Dammann gratings [3.7]. However, mechanical vibrations in the rather flexible structure set up by its scanning mechanism (a system of pulleys and elasticated ribbons), when operated at even relatively modest scan speeds, results in high image acquisition times, and it was therefore unsuitable for extensive imaging experiments. Two high-precision translation stages became the basis for a new 2-D raster-scanning system, hereafter referred to as TOAST (THz Optical Scanning Tool), which, being mechanically much more stable and rigid, could achieve much lower image acquisition times through fast scanning. The author of this thesis was involved in the development of TOAST, particularly in those aspects described in detail below.

In all experiments reported here, illumination was provided by one of two mechanically-tuned solid-state sources (Gunn diode oscillators); the two sources spanning the frequency ranges 75–100 GHz and 100-110 GHz, which together allowed for measurements to be performed over the entire W-band (75-110 GHz). Power levels from the two Gunn oscillators range from a minimum of 9.8 mW (at 110.652 GHz) to a maximum of 60.8 mW (at 88.03 GHz). The detector used was a mechanically-tuned planar zero biased (i.e. without bias making for convenient use) Schottky barrier diode. This detector is sensitive to frequencies across the W-band, can handle a maximum input power of 100 mW, and its sensitivity is quoted to be typically in excess of 550 mV/mW. A short section of standard rectangular metallic waveguide was used for free-space coupling to the detector. The waveguide used (WR10) has internal dimensions of $a \times b = (0.1 \times 0.05)$ inch, or (2.54×1.27) mm and is designed for operation at a centre frequency of 92.5 GHz [3.8]. A bare waveguide radiator with these dimensions (width a equal to twice its height b) produces a field distribution at its aperture that is described in the x direction by a truncated cosine function, and in the y-direction by a top-hat function

$$E(x,y) = \cos\left(\frac{\pi x}{a}\right) \qquad |x| \le \frac{a}{2}\,;\, |y| \le \frac{b}{2} \tag{3.1}$$

Incidentally, in terms of Gaussian beam mode analysis, such a field can be represented using a modal expansion in which maximum coupling to the fundamental Gaussian beam mode occurs for an asymmetric mode set with beam radius $(W_x, W_y) = (0.35a, 0.50b)$, which yields a maximum two-dimensional coupling efficiency to the fundamental mode of 0.88 [3.9].

The transmitter chain used in experiments is illustrated in Figure 3-2. Radiation emitted from the source is guided through a section of waveguide, through an isolator (to prevent propagation back into the source), through a variable attenuator (to control power level into the optical system), and finally through a horn antenna. The horn antenna couples the beam from a rectangular waveguide to a free-space beam. A corrugated conical horn antenna was used in experiments, which provides a circularly symmetric beam profile that is well approximated by a Gaussian function.

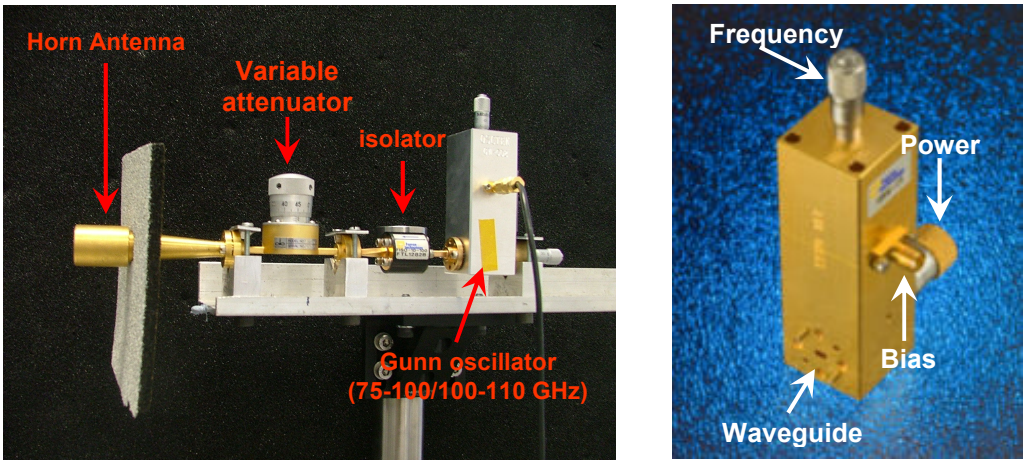


Figure 3-2. The transmitter chain (left) consists of a mechanically tuned Gunn oscillator, an isolator, a variable attenuator and a horn antenna (shown is a corrugated conical horn antenna). A choice of two wide-band, mechanically-tuned Gunn oscillators (right) were available spanning the frequency ranges 75-100 GHz and 100-110 GHz. Source frequency and power levels are set using separate micrometers.

At this point some parameters of the free-space Gaussian beam produced by a horn antenna are set out, which will needed for the design of optical elements. A free-space beam has a waist position located at the horn phase centre, a distance Δz behind the horn mouth as given by

$$\Delta z = \frac{R_h}{1+\left[\frac{\lambda R_h}{\pi W_h^2}\right]^2} \tag{3.2}$$

where the beam parameters W_h and R_h at the horn mouth are determined by the particular choice of horn type and its dimensions. The phase front radius of curvature $R_h = L$ the slant length of the horn. The beam radius W_h depends on horn type and horn mouth size (its radius r, or width a and height b) as summarised for three possible horn shapes in Table 3-1.

Horn Type	W_h
Circular	0.6435 r
Square ($b/a = 1$)	0.433 a
Rectangular ($b/a = 0.7$)	0.35 a

Table 3-1. Horn aperture beam radius W_h for circular horn antennas with radius r and for rectangular (or square) horn antennas of width a and height b [3.9].

The beam waist radius W_0 at the phase centre inside the horn is related to the beam parameters W_h and R_h at the horn mouth as follows

$$W_0 = \sqrt{\frac{W_h^2}{1+\left[\frac{\pi W_h^2}{\lambda R_h}\right]^2}} \tag{3.3}$$

The corrugated conical horn antenna used in experiments has a radius of r = 7.14 mm and a slant length of L = 64.8 mm, which yield beam parameter values at the horn mouth of W_h = 5.1059 mm and R_h= 64.8 mm. At 99.74 GHz (the nearest value to 100 GHz) the beam radius is calculated to be W_0 = 4.7067 mm at the waist position, which is located a distance Δz = 9.7361 mm behind the horn aperture.

3.2.1 The Development of TOAST

Two ball-driven linear positioning systems formed the basis of the new two-dimensional X-Y raster scanning system, referred to hereafter as TOAST (THz Optical Scanning Tool). This is a mechanically stable and smooth-running planar raster-scanned system capable of moving either a single pixel detector, or sample under test across a 2-D plane. At each sample point visited, the voltage reading from the detector, which is proportional to beam intensity, is recorded. The two linear stages were arranged with the horizontal (X-axis) stage mounted onto the table of the vertical (Y-axis) stage (Figure 3-3), which was attached to an aluminium frame constructed in the department workshop. This frame also supported two optical benches onto which optical components were mounted for different experiments and which could be realigned to accommodate the dimensions of the optical system under test. The new system was designed by Dr. W. Lanigan of NUIM.

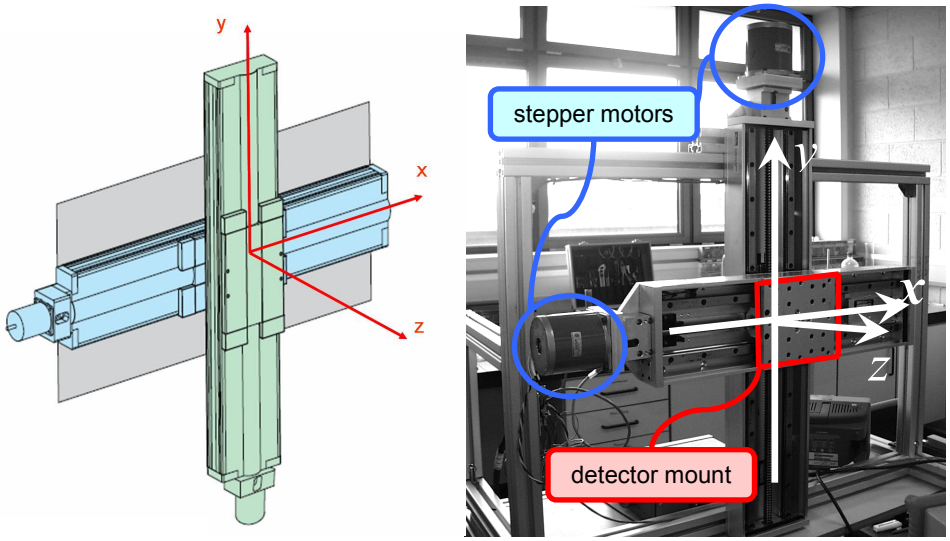


Figure 3-3. Schematic (left) and photograph (right) of the basics of the planar X-Y raster scanner.

The translation stages used were screw driven tables from the 406LN by Parker Positioning Systems[1] (Figure 3-4). Some characteristics of these models are summarised in Table 3-2. The maximum linear speed of these stages is rated at 12 in/sec (304.8 mm/sec). Linear screw speed is the product of screw lead with screw speed, which for the two stages used is a maximum of 150 mm/sec. The high speed, accuracy and smoothness characteristics of these tables mean they are typically used in high throughput, high accuracy applications such as semiconductor processing.

Model #	Axis	Max. Travel (mm)	Max screw speed (rps)	Screw Lead (mm)	positional accuracy (µm)
406012LN	X	300	60	2.5	25
406018LN	Y	450	30	5.0	48

Table 3-2. Characteristics of the two translation stages used to construct the X-Y raster-scanner.

The positioning stages were driven by hybrid stepper motors (Astrosyn[2] high performance model L709). These motors have a step angle of 1.8°, thus a full shaft rotation requires 200 pulses. The minimum linear travel (from a single pulse) of the x- and y-axis stages are (2.5mm/200) = 0.0125mm and (5.0mm/200) = 0.025mm, respectively, so twice as many pulses are required to produce the same travel in the x-direction as in the y-direction. The positioning systems were not supplied with motor coupling units or mounts, which were designed by the author and machined in the department workshop.

Figure 3-4. Photograph of two 406LN series linear screw-driven positioning systems orthogonally mounted for X-Y raster-scanning. These stages are shown with rubber bellows attached to prevent contamination of the ball screw with dirt and abrasive materials.

[1] Parker Hannifin Corporation, Electromechanical Automation Division (www.parkermotion.com)
[2] Astrosyn International Technology Ltd (www.astrosyn.com)

Both stages were equipped with limit and home photogate sensor assemblies mounted inside the working area of each stage. The home sensor provides a fixed reference position to which the table can always return while the limit sensor provides a signal when the table approaches its end of travel. Although these sensors cannot directly track position they were used to calibrate the movement of each stage. The detector – or in some experiments the sample under test – was mounted onto the scanning table (a 5 inch square plate) of the horizontally scanned (x-axis) translation stage.

All electronics and the various power supplies (for motors, their driver boards and translation stage limit sensors) were housed in a single compact casing. Motor scan control, data acquisition and real-time data display was achieved using custom-written LabView software (written by Dr. W. Lanigan of NUIM). The author was responsible for designing the motor connectors and mounts, wiring the translation stage motors to their driver boards and power supplies, and wiring position sensor power supplies and leads.

To provide the user with control over z-axis positioning (of the detector or sample) a high precision manually-operated translation stage with limited travel was attached to the scanning table. This additional facility allows for easy movement of the detector along the optical axis of an optical system, and proved particularly useful when performing reflection-mode imaging experiments.

Because of the large amount of data generated during the experimental phase of research, software was written (by the author) to allow one to quickly extract, process, and plot measured intensity data. Here we briefly describe key features of a program called DSR (Display Scan Results) that was written in the MATLAB environment to extract and display one- and two-dimensional intensity data from text files generated (upon completion of successful measurements) by both the GHOST and TOAST planar scanners.

Features of DSR are grouped according to data processing or plotting features. The latter include an extensive range of one- and two-dimensional (standard and custom-written) plot types that support *local* and *global* coordinates systems, as well as the ability to convert between plot types. Some standard image processing features of MATLAB and the Image Processing Toolbox were also included. An important feature of DSR is its ability to identify the axis tiling arrangement of individual axes within a given figure – a feature not available with MATLAB. This allows precise control of

plotting and re-plotting within individual axes, without having to specify exact axes positions.

The image processing tools included column-error correction (CEC). An unidentified problem with the data acquisition software and/or positioning hardware produced, in some scans, misalignment of columns within recorded 2-D data files (Figure 3-5). Data-processing code that was written to correct for this problem, supports both automatic and manual identification and realignment of problematic (single, alternating, or contiguous blocks of) data columns. This operation is necessary if further processing operations are to be applied to images; also low-level features, that might otherwise be obscured, are revealed after proper column realignment is performed.

Other image processing tools included in DSR are spatial and frequency filtering as well as interpolation. The spatial filtering tools found to be of most use were median filtering and adaptive noise removal filtering, both of which are features of MATLAB's Image Processing Toolbox. Interpolation was included to account for the fact that multiple intensity measurements of the same scene were recorded using different step sizes. Interpolation allows one to interpolate a measurement made on a roughly sampled grid to another made on a more densely sampled grid to allow for comparison of the two measurements, as well as for the construction of a composite image from two or more individual images.

The latest version of DSR features a command-line user-interface with the option of using either alpha-numeric keyboard input or hyper-links to navigate and select from available options. A future version would support the option to interface via dialogue boxes (with features such as command windows, buttons, drop-down menus, check boxes, etc.).

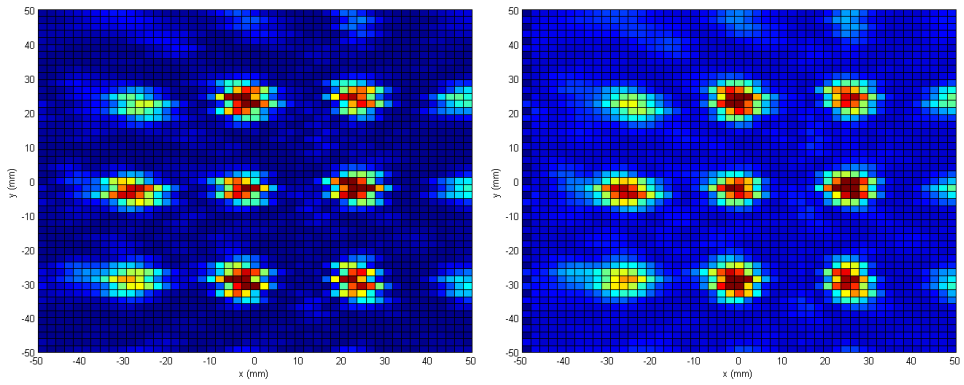


Figure 3-5. Left: Intensity measurement (from a phase grating) made with TOAST (*a*) before and (*b*) after column-error correction was performed. In this example alternate columns were shifted left/right by two places.

3.2.2 Design and fabrication of optical components

This section describes the design and manufacture of the optical components used for free-space beam guides in the various imaging experiments. Optical components were also useful for the active illumination of an object by the source beam (including those involving spatial filtering).

Although a number of optical components had been fabricated in-house during previous research projects, most were inadequate for the purposes in terms of low aberration imaging across a useful field of view. Most of these components had large angles of throw, or long focal ratios, or were not machined to a sufficiently high quality in terms of surface accuracy.

The pre-existing optical components included four 250mm focal length plano-convex lenses. These lenses were adequate for the experiments designed to measure absorption and reflection properties of various transparent materials, since only relative on-axis intensity was measured as the material being tested was inserted to fill the beam between the source and detector, which were well coupled. However when uniform, or at least Gaussian, illumination of test objects is required, these lenses prove wholly inadequate. Besides producing a non-uniform modulation of the beam, the Gaussian beam produced was too small for our needs, even for illumination purposes since we intended to perform measurements on everyday objects.

A pair of paraboloidal mirrors of focal length 150mm (with a 90° angle of throw) designed to operate at 100 GHz [3.7] was also available. The long focal ratios of these mirrors made them unsuitable for the proposed imaging experiments, which required mirrors with a short focal ratio and a lower angle of throw that could collect sufficient radiation scattered from illuminated objects and re-image without significant aberrations (including distortion). These mirrors did however find use in imaging experiments for the illumination of test objects.

A set of three 500mm ellipsoidal mirrors (with 45° angle of throw) that were optimised for experiments at 25 GHz ($\lambda \approx 12$ mm) [3.10] were also available, but were found to be problematic on several grounds. Although at 25 GHz their large collecting angles would have made them appropriate for (low resolution versions of) the imaging experiments envisaged, when operated at 100 GHz the focal ratios of these mirrors were too slow for high resolution imaging. Also generally an off-axis ellipsoidal mirror is designed to match specific radii of curvatures of incident and reflected beams. It is

therefore optimised for a particular operating frequency and its performance may not be wideband. Thus the original ellipsoidal mirrors were not ideal for operation at 100 GHz. Furthermore, two of the three mirrors were poor candidates: one still needed finishing (cutting and polishing), while the other was made from two thin aluminium blocks held together on the milling machine table that became slightly warped when released from the table, which resulted in the mirror producing poor quality images (due to a diffused point spread function). Another problem at the design stage saw rounding errors being introduced into the parameters of the ellipse on which these mirrors were based. The shape of an ellipse is defined by three parameters: the semi-major axis length a, the semi-minor axis length b and the semi-foci separation c. When used as the basis for a reflector design the values of these parameters depend upon the required focal length f, angle of throw θ and input and output beams' phase front radii of curvature R_1 and R_2. If the value of one phase front radius of curvature is known, the other can be determined using the Lens Makers Formula. It was discovered that when calculating R_2 in this way the value of R_1 was mistakenly rounded to the nearest millimetre. This actually resulted in a sufficiently inaccurate value for R_2 and hence incorrect values for parameters a, b and c, thus producing an ellipse profile unsuitable for the intended design.

Another problem with the existing ellipsoidal mirrors is that they had been machined such that their alignment required precise knowledge of the angles that incident and reflected beam would make with respect to the normal at their optical centre. Although this does not affect operation of the mirror it does make alignment of the mirrors more complicated. Alignment is greatly simplified if the normal to the mirrors' reflecting surface at the optical centre C coincides with the normal to the metal block from which the mirrors were machined. This approach was implemented when designing the new mirrors.

When conducting imaging experiment that incorporates spatial filtering it is desirable to be able to preserve the high spatial frequency content of the beam in order to resolve small-scale features of the test object. This requires the use of large aperture (i.e. fast) optics and since none of the existing optics described above satisfied this requirement new optical components were needed. It was decided to use reflective focusing elements (mirrors) since they do not suffer from absorption and reflection losses that are characteristic of refractive focusing elements (lenses). Furthermore mirrors are not subject to problems of frequency-dependent performance suffered by lenses due to

unwanted surface reflections. The choice of off-axis reflectors avoids diffraction problems that can occur with on-axis mirrors (designed for normal incidence) when the source or detector blocks part of the propagating beam.

Two sets of canonical off-axis reflectors were designed and fabricated: two 350mm focal length paraboloidal mirrors (with a 90° angle of throw) and two 500mm focal length ellipsoidal mirrors (with a 45° angle of throw). The two sets of mirrors would allow to us to assemble different Gaussian beam telescope configurations for use in the "far-field" re-imaging experiments. The near-field transmission experiments can be implemented using just one mirror to provide quasi-uniform illumination of test objects. The long focal length (500 mm) mirrors were designed for use in re-imaging experiments. The shorter focal length (350mm) mirrors were included so that sets of optics with focal lengths approximately evenly spaced between 150 mm and 500 mm would be available for future experiments. A 90° angle of throw was chosen for the off-axis paraboloidal mirrors so that compensating systems could be assembled. In such a system one mirror counteracts the distortions generated by the other [3.11].

The new mirrors were designed by the author and manufactured by David Watson in the workshop of the NUIM Department of Experimental Physics by cutting the required surface (a paraboloid or ellipsoid of revolution) into a block of cutting grade aluminium using a CNC milling machine. Ideally the new mirrors for a given focal length should be made as large as possible in order to collect sufficiently high spatial frequency content. Practical limitations however imposed an upper limit on the component size, as the maximum working area of the CNC milling machine used was approximately 300 mm × 600 mm.

Ellipsoidal Mirror Parameters

An ellipsoidal mirror is designed to match the radius of curvature of one free-space beam exactly to that of another. The surface profile of an ellipsoidal mirror is thus defined by the beam parameters: radiation wavelength λ, the radii of curvature (R_1 and R_2) and waist radii ($W_{0,1}$ and $W_{0,2}$) of the incident and reflected beams; as well as the required focusing properties of the mirror itself: its focal length f and angle of throw $\theta = 2\theta_i$, where θ_i is the angle of incidence. The ellipsoidal mirrors described here were designed to have $f = 500$ mm and $\theta = 45°$ (i.e. $\theta_i = 22.5°$).

An ellipse is the set of all points in a two-dimensional plane, the sum of whose distances from two fixed points – focal points F_1 and F_2 – is constant and greater than the distance between the fixed points (Figure 3-6). An ellipse in one of the standard positions (centred on the origin, with the foci and hence the major axis on the x-axis and the minor axis on the y-axis) is described by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{3.4}$$

where the major axis spanning the ellipse and containing F_1 and F_2 has a half-length of a – the semi-major axis length. The minor axis intersects, perpendicularly the major axis at its midpoint and has a half length of b – the semi-minor axis length. The foci are separated by a distance of $2c$.

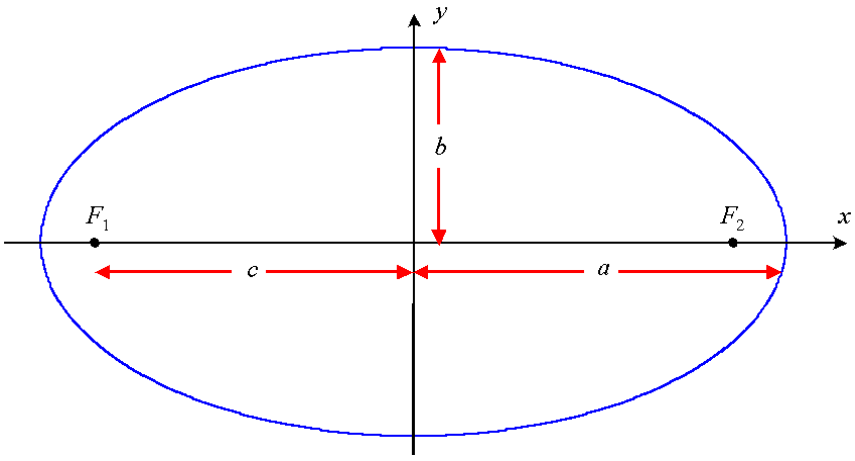


Figure 3-6. Ellipse illustrating foci F_1 & F_2, semi-major and -minor axes lengths a and b and semi foci distance c.

The ellipse geometry is chosen so as to match the phase fronts of the incident and reflected beams, so the incident and reflected radii of curvature R_1 and R_2 must first be calculated. The incident beam is at a waist position at a distance $z = f$ from the optical centre C of the mirror (Figure 3-7). The phase front radius of curvature is assumed to be infinite at the waist position (some distance inside the horn antenna used to feed the source) but initially decreases with propagation distance z (and goes through a minimum value) until in the far-field its value increases linearly with z. Between the waist position and far-field, and thus when the incident beam intercepts the ellipse at point C, the radius of curvature R_1 is given by

$$R_1 = z\left(1+\left(\frac{\pi W_0^{\,2}}{\lambda z}\right)^2\right) \tag{3.5}$$

where we set $z = f$ and $W_0 = W_{0,1}$ is the incident beam waist radius. The mirror focal length f is defined in terms of the required change in phase-front curvature (from R_1 to R_2) as

$$\frac{1}{f} = \frac{1}{R_1} + \frac{1}{R_2} \tag{3.6}$$

which gives a value for the reflected phase-front radius of curvature of

$$R_2 = \frac{R_1 - f}{fR_1} \tag{3.7}$$

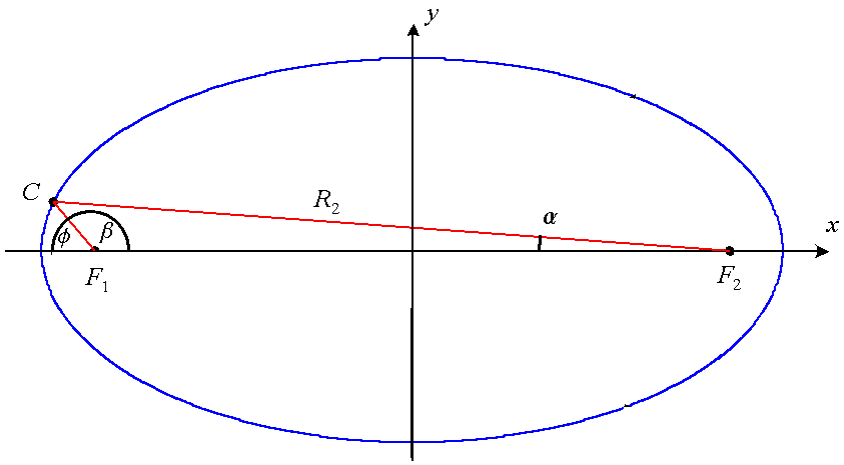


Figure 3-7. An ellipse showing the incident beam launched from focal point F_1 to the optical centre C a distance R_1 (incident beam phase-front radius of curvature) and reflected along path R_2 (reflected beam phase-front radius of curvature) to the ellipses second focal point F_2.

These mirrors will be used to focus a Gaussian beam produced by a horn antenna that couples a waveguide beam to a free-space beam. The incident beam waist radius W_0 is determined by the dimensions and type of horn antenna used to feed the source/detector as described at the start of §3.2. When used in conjunction with a horn-fed source the mirror must be positioned such that the beam waist at the source is located the focal length f away from the optical centre C on the mirror surface. Thus the distance from horn aperture to C is equal to $(f - \Delta z)$, where Δz, the distance of the phase centre from the feed-horn aperture, is given by Eq. (3.2).

Next the defining parameters (lengths a, b and c) of the ellipse are calculated using the following equations

$$c = \tfrac{1}{2}\sqrt{R_1^2 + R_2^2 - 2R_1R_2\cos\theta} \tag{3.8}$$

$$a = \tfrac{1}{2}(R_1 + R_2) \tag{3.9}$$

$$b = \sqrt{a^2 - c^2} = a\sqrt{1 - e^2} \tag{3.10}$$

where ellipticity $e = c/a < 1$ for an ellipse [3.12]. An ellipsoid of revolution is formed by revolving an ellipse about its major axis. An off-axis reflector employs just a small segment of the ellipse that is usually located well away from the axis of symmetry. This isolated segment is then revolved through a narrow angle about the major axis.

Next we must consider the rims (edges) of the mirror. The mirror must be large enough to collect sufficient power from the incident beam without too much truncation. One rule of thumb often applied to component design is to use a diameter of $4W$, where W is the beam radius where amplitude of the approximately Gaussian beam incident on the mirror drops to $1/e$ of its on-axis value [3.9]. Equivalently in the far-field of the waist position we can use the angle of divergence

$$\theta_W = \tan^{-1}\left(\frac{\lambda}{\pi W_0}\right) \tag{3.11}$$

to decide on mirror diameter. The edges of the ellipse segment (e_1 and e_2) are located at the intersection of the ellipse with two straight lines (R_{e_1} and R_{e_2}) that begin at focal point F_1 and represent the illumination cone of the incident beam as shown in Figure 3-8. The new mirrors are designed to have a surface area large enough to collect radiation from a cone of illumination whose angle γ (shown in Figure 3-8) is twice the angle of divergence θ_W.

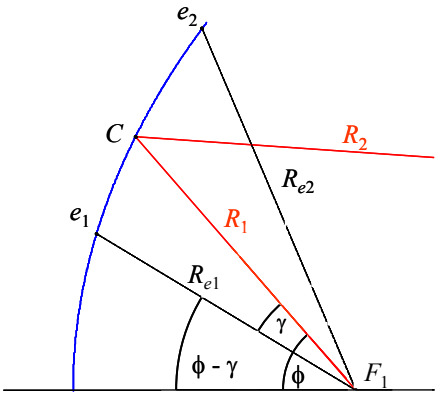


Figure 3-8. Close-up of ellipse segment that is used to form ellipsoidal mirror surface. The isolated segment is bounded by endpoints e_1 and e_2 where the ellipse intersects lines R_{e1} and R_{e2} representing the cone of illumination from a source near F_1.

After some algebra the x-coordinates of endpoints e_1 and e_2 are determined to be

$$x_e = \frac{-a^2 c\tan^2\delta \pm ab\sqrt{b^2 + (a^2 - c^2)\tan^2\delta}}{a^2\tan^2\delta + b^2} \tag{3.12}$$

where angle δ is the angle subtended by the beam edge with respect to the major axis and which has values of $(\gamma - \phi)$ for e_1 and $(-\gamma - \phi)$ for e_2. The angle γ is that subtended by

the beam edge at F_1 with respect to axis of propagation F_1C. The angle subtended by F_1C with respect to the negative x-axis is $\phi = \pi - \beta$ where the angles

$$\alpha = \sin^{-1}\left(\frac{R_1 \sin\theta}{2c}\right) \quad \text{and} \quad \beta = \sin^{-1}\left(\frac{R_2 \sin\theta}{2c}\right) \tag{3.13}$$

in Figure 3-7 specify the orientation of R_1 and R_2 with respect to the major axis. Equation (3.12) yields two points of intersection for each beam edge (e_1 and e_1' , and e_2 and e_2'). We are only interested in those two points that bound the ellipse segment containing point C since the others produce a mirror with very different focusing properties.

As already indicated, it is preferable for alignment purposes that the mirror be designed so that the normal to the mirror plane (the plane parallel to the rear face of the rectangular metal block into which the mirror surface is machined) be set parallel to the bisector of the angle of throw at the optical centre of the mirror. This is achieved by ensuring that the mirror plane is set parallel to the tangent plane at the optical centre.

Paraboloidal Mirror Parameters

Although ellipsoidal mirrors can be used to form a 4-f system often paraboloidal surfaces form a good approximation without introducing significant levels of aberrations due to phase distortions. This section describes the procedure that was used to design a pair of off-axis paraboloidal mirrors designed to have 350 mm focal length and 90° angle of throw.

A parabola is defined as the set of all points on a plane equidistant from a line called the directrix D and a fixed focal point F not on that line. One possible orientation, or standard position, of a parabola, as illustrated in Figure 3-9, is described by the expression

$$y^2 = +4px \tag{3.14}$$

where p is the distance from F to the vertex (where the parabola intersects the axis of symmetry) and from the vertex to D. A paraboloid of revolution is formed by rotating the parabola about its axis of symmetry, which in this case is the x-axis.

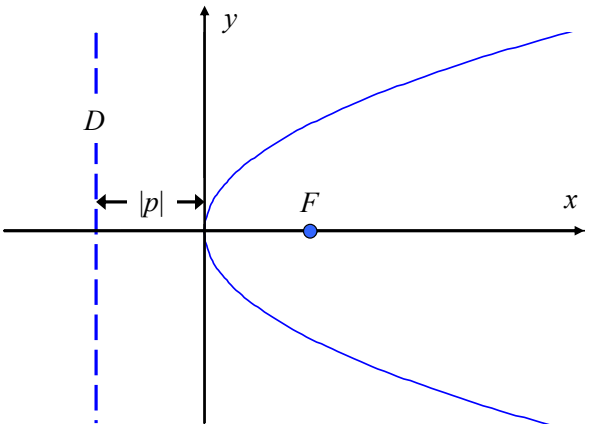


Figure 3-9. Parabola of Standard Position 1, i.e. parabola is symmetric about the x-axis and opens in the positive x-direction. The vertex (at the origin) is equidistant from directrix D and focal point F.

When used as the basis for an off-axis reflector design the paraboloidal surface profile is dependent on two sets of parameters: the beam parameters (wavelength λ, waist position and size) and the required mirror parameters (focal length f and angle of throw $\theta = 2\theta_i$, where θ_i is the angle of incidence made by an input beam at a point on the parabolic surface with respect to the surface normal at that point). The incident beam waist location is assumed to be located at the focal point F of the parabola, where the beam waist radius W_0 is determined by the specific horn antenna used at the source. After the appropriate off-axis parabolic section is determined it is revolved about the axis of symmetry to generate a three-dimensional paraboloidal surface. The mirror surface required is that part of the paraboloidal surface that lies inside the incident beam cone.

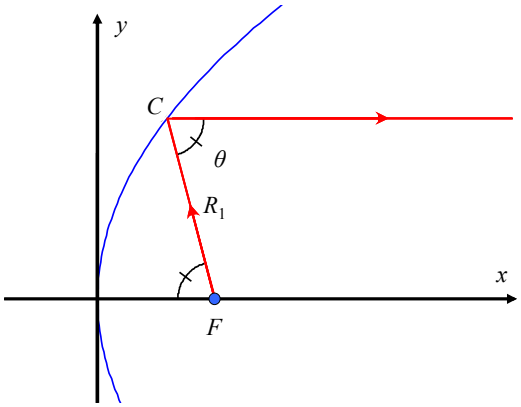


Figure 3-10. A ray launched from focal point F is reflected parallel to the axis of symmetry through an angle of throw θ at point C – the optical centre – on the parabolic surface. The separation between F and C is denoted R_1.

Consider the parabola shown in Figure 3-10. The parabola is defined by parameter p, the distance from focal point F to the vertex. A ray launched from the focal point F at an angle θ with respect to the x-axis is reflected from the point C a distance R_1 from F. The reflected ray is sent parallel to the axis of symmetry at an angle θ – the angle of throw. With the vertex of the parabola at the origin in Cartesian coordinates the optical centre C is located at the point

$$(x_C, y_C) = (p - R_1\cos\theta, R_1\sin\theta) \tag{3.15}$$

where R_1 is the phase-front radius of curvature from the input beam waist position. Substituting the coordinates for C into the defining equation of the parabola, given by Eq. (3.14) and solving for p yields values of

$$p = R_1\cos^2\theta_i \tag{3.16}$$

and

$$p = -R_1\sin^2\theta_i \tag{3.17}$$

which correspond to the distances from the vertex to focal point F (in the positive x direction) and from the vertex to the directrix D (in the negative x direction), respectively.

The rims (edges) of the off-axis section are calculated in the same way as was done previously for the ellipsoidal mirrors. First the angle of divergence γ and then the collecting cone angle (2γ) with respect to the focal point are calculated. Next two lines representing the outer beam edges are drawn from the focal point. The points at which these lines intersect the parabola are taken to be the edges of the off-axis section of parabola. These points are given by

$$x_p = \frac{p(\delta+2) \pm 2p\sqrt{\delta^2+1}}{\delta^2} \tag{3.18}$$

where $\delta = \tan(\theta \pm 2\gamma)$, the sign depending on the outer beam edge being considered.

If these coordinates are taken as the edges of the off-axis section then the line joining these points will be parallel to the tangent plane at the optical centre C. This will result in different input and output beam angles with respect to the normal of the mirror frame and make alignment difficult. Instead we require that the tangent plane at C be set parallel to the rear face of the metal block from which the mirror surface is machined. The slope of the tangent plane at point C is given by

$$\left.\frac{dy}{dx}\right|_C = \frac{2p}{y_C} \tag{3.19}$$

The tangent subtends an angle of $(\pi/2-\theta_i)$ with respect to the x-axis. The parabola is rotated through this angle to ensure that the tangent plane is parallel to the mirror frame and then translated so that the optical centre is at the origin.

3.2.3 Mirror Manufacture and Testing

Mirrors were constructed by cutting the required three-dimensional surface profile into a block of cutting grade aluminium using a computer numerically controlled (CNC) 3-axis Hurco Hawk milling machine[3]. Mirror parameters calculated in the previous section were used to generate numerically controlled (NC) code using a commercial computer aided design and computer aided manufacture (CAD/CAM) software package called Alphacam[4]. The milling machine control software uses NC code to specify cutting paths taken by cutting tools. Multiple passes were required using progressively smaller ball cutters to achieve sufficiently high surface accuracy. The first pass used a high speed, 8 mm ball-nosed, carbide cutting tool, which is designed for cutting curved surfaces at high speed. Figure 3-11 shows a 500 mm focal length mirrors being machined in the NUIM Department of Experimental Physics workshop. After cutting, each mirror was polished with increasingly fine emery papers to achieve an optically smooth surface so mirror alignment could be verified using a laser. Each mirror was designed with its optical centre offset from the physical centre of the mirror so reference points were etched into the top and bottom surfaces to aid alignment.

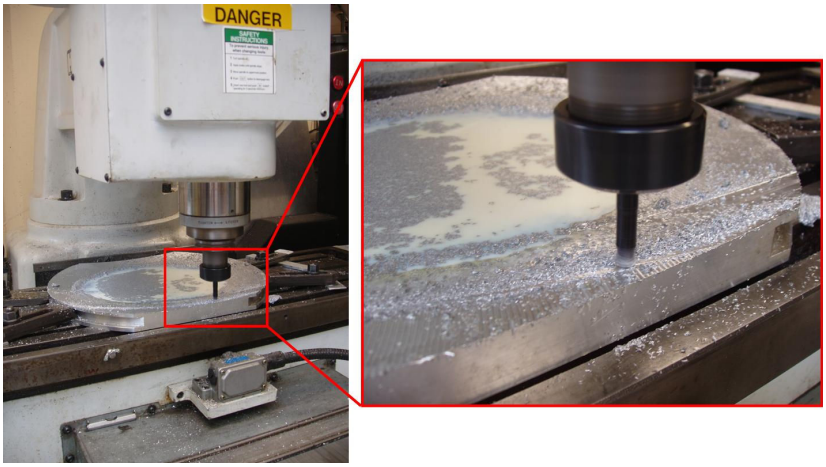


Figure 3-11. Machining of one of the 500 mm focal length ellipsoidal mirror on the CNC milling machine. Several cutting passes were made with increasingly smaller ball cutters, which are moved along pre-programmed paths which are specified by NC code.

[3] Hurco Companies, Inc.
[4] from Licom Systems Ltd. (http://world.alphacam.com/)

Before measurements of the various mirrors are presented beam pattern measurements (undertaken by the author) of one of the existing 250 mm focal length HDPE lenses are presented for comparative purposes. The four converging lenses are made from HDPE (refractive index = 1.525 [3.32]) and take the form of plano-convex lenses since they are easiest to make and require machining of only one side using a lathe.

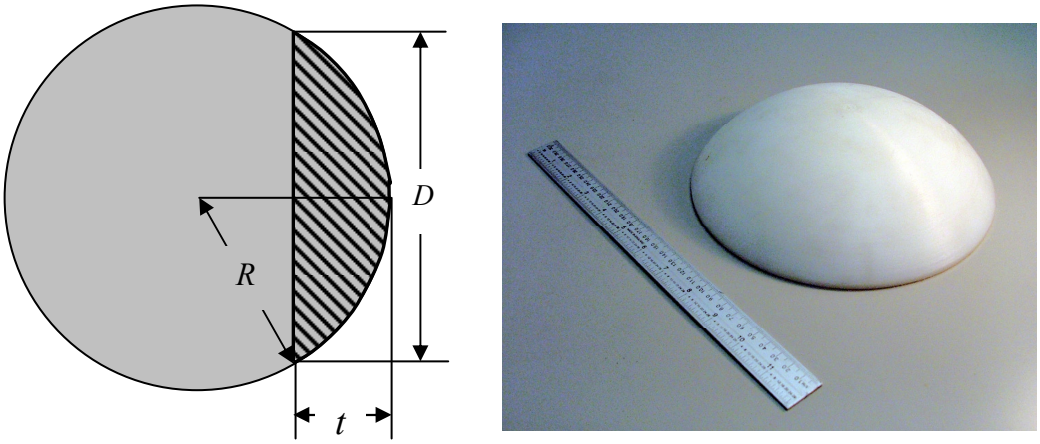


Figure 3-12. (Left) Construction used in the design of the 250mm focal length plano-convex lenses made from HDPE (shown right). Each lens has a diameter D = 220mm and thickness t = 63.5 mm.

A measurement of the incident expanding beam transmitted from the source antenna (a corrugated conical horn antenna to provide waveguide-to-freespace coupling to the Gunn oscillator at a frequency of 100 GHz) was measured at a distance of 230mm from the horn antenna phase centre - the distance where the incident beam would intercept the convex surface of the plano-convex lens [3.2]. The intensity profile of the beam at this distance has a circularly symmetric Gaussian profile with a radius of W = 46.99 mm. The measured intensity profile (Figure 3-13 & Figure 3-14) is slightly asymmetric and has an estimated radius of ~42 mm. Interference occurs across the beam pattern, indicating that standing wave effects are established between source and detector.

The measured beam intensity from a plano-convex lens (Figure 3-12) is shown in Figure 3-15. The overall shape is reasonably symmetric with a Gaussian profile of the correct size (W ~51 mm) but the intensity distribution varies dramatically across the beam. Although one possibility for the uneven intensity distribution observed is absorption losses by the lens material, particularly near the optical axis where the lens is quite thick, the dominating factor on low beam quality is most likely due to standing waves. Since the lens was measured with its curved surface facing the source standing waves established between the source and the plane surface of the lens would result in the constructive and destructive interference that are observed in Figure 3-15.

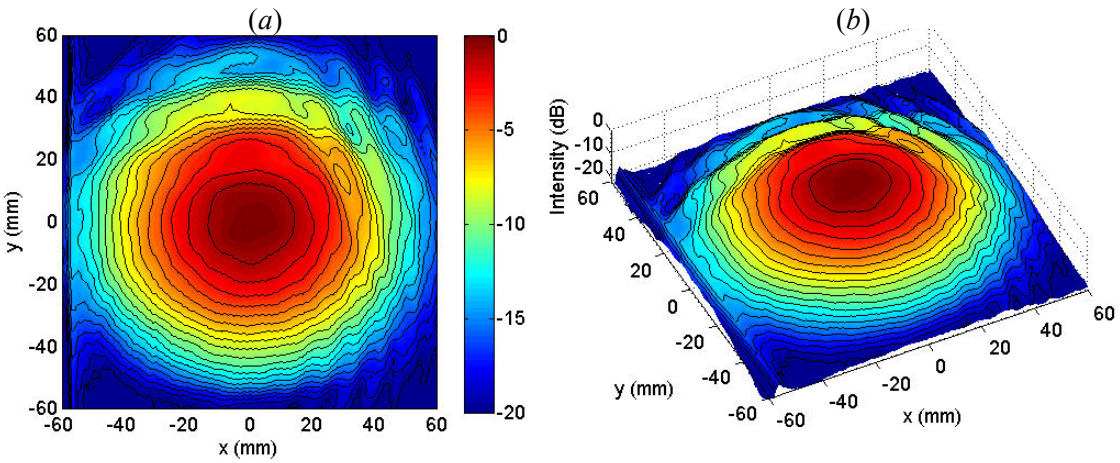


Figure 3-13. Log-scale plots of measured beam intensity at a distance of 230mm from a corrugated conical horn antenna at 100 GHz.

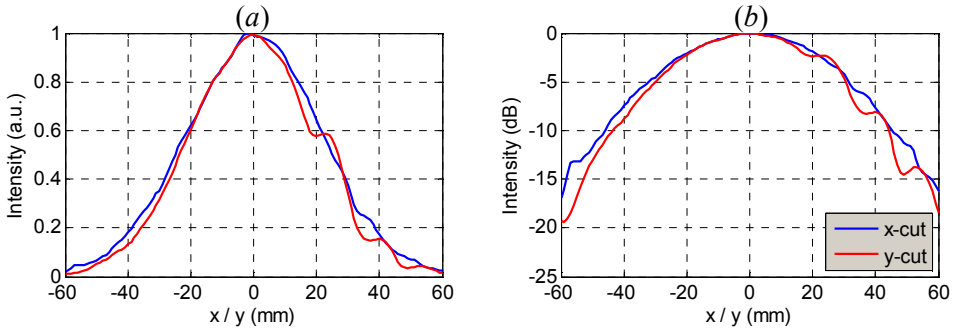


Figure 3-14. (*a*) Linear and (*b*) log-scale plots of cuts through centre of beam pattern shown in Figure 3-13 measured at 230 mm from a corrugated conical horn antenna at 100 GHz.

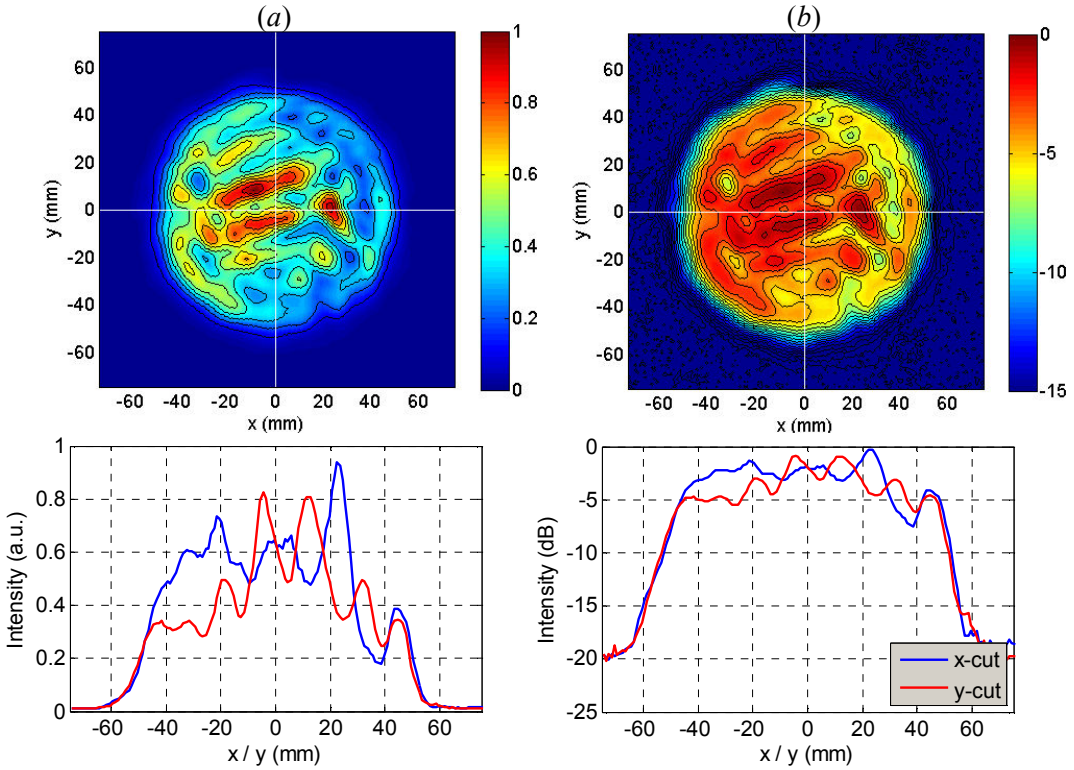


Figure 3-15. (*a*) Linear-scale and (*b*) log-scale plots of intensity measured at output plane of 250 mm focal length HDPE lens.

Ellipsoidal Mirrors

Several of the manufactured ellipsoidal mirrors are shown in Figure 3-16. Note that the pre-existing mirrors had a maximum width of 375.92 mm however the newly designed mirrors were set to have a maximum width of 399 mm. This limits the extent of the new mirror profile and may result in power loss because of the smaller surface area as well as scattering at the mirror rims.

(*a*) (*b*) (*c*)

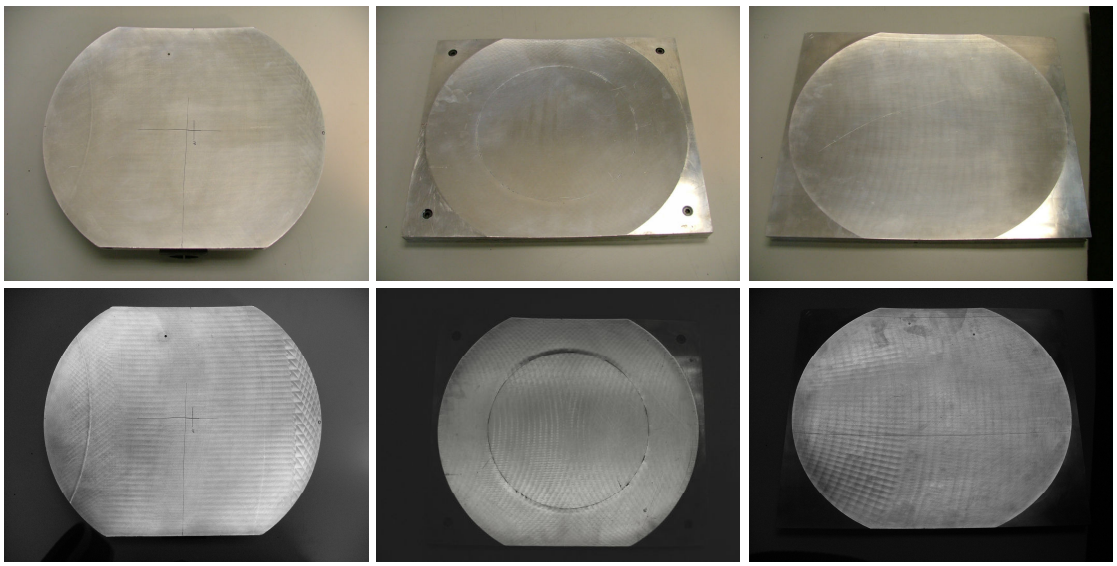


Figure 3-16. Photographs of the 500 mm focal length ellipsoidal mirrors. (*a*) One of the original mirrors after re-cutting for use at 100 GHz, (*b*) the two-section ellipsoidal mirror and (*c*) one of the newly designed mirrors, two of which were made.

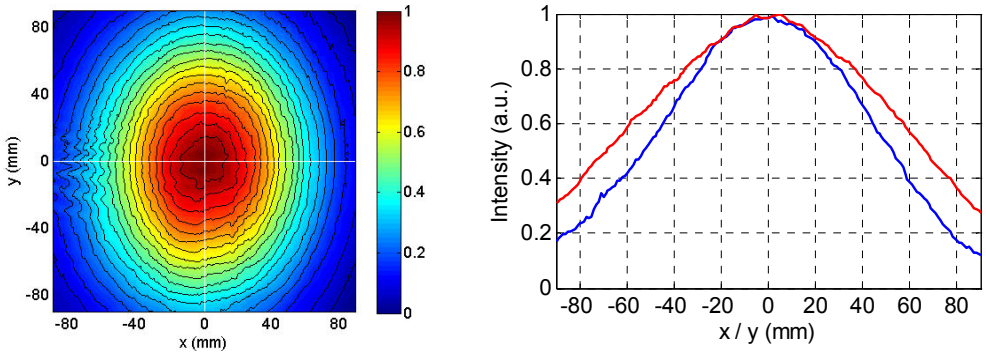


Figure 3-17 and Figure 3-18 show intensity measurements made at the output focal planes of one of the pre-existing and one of the newly designed 500 mm focal length (45° angle of throw) ellipsoidal mirrors. The horizontally aligned interference ripples (presumably due to standing waves) observed in the centre of the measured beam

pattern in Figure 3-18 do not appear in the measurement in

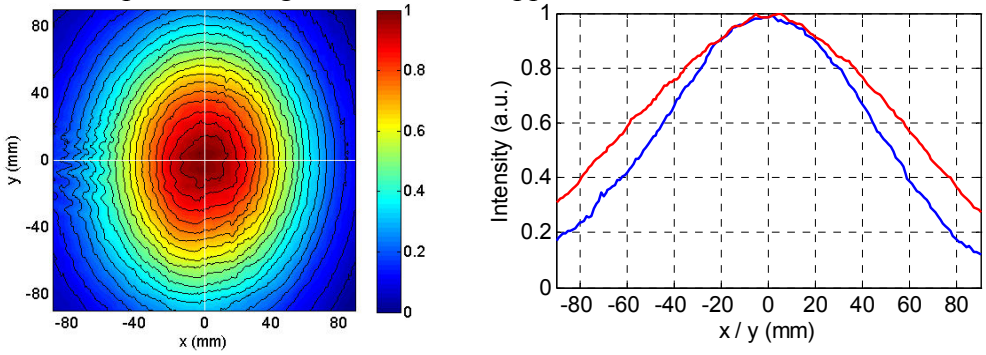


Figure 3-17.

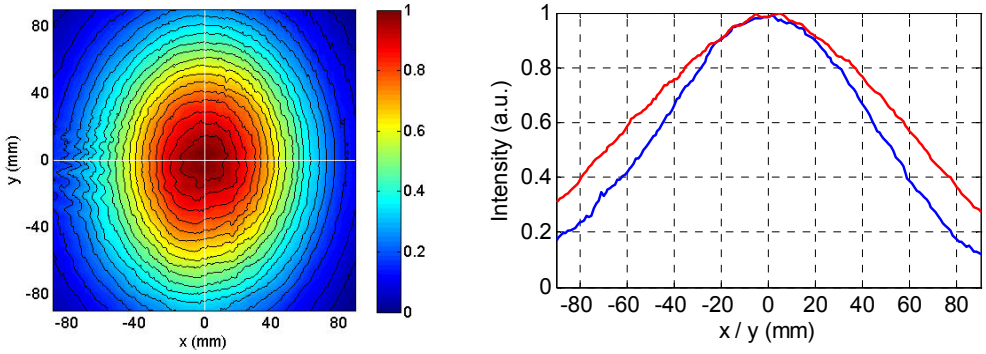


Figure 3-17. Linear-scale plots of intensity measured at the output focal plane of one of the newly designed 500 mm focal length ellipsoidal mirrors. The vertical cut through the origin is coloured red.

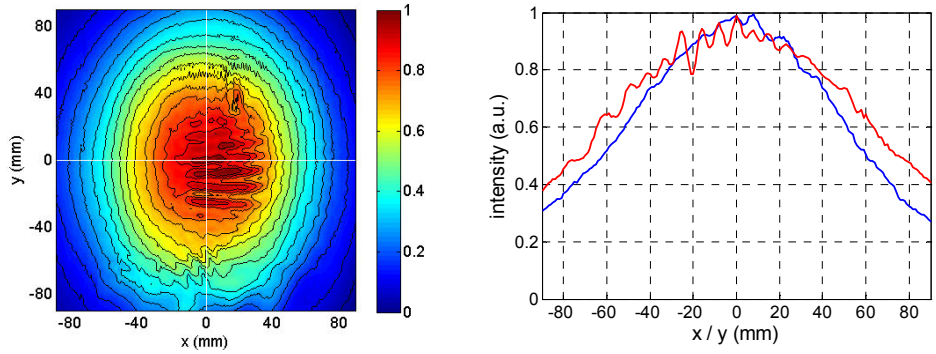


Figure 3-18. Linear-scale plots of intensity measured at the output focal plane of one of the pre-existing 500mm focal length ellipsoidal mirrors. Note the presence of horizontally aligned interference fringes, which are presumably due to unwanted reflections introducing standing wave effects into the system.

Paraboloidal Mirrors

The paraboloidal mirrors were tested by measuring the intensity at the output focal plane when illuminated with a 100 GHz source with waveguide-to-free space coupling provided by a corrugated conical horn antenna. The intensity was measured at the mirrors output focal plane, a distance *f* from the mirrors optical centre. The measured beam is expected to have a Gaussian intensity profile. Treating the mirror as an ideal thin lens, the expected beam radius at the mirrors output focal plane is given by

$$W_f = \frac{\lambda f}{\pi W_0} \tag{3.20}$$

where f is the mirrors focal length and W_0 is the waist radius of the expanding incident beam that the paraboloidal mirror is designed to collimate. At 100 GHz the incident beam from one of the corrugated conical horn antennas has an intensity profile that is well approximated by a Gaussian beam with a waist radius (at the horn phase centre) of $W_0 = 4.7053$ mm.

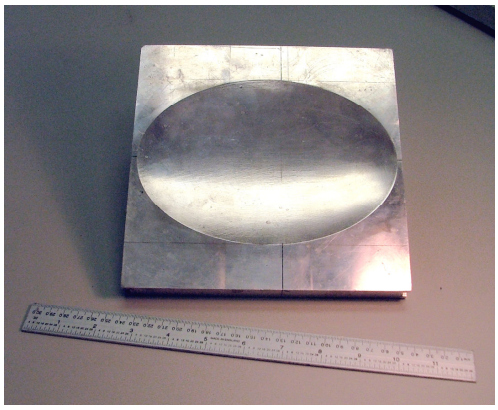


Figure 3-19. A pre-existing paraboloidal mirrors with focal length of 150 mm and 90° angle of throw.

Before testing the new 350mm focal length paraboloidal mirrors measurements were made of the existing 150mm focal length paraboloidal mirrors (Figure 3-19), which were used to illuminate test objects in transmission imaging experiments.

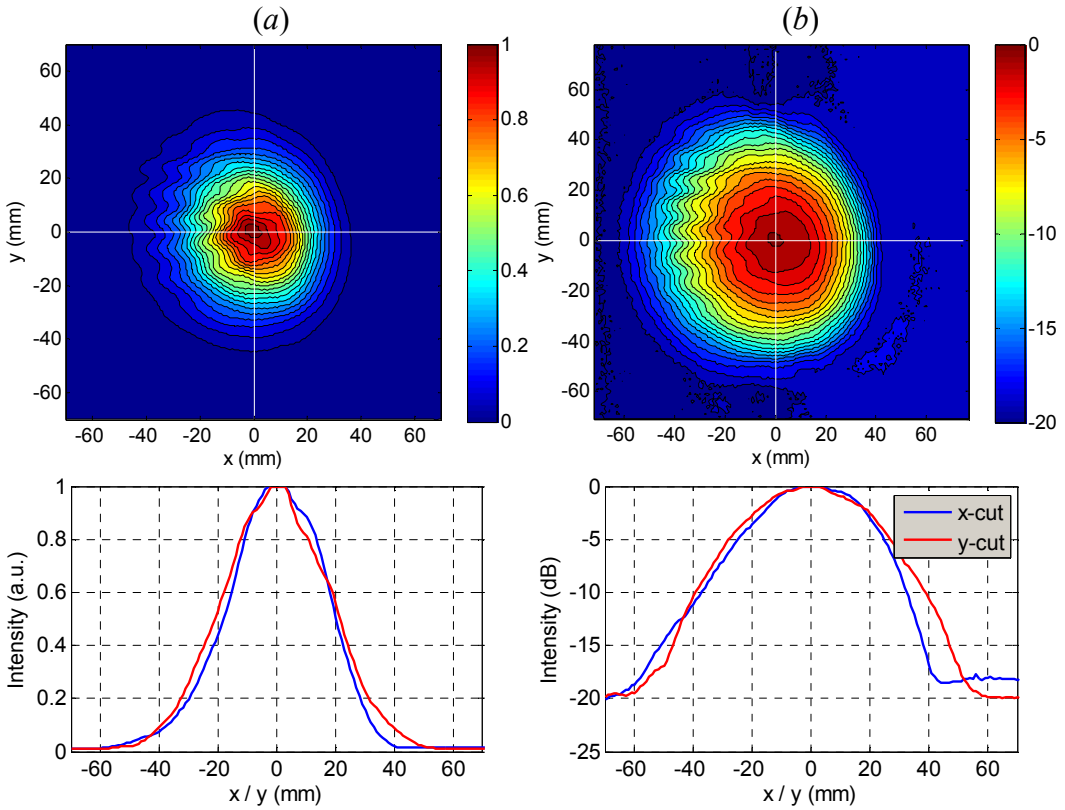


Figure 3-20. (*a*) Linear-scale and (*b*) log-scale plots of intensity measured at the output focal plane of one of the 150 mm focal length paraboloidal mirrors. The two lower plots show horizontal and vertical cuts taken though the estimated beam centre.

A 150mm focal length lens should produce a Gaussian beam with radius of 30.44 mm. Figure 3-20 shows the intensity measured at the output focal plane of one of the 150 mm focal length paraboloidal mirrors. The beam pattern is approximately Gaussian but is distorted, particularly in the horizontal direction due to the 90° angle of throw between incident and reflected beams. Notice also the appearance of horizontally aligned interference fringes. These may be due to standing wave effects due to unwanted reflections in the system. When used in imaging experiments strips of absorbing material were attached to the flat metal surface surrounding the parabolic surface in order to reduce standing wave effects.

Next the 350 mm focal length paraboloidal mirrors were tested. At 100 GHz the collimated Gaussian beam produced by a 350 mm focal length lens would have a radius of 71.03 mm. Figure 3-21 shows the intensity measured the output focal plane of one of the 350 mm focal length paraboloidal mirrors. Again the beam exhibits distortion in the horizontal direction, due to the mirrors high angle of throw. However, unlike the smaller paraboloidal mirror, no interference fringes are observed. The estimated power coupling efficiency between the measured beam intensity from the 350mm focal length mirror and a symmetric Gaussian beam of radius 71mm is 98.00%. A maximum power coupling efficiency of 98.97% was found to occur for an asymmetric Gaussian beam with radii of (W_x, W_y) = (75.4, 80) mm.

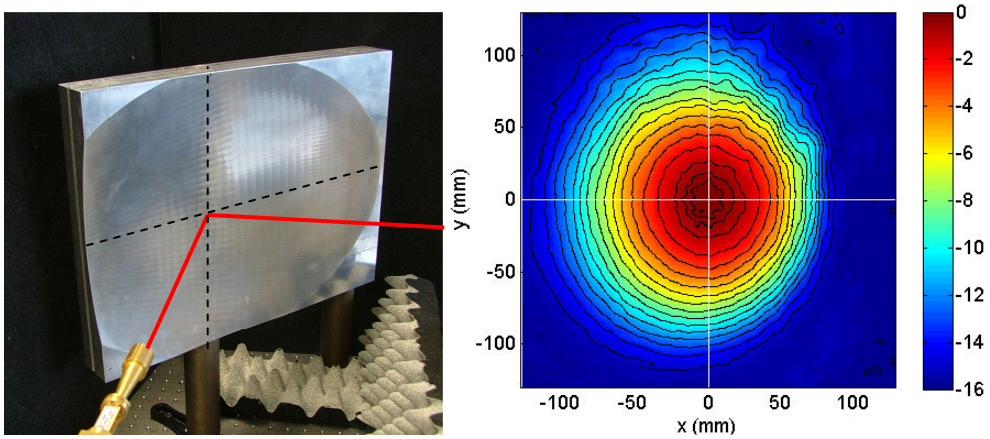


Figure 3-21. **Left:** One of the 350mm focal length paraboloidal mirrors fed by a source with a corrugated conical horn antenna. The solid red lines indicate the beam path taken from the source antenna towards the output focal plane (right of picture). **Right:** Log-scale plot of measured output focal plane intensity.

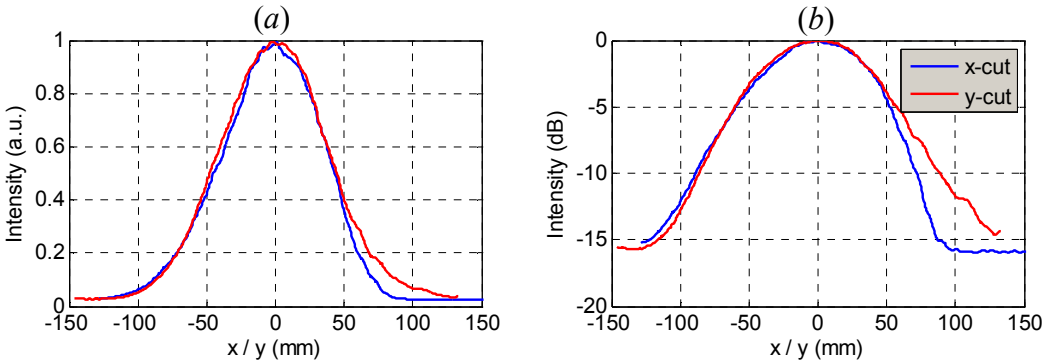


Figure 3-22. (*a*) Linear-scale and (*b*) log-scale plots of horizontal (blue curves) and vertical (red curves) cuts through the output plane intensity measurement from a single 350 mm focal length paraboloidal mirror.

Next two paraboloidal mirrors, M_1 and M_2 were arranged in a 4-*f* configuration such that a collimated beam waist was produced midway between M_1 and M_2 and the intensity at the output plane measured. The quasi-collimated beam produced by M_1 is refocused by M_2 to produce an image of the source beam at the output plane. Thus the measured intensity should be Gaussian in profile with a radius equal to that of the source beam, which in this case is a beam that is well approximated by a Gaussian beam with a radius equal to that at the phase centre of the source horn antenna, i.e. W_0 = 4.715mm. The mirrors were arranged in a compensating configuration such that the distortions generated by one (tend to) cancel those generated by the other [3.11, 3.15]. The intensity measured at the output of the 4-*f* system is found to have a coupling efficiency of ~99% to a Gaussian beam with beam radii of (W_x, W_y) = (4.67, 4.85) mm, very close to the expected beam radius W_0 (see Figure 3-23 and Figure 3-24).

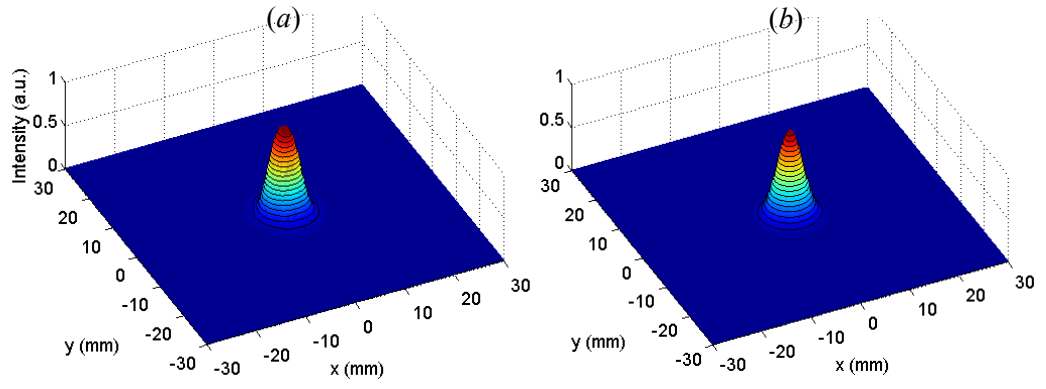


Figure 3-23. Linear-scale plot of (*a*) intensity measured at output plane of 4-*f* system (consisting of two 350mm focal length paraboloidal mirrors), which couples with ~99% efficiency to (*b*) a best-fit Gaussian with radii (W_x, W_y) = (4.67, 4.85)mm.

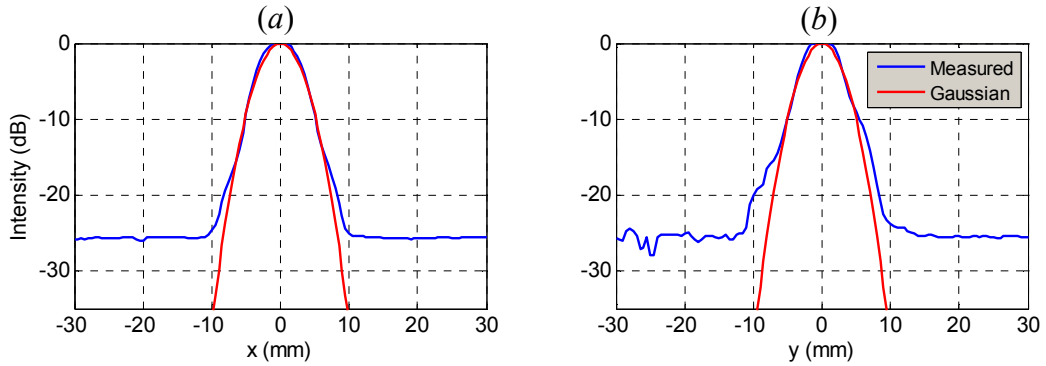


Figure 3-24. Log-scale plots of cuts through intensity measured at output plane of 4-*f* system (blue curve) and best-fit Gaussian with radii (W_x, W_y) = (4.67, 4.85)mm (red curve) in (*a*) x and (*b*) y directions.

3.3 Transmission Mode Imaging Experiments

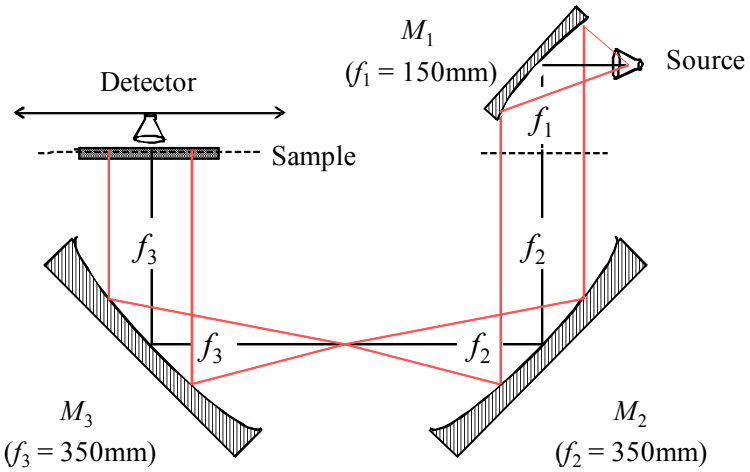


Figure 3-25. Optical arrangement used to perform both near-field and re-focused transmission mode imaging experiments. The configuration shown here (with the sample positioned in front of the detector scanning plane) is designed for near-field transmission imaging. To perform re-imaging transmission experiments the sample, or object, is placed at the dashed line between mirrors M_1 and M_2.

Both near-field and re-focused transmission imaging experiments were conducted using the system shown in Figure 3-25, which can be viewed as the combination of an illumination stage with a 4-*f* system, or Gaussian beam telescope (GBT). When no object is included in the system an image (with unit magnification) of the collimated beam produced by mirror M_1 is formed by the GBT at the output plane. The only difference between near-field and re-focused imaging experiments is in the position of the test object within the system. For near-field experiments the object under test was positioned directly in front of the output plane. When operated in this way the three-mirror system is equivalent to a system consisting of just mirror M_1, which collects and collimates the beam from the transmitting horn antenna. For re-focused imaging experiments the test object was placed at the output focal plane of mirror M_1. The GBT thus reproduces at the output plane an image of the object beam, i.e. that part of the illuminating beam transmitted through the object. A Fourier transformation of the object beam field is formed by mirror M_2 at the centre of the GBT. Thus spatial frequency filtering of the object beam field can be performed by inserting appropriate filters into the beam path at this point. Provided that the mirrors are aligned properly, in both near-field and re-focused imaging the object is illuminated by a wavefront that has a Gaussian intensity profile and a uniform phase front. The optical system was arranged

such that the output plane coincides with the scanning plane of the TOAST scanner onto which a single waveguide-fed detector was mounted. A two-dimensional intensity image of the radiation transmitted through the test object was acquired by raster-scanning the detector across the output plane. Alternatively, we could have fixed the detector in place and raster-scanned the test object in front of it.

In all transmission imaging experiments coherent radiation at 100 GHz was provided by the variable frequency Gunn oscillator as part of the transmitter chain described previously (Figure 3-2). A 150mm focal length paraboloidal mirror M_1 was used to collimate the source beam and in most of the experiments two 500mm focal length ellipsoidal mirrors were used for mirrors M_2 and M_3 in the GBT part of the optical system.

3.3.1 Near-field Transmission Imaging Experiments

The sample holder in which test objects were held consisted of a CD case with a narrow strip of clear plastic affixed on the inside (onto which objects could sit). The CD case was affixed to the vertical arms of a lens holder which was mounted onto an optical bench for stability. The system is arranged such that the sample holder is positioned as close as possible to the detector plane – the plane in which the aperture of a single detector is scanned (Figure 3-26).

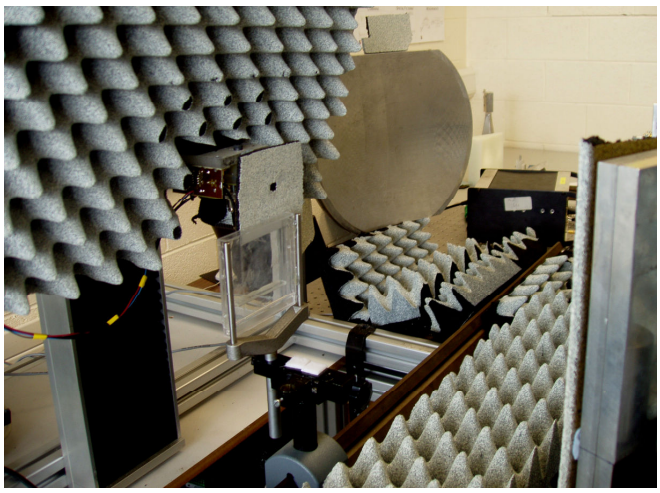


Figure 3-26. The transmission imaging system configured for near-field imaging with the sample-holder (a CD case mounted on an optical bench) positioned directly in front of the scanning plane. Mirror M_2 can be seen behind the sample holder and the edge of mirror M_3 is seen on the right.

It is necessary to keep to a minimum the distance between object and detector planes in order to reduce diffraction effects that might otherwise occur. A *z*-axis translation stage was used to minimise the distance between the waveguide mouth (used to feed the detector) and the test object. In early experiments a feed horn was used to couple radiation to the detector. Later this was replaced with a bare rectangular metallic waveguide – the face of which was machined to reduce the area perpendicular to the axis of propagation in the waveguide to reduce unwanted reflections in the system.

Before measurements of real test objects were made first low-resolution, then followed by high resolution measurements of the system with no object in the sample holder were made to ensure that i) the detector was not saturated by the unattenuated power from the illuminating source and ii) the scanned area was centred on the axis of propagation, i.e. where beam intensity is at a maximum. Figure 3-27 shows the intensity through the 3-mirror system with no test object in place.

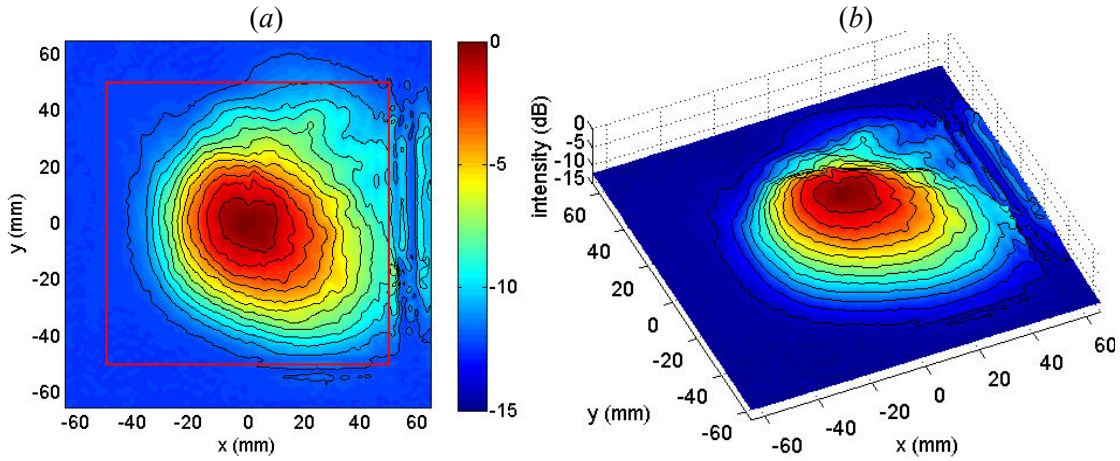


Figure 3-27. Log-scaled plots of intensity measured at the output plane of the transmission imaging system when no test object is included. Note however that the sample holder is included and the shadow of one of the vertical metal bars is seen (where the intensity falls off at x = +60mm).

Initial measurements made with the near-field transmission imaging set-up used simple geometric obstructions, such as rectangular stops and apertures. These objects were made from thin pieces of balsa wood soaked in water, which made them highly absorbing to incident radiation. In these experiments the sample holder was a CD case attached to a sheet of Styrofoam.

The first object measured was a 30mm × 20mm rectangular stop. Figure 3-28(*a*) shows the raw data in the form of intensity measured (after normalisation). The signal of interest is degraded by noise so digital image processing techniques (a combination of adaptive noise removal and median filtering) were applied to the raw intensity data to

reveal the underlying structure (Figure 3-28(*b*) & Figure 3-30). The filter size required to clean up an individual measurement was chosen by a process of trial and error.

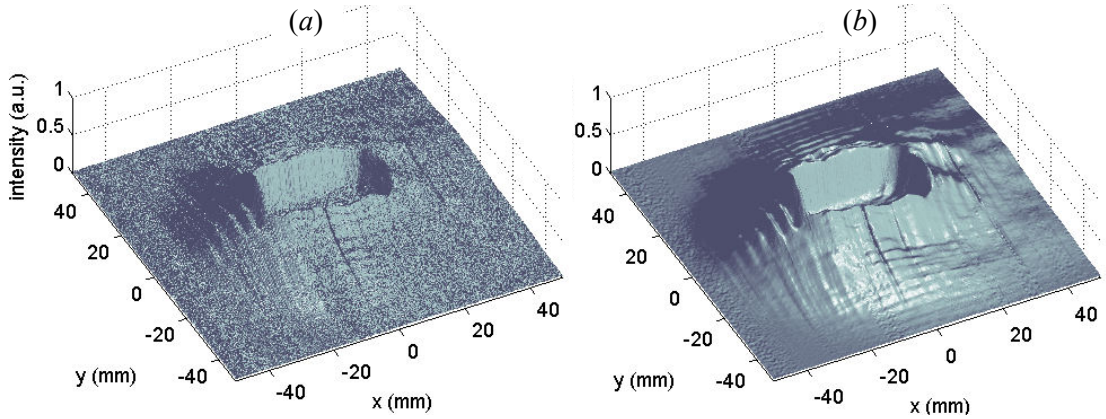


Figure 3-28. Plots of intensity as measured behind the rectangular obstruction. (*a*) shows the raw data (after normalisation) while (*b*) shows the data after digital image processing was performed.

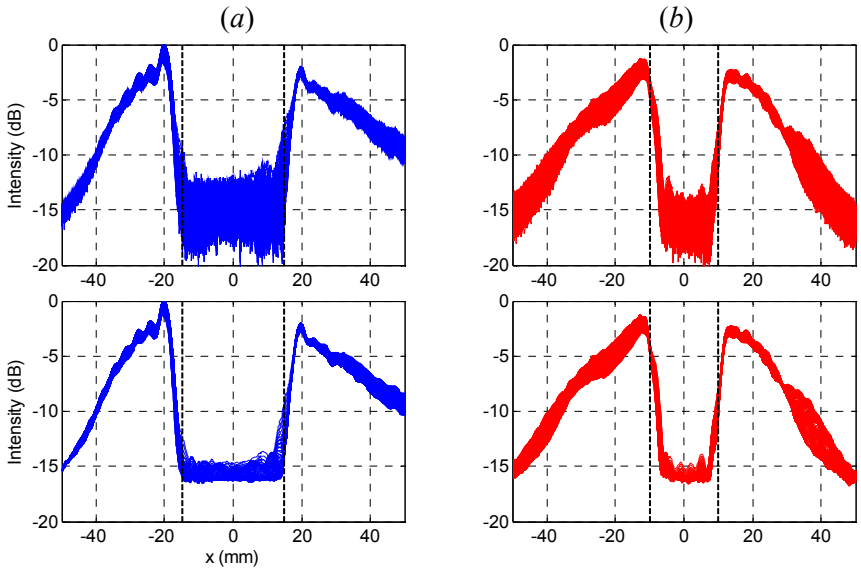


Figure 3-29. Log-scale plots of cuts taken through the intensity distributions shown in Figure 3-28 in the (*a*) x and (*b*) y directions (top) before and (bottom) after digital image processing.

The recorded image shows the source beam with a shadow in the region occupied by the rectangular obstruction. As one traverses the beam profile a smooth fall-off in intensity is observed upon entering the shadowed region (Figure 3-29 & Figure 3-31). This smooth transition in intensity is due to a combination of diffraction effects (due to the fact that a small but finite separation exists between the detector and the object) and blurring due to the finite size of the aperture used to feed the detector.

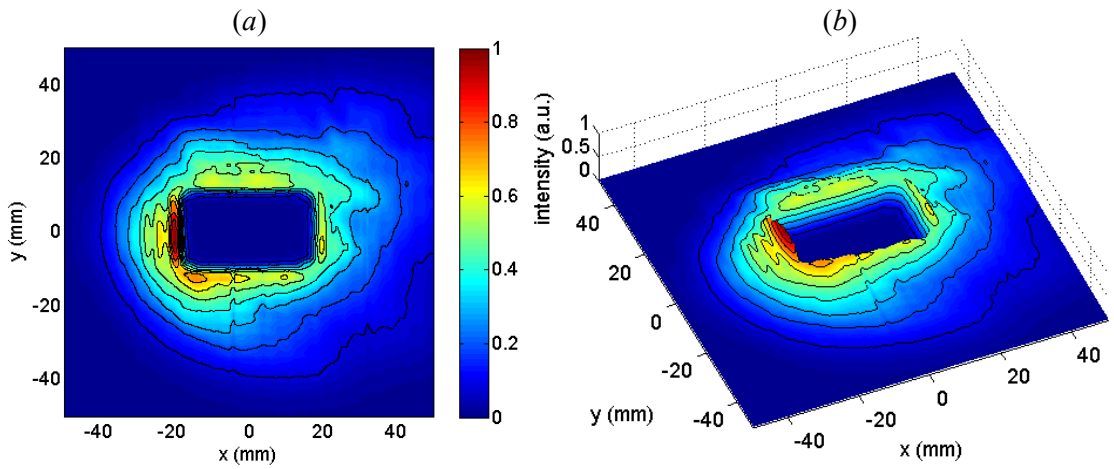


Figure 3-30. Linear-scale plots of measured intensity behind a thin rectangular piece of balsa wood that has been soaked in water. The high concentration of water causes all incident radiation to be reflected and/or absorbed by the object.

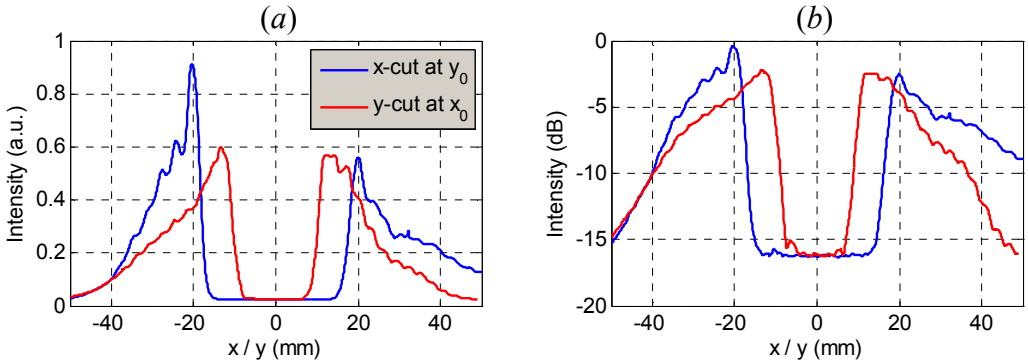


Figure 3-31. (*a*) Linear- and (*b*) log-scale plots of intensity measured behind the (water-soaked balsa wood) rectangular stop after image processing to reduce noise.

The next measurement made was of a small (3.0 mm diameter) circular stop, which was again made from water-soaked balsa wood. In the measured intensity (Figure 3-32) a peak is observed directly in the shadow of the stop similar to a Poisson spot. In other words Fresnel diffraction effects are important in forming the near-field image observed at the detector plane because of the small but finite distance between the detector aperture and the object plane.

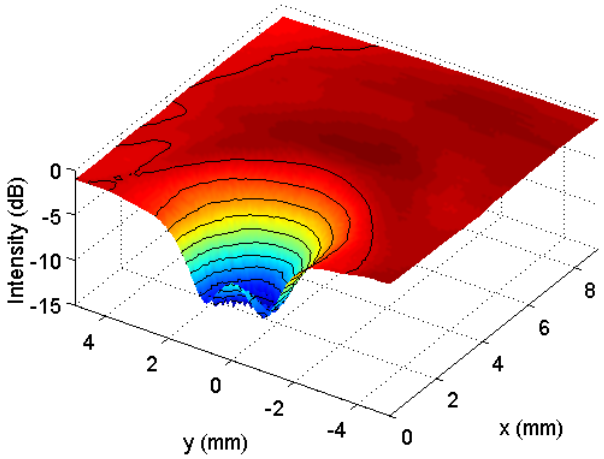


Figure 3-32. Log-scaled plot of intensity measured behind a small (3mm diameter) circular aperture. The stop is centred on the origin, where an intensity maximum is observed perhaps due to diffraction effects.

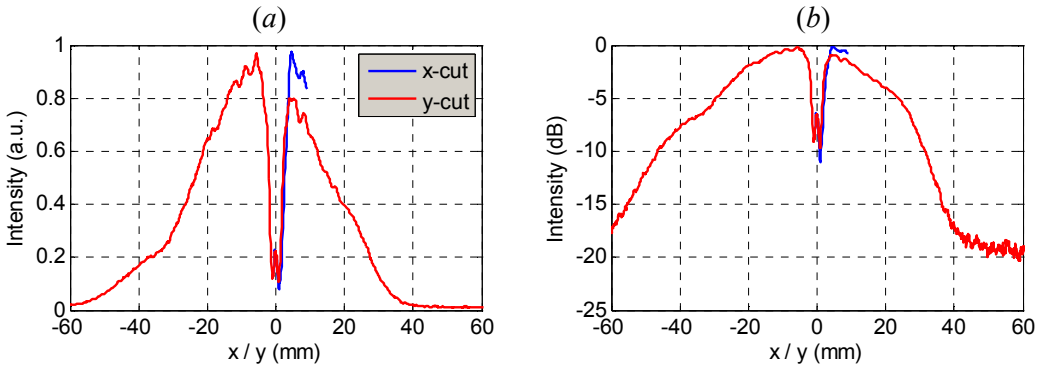


Figure 3-33. (*a*) Linear-scaled and (*b*) log-scaled plots of intensity measured behind a circular 3mm diameter stop made from water-soaked balsa wood.

In the next experiment an extended object was used to determine the maximum usable object size that could be imaged with the system. It also served to demonstrate the importance of reducing diffraction effects by minimising the object-detector separation. In this case the object consists of two narrow (~4mm) lengths of soaked balsa wood arranged into the shape of a cross. The cross was positioned with its centre at the centre of the source beam. An initial measurement was made with the object fixed to the sample-holder wall furthest from the detector. Thus the object-detector distance was approximately 5.5 mm (the thickness of plastic used in CD cases is 1.1mm and the separation between the two sides is 4.4 mm). A second measurement was then made after the object had been attached to the wall of the CD case nearest the detector, with the object-detector distance now reduced to ~1.1 mm. No changes were made to transmitter or receiver chains between the two measurements so the intensity levels can be compared (Figure 3-34). It must be noted that the second measurement was started sixteen hours after the first began by which time the balsa wood had lost some moisture.

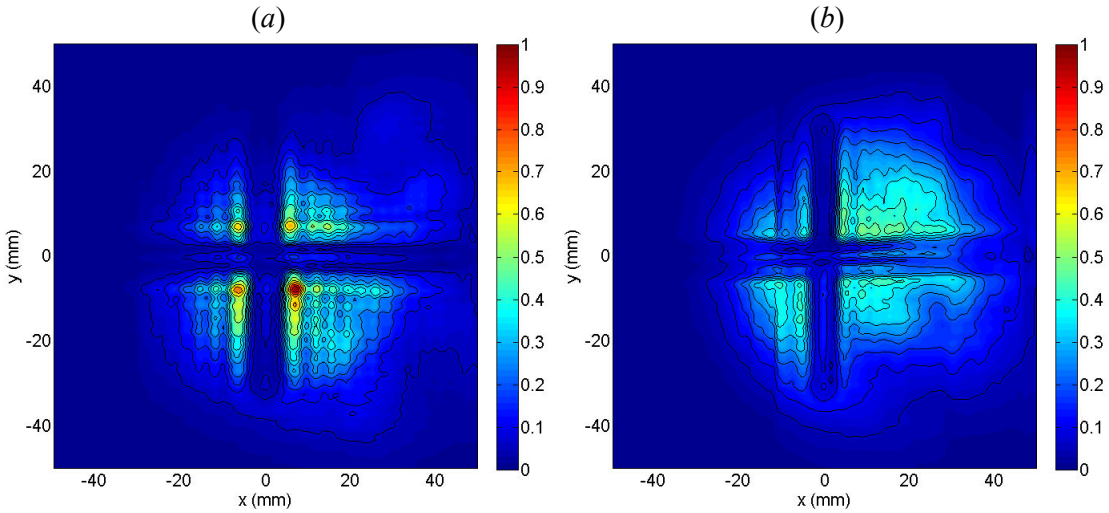


Figure 3-34. Linear-scale plots of near-field transmission intensity measurements of a cross-shaped object positioned on the optical axis for object-detector separations of (*a*) ~5.5 mm and (*b*) 1.1 mm.

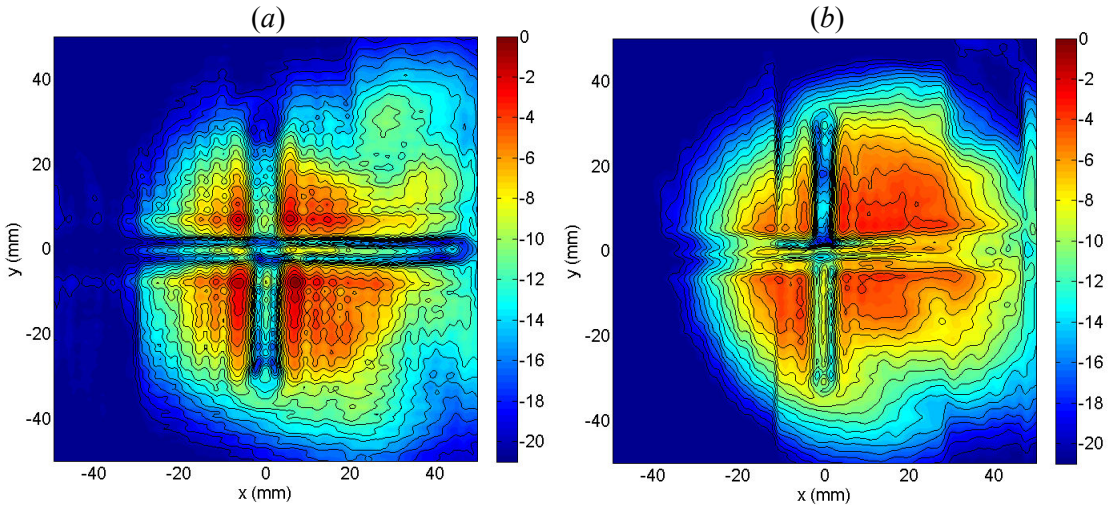


Figure 3-35. Log-scale plots of near-field transmission intensity measurements of a cross-shaped object positioned on the optical axis for object-detector separations of (*a*) ~5.5 mm and (*b*) 1.1 mm.

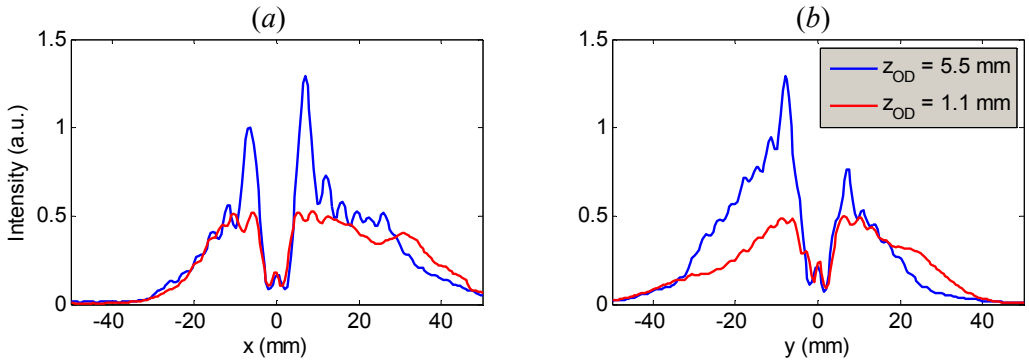


Figure 3-36. One-dimensional cuts through intensity measured behind a cross-shaped obstruction taken at the point of maximum intensity (*x*, *y*) = (7,-8) mm. Diffraction effects are more pronounced for the larger object-detector distance, z_{OD} = 5.5 mm (blue curve) than for when z_{OD} = 1.1mm (red curve). In both cases on-axis intensity level (directly behind the object obstructing the source beam) is similar.

The intensity levels in the geometric shadow of the cross are higher in the second measurement than in the first. More importantly the intensity image measured with a smaller object-detector separation is closer to the outline of the object: the shadowed region is narrower and edges are well defined.

To complement measurements the experiment was numerically simulated. The object beam field was represented as an ideal Gaussian beam with intensity nulls in regions occupied by the cross – to represent absorption by water. The object beam field was propagated to distances of 1.1 mm and 5.5 mm using Fresnel Transforms. The intensity of the simulated fields (Figure 3-37) and agree favourably with the measurements: near-field edge diffraction effects and even the appearance of intensity peaks directly behind the object – extended versions of the bright spot observed behind the circular stop (Figure 3-32).

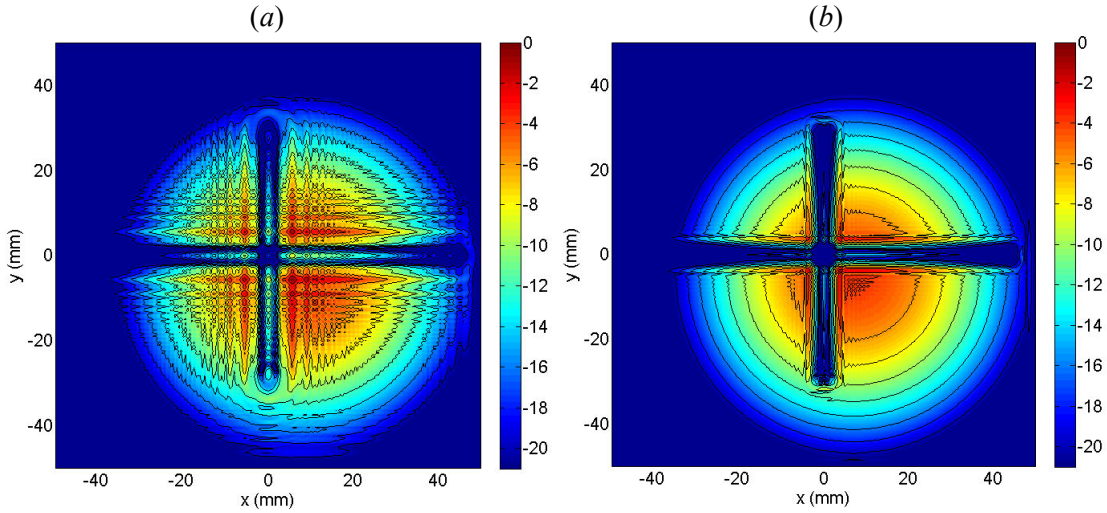


Figure 3-37. Log-scaled plots of simulated intensity at the back of a cross-shaped obstruction in a Gaussian beam for object-detector separations, z_{OD} of (*a*) 5.5mm and (*b*) 1.1mm.

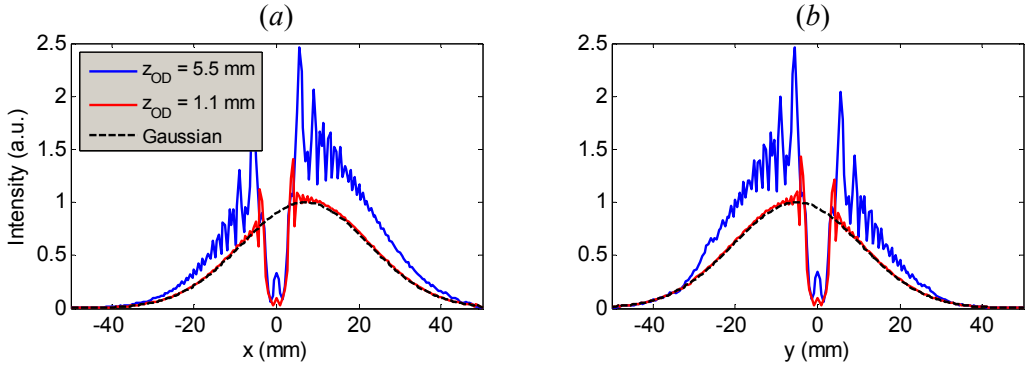


Figure 3-38. Cuts in intensity from simulated near-field transmission imaging experiment of a cross-shaped obstruction in (*a*) *x* and (*b*) *y* directions. Simulated intensity matches experimental results in that the intensity observed for an object-detector distance of z_{OD} = 5.5 mm (blue curve) is subject to significantly more diffraction effects than the intensity observed when z_{OD} = 1.1 mm (red curve).

The conclusion to be drawn from these experiments is that ideally for near-field measurements a bare waveguide should be used to feed the detector and that the object-detector feed distance should be less than a wavelength to avoid the introduction of significant Fresnel diffraction effects.

High image contrast is desirable and clearly this is only possible in regions where the object is adequately illuminated. Figure 3-39 shows multiple x and y cuts of one of the measurements made behind the cross object where it is seen that cuts taken far off-axis (further from the beam centre) have much lower contrast than those on-axis.

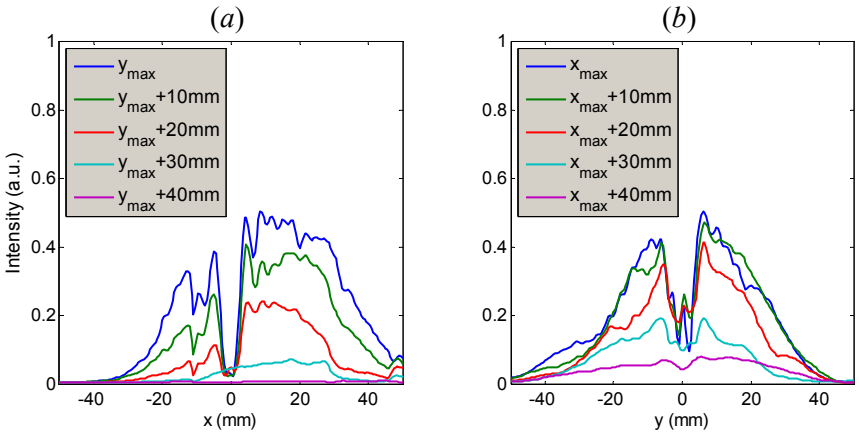


Figure 3-39. Cuts through intensity measured behind the cross-shaped aperture in (*a*) x and (*b*) y directions at separations of 10mm from the point of maximum intensity.

Contrast in measured intensity was calculated in x and y directions using Michelson's definition [3.16] of image contrast

$$c = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{3.21}$$

where I_{max} and I_{min} are maximum and minimum intensities. The contrast in the x (y) direction, c_x (c_y) was calculated by evaluating Eq (3.21) for each column (row) in a 2-D intensity array. For the intensity in Figure 3-34(*a*) indicates the estimated contrast (Figure 3-40) is seen to fall off sharply with increasing off-axis distance beyond ~30mm – the off-axis distance beyond which the illuminating intensity falls below e^{-2} of its peak on-axis value. In other words only objects no bigger than the illuminating Gaussian beam can be adequately imaged using this system. Thus maximum object size is limited to ~60mm × 60mm – a relatively small size considering that the wavelength is ~3 mm. Clearly to increase the object size that can be imaged with good contrast a larger illuminating beam is required, which calls for a longer focal length at mirror M_1.

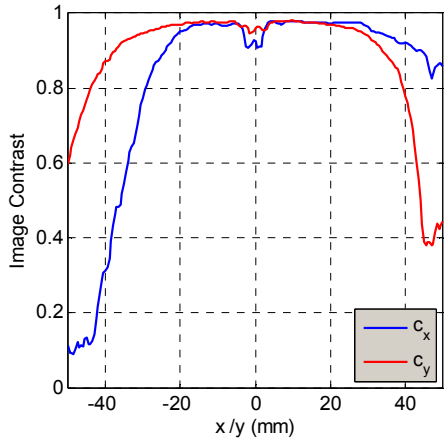


Figure 3-40. Estimated image contrast in x (blue curve) and y (red curve) directions for measured intensity of cross-shaped obstruction.

Imaging formation of objects that appear at least partially transparent to incident radiation is based on the interference of the wavefront transmitted through neighbouring regions of a sample with different refractive indices, different thicknesses, and/or absorption levels. Thus in the case of transparent objects the mechanism by which images are formed is based on the modulation of the transmitted phase-front as well as its amplitude (if some absorption occurs as well). Images of transparent objects can be obtained using absolute magnitude measurements since any phase modulation from the object (due to boundaries between regions with different refractive indices or depths) will produce interference effects at some level when the field is integrated across the collecting detector feed structure (in this case the aperture of a bare waveguide), the intensity patterns of which can be measured.

To illustrate near-field imaging of transparent objects a measurement was made of the intensity transmitted through a Dammann grating. Dammann gratings are binary phase gratings and will be dealt with in Chapter 4. For now we are merely interested in using one to demonstrate interference effects on near-field transmission imaging through transparent objects. The Dammann grating used for the current near-field measurement was made from a circular sheet of high density polyethylene (HDPE) measuring 54 mm in diameter and with a thickness of 10.3 mm. At 100 GHz HDPE is transparent to radiation and has a refractive index of 1.52. The grating profile consists of a regular arrangement of 3.0 mm deep rectangular grooves cut into one side of the circular sheet. In this experiment the grating was positioned with the flat, non-machined surface facing the detector. The intensity transmitted through the grating was measured over a 100mm × 100mm square area with a step size of 0.1 mm (Figure 3-41).

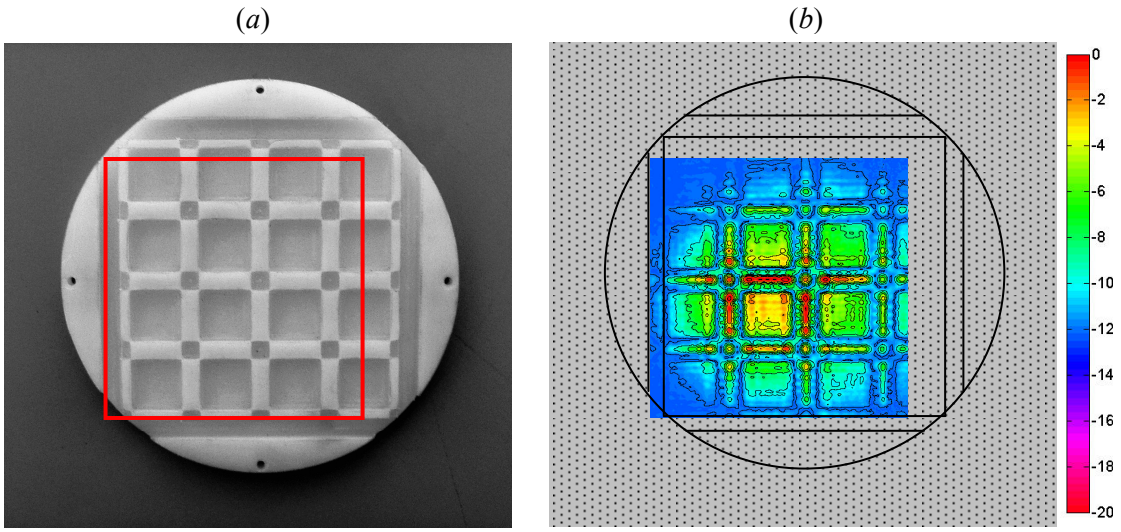


Figure 3-41. (*a*) Photograph of a Dammann grating. The red square superimposed indicates the area over which (*b*) the transmitted intensity (shown in log-scale) was measured. Note that the measured intensity is saturated so the dynamic range of measured intensity values is quite low.

The slotted grooves of the grating produce phase modulation over the surface of the grating giving rise to near-field interference and diffraction effects that are manifested as dips in intensity in regions on or close to groove boundaries. At 100 GHz the groove depth or height is equal to a single wavelength so phase shifts of 0 and π radians are imparted on the transmitted wavefront, which results in nulls in field strength as one traverses the grating. We expect zero detected power when the phase edge is halfway across the bare waveguide (or horn) feed as the phase of the fields across the boundary differ by 180°. The detector waveguide was scanned across the back of the grating, which is at a distance of 7.3 mm from the plane where phase modulation of the transmitted beam occurs. This small distance also allows extra observable interference and diffraction effects to form thus forming an intensity image revealing the structure of the transparent phase-only object.

Images of a number of biological test objects including leaves and thin slices of bacon and pork were recorded using the near-field transmission arrangement. The latter were chosen for their similarity to human flesh. These experiments take advantage of the absorbing/reflecting property of water at these wavelengths, which provides a mechanism for differentiating between regions of varying water content. The lean meat in bacon and pork samples contains more water than fatty regions, thus while some incident radiation is transmitted through fatty areas, little or no radiation can penetrate

the lean areas. Similarly the primary and secondary veins in leaves have a relatively high water content, compared to the rest of the leaf.

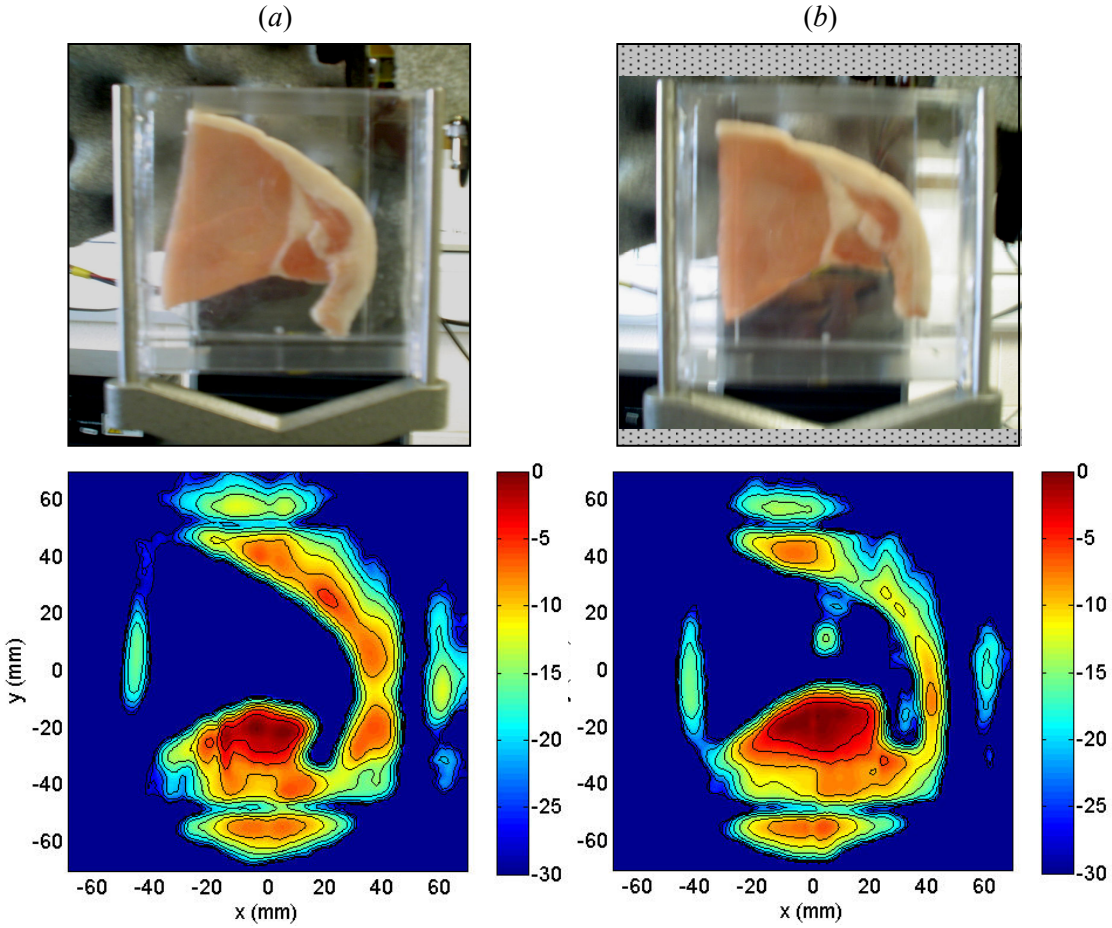


Figure 3-42. Photographs and corresponding log-scaled intensity measurements of a bacon sample (mounted in CD case sample holder). Image (*b*) was acquired 24 hours after (*a*), by which time the sample had lost enough water (due to evaporation) so that incident radiation could penetrate the fatty tissue near the centre of the scan.

Figure 3-42 shows two millimetre wave images of a piece of bacon that were measured at different times. In the first image no radiation is transmitted through the sample so only an outline of the sample is revealed. However by the time the second scan was made (24 hours later) the sample had dried sufficiently to allow some radiation to penetrate fat tissue, which has lower water content than the surrounding lean meat. Notice that image contrast decreases away from the scan centre and because the sample is larger than the size of the illuminating beam the bottom left corner of the sample is only barely distinguishable in the second image and not at all in the first.

Since only a very small amount of power is able to penetrate bacon samples in the next experiment the source attenuation was set to a minimum and the detector

sensitivity increased - to the extent that an image of the source beam with no object in place is saturated – to enable the small power levels transmitted through bacon to be detected. Figure 3-43 shows a sequence of intensity images that were measured with increasing sensitivity on the detector.

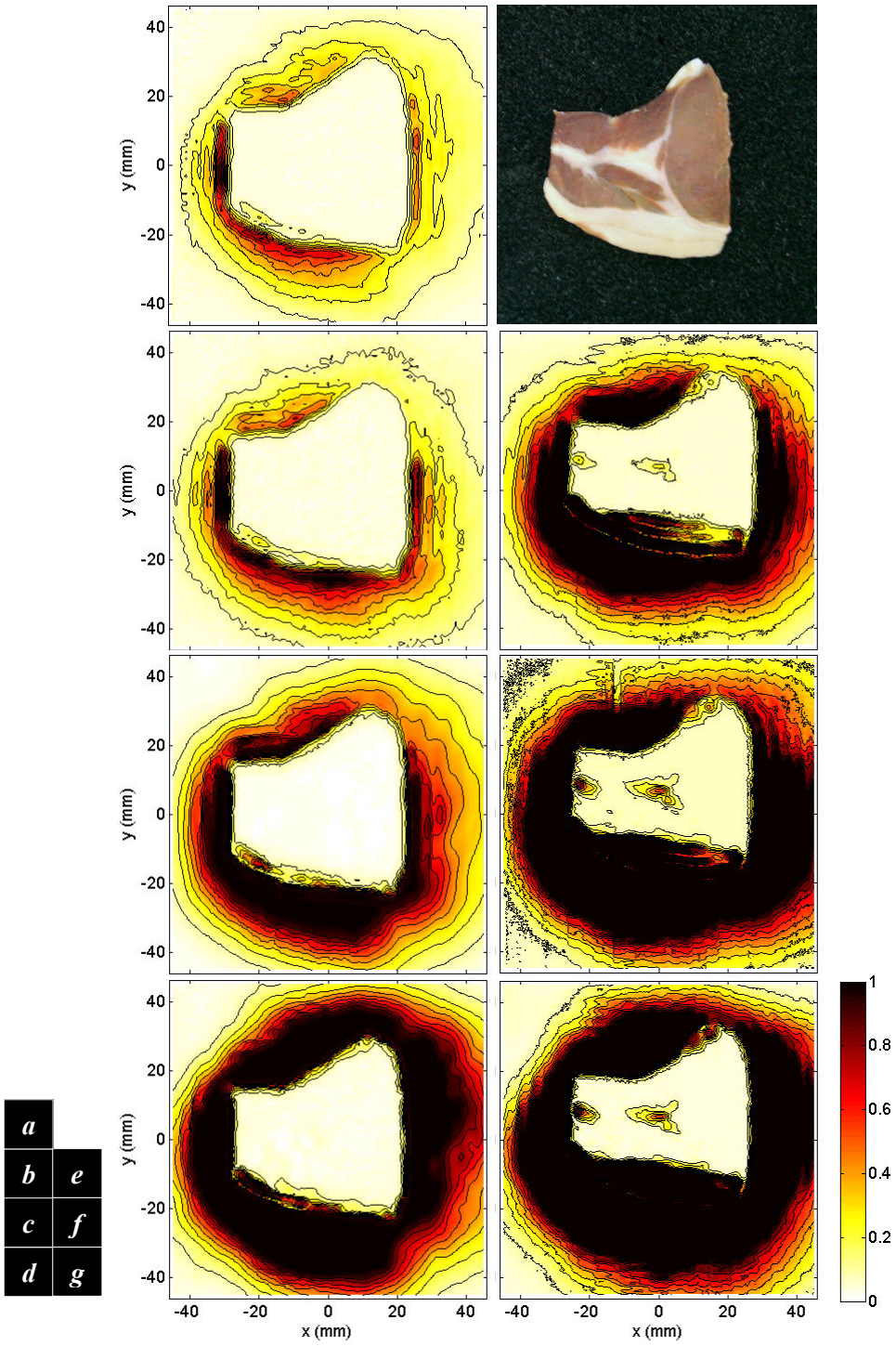


Figure 3-43. Linear-scale plots of intensity transmitted through bacon sample shown (top-right) for different settings of detector sensitivity. As sensitivity increases in scans (*a*) through (*g*) lower level features are revealed. Note that the three measurements (*e*) to (*g*) were made after the sample had been dried for 24 hours, which allows much more radiation through the strip of fat at the samples lower edge.

Figure 3-44 shows the result of another experiment in which the source and detector settings were optimised to detect low power levels. A log-scaled plot of the measured intensity shows clearly regions where transmission occurs through areas of fatty tissue. Notice that some radiation is also transmitted through the lower-right portion of the sample which contains only lean meat, presumably because water content was lost more quickly towards the samples edges. Another source of nuisance in these images can be the presence of standing waves. For example these will occur particularly when there are surfaces which give rise to partial reflections (i.e. sudden changes in refractive index). These standing waves produce modulation of the intensity pattern across the image which is not due to variations in absorption in the object itself of course, and which further complicates image interpretation.

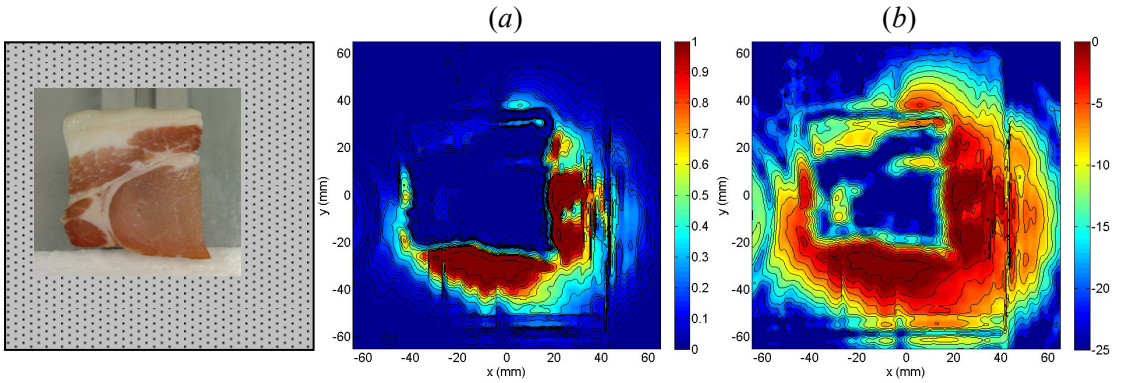


Figure 3-44. A near-field transmission image of the bacon slice (left) made with source and detector settings optimised to detect low power levels produced the intensity image shown in (*a*) linear and (*b*) log scales.

Measurements were made of a number of other objects including leaves, pieces of pork, bacon and lamb, and everyday items (key, penknife, etc.). Some of the images obtained for these objects are included in Appendix A. Most of the measurements are displayed in the format shown in Figure 3-45 with a photograph of the object, a grey-scale intensity plot (with contour lines superimposed) and a semi-transparent false-coloured intensity plot (with contours) superimposed on the photograph. The measurement shown in Figure 3-45 highlights the problems that standing waves cause. In this experiment the object was a narrow strip of bacon fat. We see that the illuminating beam passes through the object, however a drop in power is observed at the objects edges, which cannot be due to any variation in absorption by the object.

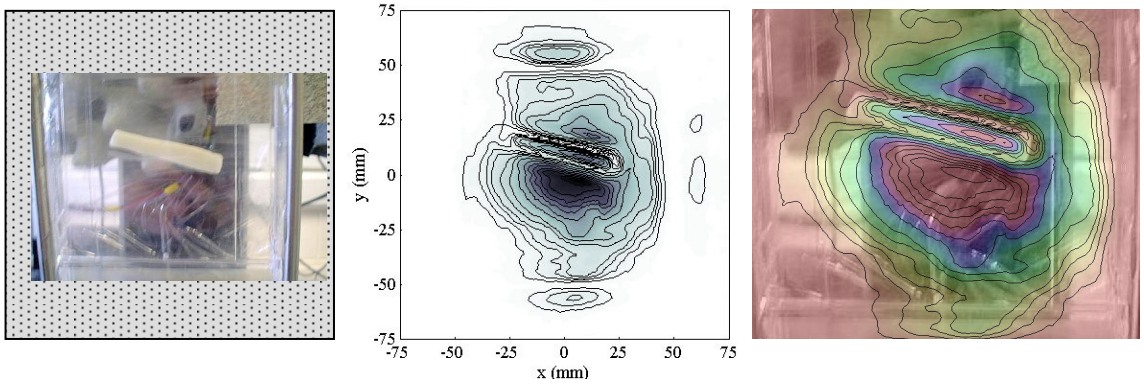


Figure 3-45. Near-field transmission intensity image a narrow strip of bacon fat. Measured intensity is displayed using a linear scale. Illuminating beam power penetrates the object, however a sharp decrease in intensity is observed at the edges of the sample.

3.3.2 Image Recovery for Near-Field Transmission Imaging

Since the near-field transmission set-up has no optics between the object and detector planes no aberrations are introduced into the object beam (transmitted from the test object). However as in all imaging systems, the intensity measured by the detector is not equal to that incident on the detector. The field at the detector is blurred by a response function, or point-spread function (PSF), of the system. Since the PSF is in this case due solely to the finite-sized waveguide aperture used to feed radiation to the detector the form that the PSF takes should be invariant with position. In systems where optical components are used to guide the beam onto the detector the PSF is due to a combination of factors and so its exact form may depend on position in the image plane.

Image recovery is used to 'deblur' a measured image that has been blurred by the point-spread function of the system. Convolution of a true signal, f with a function, h that represents the PSF produces a blurred image g given by

$$g = f \otimes h \tag{3.22}$$

where $\otimes$ denotes convolution. From the convolution theorem we have that

$$G = F \cdot H \tag{3.23}$$

where upper-case letters denote the Fourier transform of their lower-case counterparts. The process of recovering the true image from one that is blurred due to convolution by a PSF involves deconvolving the PSF from the measured field by solving for the true field f in the preceding equations, which in Fourier space is written as

$$f = \Im^{-1}\left(\frac{G}{H}\right) \tag{3.24}$$

where $\Im^{-1}$ denotes the inverse Fourier transform operator. However as well as image blur due to the PSF, the measured image quality is also degraded by the presence of noise. We assume that noise is additive – an assumption that may not be correct since multiplicative noise may also be present, but which yields simpler methods for noise removal. The measured signal is then given by

$$g = (f \otimes h) + n \tag{3.25}$$

or in Fourier space by

$$G = (F \cdot H) + N \tag{3.26}$$

where n denotes an additive noise function. When additive noise is present an estimate of the true signal is given by

$$f = \Im^{-1}\left(\frac{G - N}{H}\right) \tag{3.27}$$

Several problems arise when attempting image recovery with our system. Firstly we do not know the exact form of the PSF or the system noise. A more fundamental problem however is that our system measures only beam intensity but not its phase. This last problem is insurmountable with the equipment at our disposal. Thus the attempted image recovery reported here is approximate at best and would benefit greatly from the use of a Vector Network Analyser (VNA) to measure field amplitude and phase. The other two problems can however be tackled. The success of standard image recovery algorithms in finding a good estimate of the true signal f depends on good estimates for the point-spread function (h) and the additive noise function (n) that are responsible for image degradation. In what follows we outline the procedure for making a best guess of the point-spread function and image noise. The estimates of PSF and noise functions are then input to a standard deconvolution algorithm (blind deconvolution) in order to estimate the true intensity (before degradation by image blur and noise) from a particular intensity image recorded with the near-field transmission imaging system.

Several near-field transmission measurements were made of the intensity transmitted through a small (5 mm diameter) circular aperture cut into in a sheet of aluminium foil, that was affixed to the inside of the sample holder. Figure 3-46 shows linear- and log-scaled plots of the measured intensity – in fact the mean of five individual measurements. Apart from a noticeable dip near its centre, which was observed in all five measurements, the measured intensity has a smoothly-varying profile due to the convolution of the PSF and the top-hat field representing the beam

transmitted through the aperture. Initially it was suspected that the intensity dip was due to multiple reflections in the system (either between the source and aperture, or between the aperture and detector, or a combination of both). In the last two measurements made a sheet of absorbing material (Eccosorb) was attached to the side of the foil sheet facing the source in order to reduce unwanted reflections between source and aperture but with little effect, indicating that reflections were not affecting the field measured. A simpler explanation for the intensity dip is provided by the inclusion of near-field diffraction effects as will be shown.

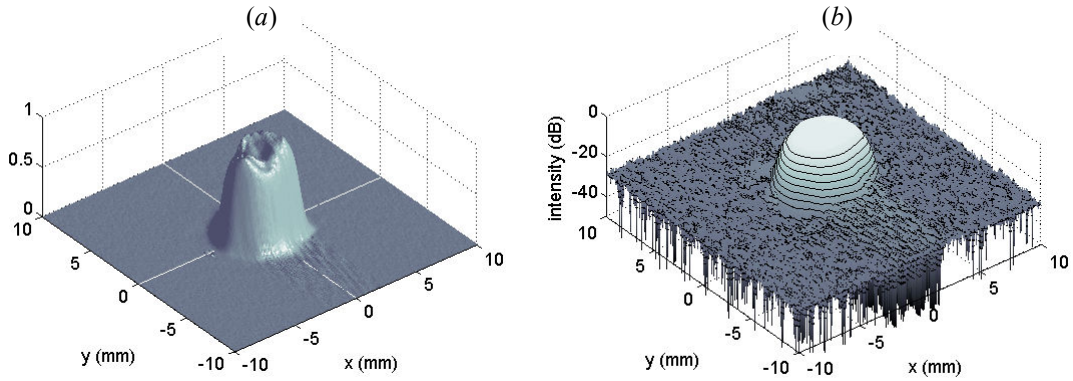


Figure 3-46. (Left) Normalised linear and (right) log-scale plots of the intensity measured behind a 5mm diameter pinhole aperture in an Aluminium foil sheet. Five measurements were made and averaged in order to reduce noise. The superimposed white lines intersect at the estimated centre of the aperture.

Estimating the point-spread function

From the measured intensity data an estimate of the systems' PSF was made. Because we expect the PSF to be invariant with position all measurements were made with the circular aperture positioned at the centre of the quasi-collimated source beam. The measured intensity was simulated using the model of image blur given by Eq (3.22), i.e. assuming no noise. The ideal intensity function transmitted through a uniformly illuminated circular aperture is described by a circular top-hat function

$$f(r) = 1, r \leq d/2 \tag{3.28}$$

for a circular aperture with diameter d. The aperture is illuminated with a quasi-collimated Gaussian beam. A 500mm focal ellipsoidal mirror produces a Gaussian beam with a waist radius of ~101mm at its focal plane, so to a good approximation the aperture can be said to be uniformly illuminated therefore Eq. (3.28) is valid.

To begin estimating the PSF we assume that h takes the form of a rectangular top-hat function – since the bare waveguide detector is also rectangular. The function h

is then convolved with the circular top-hat function *f*, representing ideal transmission through the aperture. By choosing appropriate values for the width and height of *h* we can replicate the correct roll-off in intensity observed in the measured pattern. However a top-hat PSF cannot reproduce the smooth (tapered) variations observed in the intensity patterns across the edges of the circular aperture.

An alternative useful choice of PSF is one with a Gaussian profile, as shown in Figure 3-47(*b*). We choose the size of the PSF by considering the waveguide aperture with which the PSF is associated. A fundamental Gaussian beam mode approximation to the field from a rectangular waveguide has radii of

$$(W_x, W_y) = (0.35, 0.25)a \tag{3.29}$$

where the particular waveguide used has a width of a = 2.54 mm (and height, $b = a/2$), which gives (W_x, W_y) = (0.89, 0.64) mm. The power coupling efficiency between the simulated blurred intensity pattern shown in Figure 3-47(*c*) (using the asymmetric Gaussian PSF) and the measured intensity data was estimated to be approximately 0.92.

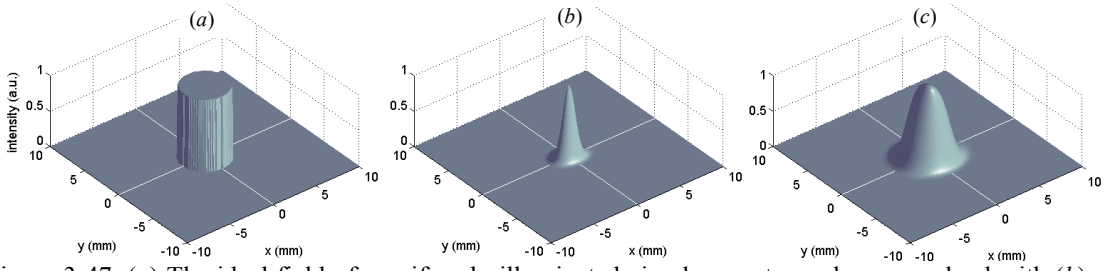


Figure 3-47. (*a*) The ideal field of a uniformly illuminated circular aperture when convolved with (*b*) an asymmetric Gaussian profile PSF yields (*c*) the blurred intensity as 'seen' by the detector.

A Gaussian-shaped PSF with slightly different values of W_x and W_y might yield improved coupling between the simulated and measured intensity data. To this end a multivariable routine was used to find values for radii W_x and W_y that provide a better fit between the measured and simulated intensity. The goal of the optimisation routine was to find radii that produce maximum power-coupling efficiency between simulated and measured intensities. The routine was initiated using the beam radii given by Eq. (3.29) and produced optimum values of (W_x, W_y) = (1.281, 0.914) mm which yielded a power-coupling efficiency of 0.94 to the measured pattern (see Figure 3-48).

A possible explanation for the discrepancy between initial guess of (W_x, W_y) – that best fits the waveguide field – and the best fit solution found by the optimisation routine is that there is a non-zero separation between the object and detector plane that results in spreading of the beam. Although every effort was made to minimise the

object-aperture distance, a small but finite spacing between the object and detector plane is evident from the presence of diffraction effects observed in other measurements. Thus the Gaussian PSF that would one might expect at the waveguide aperture may have expanded to one with larger radii at the object plane.

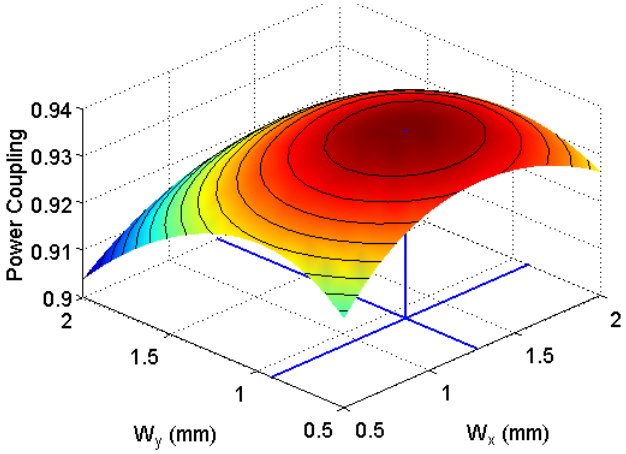


Figure 3-48. Power-coupling efficiency between measured and simulated intensity as measured behind the uniformly illuminated 5mm circular aperture. Maximum power-coupling efficiency is ~0.94 which occurs at (W_x, W_y) = (1.281, 0.914) mm.

It was estimated that object-detector separation was approximately 0.6 mm (Figure 3-49). However the ratio of radii W_x-to-W_y is approximately the same for the values that the routine began and ended with, which should not be the case as the smaller value of W_y at the waveguide aperture should expand more quickly than the larger value of W_x. This suggests that neither a top-hat nor a Gaussian function is adequate for representing the PSF (point-spread function) acting on measurements made with the present system.

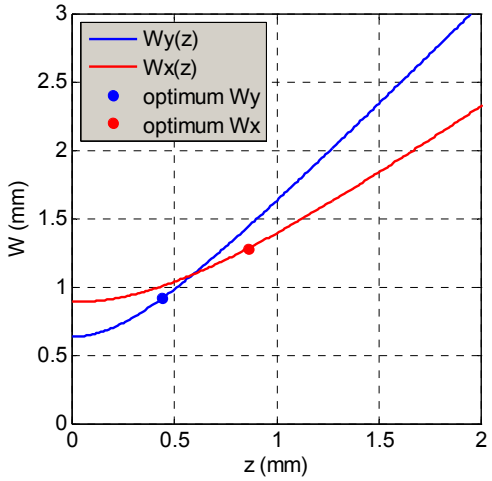


Figure 3-49. Variation of Gaussian beam radii (W_x, W_y) with distance z from the waveguide aperture. The optimised Gaussian beam radii found (by multivariable optimisation) shown by circular markers indicate that the object plane is not situated at the waveguide aperture but at some small distance, which we have taken to be ~0.6mm.

Of course a modal analysis of a rectangular waveguide field that includes only the best fundamental Gaussian beam mode is only approximate. A more appropriate estimation of the PSF would be to use a one-dimensional truncated cosine waveguide field itself (Figure 3-50). Doing so however yields a power-coupling efficiency (between the simulated and measured intensities) of ~0.93, which is slightly less than the previous value produced using an asymmetric Gaussian PSF.

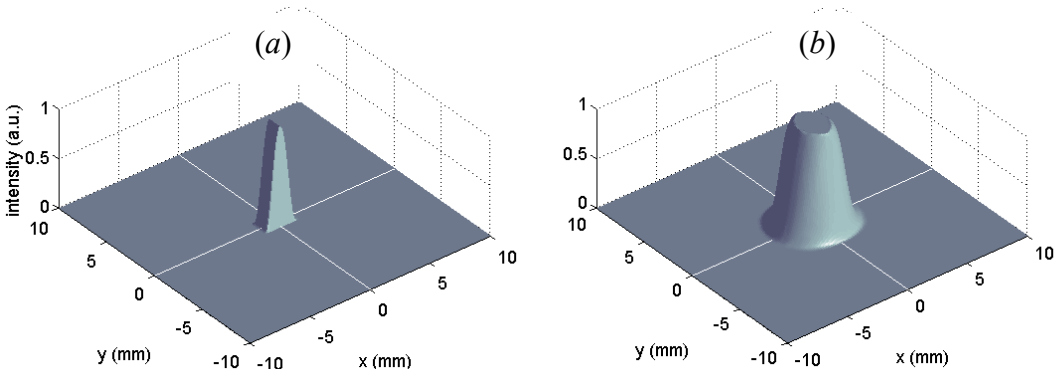


Figure 3-50. The convolution of the transmission function of a uniformly illuminated 5mm diameter pinhole aperture when convolved with (*a*) a PSF equal to the truncated-cosine field representative of the field at the rectangular waveguide aperture produces (*b*) a simulated intensity that yields a power-coupling efficiency to the measured intensity of 0.93.

The waveguide aperture field is now propagated a finite distance (a detector-object plane separation of 0.62 mm) from the waveguide and the resulting propagated field used as an estimate of the PSF. Figure 3-51 shows the simulated blurred intensity image generated using this PSF. The result is an increase in power-coupling efficiency of approximately 1%. More significantly, however the dip in intensity that was observed in the measured intensity is clearly reproduced when this PSF is used.

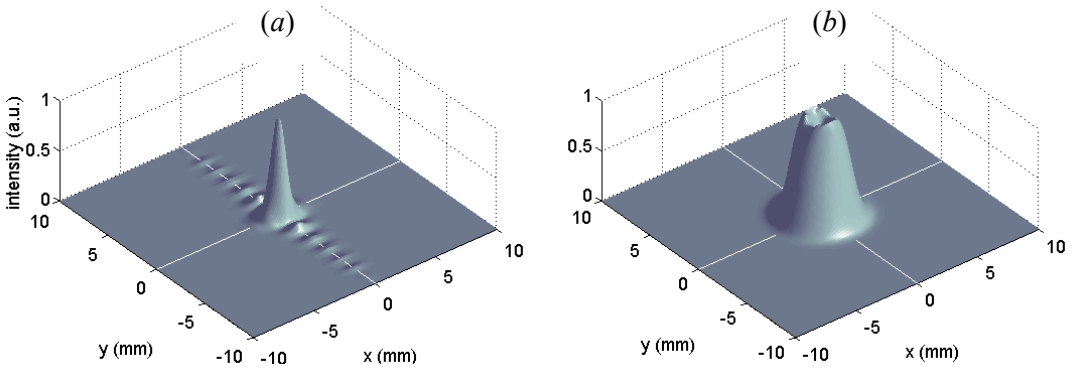


Figure 3-51. The convolution of a circular top-hat function with (*a*) a PSF derived from (truncated cosine) waveguide aperture field propagated a distance of 0.62 mm from the detector aperture produces (*b*) a simulated intensity field with increased power-coupling efficiency of 0.94 (to the measured intensity) and also manages to reproduce the central intensity dip observed in measured beam patterns.

An analysis of the variation in power coupling efficiency between measured and simulated field intensities with respect to propagation distance z was performed (Figure 3-52). The maximum power coupling was found to occur when the PSF used is equal to the waveguide aperture field that is propagated a distance of 0.62 mm from the waveguide aperture. The point is that diffraction effects must be accounted for because although the propagation distance is small, the field at the mouth of the waveguide expands rapidly with distance because the width and height of the aperture are only on the order of less than a wavelength. In fact taking a best-fit fundamental Gaussian beam mode approximation for the waveguide field we see that the bare waveguide aperture has a confocal distance of $z_c = \pi W^2/\lambda < 1$ mm, which is why the propagated waveguide field resembles a far-field pattern.

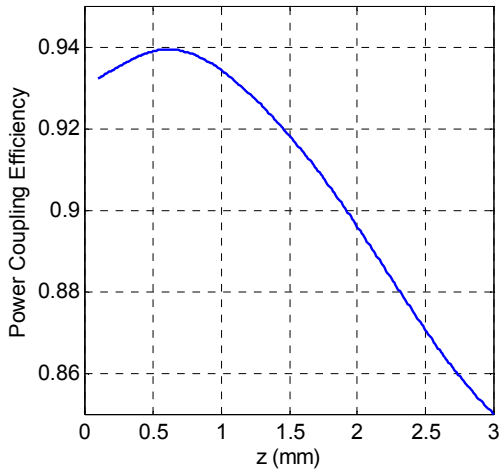


Figure 3-52. Variation of power coupling efficiency between measured and simulated field intensities behind a 5mm diameter pinhole aperture for a PSF derived from propagated truncated-cosine waveguide aperture function. Maximum power coupling occurs at a distance of 0.62 mm from the waveguide aperture. Propagation of the waveguide field was computed using a Fresnel Transform.

To reduce computational overhead only that part of the PSF (derived from a propagated waveguide field) with significant intensity need be retained. For our purposes, the lowest signal level that can be distinguished above background noise is about 30 dB so only that portion of the PSF with power over this level was retained (Figure 3-53).

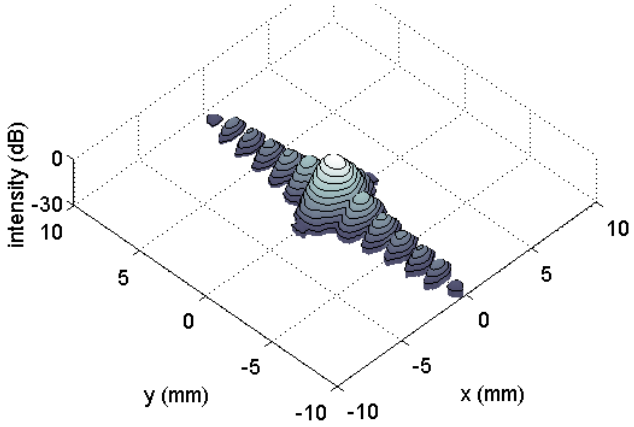


Figure 3-53. Log-scaled plot of the significantly intense (≥ 30 dB) portion of the PSF that is retained and used for deconvolution analysis.

Estimating Image Noise

The five intensity measurements were summed and used as an average measure of the intensity transmitted through the small circular aperture in order to reduce the amount of additive noise. Figure 3-54(*a*) and Figure 3-54(*b*) shows log-scale plots of a single measurement and the mean of five measurements, respectively. In order to analyse only noisy pixels in these images, regions containing the foreground signal were masked out, as shown in Figure 3-54(*c*) and histograms made of the remaining background pixels.

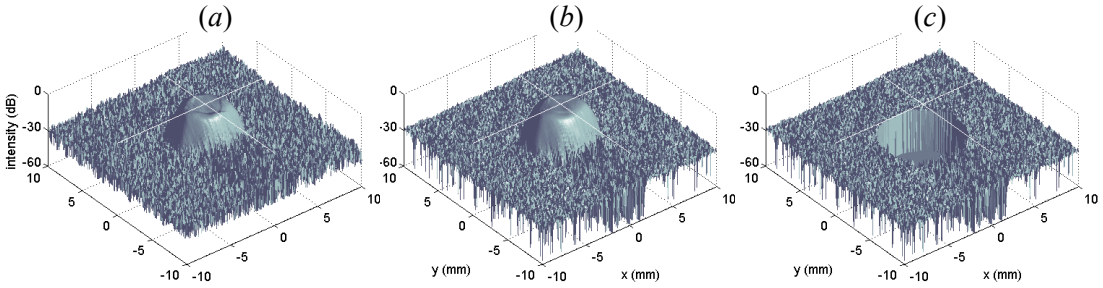


Figure 3-54. Los-scaled plots of measured intensity for (*a*) a single measurement and (*b*) the average of five measurements. The background signal in (*c*) is isolated by masking out foreground objects – the intensity through the pinhole aperture.

Histograms of background intensity pixels from a single measured image and the averaged image are shown in Figure 3-55. In both case the noisy background intensity follows a Gaussian-shaped distribution with a mean value of 1.275×10^{-3}. However the variance, σ^2 of noisy pixels from a single measurement is 1.20×10^{-3}, while that for the averaged image is only 0.65×10^{-3}, thus highlighting the value of averaging multiple sets of measured data.

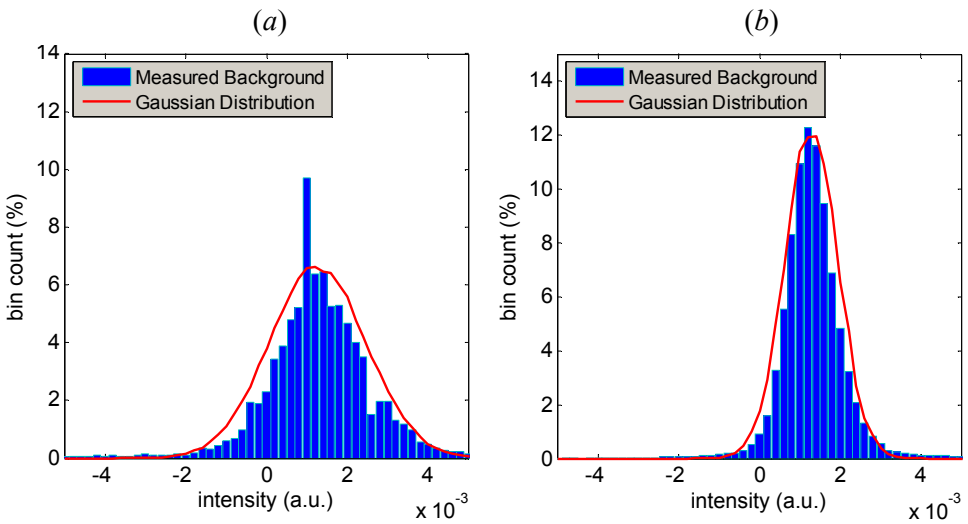


Figure 3-55. Histograms of background pixel intensity values for (*a*) a single measurement and (*b*) the average of five measurements made behind the 5mm diameter pinhole aperture. In both cases the intensity distributions are Gaussian with a mean intensity value of 1.275×10^{-3}, but the variance, σ^2 (width of the Gaussian distribution) for the averaged image is only half that of the single measurement.

The model of image degradation is completed by adding a noise signal to the blurred image created by convolving the ideal aperture transmission function with an appropriate point-spread function due to the finite detector aperture (Figure 3-56). A sequence of noisy data n that has a Gaussian distribution with a specific mean a and variance σ^2 is generated using the expression

$$n = a + \sigma R \tag{3.30}$$

where R is a normally distributed sequence of random numbers, which is multiplied by standard deviation σ.

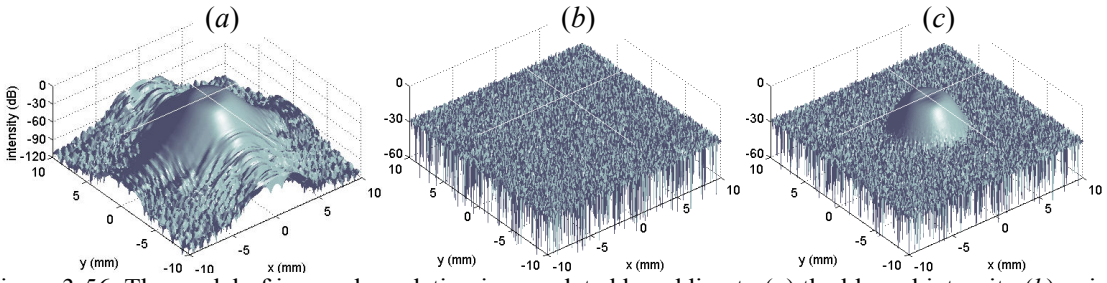


Figure 3-56. The model of image degradation is completed by adding to (*a*) the blurred intensity (*b*) noisy intensity n to produce (*c*) a blurred and noisy intensity distribution.

Now having made estimates of both the point spread function and the noise distribution that produces the image degradation observed in the measured intensity some standard deconvolution algorithms were applied to recover an estimate of the true signal f.

Deconvolution with MATLAB

The Image Processing Toolbox in MATLAB provides four deconvolution algorithms: Weiner deconvolution, regularized filtering, blind deconvolution and the Lucy-Richardson (LR) algorithm. Each function accepts as input an image in intensity, an estimate of the PSF and various optional additive noise parameters. Unfortunately none of these MATLAB functions support complex-valued input (images or point-spread functions). Thus only the intensity of the complex-valued PSF derived above could be used. All four algorithms were applied to try to recover a sharper image from the measured intensity image of the circular aperture. Deconvolution introduced ringing in the recovered images which were subsequently smoothed using a median filter. The best result was obtained using the iterative blind deconvolution algorithm (Figure 3-57), which was implemented with the syntax

```
>> [zDeblur,zPSF] = deconvblind(z,zPSF,N,0.1*var);
```

where N = 100 iterations were used. The last parameter is a damping factor which was set to 10% of noise variance. The noise-power parameter is given by the product of the noise distribution variance multiplied by the number of pixels in the image. The power-coupling efficiency between an ideal aperture transmission (top-hat) function and measured intensity was 0.83, while that for the recovered intensity is estimated to be 0.89. The recovered image, while exhibiting high frequency ringing across the aperture – but not, it must be noted, in regions where no power is expected – does have much sharper edges and resembles more closely ideal transmission through a circular aperture.

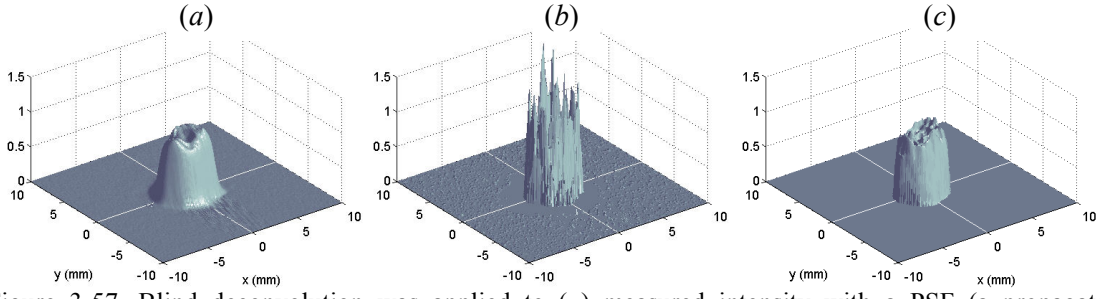


Figure 3-57. Blind deconvolution was applied to (*a*) measured intensity with a PSF (a propagated waveguide aperture field) to produce (*b*) an intensity profile with vertical sides, after which (*c*) a median filter was used to reduce high-frequency ringing.

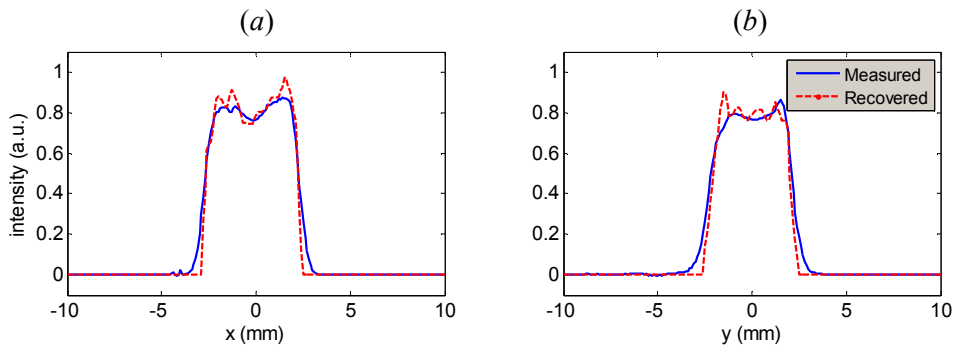


Figure 3-58. Cuts through the centre of the measured (solid blue curve) and recovered (dashed red curve) intensity images in (*a*) x and (*b*) y directions. Image recovery was done using 100 iterations of blind deconvolution.

3.3.3 Transmission Imaging with a Fourier Optics System

This section describes experimental measurements obtained with the transmission imaging system with re-focusing optics that includes the facility to perform spatial frequency filtering for edge detection of objects. Filtering spatial frequencies involves the filtering, or removal of, specific spatial frequency components from the Fourier transform of the object beam. Figure 3-59 shows the re-imaging transmission system and Figure 3-60 shows the equivalent system with lenses instead of mirrors. The latter was used as a model of the system for numerical simulations in terms of Fourier transforms and Gaussian beam mode analysis.

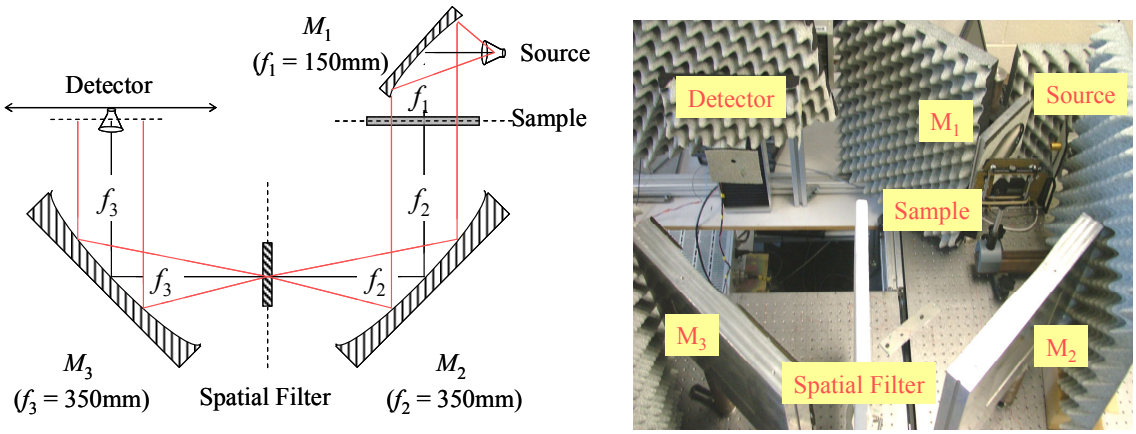


Figure 3-59. Schematic and photograph of the set-up used for edge detection experiments. The high-pass spatial filter is a circular stop made from a disc of Eccosorb with a radius $a = 2W_{SF}$, where W_{SF} = 30.44mm is the waist radius of the illuminating beam when no object (sample) is included in the system. The spatial filter was attached to a sheet of polystyrene, which is transparent to radiation at 100 GHz.

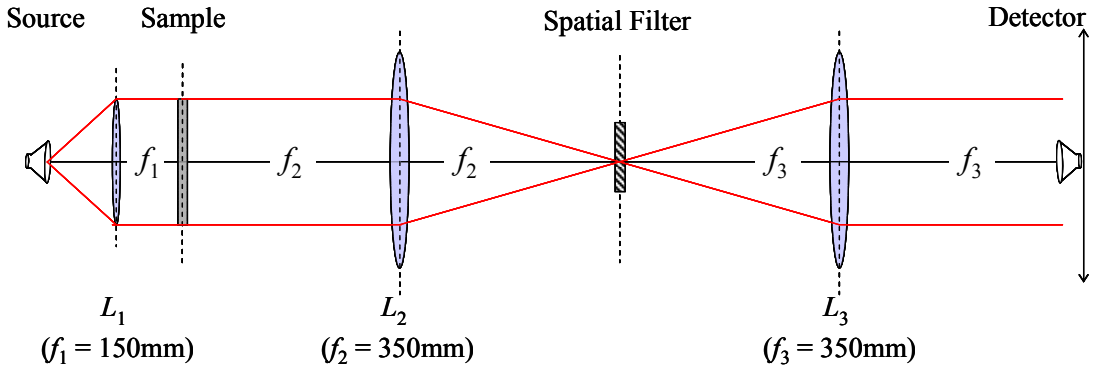


Figure 3-60. A system for spatial filtering equivalent to the one shown in Figure 3-59 but with lenses in place of the mirrors.

The re-imaging system consists of two stages: an illumination stage and a filtering stage. The amplitude and phase of the source beam are modulated by the test object. The modulated source beam, or object beam, then serves as input to the filtering stage, which consists of a pair of focusing elements arranged so as to provide a beam waist position between them. A spatial filter can be inserted at this point to modulate the amplitude and/or phase of the Fourier transform of the object beam. The final focusing element then produces a filtered image of the object beam at the output plane, where the detector is raster-scanned to record an image of the frequency-filtered intensity.

In experiments a high-pass spatial filter (i.e. a blocking filter at the centre of the Fourier plane) was used to produce edge enhanced images of the object beam field. A test of the system was performed to ensure that the high-pass spatial filter was aligned correctly and capable of filtering out the low-spatial frequencies associated with the illuminating Gaussian beam. Two measurements were made one with and one without the spatial filter included in the set-up, the results of which are shown in Figure 3-61. When the filter is omitted a (slightly asymmetric) image of the illuminating Gaussian beam is observed at the output plane, whereas when the high-pass filter is included negligible intensity is recorded at the output plane thus demonstrating the correct operation of the system.

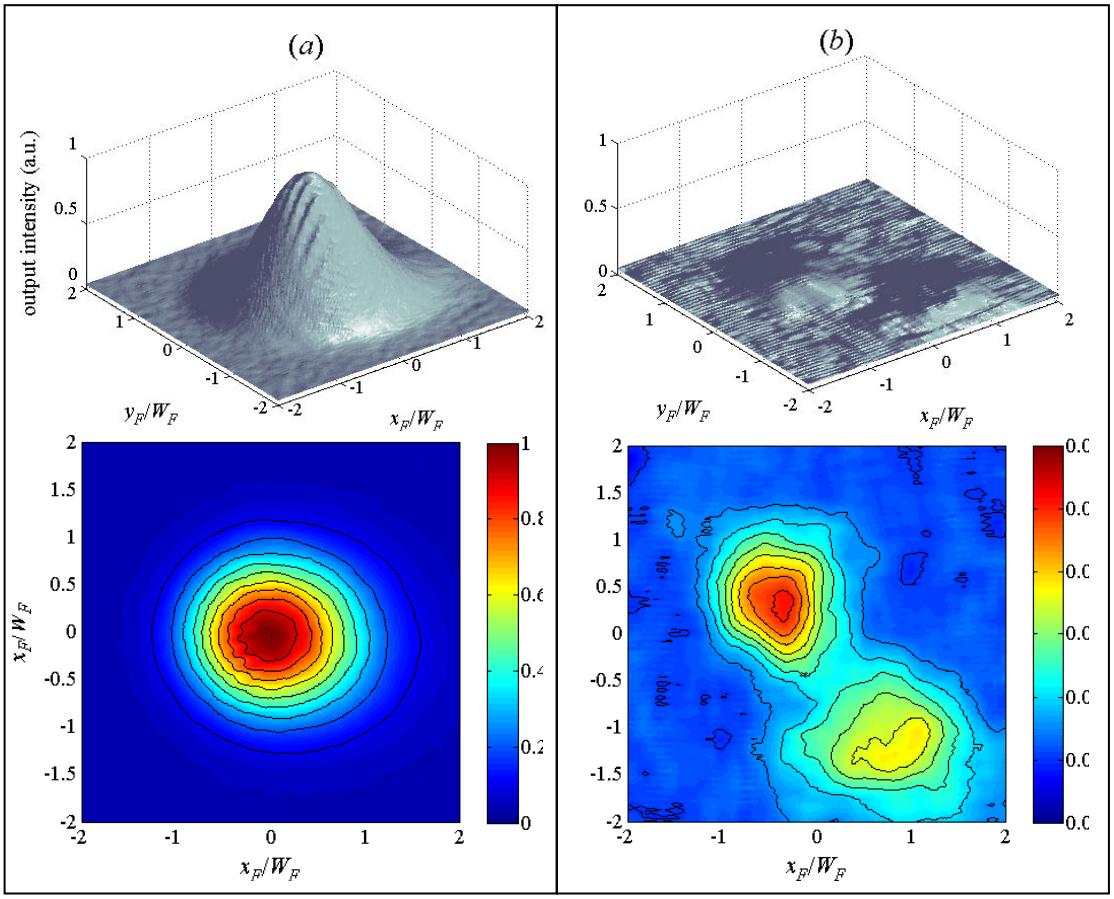


Figure 3-61. Linear scale plots of measured output plane intensity (*a*) without and (*b*) with the high-pass spatial frequency filter in place. The colour axis in the contour plot in (*b*) is scaled to show the low-level structure present in the filtered beam.

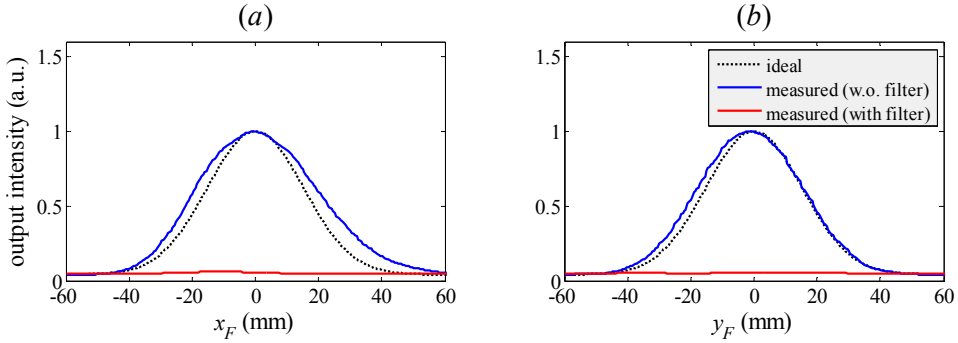


Figure 3-62. (*a*) x- and (*b*) y-cuts through measured output plane intensity with and without the high-pass spatial frequency filter in place. Notice that adaptive noise filtering was not applied to the intensity measurement made when the spatial filter is included since this technique is only applicable to measurements that contain a strong signal embedded in the noisy measurement. The best-fit Gaussian that matches the measured Gaussian intensity is asymmetric with waist radii of $(W_x, W_y) = (1.23, 1.03)W_F$, where the expected Gaussian radius is $W_F = 30.44$mm

A set of measurements were performed to examine spatially filtered imaging of an assortment of simple geometrically shaped objects: various opaque obstacles and apertures cut from pieces of Eccosorb. Ideally Eccosorb should absorb any incident

radiation so the various objects will block some portion of the illuminating beam and can be represented in numerical simulations by amplitude-modulating obstructions. Two types of measurement were performed on each test object, i.e. with and without the high-pass spatial filter in place. Both high-pass amplitude and low-pass phase filters were investigated – the latter with numerical simulations only.

The first object to be measured was a straight edge. A sheet of Eccosorb was placed in the object plane such that it obscures one half of the illuminating Gaussian beam. When no filtering is involved the output image is simply one half of the image of the illuminating Gaussian beam. Notice in Figure 3-63(*a*) that the edge appears curved because of distortions introduced by mirror M_3. When the filter is inserted the lower spatial frequencies are removed from the spectrum of the object beam and only high spatial frequencies are Fourier transformed by mirror M_3 onto the output plane. The result is image in intensity of the edge, as shown in Figure 3-63(*b*). The measured image shows a few interference fringes on either side of the straight edge with an intensity null along the position of the edge itself. This extra, unwanted filtering occurs because of truncation at the finite sized aperture of mirror M_2, which causes the highest spatial frequencies (associated with the sharp edge) being inadvertently removed from the Fourier spectrum. Thus the filtered image contains intermediate spatial frequencies from the original object beam. A numerical simulation of the system, in which the mirrors are treated as lenses and propagation was computed using Fresnel Transforms, verifies as much (see Figure 3-64).

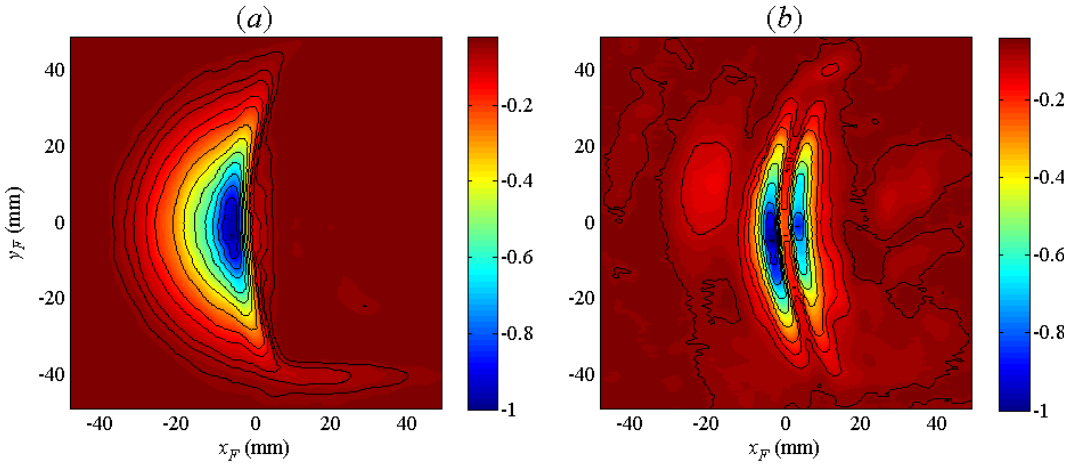


Figure 3-63. Linear-scale plots of measured output plane intensity (*a*) without and (*b*) with the high-pass filter in place. The test object is a rectangular sheet of Eccosorb that positioned so as to block the left side of the illuminating beam (as viewed in the direction of propagation from mirror M_1). Distortion introduced by mirror M_3 causes the image of the straight edge to appear curved.

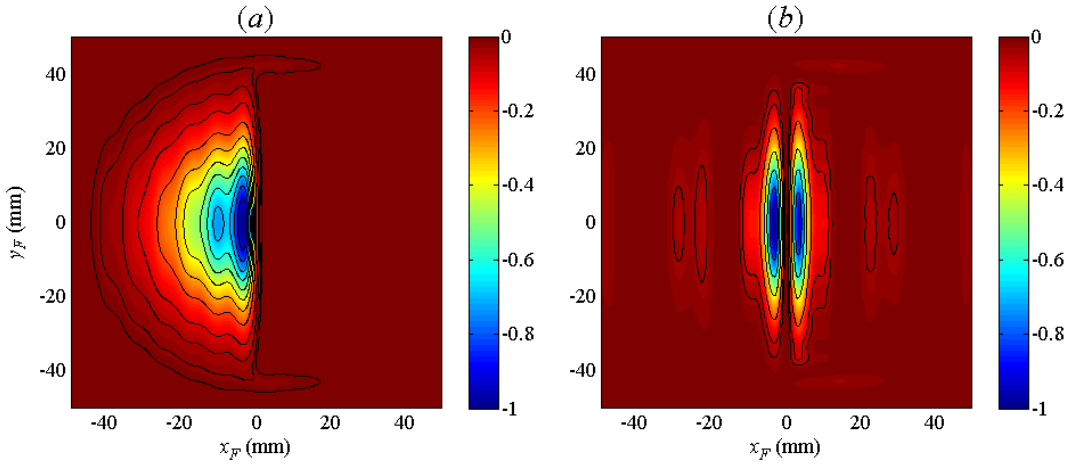


Figure 3-64. Linear scale plots of simulated output plane intensity (*a*) without and (*b*) with the high-pass spatial filter included in the set-up. Note that the system is treated as modelled as an in-line system, i.e. with (truncating) lenses instead of mirrors, so our simulations do not account for the distortion effects observed in the measured images.

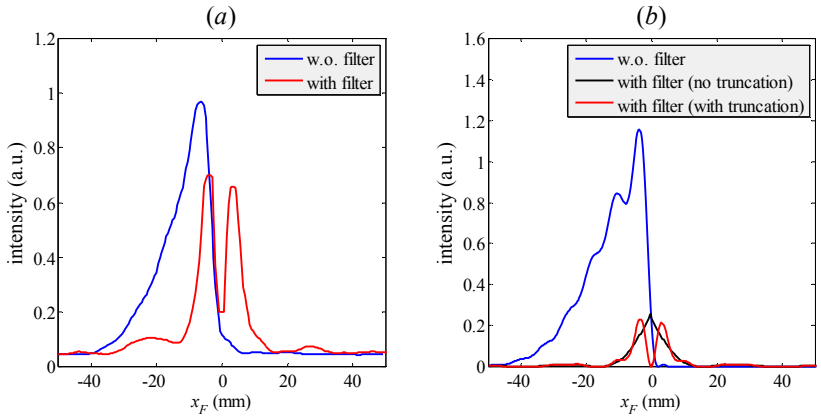


Figure 3-65. Linear-scale plots of x-cut through centre ($y_F = 0$) of (*a*) measured and (*b*) simulated output plane intensity for the vertical straight edge object. Without high-pass filtering in the shadow of the straight edge power falls off rapidly but the exact position of the edge is difficult to determine. Ideally, i.e. without truncation at mirror M_2, the system should produce an intensity peak at the position of the edge ($x_F = 0$). However truncation by M_2 filters out the highest spatial frequencies resulting in an intensity null at the edge.

In another experiment an opaque square obstacle (a piece of Eccosorb) was positioned so as to block one quadrant of the illuminating Gaussian beam for the same optical setup. The results in Figure 3-66 clearly show enhancement of the edges with spatial filtering. On comparison with numerical simulations (Figure 3-67) there is good agreement between the actual behaviour as measured and the expected behaviour from theoretical simulations, apart from distortion which is not accounted for in the model.

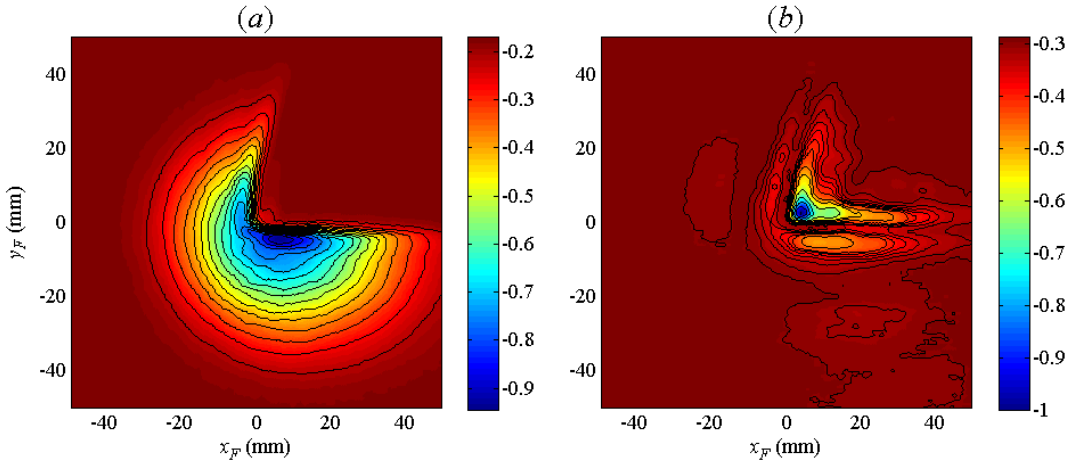


Figure 3-66. Measured output plane intensity (*a*) without and (*b*) with the high-pass filter. The test object is a square piece of Eccosorb positioned such that it blocks the lower-left corner of the beam (as viewed in the direction of propagation from mirror M_1). Again distortions result in the vertical edge being curved away from the *y*-axis.

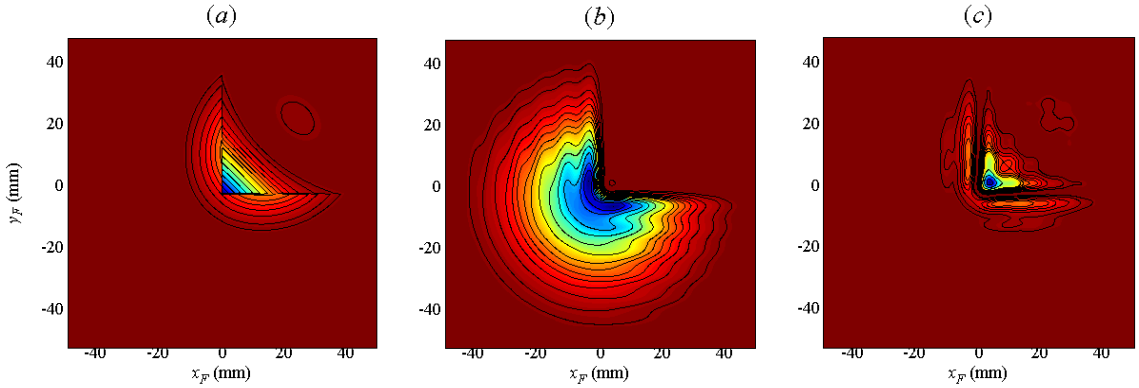


Figure 3-67. Simulated output plane intensity (*a*) for an ideal high-pass spatial frequency filtering, i.e. without truncation at mirror M_2, (*b*) with truncation included at M_2 and the high-pass filter excluded and (*c*) with truncation effects and high-pass filtering included.

In order to investigate apertures a number of measurements were made in which the object was a sheet of Eccosorb into which apertures of different shapes and sizes were cut. Figure 3-68 shows the filtered and unfiltered images of a sheet of Eccosorb with a 30mm diameter circular aperture at its centre. Again, when the high-pass filter is included, the system is most likened to a band-pass filter with intensity nulls occurring around the circumference of the circular aperture, as is most easily seen from the 1-D cuts through the centre of measured intensity in Figure 3-70.

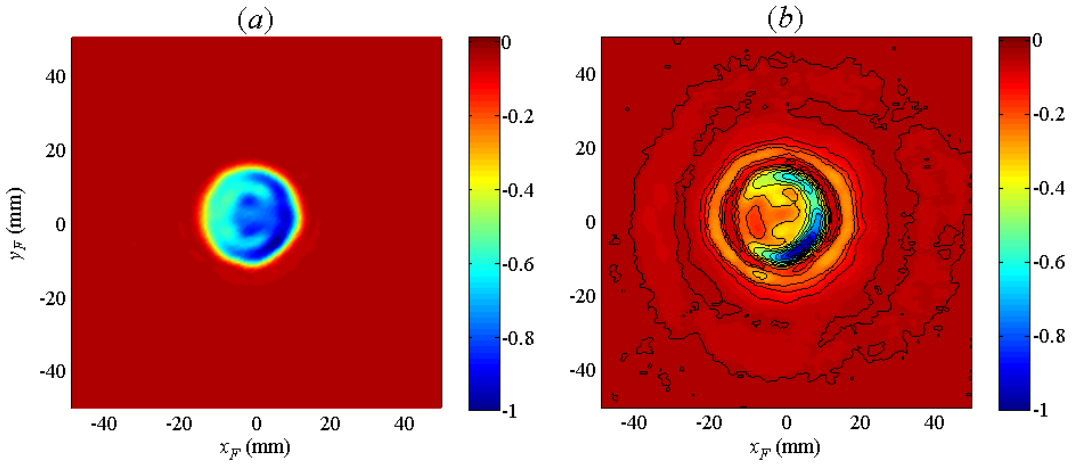


Figure 3-68. Linear-scale intensity beam pattern measurements from a circular aperture (diameter = 30mm = 10λ) (*a*) without and (*b*) with the high-pass spatial frequency filter included in the set-up. Colormap scaled to the range [0,exp(-1)].

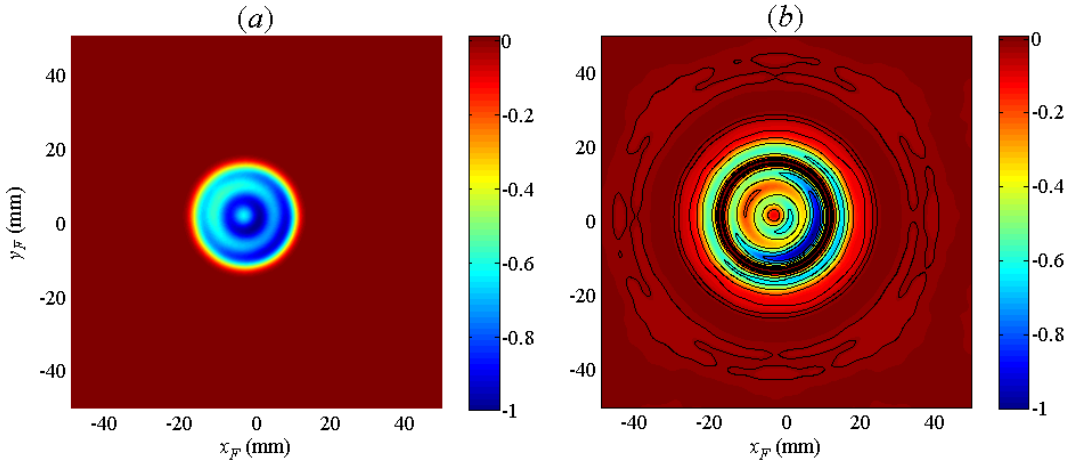


Figure 3-69. Simulated output plane intensity from the 30mm diameter circular aperture (*a*) with low-pass filtering to simulate truncation by mirror M_2 and (*b*) with band-pass filtering to simulate truncation at M_2 as well as high-pass spatial filtering. The asymmetric distribution of power observed in the measured intensity patterns in Figure 3-68 occurs because the circular aperture is offset with respect to the illuminating Gaussian beam. In order to replicate measurements, in simulations the circular aperture was defined such that its centre was located at coordinates (-3, 2) mm.

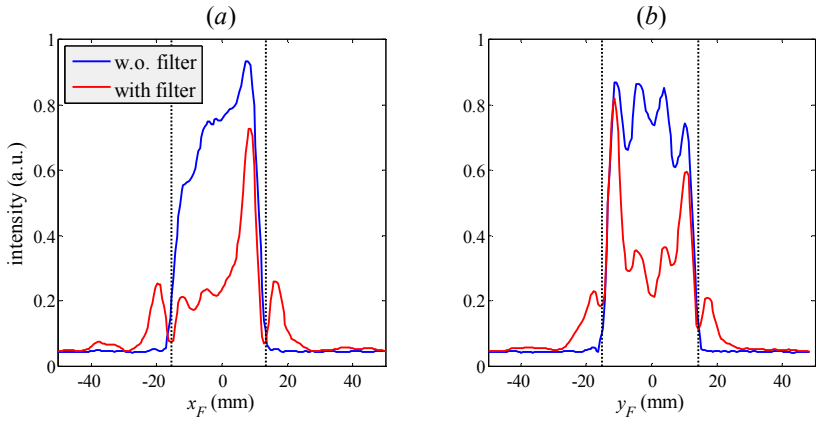


Figure 3-70. *x*- and *y*-cuts through centre of output plane intensity measurements from a 30 mm diameter circular aperture with and without the high-pass spatial filter in place.

A similar experiment where the object is a sheet of Eccosorb with a square aperture (width and height = 30mm) at its centre and placed in the beam path was also undertaken. Figure 3-71 shows the measurements of the images in intensity made at the output plane with and without spatial filtering. The asymmetric distribution of power in the 30mm×30mm square region observed in the measured output intensity distribution was due to a slight misalignment of the square aperture with respect to the centre of the illuminating Gaussian beam. If the illuminating Gaussian is defined with its centre at the point (+4,-4) mm in the object plane the resulting simulations of filtered and unfiltered output plane intensities (Figure 3-72) are comparable with the measurements.

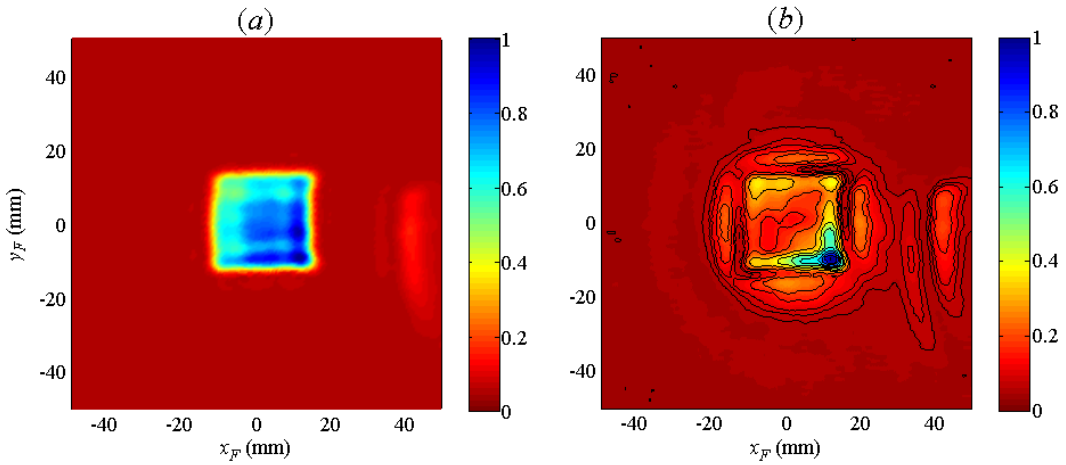


Figure 3-71. Linear-scale plots of measure output plane intensity from the square aperture (*a*) without and (*b*) with the high-pass spatial filter in place.

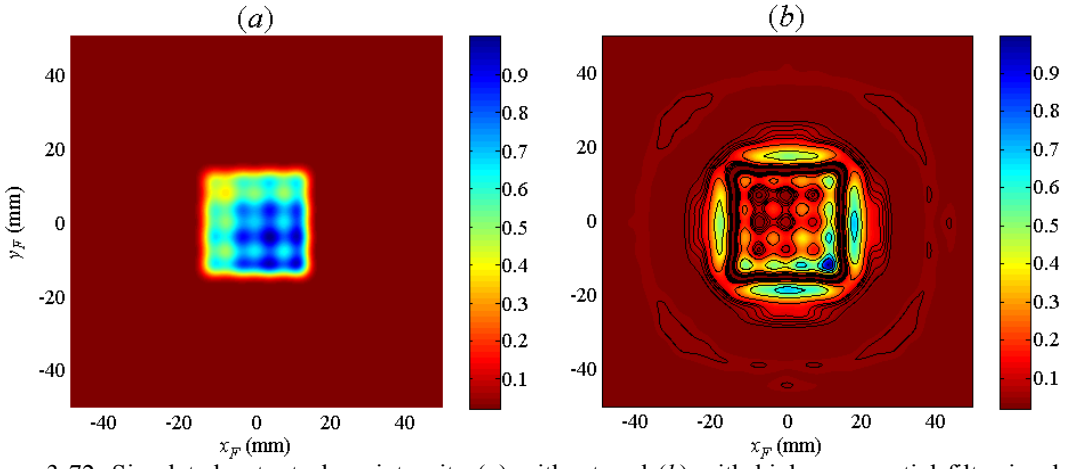


Figure 3-72. Simulated output plane intensity (*a*) without and (*b*) with high-pass spatial filter in place. The illuminating Gaussian is defined with its centre at the point (x_S, y_S) = (+4,-4) in the object plane.

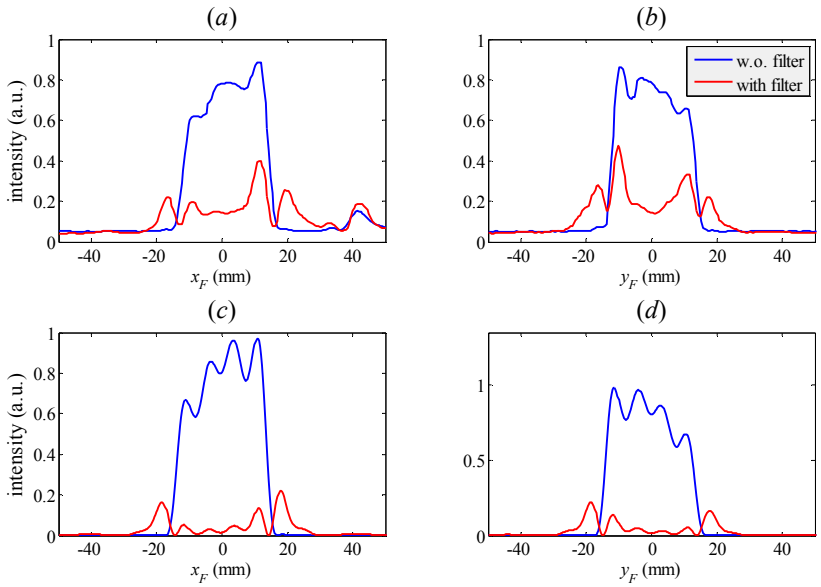


Figure 3-73. (Left) x- and (right) y-cuts through the centre, i.e. at the point (x_F, y_F) = (0, 0), of the (top) measured and (bottom) simulated output plane intensity patterns for a square aperture of width 30mm.

Before continuing we consider the use of a phase-filter instead of the amplitude filter used so far. Instead of blocking the low spatial frequency components of the object beam spectrum we apply a linear phase shift to this part of the spectrum. The result is that the image of the part of the beam associated with this part of the spectrum, i.e. the illuminating Gaussian beam, is reproduced off-axis at the output plane. Meanwhile the part of the beam associated with the non-phase shifted part of the spectrum, i.e. the edges of the object, are formed on-axis as usual. Thus the low-and high-frequency terms of the object beam can be spatially separated at the output plane. This allows one to simultaneously produce an edge-enhanced image of the object beam, as well as an image of the illuminating beam itself. Figure 3-74 shows the result of a simulation in which the phase filter is of the form

$$\phi(r) = \begin{cases} \exp[-i2\pi x/\Delta] & r \le a \\ \exp[+i2\pi x/\Delta] & r > a \end{cases} \tag{3.31}$$

where in the example shown radius $a = 3W_{SF}$ and $\Delta = 1.5$. In practise a phase filter would most likely consist simply of a single circular region with a linear phase shift by simply inserting a wedge of refractive material with the appropriate thickness to achieve the required linear phase shift on the low-frequency components, i.e. the first line of Eq. (3.31). The external linear phase shift, i.e. the second line of Eq. (3.31), is only included in our simulation for clarity (to produce two well-separated images on either side of the optical axis at the output plane). A more sophisticated device, made from a number of

concentric rings each with a different linear phase shift, would enable one to separate the object beam into multiple images, each containing power in a different spatial frequency band.

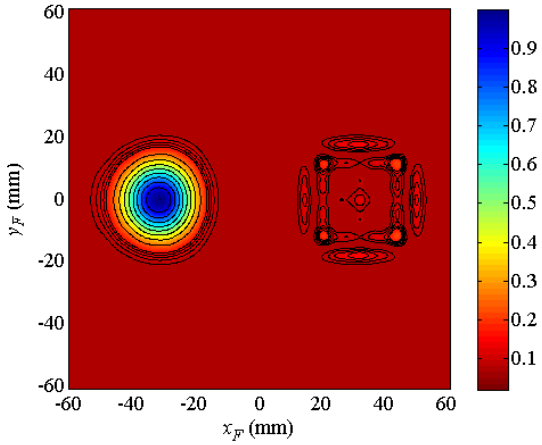


Figure 3-74. Simulated output plane intensity pattern when spatial frequencies are filtered using a linear phase filter in order to separate the object beam into two parts: one containing low spatial frequencies (the illuminating Gaussian beam) and the other containing high spatial frequencies (the edges of the object, which in this case is a square aperture).

Figure 3-75 shows the output plane intensity recorded for a square obstacle (Eccosorb of width 30mm) in the path of the illuminating beam. Again there is slight misalignment between the object and the illuminating beam. Also a small amount (a region of ~6mm in width) of the back surface of the square piece of Eccosorb was missing which resulted in some power being transmitted through this region and appearing with higher intensity in the filtered image. In the simulation of this experiment we define the Gaussian with waist radii of $(W_{Sx}, W_{Sy}) = (1.03,1.23)W_S$, where $W_S = 30.44$mm and with its centre at $(x_S, y_S) = (+3,-3)$mm at the object plane.

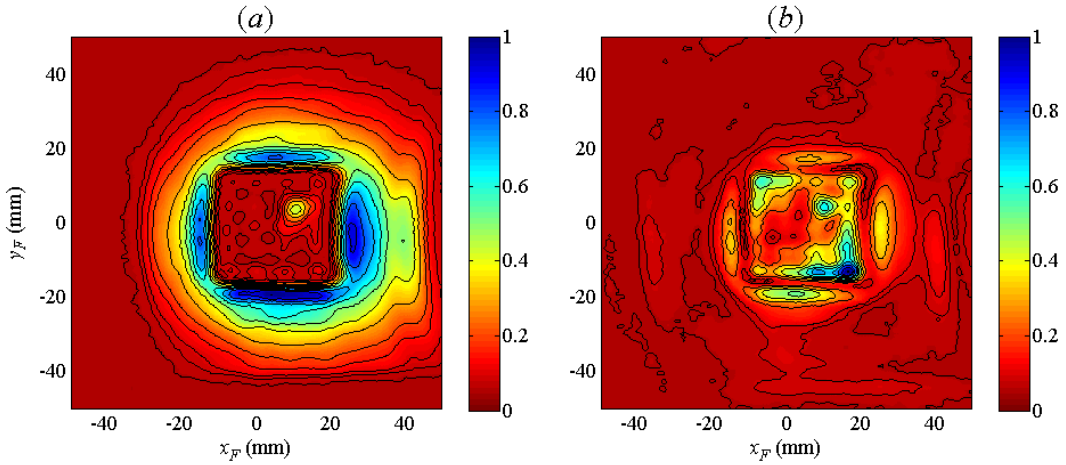


Figure 3-75. Linear-scale plots of measure output plane intensity from the square obstacle (*a*) without and (*b*) with the high-pass spatial filter in place.

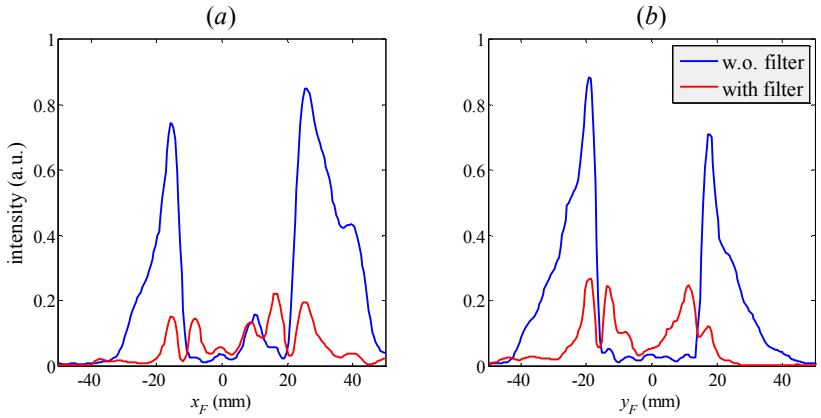


Figure 3-76. x- and y-cuts through measured output plane intensity with and without spatial frequency filtering.

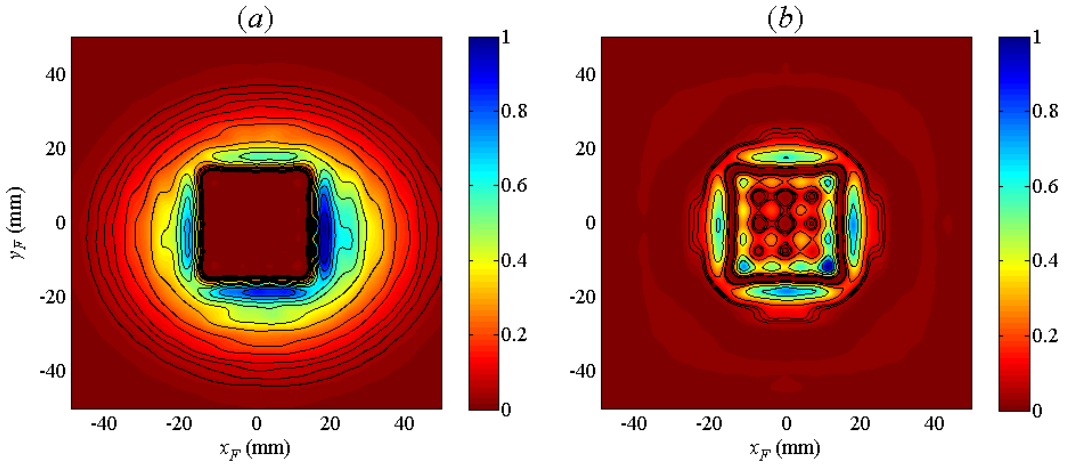


Figure 3-77. Simulated output plane intensity (*a*) without and (*b*) with high-pass spatial filter in place. The illuminating Gaussian is defined with its centre at the point $(x_S, y_S) = (+3,-3)$ in the object plane.

3.3.4 Modelling Spatially Filtered Imaging with GBMA

The numerical simulations thus far described of spatially filtered imaging have been undertaken using Fourier transforms. Next we develop a GBM model of the system. A scattering matrix approach [3.17] allows one to develop a model for the system that once in place can be re-used for a variety of test objects, provided the same basis mode set can adequately represent those object fields. To demonstrate this method the re-imaging transmission system was modelled using a mode set capable of reproducing obstacles that have a minimum feature size of 5 mm × 5 mm, which allows us to define a square aperture, or stop with the same dimensions. This modal description then allows us to model larger objects, such as one of the 30 mm × 30 mm square apertures or stops that were used in experimental measurements, without having to recalculate the relevant scattering matrices. Furthermore the object beam need not be centred on the origin so

we can test what effects repositioning the object with respect to the illuminating beam has on image formation at the output plane.

We begin by constructing the object beam: as an example we consider a Gaussian amplitude distribution (representing the illuminating beam) with a 5mm×5mm stop at its centre. Next the mode-set parameters (number of modes and fundamental beam waist radius W_0) are selected. For this example the highest-order mode indices were set to $m_{\max} = n_{\max} = 47$ and the beam mode scaling factor was set to $W_{0x} = W_{0y} =$ 8.875mm. Next the Gaussian beam mode coefficients, A_{mn} for the object beam were calculated. Because the reimaging system includes a spatial filter, mode scattering is evaluated in the same way as is done for truncation (see Chapter 2). Three scattering matrices are required: two to account for truncation at mirrors M_2 and M_3 and one to account for the on-axis blockage or phase transformation at the spatial filter. Once the three scatter matrices have been computed the scattered mode coefficients and subsequently the field distributions at various planes in the system can be calculated. Because the initial object beam is symmetric the modal decomposition will include power in only the symmetric (even-numbered) modes. However when calculating the scatter matrices all modes (even- and odd-numbered modes up to $m_{\max}$ and $n_{\max}$) were included, so as to allow the simulation of object beams that have arbitrary symmetry.

Since propagation through the system is now known, the initial test object can be replaced and the scattering matrices used to propagate the field of a different object beam through the system. For example Figure 3-78 shows the simulated output plane intensity produced when a square 30 mm × 30 mm square aperture is placed at the object plane (with and without high-pass spatial filtering). To replicate the results from experimental measurements, the aperture was simulated with its centre at the point (-8 mm, +3 mm), i.e. misaligned with respect to the illuminating beam. Beam propagation was computed using both Fourier transforms and GBMA with the scatter matrix approach.

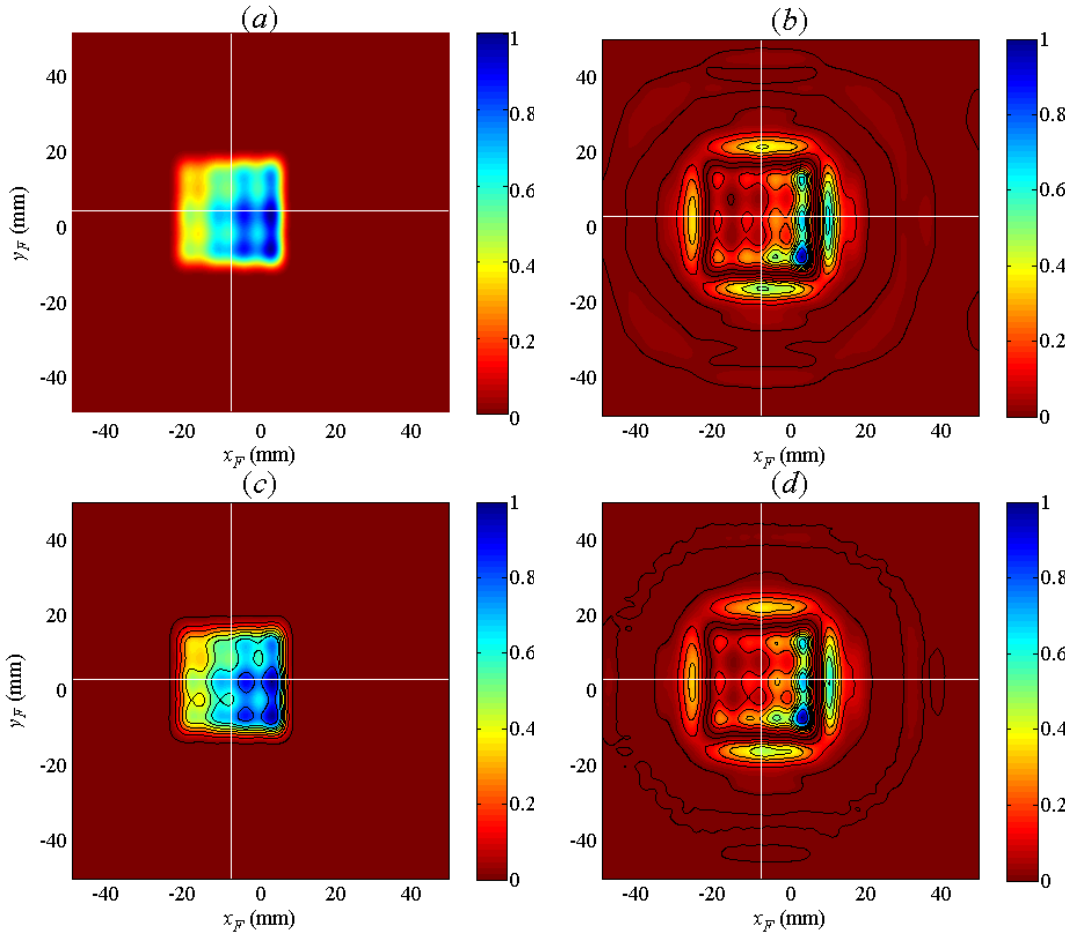


Figure 3-78. Results of numerical simulations using (top) Fresnel transform and (bottom) Gaussian beam modes (using a scattering matrix approach) to simulate propagation through the re-imaging transmission system. Here the object is a square 30mm × 30mm aperture centred at (x_S, y_S) = (-8,+3)mm in the object plane. (*a*) and (*c*) show the simulated output plane intensity when the spatial filter is omitted (truncation effects are included); (*b*) and (*d*) show output intensity when high-pass spatial filtering is included.

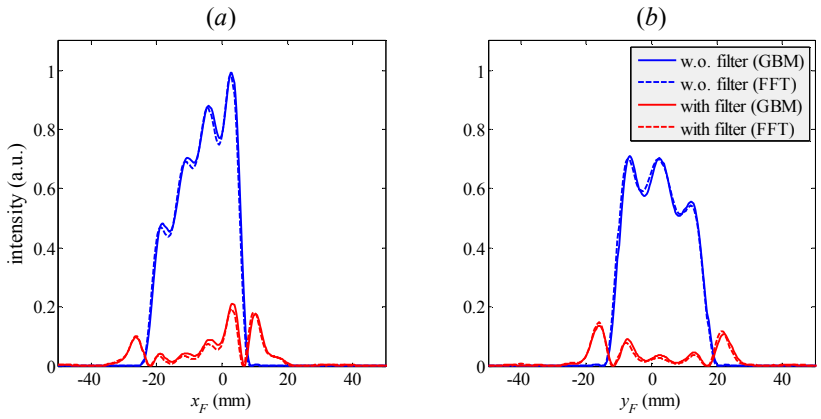


Figure 3-79. x- and y-cuts through the (GBM and FFT) simulated output plane intensity plots for a square 30mm×30mm aperture centred at (x_S,y_S) = (-8,+3) in the object plane which is illuminated with a collimated Gaussian beam centred on the optical axis.

A scattering matrix approach is versatile because the scattering matrices that describe the Fourier optics system (including the filter) need only be calculated once. They are then available for use in the analysis of the imaging of different illuminated objects. One problem that was encountered was high computational cost, in terms of the time required to calculate the scattering matrices. In the above example a modest number of modes (48×48) were used yet it took eight days to calculate the three scattering matrices on a desktop computer using straightforward overlap integrals to determine the scattering matrix elements. However once these were calculated it then took only 2-3 minutes to calculate the mode coefficients for a new object beam and propagate the beam through the system to the output plane including truncation effects and spatial filtering along the way. The alternative method, of calculating the scattered mode coefficients at each aperture (at mirrors M_2 and M_3 and at the spatial filter), would take significantly longer and would have to be repeated every time a new object was used. A cleverer approach to computing the scattering matrices using singular-value decomposition (SVD) could reduce computational overload. Computational overhead could also be further reduced by omitting weakly contributing modes. Although Fourier transforms do outperform the GBM approach in terms of computational speed they cannot be used to account for truncation effects at non-Fourier or image planes. For example we could not include truncation at mirror M_3 using Fourier transforms. Furthermore, at present in our model of the system mirrors M_2 and M_3 are treated as truncating lenses. The actual system involves mirrors which introduce distortion effects as well as truncation and although we have not done so here our GBM model can be adapted to incorporate distortions and other aberrations that would are introduced by off-axis mirrors (as described in [3.18]). The transmission re-imaging system was also simulated [3.22] using the software package MODAL (Maynooth Optical Design and Analysis Laboratory) that was developed in the Department of Experimental Physics at NUI Maynooth. MODAL can perform beam propagation through an optical system in terms of Gaussian beam mode analysis, Fresnel integrals as well as physical optics (PO) techniques. It allows one to include focusing elements such as off-axis reflectors and so takes into account both truncation and distortion effects.

3.3.5 Examples of Spatially Filtered Imaging of Real Objects

Next we present the results of imaging experiments with real objects as well as results of simulations using GBMA. First we consider opaque objects with complicated shapes, before moving onto a transparent object that imparts a phase-only modulation on the object beam. Finally we present measurements of spatially filtered imaging of some biological samples: two leaves and a thin slice of meat with fatty tissue.

In the first experiment the objects are two small opaque metal objects: a key and a ring. These were arranged close to each other, at the object plane as shown in Figure 3-80(*a*). Both objects are approximately 1mm in thickness and so can be treated as infinitely thin perfect reflectors. Since both objects are opaque to incident radiation so they cause an amplitude modulation of the object beam.

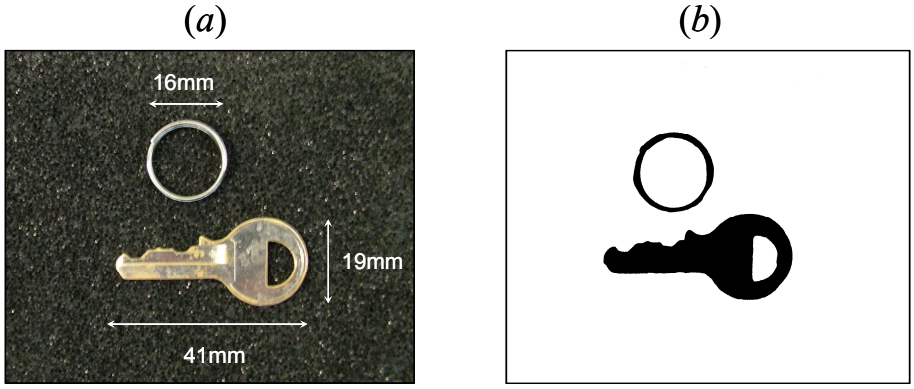


Figure 3-80. The photograph (*a*) of the two metallic test objects, a small key and ring, was used to create a binary mask (*b*) representing the transmission function imposed on the illuminating beam. The key has a length of 41 mm and a width of 19 mm. The ring has inner and outer diameters of 16 mm and 17 mm, respectively.

Two images were recorded by scanning the detector with TOAST at 100 GHz with and without high-pass spatial filtering (Figure 3-81). Discontinuities in the image of the object beam are enhanced when the high-pass spatial filter is included. Because the illuminating beam is relatively small compared to the size of the objects, high contrast is only obtained towards the centre of the arrangement, where illuminating beam intensity is at a maximum. Accurate imaging of larger objects would require a larger illuminating beam. For example it would have been better had the illuminating beam been collimated using one of the 350 mm focal length mirrors (instead of the 150 mm focal length mirror) as this would have provided a Gaussian beam with a radius of ~71mm. However

when the experiment was conducted the two mirrors of focal length 350 mm were needed to form the Fourier optics imaging arrangement as no other large fast mirrors (i.e. with short focal ratios) were available to collect scattered radiation from the object beam and the filtered beam. Despite the fact that illuminating beam power decreases rapidly with distance from the optical axis, the spatially filtered image even reveals the parts of the edges of the plastic CD case that was used as a sample holder.

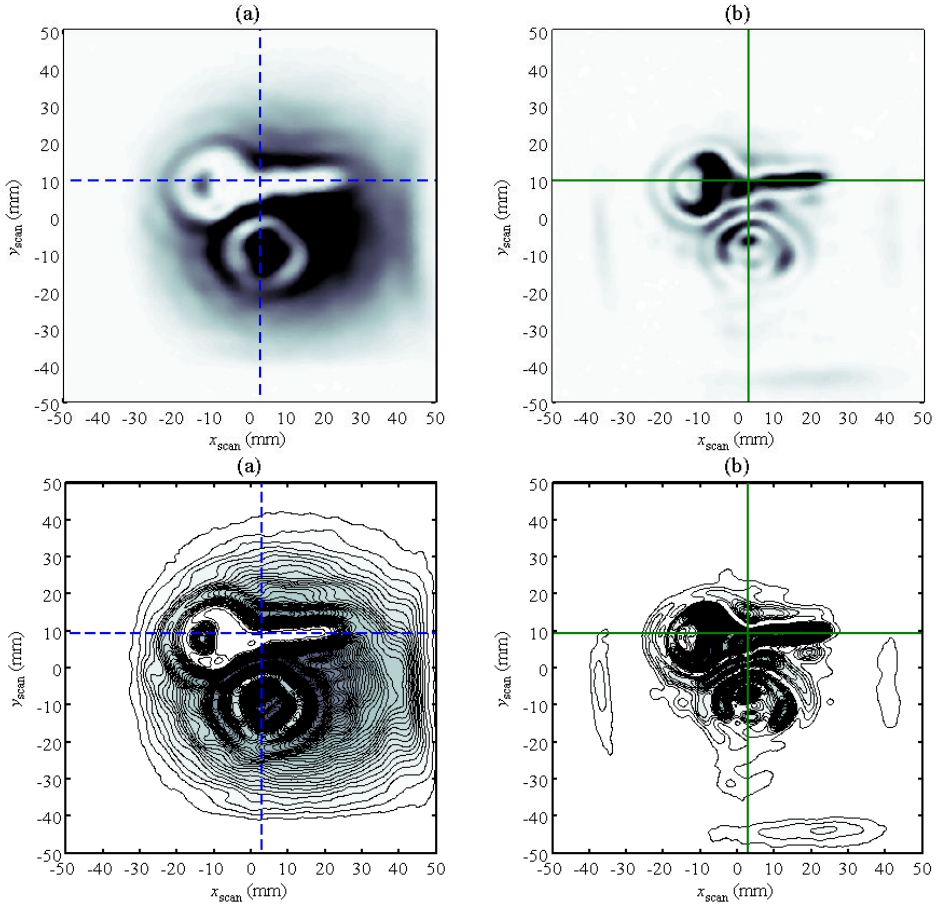


Figure 3-81. Measured image patterns at the output plane in intensity for the small metal key and ring as object (*a*) without and (*b*) with the high-pass spatial filter included in the system. Despite the decrease in power away from the optical axis (the origin) discontinuities in the beam are enhanced even in regions where the illuminating beam is relatively weak. For example edges of the plastic CD that was used as a sample holder are clearly visible in the lower and on the left and right of the image in (*b*).

A computational simulation of this experiment was performed using Gaussian beam mode analysis with code developed by the author. The binary mask shown in Figure 3-80(*b*) that represents the amplitude modulation of the object beam was scanned (row-by-row and column-by-column) to calculate the smallest feature size in the object beam, which were determined to be δ_x = 0.9 mm and δ_y = 0.6 mm, in x and y directions, respectively. The object plane has dimensions of $4(W_S \times W_S)$ = 60 mm × 60 mm, where

W_S is the radius of the source beam at the object plane. In order to accurately reconstruct the object beam a set of Gaussian beam modes must be chosen that can reproduce features with a minimum size of $\delta_x \times \delta_y$. Given the width and height of the object plane and the minimum feature size, this required a mode-set with highest-order mode indices of m_{max} = 400 and n_{max} = 267. The relatively large obstacles (circular and rectangular apertures and blocks) of the previous experiments required relatively few modes which meant a scattering matrix approach could be used to analyse truncation and spatial filtering of the object beam as it propagated through the system. The much larger number of Gaussian beam modes (401 × 268 = 107,468) needed to describe the current object beam results in prohibitive computational times needed to calculate the various scatter matrices (at mirror M_2, the spatial filter and mirror M_3). Thus we resorted to a step-and-stop method: propagating the beam a finite distance, truncating the resultant field with an appropriate binary mask (representing the aperture of a mirror or the spatial filter) before propagating to the next element. Because of the larger number of modes involved, when propagating to the various planes a pre-processing step (a thresholding operation) was employed to exclude from the modal summation any weakly contributing modes. Thus, for example when calculating the ideal output field (without truncation or spatial filtering) the modal summation involved only ~35% (37,917) of the total number of modes.

Figure 3-82 shows the GBM-reconstructed image of the beam that would be formed at the output plane with and without the high-pass spatial filter in place.

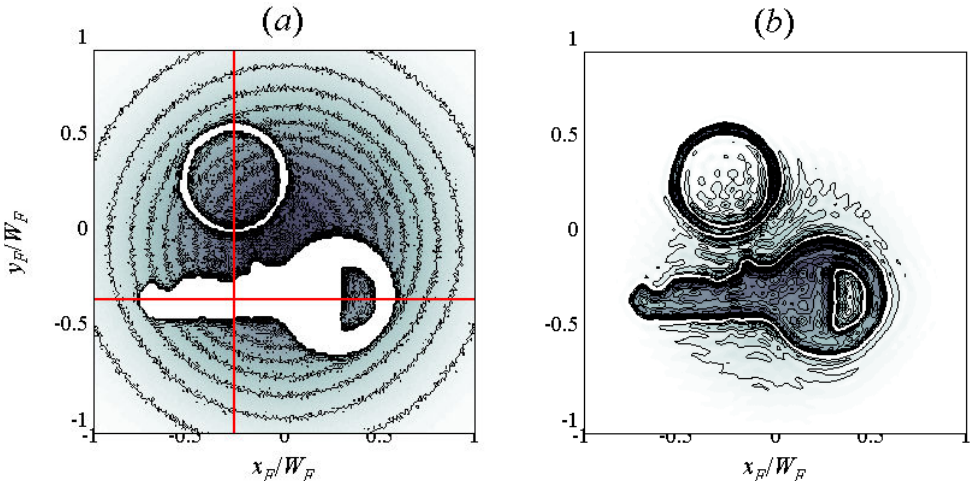


Figure 3-82. GBM-reconstruction of output plane intensity (*a*) without and (*b*) with high-pass spatial filtering involved. The red lines superimposed on (*a*) show where (at x_F = -0.26W_F and y_F = -0.3W_F) horizontal and vertical cuts are taken from as displayed in Figure 3-83.

Note that in our simulations mirrors M_2 and M_3 were modelled as 500mm focal length ellipsoidal mirrors, instead of the actual 350mm focal length parabolic mirrors that were used in experimental measurements. Notice that whereas image contrast decreases with increasing off-axis distance in the image formed without spatial filtering, this decrease in image contrast is not so apparent in the spatially filtered image.

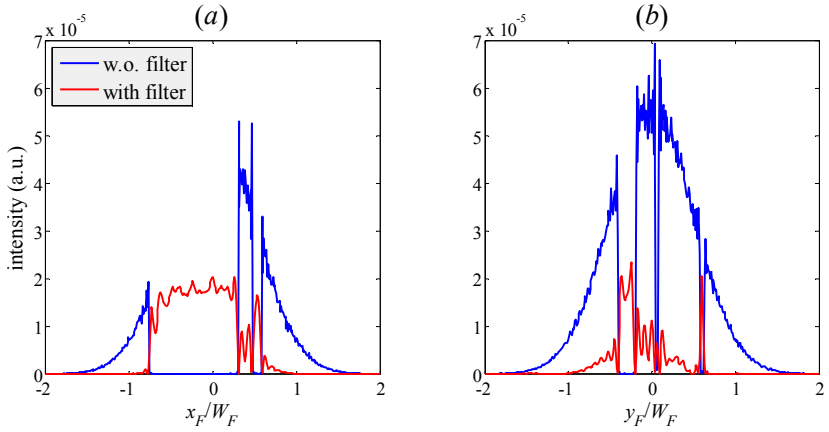


Figure 3-83. (*a*) x- and (*b*) y-cuts through the GBM-simulated output plane intensity with and without the high-pass spatial filter in place.

Although the simulation does produce images with features similar to those observed in the experimentally measured images, there are significant differences. Most notably, the edges in the simulated intensity patterns appear sharp, whereas those in the measured images are quite blurred (Note that recorded intensity images were not deconvolved with the PSF of the waveguide probe so we do not expect to see sharp edges). In order to reconcile the experimental measurements with simulated images the computational model must be made more realistic, which can be done by taking into account truncation at the finite apertures of the optical elements in the Fourier optics part of the system: mirrors M_2 and M_3. These mirrors were designed to be capable of collecting radiation from an undistorted Gaussian beam. However the small features and sharp edges of the object beam means diffraction will cause the object beam to have spread into a large area by the time mirror M_2 is encountered. The finite-sized aperture of M_2 results in the loss of information contained in the outer regions of the beam. Mirrors M_2 and M_3 thus effectively act as low-pass filters: they remove some of the high spatial frequency content of the beam. A more accurate model of the system must therefore include truncation effects at mirrors M_2 and M_3. Figure 3-84 shows the simulated output plane intensity (with and without high-pass spatial filtering) when truncation at mirrors M_2 and M_3 is included.

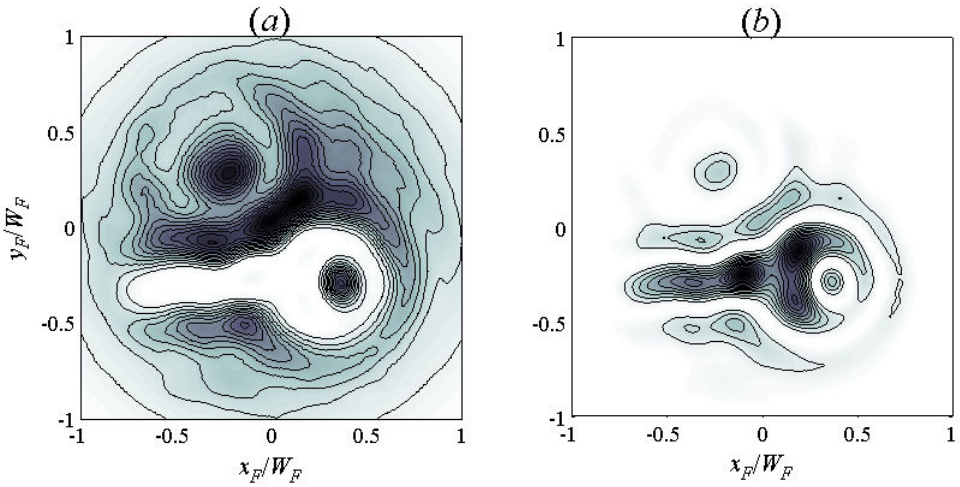


Figure 3-84. GBM-reconstructed output plane intensity (*a*) without and (*b*) with high-pass spatial filtering. The system is now modelled to include truncation effects at mirrors M_2 and M_3.

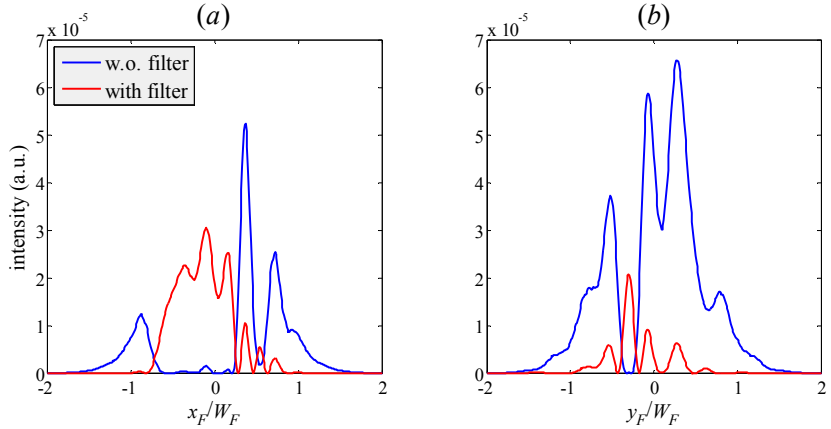


Figure 3-85. (*a*) x- and (*b*) y-cuts through the output plane intensity with and without the high-pass spatial filter in place and with truncation effects at mirrors M_2 and M_3 included.

The impact that including truncation effects by mirrors M_2 and M_3 has on the numerical model is dramatic with there now being very close agreement between simulated and experimentally measured output images. Clearly truncation effects reduce the resolution of the system by filtering out high spatial frequencies. However, it is not yet clear at what point in the system these effects are most problematic: at M_2 or M_3. An understanding of this problem is required if one wishes to re-design the system so as to achieve higher resolution. To this end we now examine how the distribution of power between mode coefficients changes due to truncation by the mirrors M_2 and M_3 and the spatial filter. Maps of mode coefficients are shown on a log-scale and only even-numbered mode coefficients are displayed because, apart from a few low-order modes, the majority of odd-numbered modes contain no power.

Figure 3-86 shows the intensity of mode coefficients A_{mn} before and after truncation by the high-pass spatial filter, which causes a redistribution of power between mode coefficients.

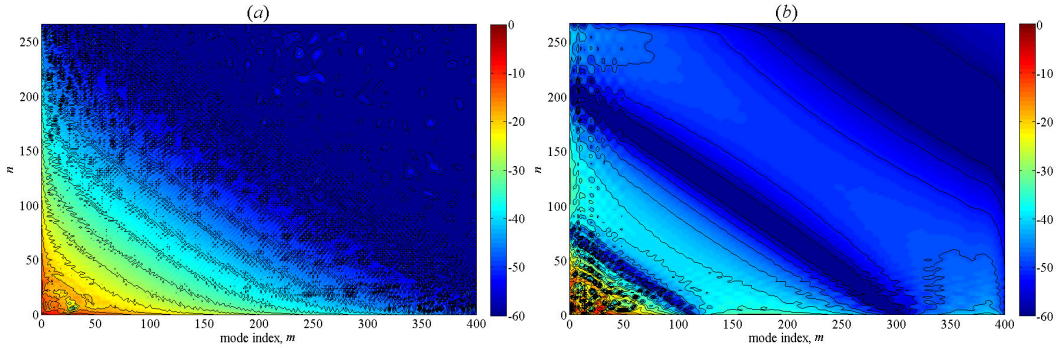


Figure 3-86. Mode coefficient intensities $|A_{mn}|$ (*a*) before and (*b*) after spatial filtering.

Figure 3-87 shows the intensity of A_{mn} when (*a*) truncation at mirror M_2 is included and (*b*) when truncation at both M_2 and M_3 is included. The gross structure of the mode map after truncation by M_2 alone is similar to that for the object beam without truncation (Figure 3-86(*a*)), but with the finer details smoothened out. There is little difference between the two mode maps in Figure 3-87, which indicates that the majority of blurring of the output plane intensity seen in Figure 3-84(*a*) is due mainly to truncation effects incurred at mirror M_2.

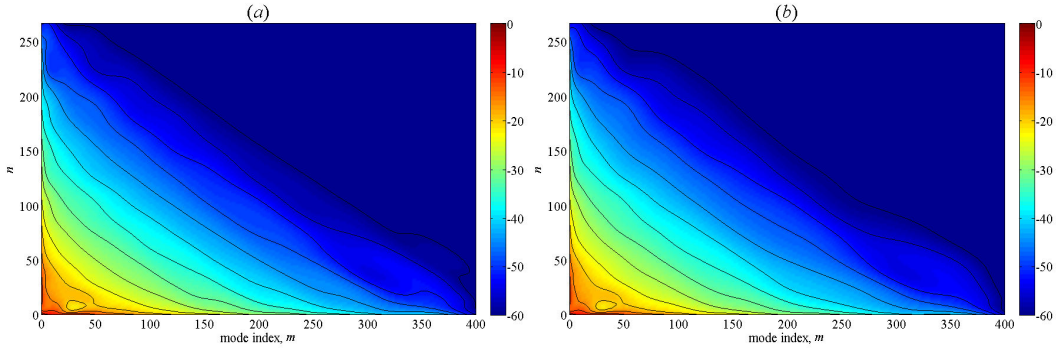


Figure 3-87. Maps of mode coefficient intensity (*a*) with truncation at mirror M_2 and (*b*) with truncation at mirrors M_2 and M_3. The high-pass spatial filter is not included in this model.

Figure 3-88 shows the simulated intensity distribution at the spatial filter plane with and without truncation at mirror M_2 included. The introduction of mirror M_2 results in the loss of all high spatial frequency content beyond a radius of ~16 W_{SF}, where W_{SF} is the radius that the illuminating beam would have at the spatial filter were no object included in the system. Thus M_2 effectively acts as a low-pass spatial filter, the size of which limits the resolution that one can achieve with the system. The combination of

the high-pass spatial filter and truncation at mirror M_2 could therefore be modelled equivalently as a single band-pass spatial filter.

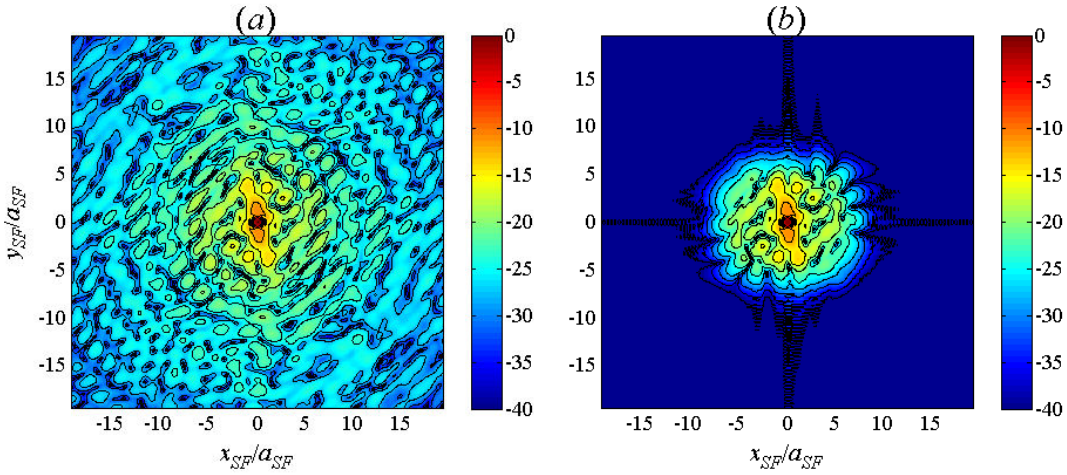


Figure 3-88. Calculated beam pattern intensity at the spatial filter plane (*a*) without and (*b*) with truncation effects at mirror M_2 included. The radius of the spatial filter is $a_{SF.} = 2W_{SF}$.

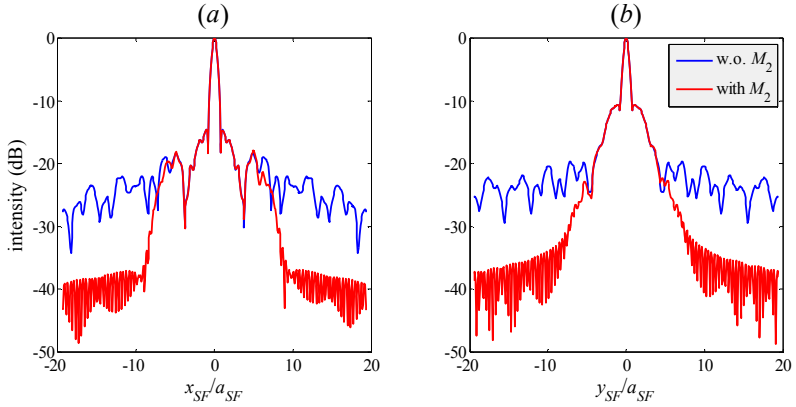


Figure 3-89. (*a*) x- and (*b*) y-cuts through the spatial filter plane beam pattern intensities of Figure 3-88.

Figure 3-90 shows mode coefficient intensities when the spatial filter is included. The two plots correspond to mode coefficients when truncation is included at (*a*) mirror M_2 only and (*b*) both M_2 and M_3. Little difference is observed before and after truncation by mirror M_3 because so much of the beam has already been truncated by mirror M_2.

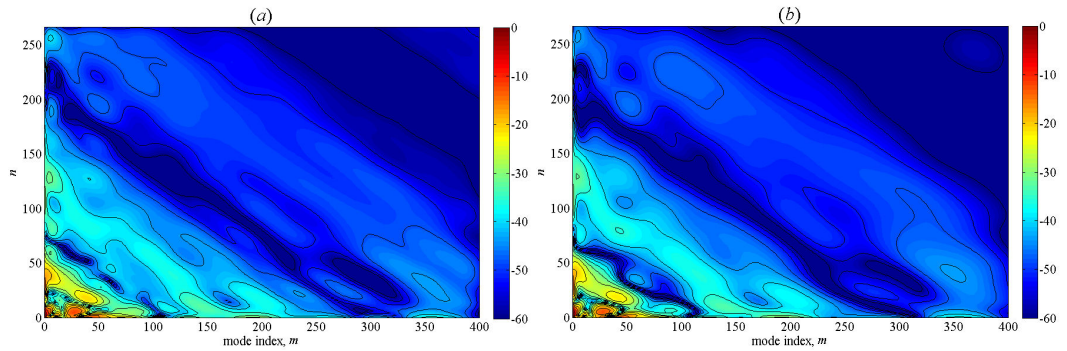


Figure 3-90. Mode coefficient intensities for the system including the spatial filter (*a*) with truncation at mirror M_2 only and (*b*) with truncation at M_2 and M_3.

The imaging of real transparent objects was also investigated; specifically a piece of high density polyethylene (HDPE) of uniform thickness that was cut into the shape of the letter 'R' (Figure 3-91). The uniform thickness of the object (t = 4 mm ≈ 1.1λ at a frequency of 100 GHz) means that a constant phase modulation of $t(n-1)2\pi/\lambda$ = 1.38π is imparted on the portion of the illuminating beam transmitted through the object, since the refractive index of HDPE at 100 GHz is taken to be, n = 1.52.

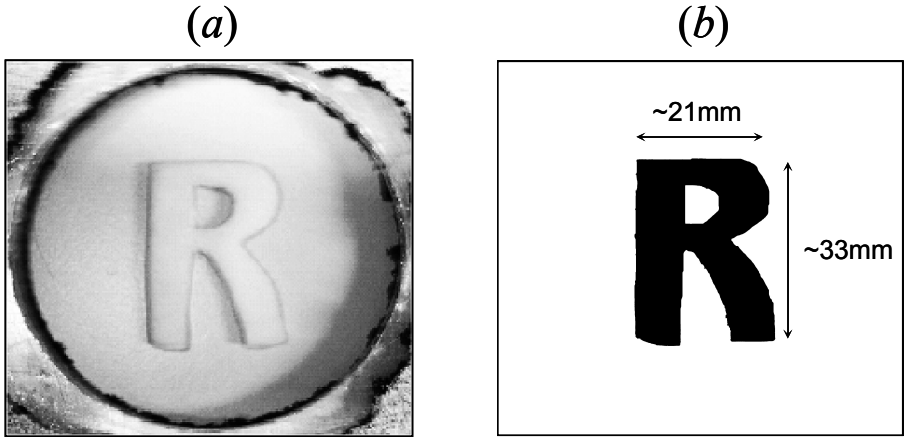


Figure 3-91. (*a*) The transparent HDPE object (4 mm thick) and (*b*) the binary phase mask used to represent the phase-modulation that it imparts on the illuminating beam.

An alternative method for imaging transparent objects is to use a 4-*f* correlator for phase contrast illumination of the object by including a quarter-wavelength plate (or dot) at the Fourier plane – the spatial filter plane. However, as will be seen, the limited size of the focusing elements provided an adequate means of performing phase contrast imaging. Since the object does not (ideally) modulate beam intensity we would expect that the image formed at the output plane (when the spatial filter is omitted) should be just an image of the illuminating Gaussian beam. The beam pattern measurements that were made of the HDPE object with and without the high-pass spatial filter in place are shown in Figure 3-92. Interestingly the edges of the object are revealed even when the spatial filter is omitted. In this case edge enhancement was found to occur due to truncation at mirror M_2, equivalent to the use of a low-pass spatial filter between mirrors M_2 and M_3. The high spatial frequencies associated with the edges of the object are lost from the beam as it is reflected from M_2, thus the image formed at the output plane contains less power at the location of the object edges. When the spatial filter is then inserted into the system it filters out the spatial frequencies associated with the illuminating beam. Thus the finite-sized optics mean that the high-pass spatial filter is

not necessary to perform edge detection of transparent objects. However its inclusion helps to improve image contrast.

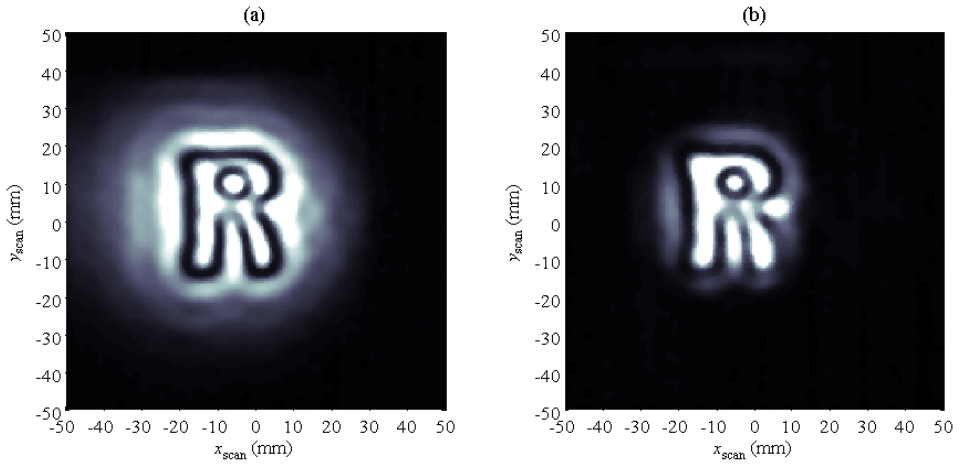


Figure 3-92. Measured output plane intensity of HDPE in the shape of the letter 'R' (*a*) without and (*b*) with the high-pass spatial filter in place. The edges of the object are reproduced in both measurements. In fact, more correctly the edges are not reproduced in both images because the high spatial frequencies are filtered out of the object beam by the aperture at mirror M_2. The only benefit of including the high-pass filter is to remove power associated with low spatial frequencies of the illuminating beam, thus providing better contrast in (*b*). The colour axis is scaled to the range $[0, e^{-1}]$.

A GBM-simulation of the Fourier Optics re-imaging system with the HDPE 'R' as object was performed. This time mirrors M_2 and M_3 were specified with the parameters of the 350mm focal length parabolic mirrors that were used in the experimental measurements. The smallest feature size in the object beam (a Gaussian amplitude distribution with R-shaped constant phase modulation) was determined to be $(3\times2)\Delta x$, where the object plane sample rate is $\Delta x = \Delta y = 0.3044$mm. This leads to a mode set with parameters $m_{max} = 267$ and $n_{max} = 400$. Since the object beam possesses neither even or odd symmetry all modes were used in the GBM decomposition of the object beam. Figure 3-93 shows the intensity and phase distributions of the GBM-approximated object beam. The intensity distribution exhibits sharp ringing reflecting the sudden underlying phase modulation imparted on the object beam by the transparent obstruction. Ideally no structure due to the phase modulation should be evident in the intensity of course so our GBM-approximation of the object beam is not completely ideal and a more accurate approximation would require more modes to be used. However for demonstrative purposes the current modal description is entirely adequate as there is a 99.98% correlation between the object beam intensity and its GBM-approximation.

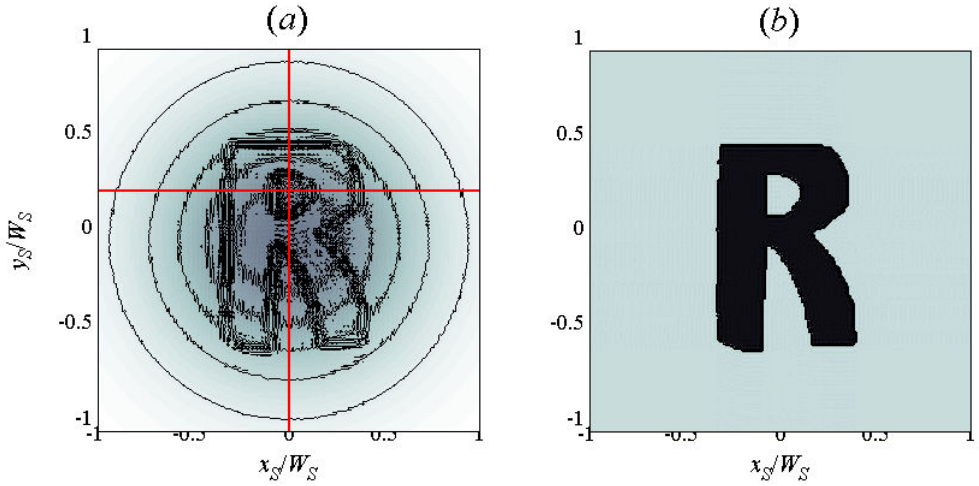


Figure 3-93. (*a*) Intensity and (*b*) phase distributions of the GBM-approximated object beam field.

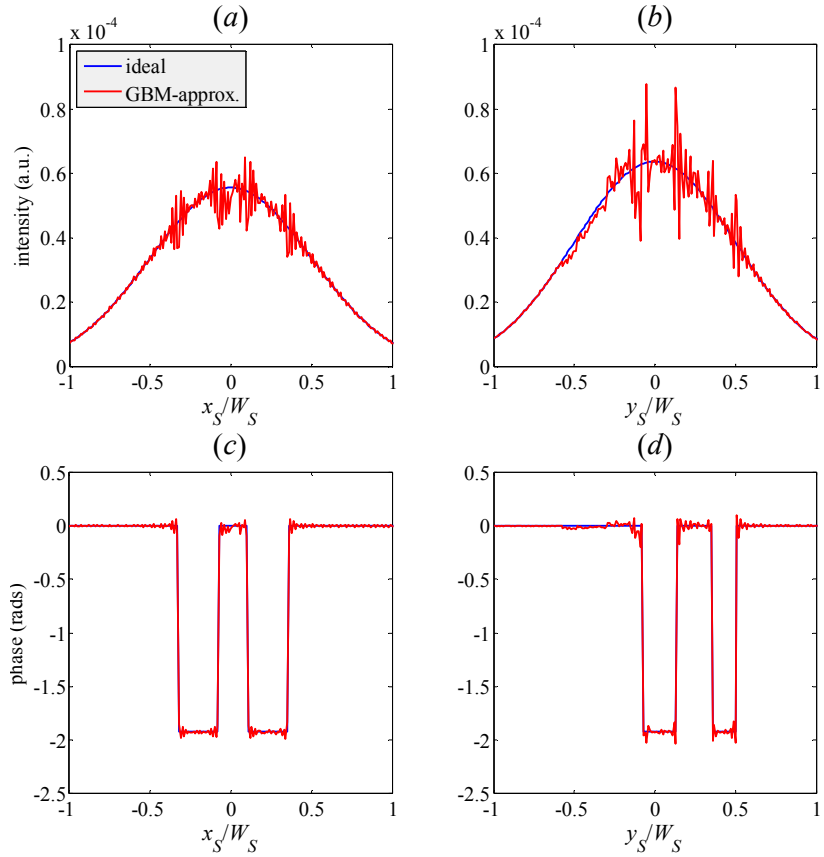


Figure 3-94. (*a*) x- and (*b*) y-cuts through (top) intensity and (bottom) phase of the ideal and GBM-approximated object beam field for the HDPE 'R'.

Figure 3-95 shows the simulated intensity formed at the output plane of an ideal system (i.e. without truncation effects included at mirrors M_2 and M_3) with and without the high-pass spatial filter included. There is 96.03% correlation between the beams at the object plane and output plane. When the spatial filter is included an image is formed with a clear outline at the location of images of the dislocations in phase.

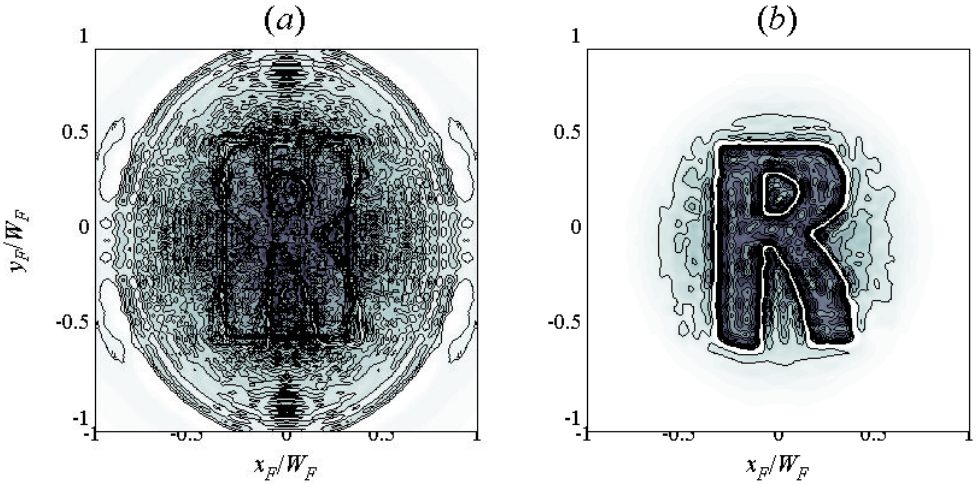


Figure 3-95. Output plane intensity patterns from the GBM simulation (*a*) without and (*b*) with the high-pass spatial filter included in the system. The additional structure in (*a*) compared to the reconstructed object beam in Figure 3-93(*a*) is explained further on.

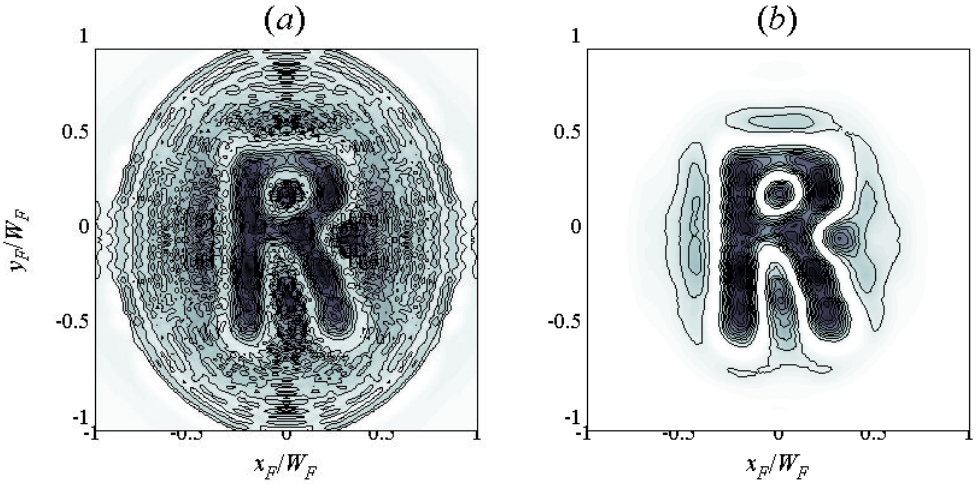


Figure 3-96. Negative (linear-scale) contour plots of output plane intensity pattern from the GBM simulation with truncation at mirrors M_2 and M_3 include (*a*) without and (*b*) with the high-pass spatial filter included in the system.

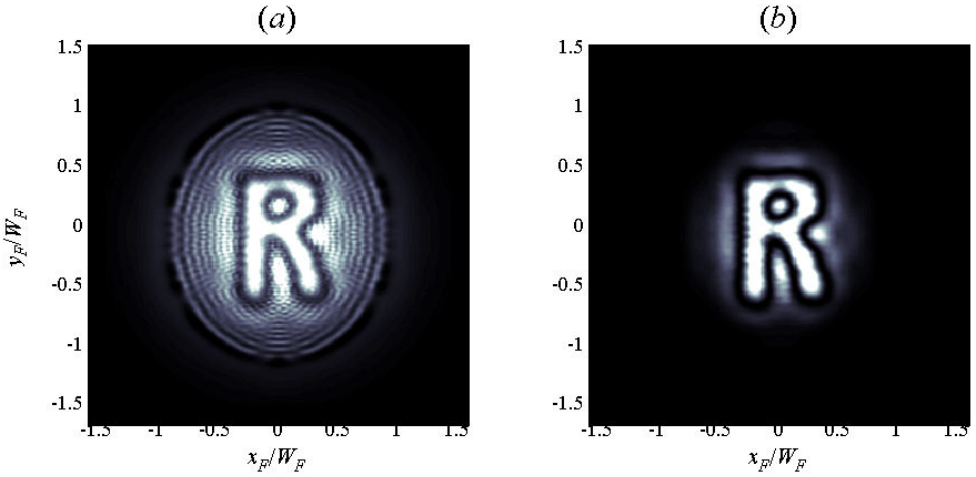


Figure 3-97. False-coloured plots of simulated output plane intensity using GBMA with truncation at mirrors M_2 and M_3 included (*a*) without and (*b*) with high-pass spatial filtering.

Once again the accuracy of the GBM model is improved when truncation at mirrors M_2 and M_3 is included, the results of which are shown using contour plots (Figure 3-96) and false-coloured plots (Figure 3-97) for comparison with measured patterns (Figure 3-92). Clearly the fact that the system reproduces an edge-enhanced image of the transparent object when the spatial filter is omitted is explained by the truncating effects of mirror M_2. As before, the combination of truncation at M_2 and at the high-pass filter is equivalent to a single band-pass filter, as shown by plots of the intensity pattern at the filter plane (Figure 3-98). Combining the truncation effects of the two elements into a single band-pass filtering operation reduces computational overhead since the mode coefficients need only not be evaluated at mirror M_2. Furthermore it would also allow the system to be analysed using Fourier methods as well. The single band-pass filter would consist of a small circular stop (representing truncation by the high-pass filter) of radius $a = 2(W_{SF})$, within a projection of the aperture function $p(x, y)$ representing the extent of mirror M_2. Since M_2 is illuminated with a collimated wavefront and we are interested in the field at the spatial filter plane, which is at the back focal plane of M_2, the projected aperture function $P(x_{SF}, y_{SF})$ at the spatial filter plane is equal to the aperture function $p(x, y)$ at M_2. The extent of one of the 350 mm focal length parabolic mirrors is represented by a circular aperture function of radius $r_2 = 142.37$ mm. Thus when truncation effects are included at mirror M_2, power at the spatial filter plane only resides in the geometric shadow of the aperture of M_2, i.e. within a circle of radius r_2, as shown in Figure 3-98

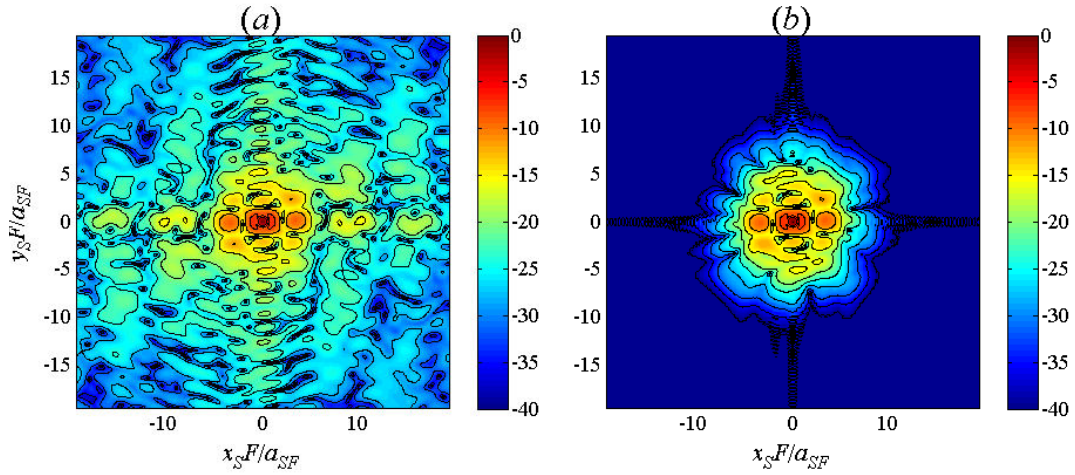


Figure 3-98. Beam pattern intensity at the spatial filter plane, i.e. the spectrum of the object beam (*a*) without and (*b*) with truncation effects at mirror M_2 included.

In simulations of the optical system using GBMA, for the computation of mode coefficients the spatial filter plane must be treated as having finite extent. The size of the spatial filter plane was arbitrarily defined as having a width and height equal to three times the diameter of mirrors M_2 and M_3. However, depending on the object, some proportion of scatted power may exist outside this area. Ideally the extent of a plane should be large enough to contain all features of the highest-order mode since clearly all power in a GBM-approximated field is confined to lie within an area of that size. However in order to maintain reasonable resolution at the spatial filter plane and simultaneously keep array sizes to a minimum (for computational efficiency), a smaller area was chosen. We now use Fourier transforms to examine how our particular choice of the border of the finite spatial filter plane affected the GBM results obtained.

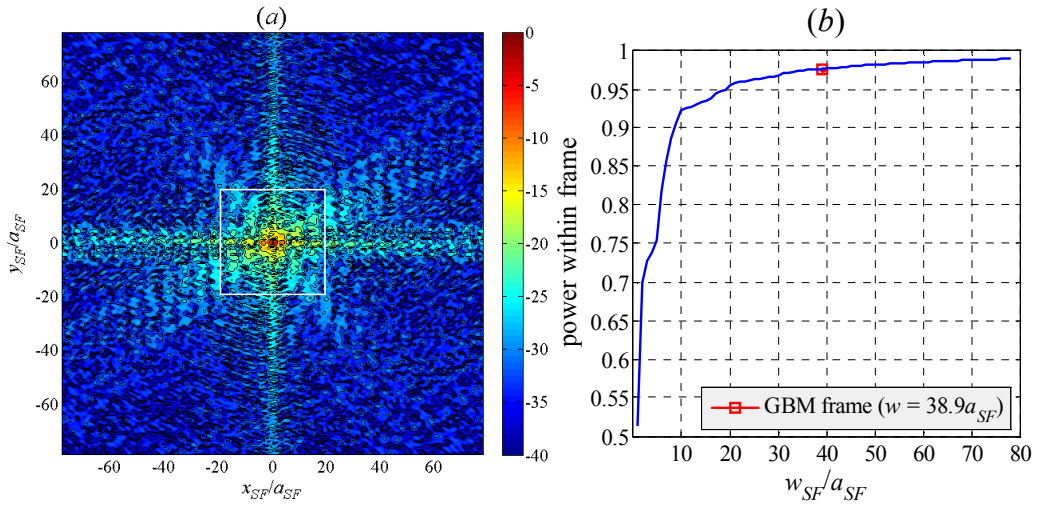


Figure 3-99. (*a*) (Log-scaled) plot of spectrum intensity of the object beam (illuminated HDPE 'R') calculated using Fourier transforms. (*b*) Power within square frames at the spatial filter plane as a function of frame width w_{SF}. Approximately 97.6% of spectrum power is contained within the spatial filter frame used in GBM simulations – represented by the white square in (*a*).

Figure 3-99(*a*) shows the spectrum intensity of the object beam that was calculated using Fourier transforms, i.e. the intensity distribution at the spatial filter plane. The white square superimposed on the spectrum intensity represents the extent of the spatial filter plane used in GBM simulations, whose width is $w_{SF} = 38.9a_{SF}$, where the high-pass spatial filter radius $a_{SF} = 2W_{SF}$. Clearly some power exists outside this boundary. It was estimated that most (approximately 97.6%) of power in the spectrum is contained in this region, as shown in Figure 3-99(*b*). Thus the finite area of the spatial filter plane used in GBM simulations should have little impact on the simulated output plane image.

Fourier transforming the spectrum of the object beam produces an image of the object beam (at the output plane). Figure 3-100 shows the image obtained with and without truncation of the Fourier spectrum with the square boundary used in GBM simulations. When the finite-sized edge is used some of the high spatial frequencies associated with the phase discontinuities at the edges of the object are filtered out, which results in intensity nulls at corresponding positions in the output plane image, which explains the appearance of additional structure observed in the GBMA simulation of the unfiltered output plane intensity (Figure 3-95). Thus despite there being only a small amount of power lost from the spectrum due to the truncating effect of using a finite spatial filter plane, the high spatial frequency components that are lost are clearly necessary to reproduce an image of the object beam.

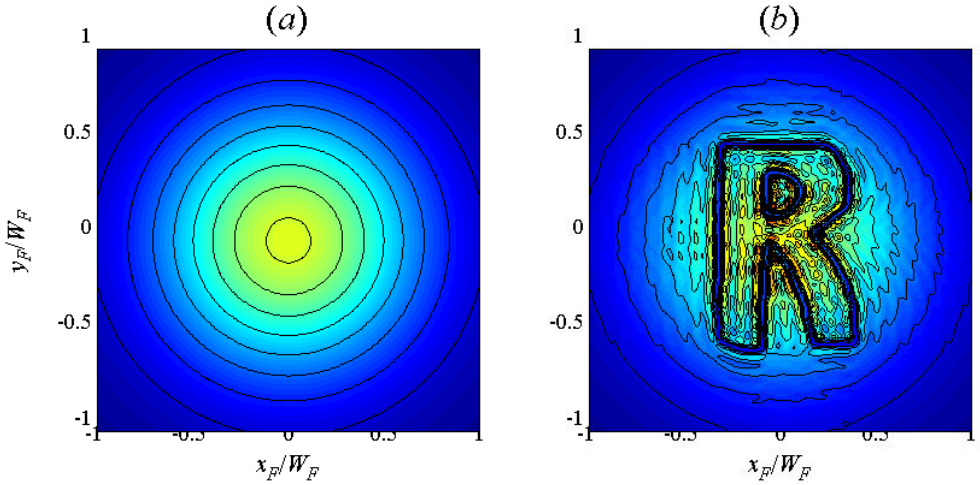


Figure 3-100. Fourier transform intensity of the object beam spectrum (*a*) without and (*b*) with truncation of the object spectrum beyond the edges of the spatial filter plane used in GBM simulations.

In conclusion whereas Fourier transforms can be easily applied to model propagation through an ideal optical system with no truncation effects, a Gaussian beam mode approach will inherently include truncation effects and extra care must be taken (by defining planes that are large enough to include all frequency components) if one is to use GBMA as an alternative to Fourier transforms. This inherent difference is illustrated in Figure 3-101, which shows the output intensity and phase distributions that result from taking a Fourier transform of the Fourier spectrum in Figure 3-99(*a*) after it has been passed through the high-pass spatial filter. Clearly the resulting distributions are highly unrealistic and in no way agree with the experimentally obtained intensity image or with the GBM-simulated results. However, only a slight modification of the Fourier transform model is required to include truncation effects (for this system at least) by

simply including a low-pass spatial filter representing the truncating aperture function of mirror M_2. The resulting output plane intensity distributions shown in Figure 3-102 is in good agreement with both experimental and GBM simulated results.

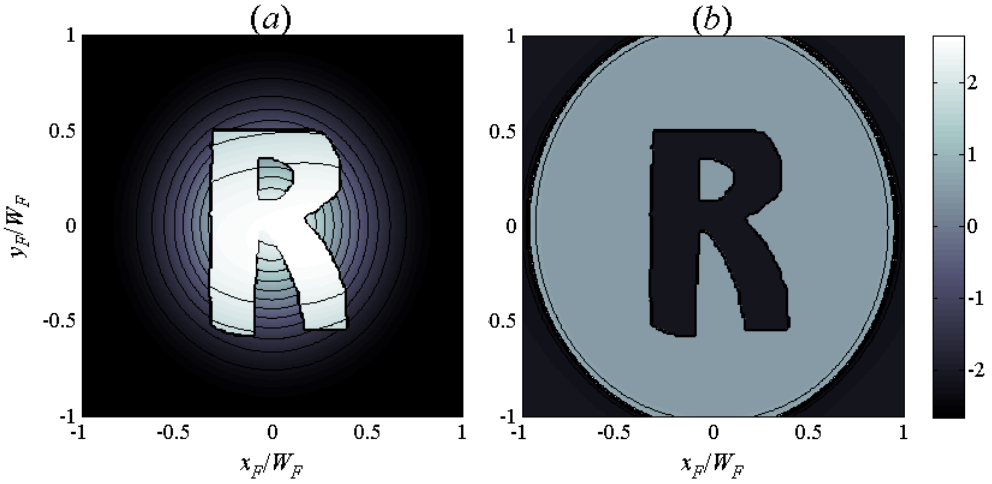


Figure 3-101. Simulated output plane (*a*) intensity and (*b*) phase distributions (when high-pass spatial filtering is included) computed by taking the Fourier transform of the object spectrum shown in Figure 3-99(*a*).

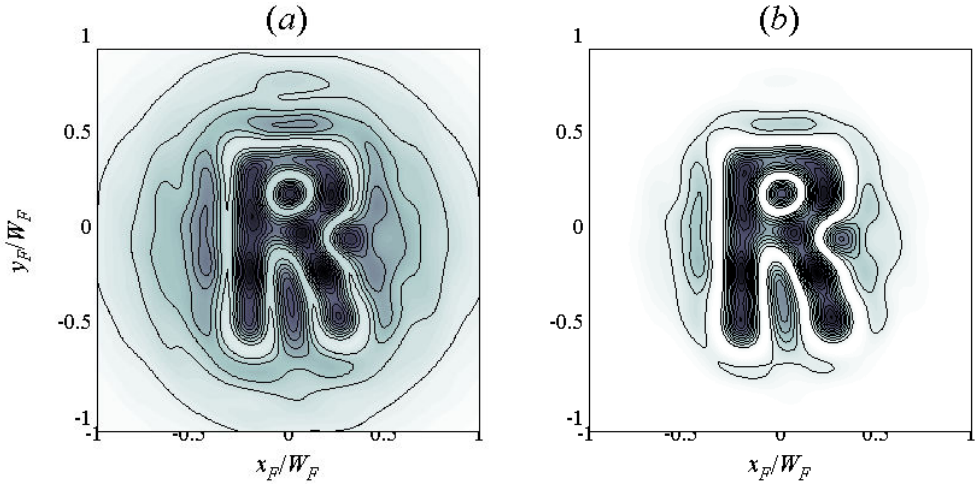


Figure 3-102. Output plane intensity pattern calculated using Fourier transforms with (*a*) a low-pass spatial filter and (*b*) a band-pass filter at the back focal plane of mirror M_2.

In this section Gaussian beam mode analysis was used to accurately simulate beam patterns that have quite complicated structure. Typically GBMA is used for analysis of beam patterns with much simpler profiles. In those situations it proves to be a computationally efficient tool since accurate beam analysis can be achieved using only a small number of modes. However, as has been shown here, for more complicated beam profiles this advantage no longer exists and gives way to increased computational overhead (higher execution time and memory requirements). Thus Fourier techniques may be more suited to the efficient analysis and propagation of complex beam patterns.

3.4 Reflection-Mode Imaging Experiments

As the previous experiments demonstrated, millimetre wavelength radiation can only penetrate a short distance through water-laden samples, so transmission mode imaging may only find use in imaging thinly sliced samples, for example in *ex vivo* histological examinations. To pursue *in vivo* imaging one must resort to measuring the radiation that is reflected or scattered from near-surface layers of the object under test.

3.4.1 Near-Field Reflection Imaging

The reflection mode imaging experiments were performed using a near-field arrangement (Figure 3-103). A small area on the object directly in front of the source (transmitter Tx) is illuminated with coherent radiation and the reflected radiation intensity is measured using a detector (receiver Rx). Whereas in transmission mode, image acquisition was performed by raster-scanning the detector, in reflection mode both source and detector were fixed in position and the object raster-scanned (Figure 3-104). The measured magnitude of reflected radiation gives structural information of an illuminated object. For objects with reflective or absorptive surfaces the intensity of reflected radiation can be used to provide a map of object surface. However if samples are partly transparent, clearly reflected radiation includes that which is reflected and scattered from different depths within the object as well as surface reflections. The challenge is then to interpret the image and only by examining and mapping known objects can knowledge be gained about how to undertake such a procedure.

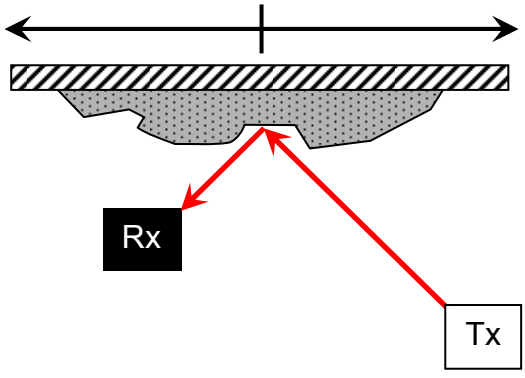


Figure 3-103. Schematic of the near-field reflection imaging set-up. The test object (grey region) is mounted onto the positioning table of the X-Y raster scanner. At any point in the scan only a small area on the object in sight of the source (or transmitter Tx) is illuminated. Radiation reflected from that area on the object is then collected by the detector (or receiver Rx).

The advantage of the near-field reflection set-up shown in Figure 3-103 over its transmission counterpart is that object illumination is independent of optical components. In transmission mode the maximum size of test object that could be imaged with adequate contrast was determined by the parameters of the mirror that collimated the source beam. On the other hand the reflection set-up (Figure 3-103) uniformly illuminates the entire scene. Thus apart from the limitations imposed by the maximum travel permitted by a positioning system there is no upper limit on object size that can be imaged. Furthermore image contrast is invariant with position.

In all experiments the test object was mounted onto a piece of absorbing material that was attached to the scanning table of the X-Y scanner. The absorbing material was needed to prevent radiation reflected from the mount from entering the detector. If this background radiation (from the mount) is not absorbed the resulting image will have poor contrast between foreground (object) and background signals.

3.4.2 Experimental Arrangement

Several variations of the near-field reflection configuration were tested to find which would yield best image quality, i.e. highest resolution and least confusion. In the first set-up, shown in Figure 3-104 both the source and detector were fed with corrugated conical horn antennas via short rectangular metallic waveguide sections. The source and detector were set at different distances from the object plane to reduce cross-talk between the two, with the source placed furthest from the object plane to minimise diffraction effects (between illuminated object and the detector).

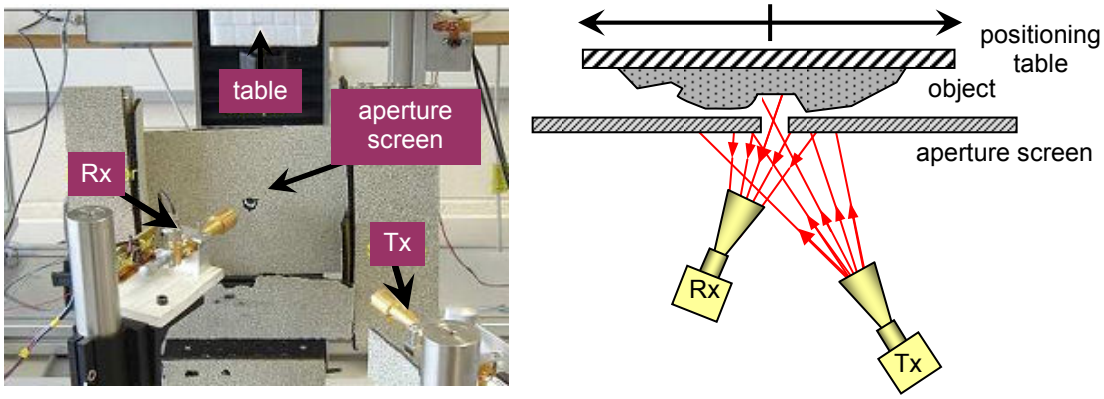


Figure 3-104. Initial near-field reflection imaging arrangement. The source Tx and detector Rx are fed by corrugated conical horn antennas. Over illumination of the object was reduced by including an opaque screen (of absorbing material) with a small circular aperture at its centre between the object and the source-detector combination.

Initial images obtained with this version of the near-field set-up had poor resolution, as would be expected, for two reasons. Firstly images were blurred because the signal received by the detector from the illuminated point on the object was convolved with the point-spread function of the horn antenna – a Gaussian beam. Secondly, although the horn antennas are designed to be directional the illuminating beam is still Gaussian in profile so instead of each scanned point on the object being illuminated with a pencil-like beam, rather illumination is provided by a spreading Gaussian beam. The resulting over-illumination of each point means that although the detector may be aimed at a single point on the object it receives the radiation reflected from the entire area illuminated by the Gaussian beam from the source horn.

In an attempt to resolve this particular problem an opaque screen with a small aperture at its centre was positioned vertically between the source-detector combination and the object plane in order to reduce object illumination to only the area directly behind the aperture (Figure 3-104). With the aperture screen in place resolution is thus determined by the aperture diameter, which was 10 mm. While the aperture does restrict the detectors field of view of the object of course any radiation reflected from the screen itself is also coupled to the detector. The aperture screen was thus made from a sheet of Perspex with absorbing material attached to the side facing the source. It was later discovered during tests made of various absorbing materials [3.22] that the particular absorbing material used (Eccosorb) is not optimised for use at 100 GHz and in fact produces non-negligible reflection at this frequency. Thus unwanted reflections from the over-illuminated screen caused a constant-level signal to be superimposed on the object signal causing standing wave effects. Furthermore the inclusion of the aperture does not solve the problem of image blur, since resolution is now ~10 mm.

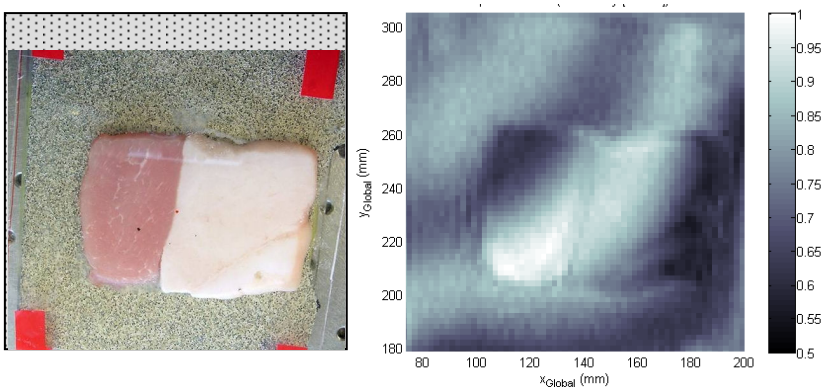


Figure 3-105. Single near-field reflection image of strips of pork fat and flesh obtained using the near-field reflection set-up in which the source and detector are fed by corrugated conical horn antennas.

The measurement of intensity reflected from strips of pork fat and flesh laid side-by-side made using this arrangement is shown in Figure 3-105. The pork samples were secured to the positioning table with cling-film. Outlines of the pieces of pork fat and flesh can be clearly seen and low intensity is observed along the boundary between fat and flesh. Notice however, the appearance of bright and dark intensity fringes across the image in Figure 3-105.

To investigate these fringes further a measurement was made of a much simpler object: a thin square (60mm × 60mm) aluminium plate mounted onto the absorber to minimise radiation reflected back into the detector from the region surrounding the plate. The measured intensity from the metal plate (Figure 3-106) shows a noticeable intensity variation across the plate with the maximum and minimum intensity levels occurring at the lower-right and upper-left corners of the plate, respectively, which indicates that the metal plate is not parallel to the scanning plane.

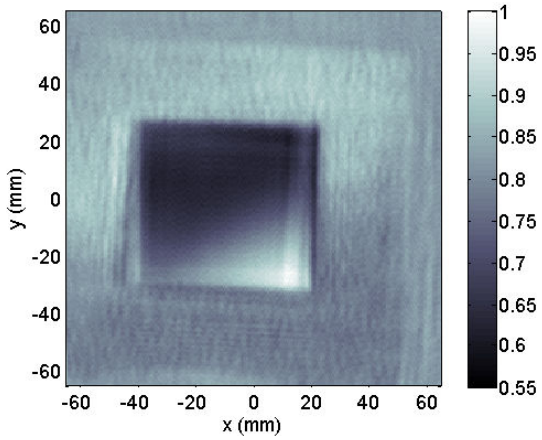


Figure 3-106. Image of near-field reflection reflected intensity from a square (60mm × 60mm) aluminium plate. The intensity variation across the metal plate indicates a tilt with respect to the scanning plane.

An important observation to be made from Figure 3-106 is that the minimum measured intensity that occurs near the upper-left corner of the metal plate is lower in value than the intensity surrounding the plate. However we would expect that because the plate was mounted on an absorbing material the background intensity should be lower in value than anywhere in the region occupied by the metal plate. This peculiarity suggests that the system is subject to standing wave effects.

3.4.3 Standing Wave Effects

Standing waves are a problem inherent in any coherent quasi-optical system. They can occur between a source and the reflective surfaces of subsequent optical components when a beam of radiation is reflected or partially reflected along the path of propagation. Understanding standing wave effects is a major focus of ongoing work by the THz Optics Group at NUI Maynooth [3.18, 3.19, 3.20, 3.21].

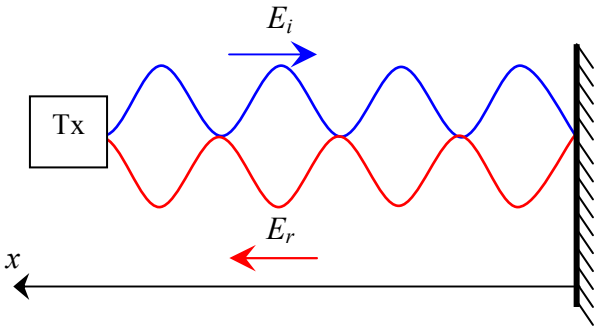


Figure 3-107. Formation of a perfect one-dimensional standing wave between a source Tx and the reflective surface to its right. Nodes occur at points along x where incident and reflected waves interfere destructively to cancel. Antinodes occur halfway between nodes, where incident and reflected waves interfere constructively to produce maximum amplitude of $2E_0$, where $|E_i| = |E_r| = E_0$.

In two dimensions the presence of standing waves result in the appearance of a series of destructive (dark) and constructive (bright) interference fringes. If a screen is placed at the source dark (bright) fringes will occur at points where the screen intersects nodes (antinodes) in the standing wave (Figure 3-107). Successive nodes are separated by a half wavelength, as are successive antinodes, while the separation between a node and the next anti-node is a quarter wavelength. The separation between adjacent bright and dark fringes thus corresponds to a change in reflective surface distance (or height) of a quarter wavelength. It may therefore be possible to infer from the positions of bright and dark standing wave fringes useful information on the local surface profile of the reflecting object.

To establish whether standing waves were indeed responsible for the bright and dark fringe patterns observed in near-field reflection images another measurement of the intensity reflected from the square aluminium plate was made after the plate had been moved slightly further from the aperture screen. If the bright and dark fringes seen in Figure 3-106 are due to standing waves then a change in object-source distance will cause a shift in the fringe pattern. The object-source distance was increased by a quarter-wavelength and the recorded image, shown in Figure 3-108(*b*), has a similar

intensity variation across the plate, but with the positions of minimum and maximum intensity now reversed. The dependence of fringe pattern position on object-source distance thus verifies the presence of standing waves in the system.

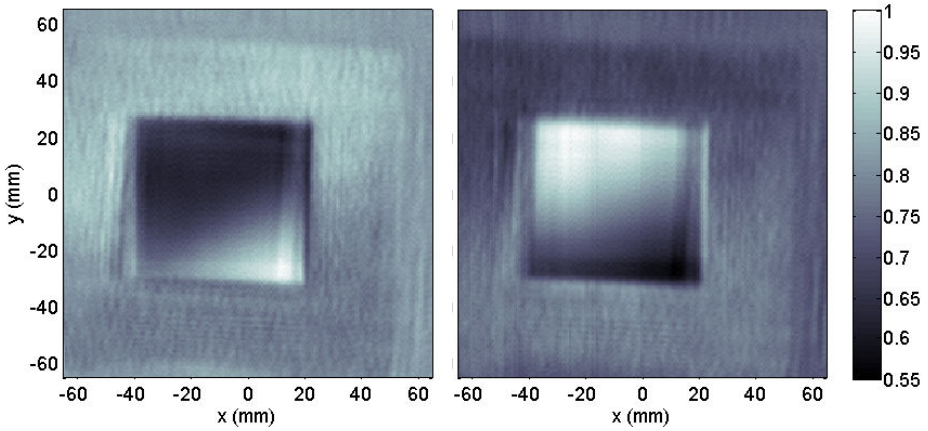


Figure 3-108. Two images of a square metal plate that were obtained for object-aperture distances of (*a*) z_A and (*b*) $z_B = z_A + \lambda/4$. Standing waves in the system mean that the positions of minimum and maximum intensity are dependent on the distance from the source to the object.

The second image was obtained with the object-aperture distance changed by a quarter-wavelength resulting in a reversal of positions of bright and dark interference fringes. The composite image in Figure 3-109 was constructed by adding the two images in Figure 3-108 together so that bright fringes of one cancel with dark fringes of the other. In practise since the two images are recorded with the object at slightly different distances from the source the recorded wavefront intensities will not be exactly equal due to diffraction effects so the fringe patterns may not cancel exactly (as is seen here).

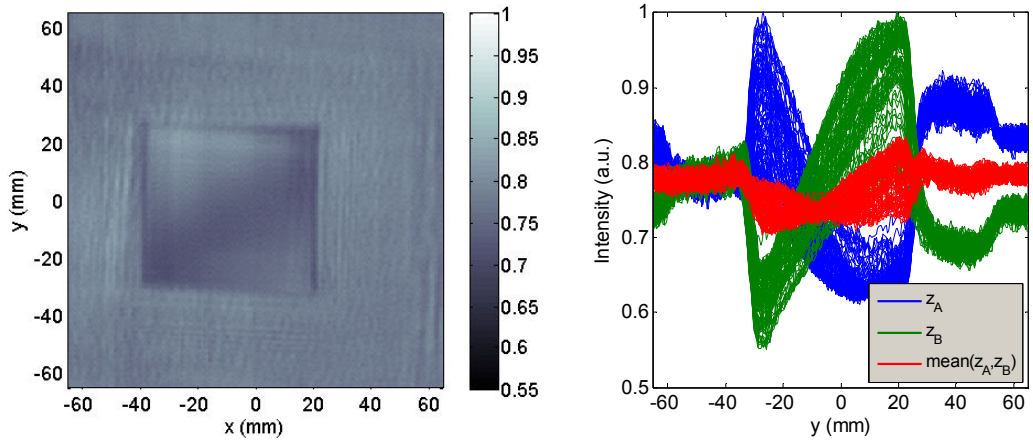


Figure 3-109. (*a*) composite image of the square metal plate obtained by summing images A and B from Figure 3-108 recorded with object-aperture distances of $z_A = (½)\lambda = 1.5$mm and $z_B = (¾)\lambda = 2.25$mm. (*b*) Horizontal cuts through the composite and individual images show cancelling of the z-dependent constructive and destructive interference minima and maxima.

A procedure to lessen the effects of standing waves on images was developed and routinely applied when making near-field reflection measurements. Four reflected intensity images were recorded with a different object-source distance used for each. Images I_A and I_C were recorded at object-source distances of $z_A = z_0$ and $z_C = z_0 + (½)\lambda$, while images I_B and I_D were recorded at object-aperture distances of $z_B = z_0 + (¼)\lambda$ and $z_D = z_B + (½)\lambda = z_0 + (¾)\lambda$. Any two images recorded at object-source distances that are separated by a half-wavelength will have matching constructive and destructive fringe patterns. A composite image in which the bright fringes from one image cancel with the dark fringes of the other can be formed from any two images recorded with object-source distances that are separated by a quarter-wavelength, i.e. from I_A and I_B, or I_B and I_C, or from I_C and I_D. Alternatively all four images can be used to create a composite image with a much reduced noise level as follows

$$I = ½ (I_A + I_C) + ½ (I_B + I_D)$$

This procedure of recording four reflected intensity images and creating a composite image was used for most of the near-field reflection measurements. A quarter-wavelength object-source distance separation of ~0.75 mm was needed for measurements made at 100 GHz.

Figure 3-110 shows the result of an experiment designed to further illustrate the effects of standing waves on reflection images. The test objects were two coins: a 1 Punt coin and a smaller 1 Euro coin – the latter being thicker than the former by approximately a quarter-wavelength at 100 GHz. Two measurements were made of the coins, with the object-source distance used in the two scans differing by a quarter-wavelength. In the first image obtained the intensity in the region occupied by the smaller coin is much lower than that across the larger coin, indicating that the surfaces of the smaller coin is a quarter-wavelength closer to the source than that of the larger coin. However in the second image, obtained after the two coins were moved a distance of $\lambda/4$ away from the aperture screen, the situation is reversed and the image of the larger coin appears brighter.

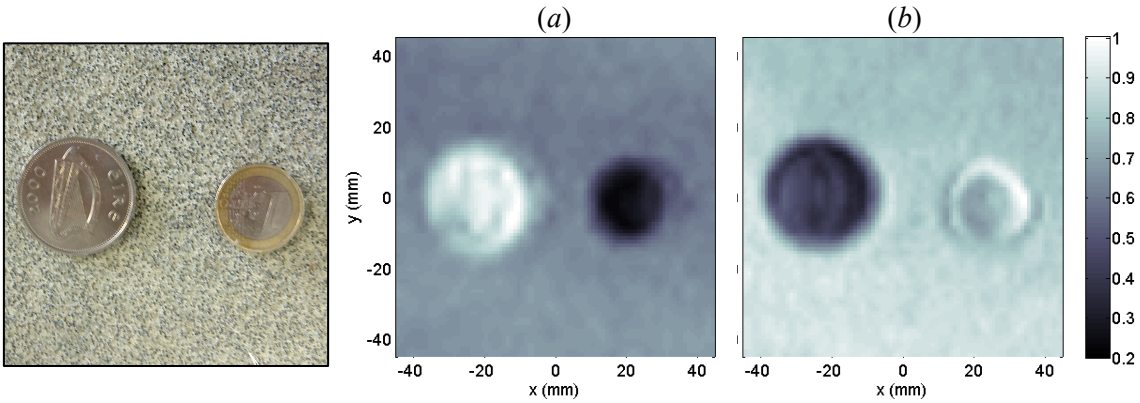


Figure 3-110. Measured near-field intensity reflected from two coins. Images (*a*) and (*b*) were recorded for object-source distances separated by a quarter-wavelength. The reversal in positions of bright and dark fringes across the two coins illustrates that the coins differ in thickness by about the same value, i.e. 0.75 mm.

3.4.4 Preliminary Near-Field Reflection Imaging Results

Some images of various test objects obtained with the early near-field reflection imaging system are now presented.

Figure 3-111 shows a composite image of strips of pork fat, flesh and skin. By combining images recorded at different distances the background intensity level is nearly constant and contains no bright and dark fringes. Figure 3-112 shows a composite image taken of a piece of pork in which a hole was cut to reveal the flesh below. An outline of the triangular shaped incision is visible in the measured intensity. Figure 3-113 shows two images taken from two sets of measurements of a single leaf: one set when the leaf was fresh and therefore had high water content and the other after the leaf had been allowed to dry out. The image of the leaf obtained when it was fresh shows bright and dark regions. These intensity variations may correspond to regions with different water content, such as veins as they are absent in the image obtained when the leaf is dry. Figure 3-114 shows two images taken of a large leaf and a small leaf. Again the first image was made when the leaves were fresh and the second after the leaves had been allowed to dry out. The larger leaf is an ivy leaf and it appears very similar in both images in comparison to the smaller leaf which is barely visible in the second image indicating that it has lost significantly more water in the intervening period.

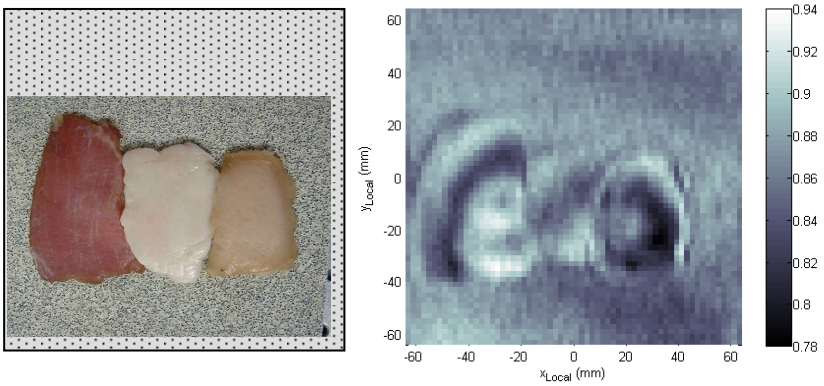


Figure 3-111. Composite near-field reflected intensity image of strips of pork fat, flesh and skin.

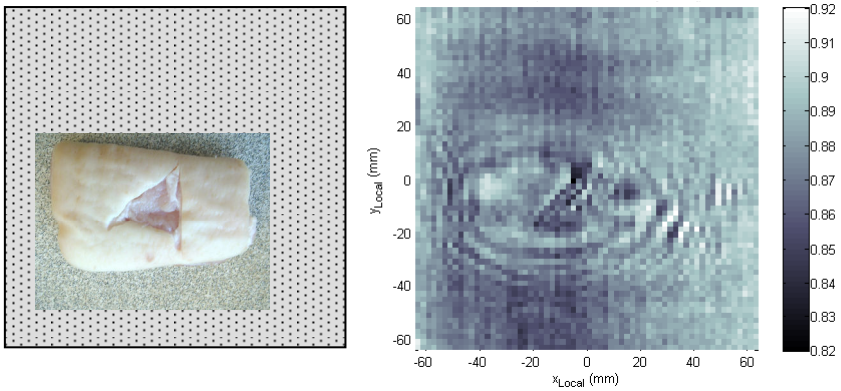


Figure 3-112. Composite image of pork with a triangular incision cut into the skin revealing flesh below.

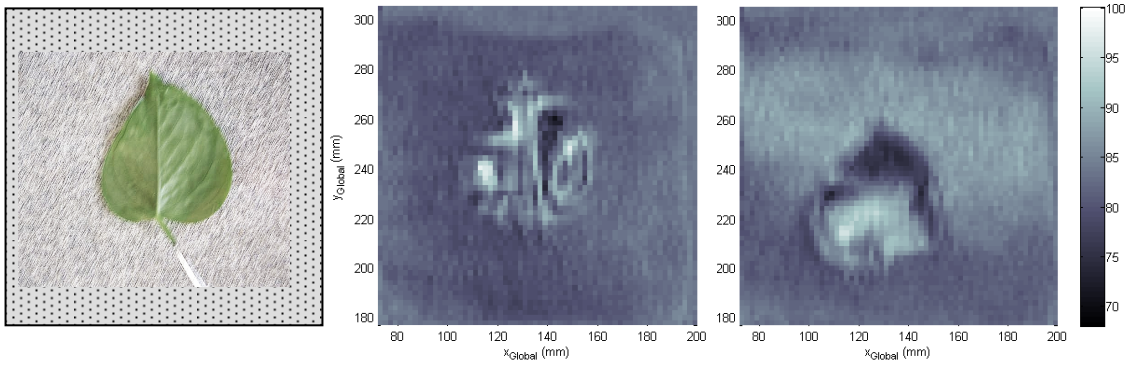


Figure 3-113. Composite images of a leaf when (left) it is fresh and (right) is dry.

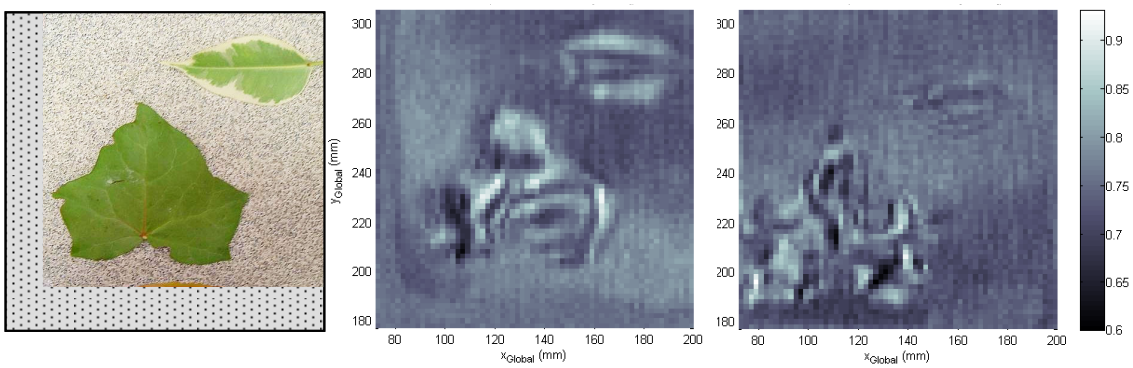


Figure 3-114. Composite images of two leaves obtained (left) before and (right) after being allowed to dry. Contrast between the small leaf and the background is much lower in the latter since more radiation penetrates the dry leaf and is reflected from the backing material (Eccosorb).

A feature common to all of the near-field reflection images seen so far is the low contrast between foreground objects and the background material on which they sit. In these early experiments Eccosorb was used to absorb unwanted background reflections. However this type of absorber is not designed for operation at 100 GHz, but rather is optimised for use at 40 GHz. Indeed at 100 GHz Eccosorb was found to be quite reflective. Use was made of the near-field reflection imaging system to experimentally investigate the reflective properties of a number of different absorbing materials at 100 GHz [3.22]. Of the different absorbers that were tested a pyramidal tessellating THz Radar Absorbing Material (RAM) tile from Thomas Keating Instruments[5] was found to offer the best performance at 100 GHz. At this frequency the TK RAM has a quoted reflectivity of -40 dB at normal incidence. A square THz RAM tile, like that shown in Figure 3-115, was used in all subsequent experiments to absorb unwanted background reflections.

Figure 3-115. One of the 10cm×10cm pyramidal tessellating Thomas Keating Radar Absorbing Material (RAM) tiles that was used as an absorber in later near-field reflection experiments.

A measurement of the square metal plate was made after the TK RAM was attached to the positioning table. The resulting image (Figure 3-116) registers a much weaker background signal than before, thus yielding greater image contrast between foreground and background signals.

[5] Thomas Keating Ltd., UK (www.terahertz.co.uk)

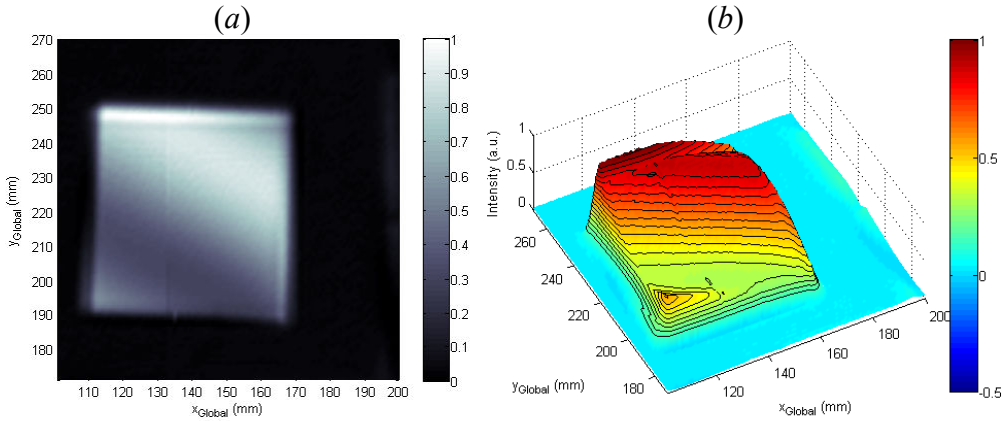


Figure 3-116. Measured reflected intensity from a square aluminium plate obtained after the TK RAM was used in place of Eccosorb as a backing material to absorb unwanted background reflections.

With the new absorber in place measurements were made of various objects. The results of a measurement made of bacon samples (fat and flesh) are shown in Figure 3-117. The image contrast is vastly improved compared to the images obtained when Eccosorb was used as an absorber. The return signal from the piece of bacon flesh is much higher than that from the fatty tissue, since the incident radiation penetrates more easily the latter before it is absorbed. In fact the signal level reflected from the fatty tissue was so weak that it was necessary to increase detector sensitivity to the extent that the signal reflected from the bacon flesh was saturated.

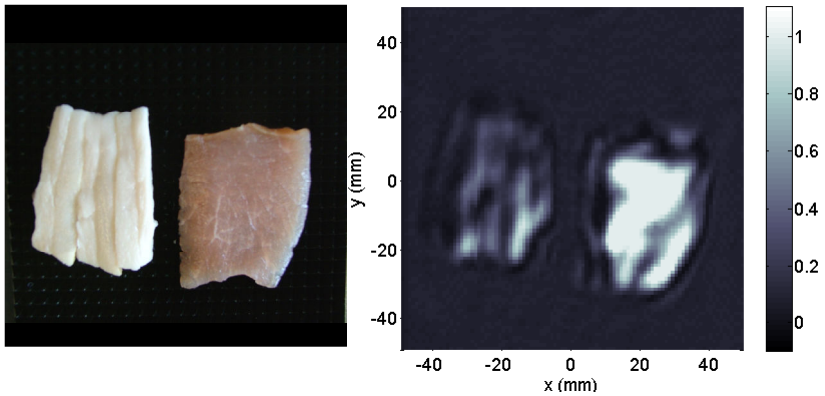


Figure 3-117. Measured near-field reflected intensity from samples of bacon fat and flesh that was made after the TK RAM absorber had replaced the Eccosorb absorber.

3.4.5 Improved Near-Field Reflection Imaging Results

The main problem with the set-up used thus far was that of image blur. This and other problems encountered with the first version of the near-field reflection set-up were addressed by simply removing the horn antennas from both the source and detector –

see Figure 3-118. Firstly, image blur was greatly reduced since the point-spread function of a bare waveguide is much smaller than that of a horn antenna. The problem of over-illumination was also solved since, although a horn antenna produces a much more directional beam than that from a bare waveguide, its absence meant that the source-detector combination could be positioned much closer to the object plane. The reduced object-source distance meant that the source could illuminate a much smaller area and the detectors field of view would also be reduced. This modified set-up made the aperture screen redundant so it was removed, which would also have eliminated any unwanted reflections from it. The source and detector were positioned adjacent to each other to reduce diffraction effects. Further improvement might be achieved if the source and detector were properly isolated from each other, perhaps by inserting an absorbing or reflecting plate between the two to reduce cross-talk.

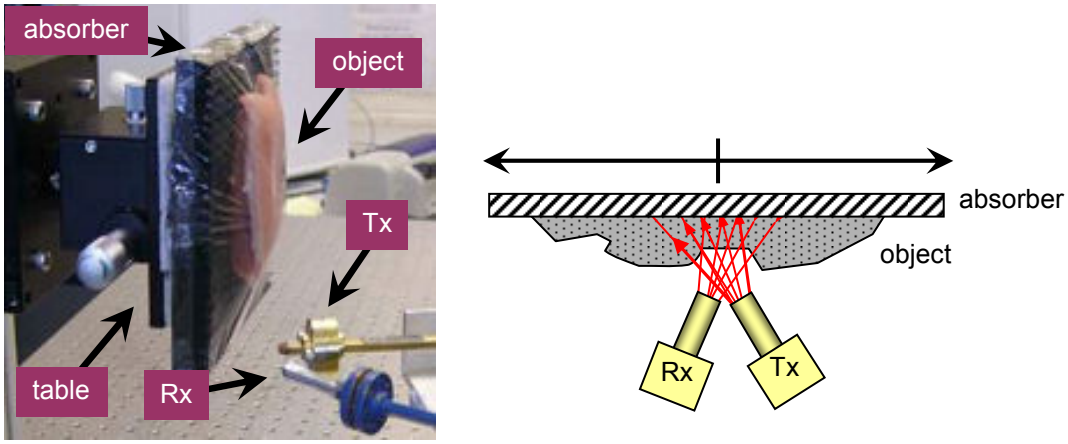


Figure 3-118. Modified near-field reflection imaging arrangement. Source and detector are fed by bare waveguides only and are placed side by side. Aperture screen has been removed. The test object is mounted on absorbing material to reduce unwanted background reflections.

A log-scale plot of measured intensity reflected from a leaf is shown in Figure 3-119. Much more detail of the leaf's structure is revealed with the greater resolution offered by the smaller aperture of the bare waveguide that is now used to feed the detector in the near-field reflection imaging system. Notice the regions of high intensity that correspond to positions of primary and secondary veins of the leaf, as we would expect to see since these have a higher water content than the rest of the leaf, which reflects the incident radiation back into the detector.

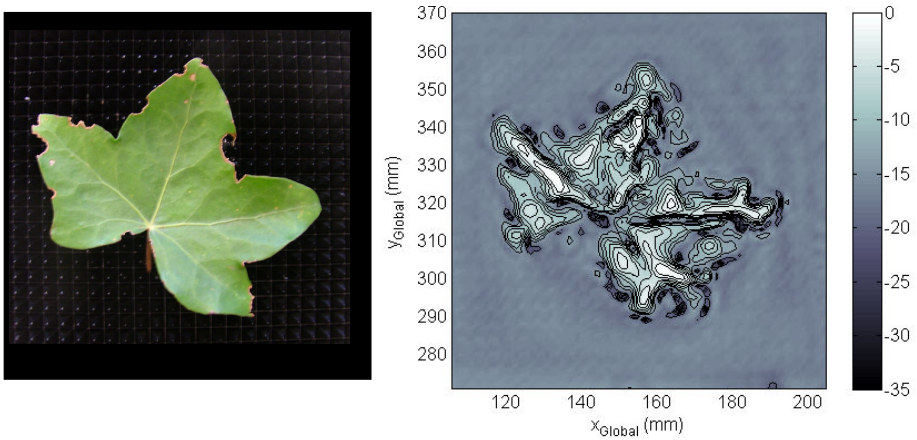


Figure 3-119. Log-scaled intensity reflected from a leaf recorded using the set-up in which both source and detector were fed with bare waveguides. Colour-axis is scaled to emphasise the foreground signal.

As was seen in images of the metal plate because standing wave effects are present the foreground signal can contain intensity levels with values less than the background signal level. In order to isolate the foreground and background signals we must identify those image pixels with intensity values close to the background intensity level. The remaining pixels correspond to the foreground signal. The contour-type plot shown in Figure 3-119 was created by excluding all contours at intensity levels immediately above and below the background signal level. We can also effectively remove the foreground signal from the image by scaling the log-scale intensity plot so that its limits are equal to the minimum and maximum intensity values of the background signal.

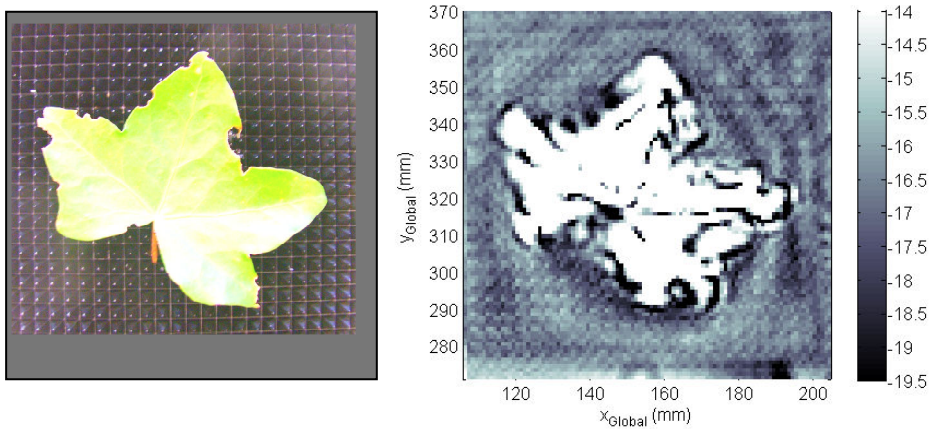


Figure 3-120. Log-scaled plot of reflected intensity from a leaf. The colour-axis has been scaled so as to emphasise the background signal, which reveals the regular structure of the pyramidal TK RAM tile.

The log-scale intensity plot shown in Figure 3-120 was created by scaling the colour-axis so that it spans the range -19.5 to -14 dB, within which the background signal lies. By scaling the measured intensity in this way the structure of the intensity reflected from the background is revealed. In this case, where the leaf was mounted on the TK

RAM tile, the scaled intensity image reveals the pyramidal structure of the absorbing tile itself.

Next the improved near-field reflection imaging system was used to make measurements of various samples of bacon and pork slices, which are shown in Figure 3-121 to Figure 3-126. In each of these figures the amplitude of the measured signal intensity is plotted to reveal low-level features. In most cases only a single image was recorded and hence standing wave effects occur.

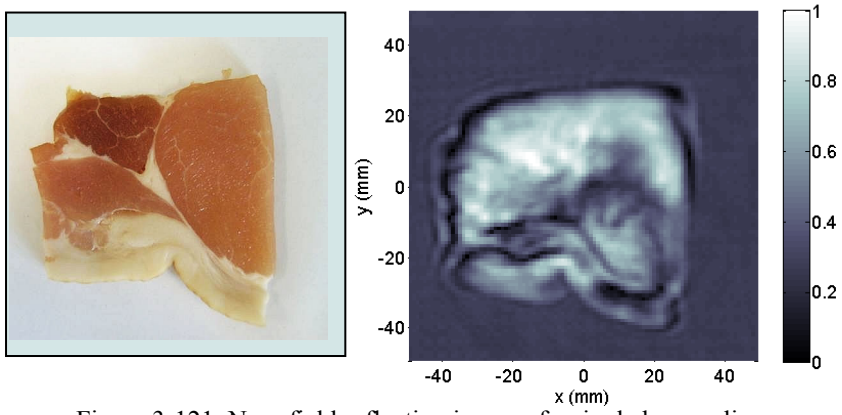


Figure 3-121. Near-field reflection image of a single bacon slice.

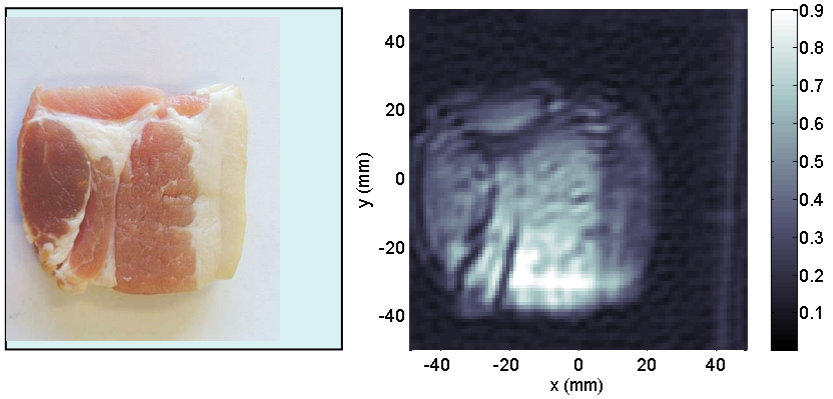


Figure 3-122. Near-field reflection image of three bacon slices laid one on top of the other.

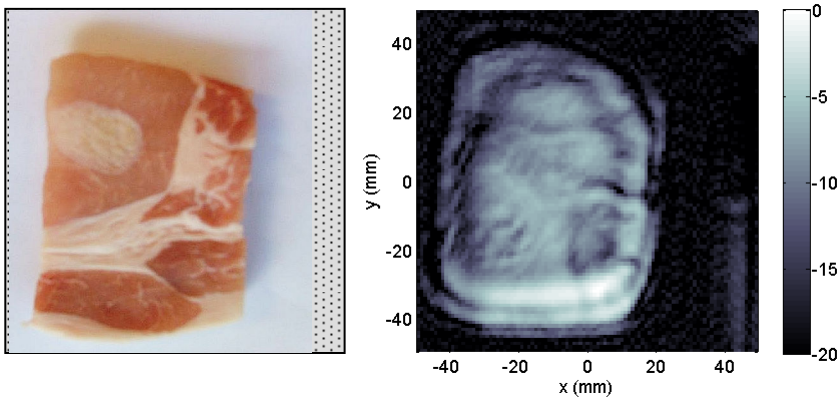


Figure 3-123. Near-field reflection image of a piece of pork with a small burnt area.

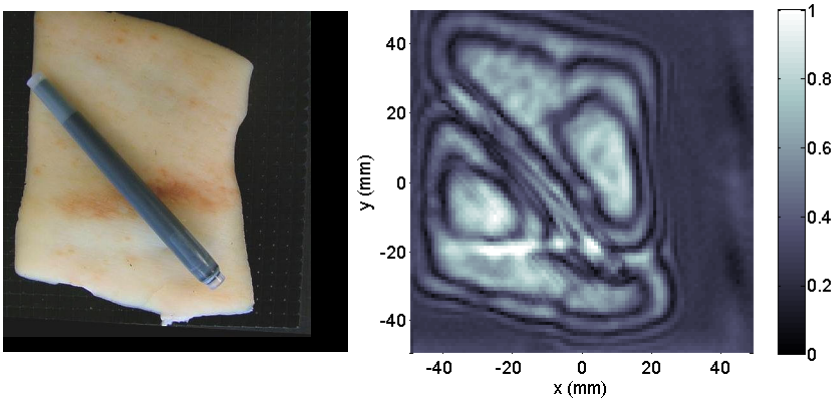


Figure 3-124. Near-field reflection image of a piece of pork with a fountain pen placed on top.

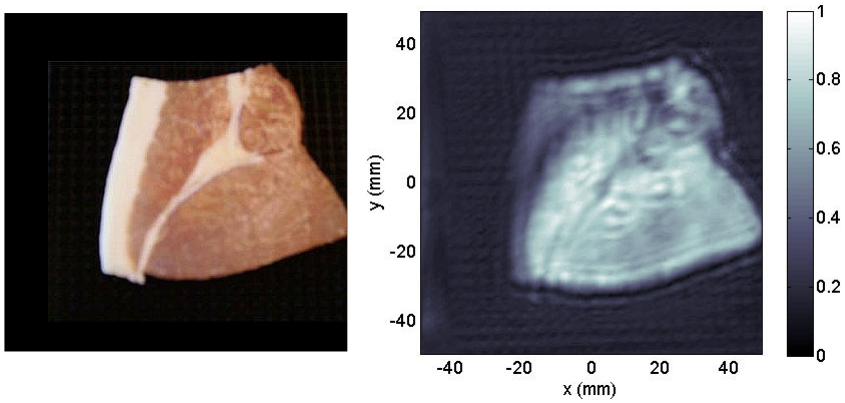


Figure 3-125. Near-field reflection (linear-scale amplitude) image of a single bacon slice, composed from four measurements made at object-source distances that were separated by a quarter-wavelength.

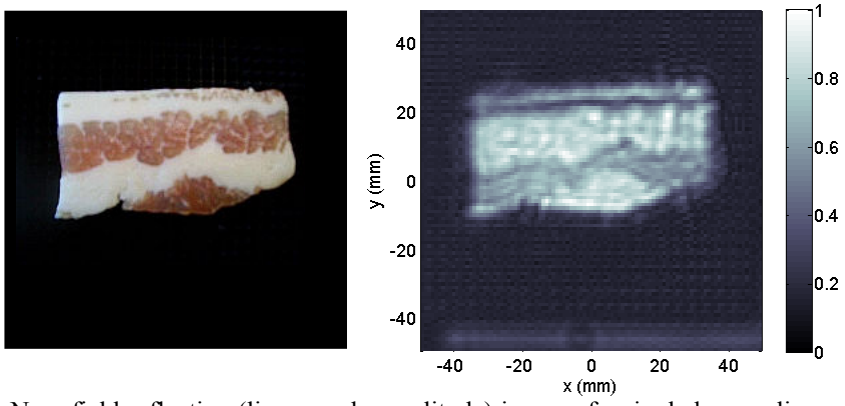


Figure 3-126. Near-field reflection (linear-scale amplitude) image of a single bacon slice composed from four measurements made at object-source distances separated by a quarter-wavelength.

An attractive property of mm-wave and THz radiation is its ability to penetrate fabrics – a property which has driven research into the development of THz imaging capabilities for security screening applications. Another potential application is in medical imaging as it would permit examination of tissue through wound dressings, thereby eliminating the need to expose damaged tissue and risk infection. To simulate such a situation an incision was made in a piece of pork, exposing muscle tissue beneath the skin. The

reflected intensity from the pork sample before and after an incision was made and with and without dressings (approximately 5mm thick) was measured (Figure 3-127). Apart from changes in fringe patterns positions there is very little difference between the images obtained with and without dressing in place.

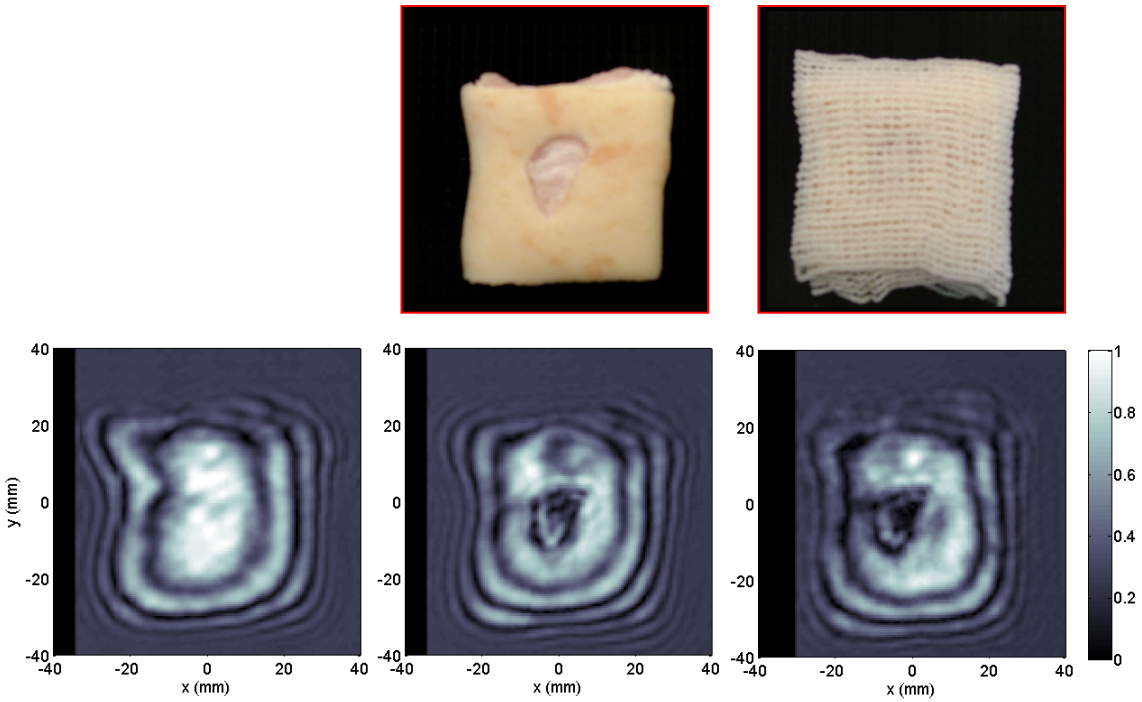


Figure 3-127. Simulated wound analysis. The lower row contains linear-scale plots of reflected signal amplitude (left) from the pork before an incision was made, (centre) after the incision was made and (right) through dressings.

3.5 Chapter Conclusions

The near-field transmission imaging experiments reported here were performed by illuminating the object under test with a collimated beam. An alternative test arrangement in which the illuminating beam is focused to a beam waist at the object surface may provide superior results. Firstly image contrast would be constant across the test object and secondly the acquired images would exhibit higher dynamic range since the intensity from a quasi-focused beam is greater than the intensity at any position within a wide collimated beam.

3.5.1. The Influence of Water in Samples on Imaging

Transmission imaging experiments of dry, non-biological samples yielded good contrast between internal structures in the object, e.g. between regions with different or varying

refractive indices. However, as the results of transmission imaging experiments in §3.3 show, the presence of even small amounts of water in a sample results in significant absorption and reflection of incident radiation. Transmission imaging experiments of biological samples were thus limited in usefulness by high water content. Thus imaging of samples with high water content is useful in as far as an outline of the image is produced. However information on the internal structure cannot be acquired, the resulting image being closer to a binary map delineating between regions of high/low water content.

This is not to say that obtaining high quality transmission images with THz radiation is not possible. Many good examples are to be found in the literature, however usually these were made under very carefully controlled conditions. Transmission imaging can be applied successfully provided a sufficient amount of radiation passes through the sample without significant absorption and/or reflection, i.e. we require appropriately low water content. Transmission imaging of freeze-fried samples (a process which can reduce residual humidity in the sample to between 1% and 4%) have been reported [3.23] in which images of biological samples with much higher image contrast than was possible from either fresh or frozen samples were acquired. However, as well as being costly and time consuming – after freezing, heat must be added slowly in order to avoid melting or structural deformations in the sample – the freeze drying process cannot differentiate between water and other chemicals capable of sublimation and these are also removed thus changing the nature of the test object [3.24].

Besides freeze-drying, the simplest way to reduce water content in the optical path of a propagating beam is to reduce the optical path length by reducing sample thickness sufficiently. Practical applications of transmission-mode imaging with THz radiation may thus be limited to use as a complimentary imaging modality in histological examinations. However even in the experiments reported here that were performed on narrow (a few millimetres) samples radiation was unable to penetrate the object and reach the detector. As far as the application of THz imaging techniques to samples *in vivo* is concerned, due to high water content THz imaging will be applicable to the examination of regions very close to (within a few millimetres of) the surface and then only in reflection mode – thus explaining the great interest in exploiting THz imaging techniques to identification and study of epithelial (skin) cancer as well as the possible applications in wound analysis (especially useful through dressings that are opaque in the visible). In reflection mode, water contained in samples cause incident

radiation to be reflected back to the detector, thus revealing the structure of the water-laden surface. In contrast objects or regions of an object with low water content allow radiation to transmit and scatter through the object. Provided the transmitted radiation is not permitted to return to the detector the resulting image shows low intensity, as shown for example in Figure 3-117, where the mainly transmitting piece of fat appears much darker than the absorbing/reflecting piece of lean meat.

In the transmission imaging experiments performed image resolution was limited by the finite detector size used, which resulted in blurring of the signal beam reaching the image plane. Attempts to perform image recovery, or restoration, with deconvolution were hindered by the fact that our scanning system was capable of recording only beam intensity, which meant that the estimated point spread function (PSF) necessary for deconvolution was inaccurate. In order to perform more accurate image recovery one must have full knowledge of the complex PSF, which of course implies that signal phase must also be obtained. The recent acquisition of a Vector Network Analyser (VNA) by the THz Optics group at NUI Maynooth makes this a possibility and future experimentation that takes full advantage of the ability to record both signal phase and amplitude would allow one to produce sharper images. Alternatively for systems in which it is not possible to obtain phase directly by measurement, one could resort to phase retrieval methods (ref. to chapter 4 & 5). Phase retrieval techniques are used to recover an estimated signal phase from just two intensity measurements of the signal (recorded at two different planes in the optical path of the system).

After it was optimised for best resolution, the reflection mode imaging system yielded very promising results that revealed in great detail the surface structure of objects under test. Unfortunately (but also perhaps fortuitously) standing wave effects dominate when imaging thick objects (greater than half a wavelength). These effects can be eliminated by averaging a number of images obtained at different distances. Alternatively it may be possible to extract from such images information on object surface height by analysing the positions in the image where constructive and destructive interference occur.

The reflection imaging arrangement used provides a map of the dominant reflecting surface, i.e. weakly reflecting surfaces (between source and strongly reflecting surface) are obscured or lost. Future system design should include capabilities for depth (z-axis) resolution. This could be achieved by scanning, at specific depths, the

object with a probe-beam. The probe-beam depth could be controlled by varying the separation between source and receiver beams. Such a system would of course require a tightly focused/collimated probe-beam in order to maintain spatial resolution with increasing depth. This might be achievable with conventional optics (although this could be problematic because of the short confocal distance associated a tightly focused beam), or may call for the use of alternatives such as axicons (to produce a narrow Bessel beam [3.25, 3.26]) or beam-shaping phase plates. Alternatively, the use of heterodyne techniques would allow one to reference phase information which would effectively allowing one to control scan depth [3.27].

3.5.2. Single-Pixel versus Multi-Pixel Imaging Systems

While TOAST is certainly faster and more stable than GHOST, both are inherently slow because images acquisition is done by a serial, or raster-scan process, whereby the single detector (pixel) is scanned bidirectionally across a plane, stopping at each sample point to record relative intensity before moving to the next point to create a two-dimensional image of the scene. Image acquisition time in single-pixel systems is proportional to $1/\Delta^2$, where sample spacing Δ must be small (below minimum feature size) to improve image resolution and for the successful application of standard image processing techniques to raw data (e.g. to allow one to apply noise reduction techniques without significantly degrading features of the underlying object signal). Image acquisition times with TOAST range from approximately one hour for low resolution scans (with a step size of 1mm) to ~15 hours for high resolution scans (with a step size of 0.1mm). An acquisition time of 15 hours is far from real-time imaging so the next important objective would be to significantly reduce scan-times. Another type of single-pixel system involves serial scanning of two plane mirrors mounted on independently controlled and orthogonally aligned motors allowing for control of the azimuth and elevation of the signal beam [3.28] onto the detector. Such a scanning system has been demonstrated at THz frequencies for stand-off imaging applications (airport security screening, etc.) and has yielded a scan-rate of 2 frames per second [3.29].

The ideal THz imager is a camera-like device capable of real-time image acquisition without mechanical scanning. Although a few real-time systems have been devised [3.30], as yet, these are not commercially available. An intermediate step between a single-pixel system and a fully two-dimensional detector array, or focal plane

Chapter 4.

The Design and Experimental Investigation of Regular Phase Gratings

4.1 Introduction

In this chapter we discuss the concept of the diffractive phase element (DPE), a highly efficient wavefront transforming optical element that maximises throughput. We concentrate on the use of DPE's for generating arrays of equally intense far field Gaussian beams, such a DPE being referred to as a diffractive beam-splitter.

We begin by examining one of the earliest solutions investigated to this problem: the periodic binary-level Dammann grating (DG), the theory of which is presented in §4.2. The design of Dammann gratings can be treated as a multivariable optimisation problem and is discussed in §4.3. In §4.4 symmetry considerations used reduce the number of parameters needed to describe each period of these elements are discussed. Typically Fourier analysis (using a fast Fourier transform) is used to model diffraction from such a phase grating, but in §4.5 we shows how Gaussian beam mode analysis can be applied to model DG's. Practical considerations such as how the required phase modulation is encoded into a physical medium at sub-millimetre wavelengths is discussed in §4.6.

Finally §4.7 describes experimental measurements that were made of two Dammann gratings, which were tested using the measurement system and optical components (thin lenses and ellipsoidal and paraboloidal mirrors) described in Chapter 3. The optical design software package MODAL was used to accurately simulate the experimental testing of these gratings. The high degree of similarity between experimental and simulated data in turn provided a good verification of MODAL itself.

4.1.1 Diffractive Optical Elements

In the most general terms a diffractive optical element (DOE) is any optical element that imparts a wavefront transformation on an incident wavefront. The design freedom offered by diffractive optics permits the realisation of diffractive versions of classical refractive counterparts such as focusing elements (lenses and mirrors) that perform a 1-to-1 mapping of the incident wavefront. More importantly, diffractive optics offers the possibility of realising components with optical functions, for which no classical refractive counterparts exist. The most familiar DOE is the amplitude diffraction grating, which performs a 1-to-N mapping of an incident wavefront. A wavefront transformation of this type, in which one input beam is divided into a discrete number

of output beams, is referred to as *beam splitting*. The more general case of a DOE that transforms a wavefront into a continuous signal, (other than the one that would be generated were the DOE absent) is referred to as *beam shaping* and has found applications in the correction of aberrated laser beams as well as for converting the profile of a laser beam from Gaussian to, for example super-Gaussian and top-hat beams as well as to higher-order laser modes [4.1,4.2]. The grouping of wavefront transformation into either beam splitting or beam shaping offers a convenient means of DOE classification.

DOE's such as diffraction gratings, computer generated holograms (CGH's) and Fresnel zone plates produce the required far field intensity by modulating the amplitude of the incident wavefront. However when power is limited, as it is at THz wavelengths, amplitude-modulating devices prove expensive in terms of throughput. A diffractive phase element (DPE) that modulates only the phase of the incident wavefront is (ideally) transparent at the wavelength of the incident radiation and so suffers from only low throughput losses, thus making it a more efficient device.

4.1.2 Beam-Splitting with Diffractive Phase Gratings

In this thesis we have concentrated on the design, analysis, fabrication and testing of beam-splitting DPE's. One solution to the beam-splitting problem is to use semi-transparent plates, or foils, where each foil splits the incident beam into a transmitted and a reflected beam (perpendicular to the direction of propagation) [4.3]. While this approach is suitable for generating small regular beam arrays, more complicated beam geometries and larger numbers of beams necessitate ever more intricate foil arrangements. A DPE provides an elegant, single-element solution to the beam-splitting problem. Beam-splitters produce periodic beam arrays and are thus referred to as diffractive phase gratings since their operation is based on that of the familiar amplitude diffraction grating.

Phase gratings are a variation of the well-known amplitude diffraction grating, which has been developed and applied for over two centuries in the visible part of the electromagnetic spectrum, primarily for high-resolution spectroscopy. However, to date the use of diffraction gratings at longer wavelengths has been relatively limited. Early applications at millimetre and sub-millimetre wavelengths were primarily for wavelength determination. One of the earliest microwave systems incorporating a

diffraction grating was developed for spectroscopy of the ammonia molecule at a wavelength of ~1.1cm [4.4]. Diffraction gratings have also been utilised in systems developed to analyse radiation from plasma fusion reactors [4.4]. With the advent of heterodyne arrays in recent years phase gratings, with their high throughput and beam-splitting abilities, have found an important role in the sub-millimetre range as local oscillator multiplexers [4.5,4.6]. In such systems these passive quasi-optical devices can be used as efficient beam splitters to match the signal beam of a single local oscillator source to an array of detector devices. Beam-splitting phase gratings also prove highly efficient as *N*-to-1 multiplexing devices for the coherent summation of He-Ne lasers [4.7]. In this case the beam-splitting device is operated in reverse (i.e. as a beam combiner) with the laser beams positioned at the locations of the signal orders of the grating [4.8]. Another application for beam-splitters is in optical digital computing, where arrays of spots are required to provide optical power supply beams for arrays of logic devices [4.9].

A diffraction grating is defined as any regular array of diffracting elements or obstacles that has the effect of producing periodic alterations in the amplitude and/or phase of an incident wavefront. The periodically modulated electromagnetic wave generates a set of waves, called diffraction orders, which in the far field propagate in discrete directions. The simplest arrangement is a multiple slit configuration consisting of a set of parallel lines on a plane surface, the separations of which must not be much greater than the order of magnitude of the wavelength of light being used in order to produce well separated, so called far-field diffraction order fringes. If the ratio of slit width a to slit separation d is small, the intensity of the light on a screen in the far field behind the grating has a series of very well separated narrow maxima to either side of a primary, or principal, maximum, the intensities of which fall off slowly with increasing angle. These are the diffraction orders produced by the grating, whose positions depend on the grating period d and the wavelength of the incoming radiation and is summarised by the so-called *grating equation*,

$$\sin\theta_d = \sin\theta_i + \frac{n\lambda}{d}, \qquad n = \{0, \pm1, \pm2, \pm3, \ldots\} \tag{4.1}$$

where θi and θ_d are the angles subtended, with respect to the gratings' normal, of the incident and diffracted beams respectively and each value of n corresponds to an individual *diffraction order*, the $n = 0$ or zeroth order corresponding to the undeflected $\theta_d = \theta i$ position. If the grating is illuminated by an incident wave along the gratings'

normal ($\theta i = 0$) then the above equation is reduced to the well-known grating equation for normal incidence

$$n\lambda = d \sin\theta, \quad n = \{0, \pm1, \pm2, \pm3, \ldots\} \tag{4.2}$$

The far field diffraction pattern also includes regions occupied by secondary, or subsidiary, maxima that appear as extremely faint lines between the principal maxima, especially close into the main diffraction orders.

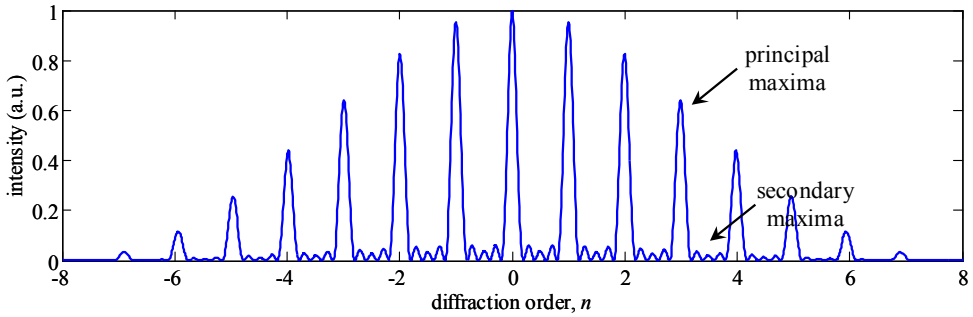


Figure 4-1. The far field intensity from an amplitude diffraction grating with $N = 5$ narrow slits consists of an array of bright principal maxima, which are separated by $N–2 = 3$ weak secondary maxima. Here $\lambda = 3$mm, slit separation $d = 75$mm and slit width, $a = 3.6$mm. The first diffraction order ($n = 1$) is located at an angle of ~2.3°.

An important consideration for a diffraction grating is its efficiency, in terms of the fraction of incident power that is directed into the desired output beams. Equations (4.1) and (4.2) show that several diffraction orders can propagate simultaneously from the grating (essentially a result of the array theorem, of course [4.22]). For beam-splitting applications not all of these orders are needed. Instead we require that the limited incident power be redirected into only a small number of the diffraction orders – the *signal* orders – and that the power in the remaining *parasitic* orders [4.10] be suppressed.

In the more general case a one-dimensional periodic diffraction grating is defined as consisting of a number of parallel linear structures formed in a plane. The repeated structure is not limited to slits but can include grooves in a metal plate, apertures in a sheet, or variations in thickness of a dielectric material of suitable refractive index. The first of these is an example of a reflection phase grating, while the latter two are transmission gratings the first being an amplitude grating and the latter a phase grating. In the following discussion we refer to the repeated structure of the grating as "*grooves*", irrespective of the actual shape of the structure. The grating period d denotes the spacing over which the groove pattern is repeated. As will be shown it is

the precise form of the repeated groove pattern that determines the relative intensity of the principal diffraction orders produced by a diffraction grating. Thus the problem of beam-splitting is to find appropriate periodic groove pattern that maximises power in the required signal orders, while simultaneously suppressing power in parasitic orders.

4.2 Theory of Dammann Gratings

Some of the theory outlined in this section is also presented in [4.18]. First described by Dammann *et al* in [4.11] and subsequently in greater detail in [4.12], Dammann gratings (DG), as they have come to be known are binary-level phase gratings. Although originally developed for use at visible wavelengths to produce multiple images of equal intensity of a single input image they also make efficient beam-splitting devices and can be designed for any given wavelength. For example they have been incorporated in long-wavelength millimetre–wave quasi-optical systems to generate one- or two-dimensional arrays of regularly spaced diffraction spots for feeding a single local oscillator (LO) source to an array of horns [4.23]. At NUI Maynooth phase gratings have been designed to operate in the millimetre and terahertz wavebands, for instance at a centre frequency of 100 GHz (or 0.1 THz) [4.24,4.25,4.26,4.27]. In the visible (and near infrared) binary phase gratings have also been developed as cost effective solutions to providing star couplers in fiber-optic networks for conveying light from a single input port to N output ports [4.13,4.14].

The motivation behind Dammann's work was the limitations inherent in technology available at the time. Various holographic techniques (based on commonly recorded holograms of arrays of light sources) had been proposed for multiple imaging. In reconstruction a single object beam illuminates the hologram and, instead of an array of point light sources, an array of images of the incident object beam is generated. A major drawback inherent in this method is that 1) the image array is generated off-axis, an unfavourable arrangement that results in aberrations in the multiplied images and 2) the efficiency is relatively low due to low reconstruction efficiency of the commonly recorded thin holograms [4.11].

One solution to overcome these difficulties is to use in-line, i.e. on-axis, phase-only holograms. The brightness distribution of the output is determined by the intensity of the Fourier transform of a single groove, e.g. a simple rectangular 1-D groove

produces a sinc function. As such each groove can be considered as an elementary in-line Fourier transform hologram and the grating as a multiple phase hologram consisting of a two-dimensional array of small elementary holograms. It was pointed out that the groove shape needed to generate the required output pattern could be determined using an inverse Fourier Transform. At the time, however, it was considered that this would result in groove shapes too complex to be realized in practice [4.11] and with low efficiency (presumably due to off-axis scattering). Since this technology was intended for application at visible wavelengths the manufacture of such structures, with groove depths on the scale of a micron or so, would indeed have proved impossible at the time. Dammann therefore restricted study to binary-level phase-only structures, which could be constructed relatively easily using lithographic techniques developed for the semiconductor industry.

Thus Dammann gratings consist of a regular array of milled slots or grooves of equal depth in a transparent dielectric material of suitable refractive index (for a transmission grating) or, alternatively in a reflecting surface (for a reflection grating). The DG is therefore effectively a binary optical element that subjects a beam incident on the grating to a phase-only modulation with two possible phase shifts, or delays (typically chosen to be 0 and π radians), due to relative path length differences imposed by the surface grooves. A delay of one wavelength corresponds to a phase shift of 2π similarly a phase shift of π radians results in path lengths through the grating modulated by discrete steps of half a wavelength.

As mentioned above phase gratings can be designed as transmission or reflection gratings. Transmission gratings offer potentially high coupling efficiency with low attenuation loss. Ideal materials for use at sub-millimetre wavelengths include quartz (which exhibits low absorption losses [4.16] and has a refractive index of 2.0) and high-density polyethylene (HDPE) (with a refractive index of 1.52) that is inexpensive by comparison and easier to machine. A possible limitation to performance of transmission phase gratings is the presence of standing waves within the structure along with, and related to, reflections from the air-dielectric interface – an issue previously investigated by Trappe [4.18]. This problem is avoided in reflection gratings, although these too present their own difficulties, foremost amongst which is that the grating must be designed for operation in an off-axis configuration. At visible wavelengths phase gratings have also been demonstrated by encoding the phase modulation into phase-only

liquid-crystal spatial light modulators [4.15,4.17], which allows the phase modulation to be changed in real-time.

Besides restricting solutions to only binary phase functions, the other main simplification in Dammann's method is that 2-D transmission functions are restricted to those that are separable into two 1-D functions of the transverse spatial coordinates x and y, i.e. of the form

$$t_G(x, y) = t_1(x) \cdot t_2(y) \tag{4.3}$$

where $t_1(x)$ and $t_2(y)$ individually generate, on the output Fourier plane, 1-D arrays with M and N beams respectively. Thus the 2-D diffraction envelope $T_G(u, v)$, given by the Fourier transform of $t_G(x, y)$, can then also be decomposed into the spatial angular frequencies u and v as

$$T_G(u, v) = T_1(u) \cdot T_2(v) \tag{4.4}$$

where $T_1(u)$ and $T_2(v)$ are the 1-D Fourier transforms of $t_1(x)$ and $t_2(y)$ respectively. The problem of finding a 2-D phase-only transmission function of the form

$$t_G(x, y) = e^{-i\phi_G(x, y)} \tag{4.5}$$

is thus reduced to two 1-D problems. The separable 2-D phase function is

$$\phi_G(x, y) = \phi_1(x) + \phi_2(y) \tag{4.6}$$

so Eq. (4.5) becomes

$$t_G(x, y) = e^{-i\phi_1(x)} \cdot e^{-i\phi_2(y)} \tag{4.7}$$

If the 2-D grating transmission function $t_G(x, y)$ is derived from two separable functions then naturally such a function can only generate rectangular $M \times N$ beam arrays.

Since for a Dammann grating the phase is restricted to values of 0 or π radians, following from Eq. (4.7) the only values permitted to transmission functions $t_G(x, y)$, $t_1(x)$ and $t_2(y)$ are ± 1, as tabulated in Table 4-1. Although typically the phase levels for Dammann gratings are assigned values of 0 and π, the far field intensity is invariant to the absolute phase [4.48] so only the relative phase values (the phase difference $\Delta\phi$ between the phase levels) are of consequence. Thus a DG with phase levels $\pi/2$ and $3\pi/2$ (or any other two values separated by π radians) produce the same phase modulation, as illustrated in Figure 4-2.

$\phi_1(x)$	$\phi_2(y)$	$\phi_G(x, y) = \phi_1(x) + \phi_2(y)$	$t_1(x)$	$t_2(x)$	$t_G(x, y) = t_1(x) \cdot t_2(y)$
0	0	0	+1	+1	+1
0	π	π	+1	−1	−1
π	0	π	−1	+1	−1
π	π	2π = 0	−1	−1	+1

Table 4-1. The four combinations of values for the 1-D phase functions $\phi_1(x)$ and $\phi_2(y)$ and the resultant real-valued 1-D and 2-D transmission functions $t_1(x)$, $t_2(y)$ and $t_G(x, y)$.

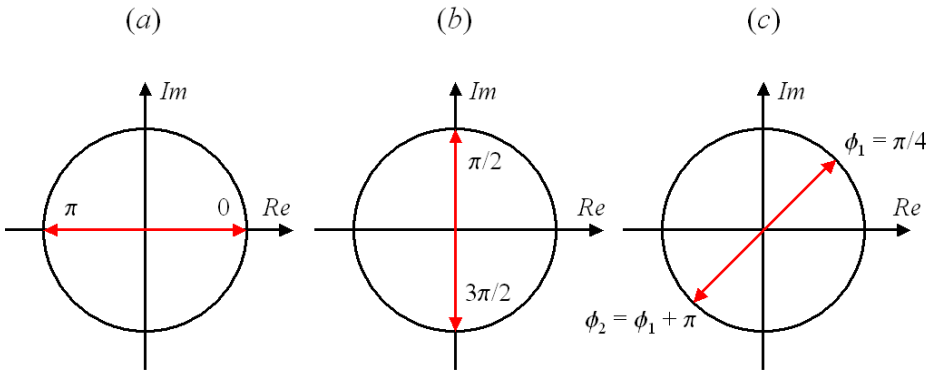


Figure 4-2. Argand Diagrams showing the equivalence of different pairs of phase level values (*a*) {0,π} and (*b*) {π/2, 3π/2} and (*c*) {π/4, 5π/4} for a binary-level (Dammann) phase grating. In each case the phase level difference Δϕ equals π radians. The only difference is in the form of the transmission function $t(x, y)$, i.e. in (*a*) $t(x, y)$ is purely real, in (*b*) it is purely imaginary and in (*c*) it has both real and imaginary components.

The reason for setting the phase difference $\Delta\phi = \pi$ when designing a 2-D binary phase element, such as a DG, is that when two 1-D solutions are used to form a 2-D grating surface, the phase levels can be reduced modulo 2π and the binary nature of the phase surface maintained [4.48].

Besides the phase difference, Δϕ, the only free parameters of a binary-level phase function are the locations of transition points: points on the grating where the phase function steps between the two permitted values ϕ_1 and $\phi_2 = \phi_1 + \Delta\phi$. Figure 4-3 shows the cross section of a 1-D binary phase function $\phi(x)$ defined by phase-levels 0 and π, and a set of transition points $\{x_1, x_2 \ldots x_5\}$, and the corresponding binary (real-valued) transmission function $t(x)$.

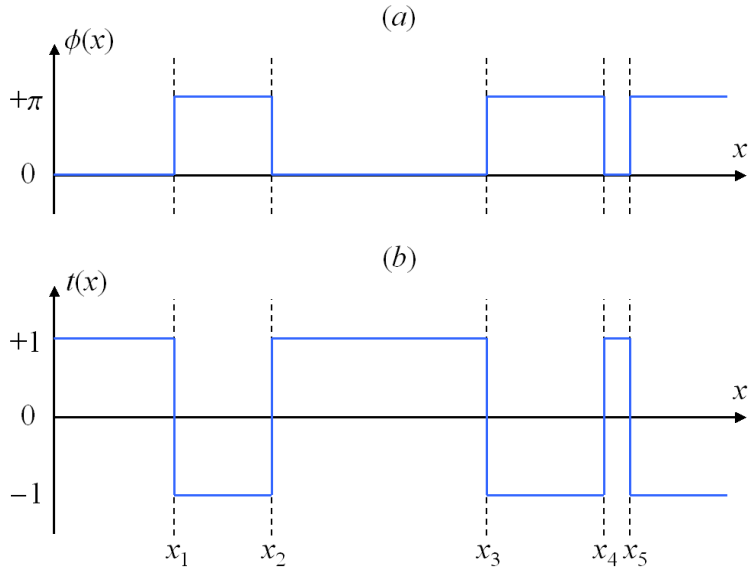


Figure 4-3. The 1-D binary-phase function $\phi(x)$ in (*a*) is defined by the set of five transition points $x_t = \{x_1, x_2 \ldots x_5\}$ and the phase delays $\{\phi\} = \{0, \pi\}$, resulting in phase steps of $\Delta\phi = \{0, \pi\}$; and (*b*) the corresponding real-valued binary transmission function $t(x)$.

The phase $\phi(x)$ of a Dammann grating with 4 identical cells is shown in Figure 4-4, where the phase profile of each cell $\phi_{cell}(x)$ is that shown in Figure 4-3.

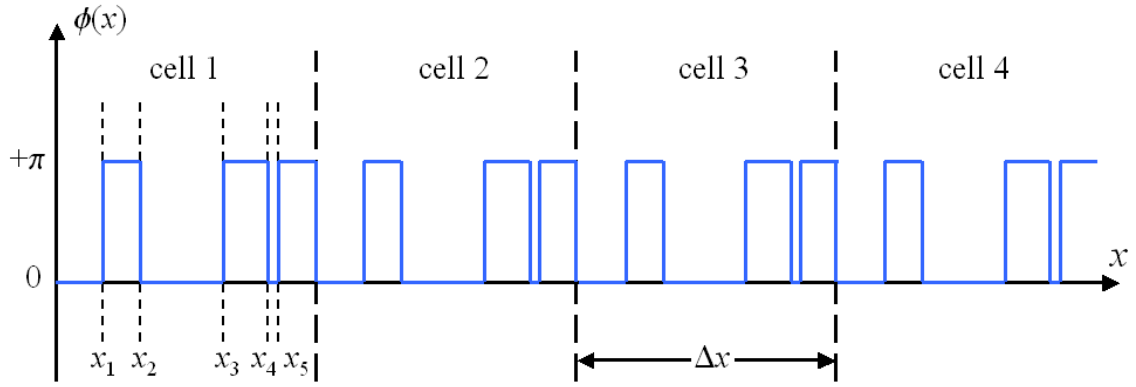


Figure 4-4. Cross-section of a 1-D binary-level phase grating with period Δx, consisting of 4 repeated unit-cells. The unit-cell phase function is that of Figure 4-3(*a*).

4.2.1 The Diffraction Envelope from a Dammann grating

As with an amplitude diffraction grating the diffraction envelope $T(u, v)$ of a Dammann grating is given by the Fourier transform of the transmission function $t_{cell}(x, y)$ of the unit cell. Typically this calculation is performed with a discrete Fourier transform (FT), usually with a Fast Fourier transform (FFT) algorithm. Here we present a quasi-analytical means of calculating $T(u, v)$ from a DG unit cell by invoking useful theorems from Fourier transform theory (such as the convolution theorem, shift theorem, etc.) that

does not require computing the FFT of the discretely sampled unit cell. This approach is especially suitable for analysis of phase gratings since we are not necessarily interested in obtaining the far field distribution at all points on a plane, but rather at only a small number of discrete positions: the diffraction orders, where the diffraction envelope $T(u)$ coincides with the maxima of the interference term $A(u)$ given by the Array Theorem. However, it will be shown how this approach can be extended to allow the evaluation of the FT of any discretely sampled function, as an alternative to using a FFT algorithm.

The unit cell transmission function $t_{cell}(x)$ of a Dammann is restricted to values of ±1 and so can be expressed as the sum of two binary-valued functions

$$t_{cell}(x) = t_{+1}(x) + t_{-1}(x) \tag{4.8}$$

where the non-zero parts of $t_{+1}(x)$ corresponds to regions of the unit cell where $\phi(x) = 0$; and the non-zero parts of $t_{-1}(x)$ corresponds to regions where $\phi(x) = \pi$, where $t(x) = 0$ corresponds to no transmission of course. Figure 4-5 shows this for the unit cell of a one-dimensional DG solution designed to produce five equi-intense diffraction orders with four transition points $x_t = \pm\{0.019, 0.368\}\Delta x$.

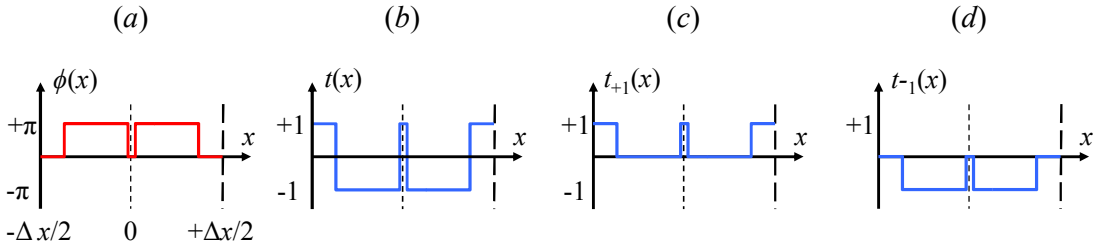


Figure 4-5. (*a*) A binary phase function $\phi(x)$ with the phase levels 0 and π corresponds to (*b*) a transmittance function $t(x)$ with values of ±1. $t(x)$ can thus be expressed as the sum of two functions (*c*) $t_{+1}(x)$ and (*d*) $t_{-1}(x)$ with values of [0, +1] and [0, −1], respectively.

If the unit cell is defined by M transition points then clearly $t_{-1}(x)$ and $t_{+1}(x)$ can be represented as summations of $(M+1)$ appropriately positioned and scaled rectangular functions, each function representing a single constant-valued phase segment of the unit cell. For example, the basis cell of the 5-order DG is defined by a total of $M = 4$ transition points, so its transmission function can be represented by a summation with $(M+1) = 5$ terms. The three segments of $t_{+1}(x)$, see Figure 4-6(*c*), are represented by the rectangular functions (as defined by *Gaskill* in [4.49])

$$f_1(x) = rect\left(\frac{x-x_{f_1}}{d_1}\right),\ f_3(x) = rect\left(\frac{x-x_{f_3}}{d_3}\right),\ f_5(x) = rect\left(\frac{x-x_{f_5}}{d_5}\right)$$

and the two segments of $t_{-1}(x)$, see Figure 4-6(b), are represented by

$$f_2(x) = rect\left(\frac{x-x_{f_2}}{d_2}\right),\ f_4(x) = rect\left(\frac{x-x_{f_4}}{d_4}\right)$$

where d_i and x_{f_i} (= $x_t(i)$ + $d_i/2$) denote the width and centre of the i^{th} rectangular function. Each rectangular function is weighted by the value of $e^{i\phi(x)}$ at that point. Denoting the phase associated with the i^{th} rectangular function by ϕ_i gives

$$t_{cell}(x) = \sum_{i=1}^{M+1} e^{i\,\phi_i} f_i(x) \tag{4.9}$$

Since the phase of a Dammann grating has values of 0 or π the rectangular functions are weighted by either $e^{i\pi} = -1$, or $e^{i0} = +1$ so Eq. (4.9) becomes

$$t_{cell}(x) = \sum_{i=1}^{M+1} (-1)^i f_i(x) \tag{4.10}$$

where the order in which positive/negative signs are assigned is irrelevant (since a change of sign to all terms simply inverts the grating pattern). In other words the transmission function for a DG is given by a sum of positive and negative rectangular functions. For the 5×5 DG we have

$$t_{cell}(x) = -[f_2(x) + f_4(x)] + [f_1(x) + f_3(x) + f_5(x)] \tag{4.11}$$

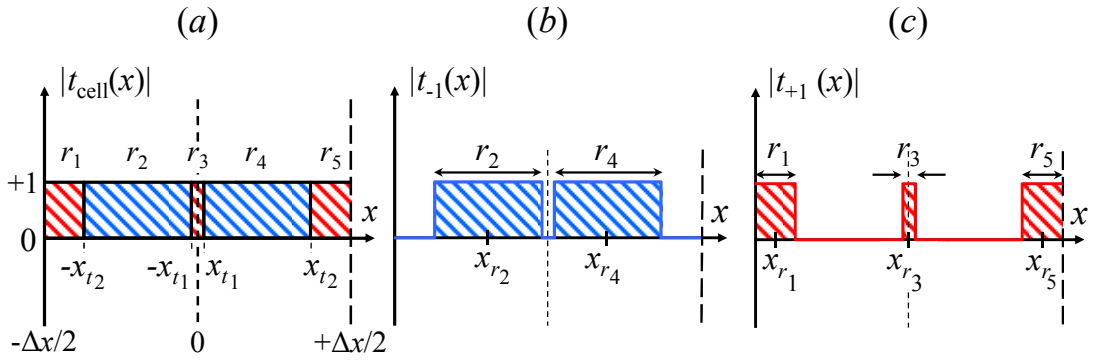


Figure 4-6. (a) The unit cell transmission function $t_{cell}(x)$, defined by the transition points $\pm x_{t1}$, $\pm x_{t2}$, can be represented by a summation of five rectangular functions (of width d_i and centred at x_{fi}) shown in (b) and (c). The red and blue functions represent the π-valued and 0-valued phase regions of unit cell.

The diffraction envelope $T_{cell}(u)$ from the unit cell has the form

$$T_{cell}(u) = T_{-1}(u) + T_{+1}(u) \tag{4.12}$$

where $T_{-1}(u)$ and $T_{+1}(u)$ are the Fourier transforms of $t_{-1}(x)$ and $t_{+1}(x)$ respectively. The shifting property of Fourier transforms [4.49]

$$\Im\{f(x\text{-}a)\} = F(u)e^{-i2\pi au}$$

implies that

$$rect\left(\frac{x-x_f}{d}\right) \rightarrow |d|\text{sinc}(du)\, e^{-i\,2\pi x_f u} \tag{4.13}$$

Thus the two components of the diffraction envelope evaluate to

$$T_{-1}(u) = -\left[|d_2|\text{sinc}(d_2u)\, e^{-i\,2\pi x_{f_2} u} + |d_4|\text{sinc}(d_4u)\, e^{-i\,2\pi x_{f_4} u}\right] \tag{4.14}$$

$$T_{+1}(u) = +\left[|d_1|\text{sinc}(d_1u)\, e^{-i\,2\pi x_{f_1} u} + |d_3|\text{sinc}(d_3u)\, e^{-i\,2\pi x_{f_3} u} + |d_5|\text{sinc}(d_5u) e^{-i\,2\pi x_{f_5} u}\right] \tag{4.15}$$

a summation of phase-shifted sinc functions. The transition points for the 5-order DG are located at $x_t = \{-0.368, -0.019, +0.019, +0.368\}\Delta x$ so the phase discontinuities in the basis cell occur at $x_d = \{-½, x_t, +½\}\Delta x = \{-0.5, -0.368, -0.019, +0.019, +0.368, +0.5\}\Delta x$, i.e. the transition points plus the unit cell endpoints (at $x = \pm\Delta x/2$). The width d_i of the i^{th} constant-valued phase segment (rectangular function) is thus

$$d_i = x_{d_{i+1}} - x_{d_i} \tag{4.16}$$

Due to the reflection symmetry properties of this set of transition points ($d_4 = d_2$, $d_5 = d_1$, $x_{f_4} = -x_{f_2}$, and $x_{f_5} = -x_{f_1}$) and since $f_3(x)$ is on-axis ($x_{f_3} = 0$) so its corresponding Fourier plane sinc function is not phase-shifted, Eq.'s (4.13) and (4.14) become

$$T_{-1}(u) = -2|d_2|\text{sinc}(d_2u)\cos(2\pi x_{f_2} u) \tag{4.17}$$

$$T_{+1}(u) = +2|d_1|\text{sinc}(d_1u)\cos(2\pi x_{f_1} u) + |d_3|\text{sinc}(d_3u) \tag{4.18}$$

Finally, evaluating $|T(u)| = |T_{-1}(u) + T_{+1}(u)|$ at $u_n = \pm n/\Delta x$, for $n = 0, 1, 2$ yields the amplitude of the diffraction envelope at the positions of the five central diffraction orders. Figure 4-7 shows the far field amplitude distribution $|E_F|$, i.e. the diffraction envelope $|T(u)|$ from the unit cell for this particular 5-order DG solution, which we will refer to as solution S_1. The diffraction order intensities $I_n = |T(u_n)|^2$ for $n = \{0, \pm1, \pm2\}$ are $I_0 = 0.1568$, $I_{\pm1} = 0.1550$, $I_{\pm2} = 0.1539$. These are almost identical so we can say that the five central diffraction orders are of approximately equal intensity, as required.

Another DG solution (which we will refer to as solution S_2) that produces five equi-intense diffraction orders is defined by the transition points $x_t = \pm\{0.132, 0.481\}\Delta x$ and its diffraction envelope is also shown in Figure 4-7. Although the diffraction envelope from solutions S_1 and S_2 are quite different, at the locations of diffraction orders their intensities I_n are equal. Thus the unit cells defined by these two sets of transition points when repeated periodically in a grating will produce (nearly) identical patterns of five equi-intense diffraction orders. In practise the finite extent of a diffraction grating (i.e. the finite number of basis cells) results in the appearance of

secondary maxima between the diffraction orders. Since the diffraction envelope also determines the relative intensity of these maxima slightly different far field intensity patterns will be observed from gratings derived from solutions S_1 and S_2.

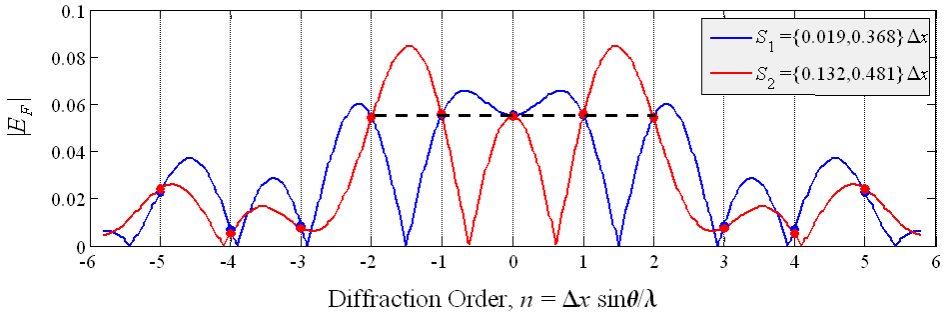


Figure 4-7. Normalised (integrated intensity equals unity) amplitude distribution of the far field diffraction pattern (the diffraction envelope $|T(u)|$) from the unit cell of the two 5-order DG solutions S_1 and S_2. When the unit cell is repeated periodically (as in a grating) the diffraction orders are weighted by the value $T(u)$ at the discrete positions $u_n = n\Delta u$. Solutions S_1 and S_2 will produce an array of 5 equi-intense diffraction orders, as indicated by the black dashed line through the two diffraction patterns at their points of intersection with the diffraction orders $n = \{0, \pm1, \pm2\}$. These two solutions are equivalent since the diffraction order weighting (the value of $|T(u)|$ at integer-valued n) is identical for S_1 and S_2, as indicated by the circular markers.

Extending the approach described above to its logical conclusion leads one to a means of computing the discrete Fourier Transform (DFT) of a given complex-valued field $f(x)$ (Note: in this context $f(x)$ refers to any arbitrary function, and not necessarily rectangular functions defined above). For a discretely sampled function $f(x)$ with N samples, each sample $f(x_i)$ can be treated as a single appropriately weighted rectangular function, shifted in x by its position x_i and with a scaling factor equal to the sample rate Δx, i.e. $rect(x–x_i/\Delta x)$. Thus $f(x)$ can be represented as a summation of N spatially-shifted rectangular functions, each of which is weighted by the value of $f(x)$ at the point x_i, i.e.

$$f(x) = \sum_{i=1}^{N} f(x_i) \cdot rect\left(\frac{x-x_i}{\Delta x}\right) \tag{4.19}$$

The Fourier transform $F(u)$ of $f(x)$ is then equal to the sum of N phase-shifted sinc functions

$$F(u) = \sum_{i=1}^{N} f(x_i) \cdot |\Delta x|\mathrm{sinc}(\Delta x u)e^{-i\,2\pi x_i u} \tag{4.20}$$

For a constant sample spacing Δx

$$F(u) = |\Delta x|\mathrm{sinc}(\Delta x u) \sum_{i=1}^{N} f(x_i)\, e^{-i\,2\pi x_i u} \tag{4.21}$$

and writing $e^{-i2\pi u} = \alpha$

$$F(u) = |\Delta x|\text{sinc}(\Delta x u) \sum_{i=1}^{N} f(x_i)\, \alpha^{x_i} \tag{4.22}$$

and in two-dimensions

$$F(u, v) = |\Delta x \Delta y|\text{sinc}(\Delta x u, \Delta y v) \sum_{i=1}^{M} \sum_{j=1}^{N} f(x_i, y_j)\, \alpha^{x_i} \beta^{y_i} \tag{4.23}$$

where $\beta = e^{-i2\pi v}$ and the sample rate Δy is also constant. If the discretely sampled spatial frequency coordinates (u, v) are required at discretely sampled points $(u_{i'}, v_{j'})$, the summation can be computed using matrix multiplication, which would reduce execution time, as

$$F(u_{i'}, v_{j'}) = |\Delta x \Delta y|\text{sinc}(\Delta x u_{i'}, \Delta y v_{j'}) \sum_{i=1}^{M} \sum_{j=1}^{N} f_{ij}\, \alpha_{ii'} \beta_{jj'} \alpha^{x_i} \beta^{y_i} \tag{4.24}$$

where $f_{ij} = f(x_i, y_j)$, $\alpha_{ii'} = e^{-i2\pi(u_i x_i)}$ and $\beta_{jj'} = e^{-i2\pi(v_j y_j)}$

When using a FFT the same number of sample points is used at the input and output planes. Thus in order to increase output plane resolution the input plane must be padded with trailing zeros (as explained in Appendix A.2), leading to high computational overhead. The approach outlined above for computing the DFT is more flexible since no restrictions are placed on the number of samples at the output plane. Its major shortcoming however is its slow execution speed compared to FFT, especially for two-dimensional calculations where execution times become quite prohibitive.

4.2.2 Evaluating phase grating performance

When designing a particular phase grating one must be able to quantify its performance in a meaningful way. A figure of merit is used to evaluate phase grating performance. Since phase freedom is used in the design of phase gratings, values for merit figures are calculated from the far field intensity produced by a particular grating solution.

The choice of which figure of merit to use depends on the problem at hand. The design of a beam-shaping element requires that the far field intensity generated by the grating, the "trial" image, match as closely as possible the intended "target" image. Thus, evaluating the performance of beam-shaping phase gratings involves calculating a correlation between the target and trial images.

In beam-splitting applications the grating is typically required to generate an array of equi-intense diffraction orders, while simultaneously ensuring that the maximum amount of transmitted energy is diffracted into those orders. Thus, when evaluating the performance of beam-splitters two commonly used merit functions are linear diffraction efficiency η and the standard deviation σ between diffraction order intensities. A high quality beam-splitting element is one that diffracts most of the transmitted power into the set of signal orders with minimum deviation in intensity between the diffraction orders.

Diffraction Efficiency and Beam Uniformity

The one-dimensional 'linear' diffraction efficiency, η is defined as the ratio of radiant flux in the equally bright central diffraction orders to the total radiant flux incident on and therefore transmitted through the grating, and is defined by Dammann [4.11] as

$$\eta_1 = \frac{\text{radiant flux within signal orders}}{\text{total radiant flux through the grating}} \leq 1 \tag{4.25}$$

The efficiency of a two-dimensional phase grating that is separable as two 1-D grating functions is then given by

$$\eta_2 = \eta_x \cdot \eta_y \tag{4.26}$$

where η_x and η_y are the diffraction efficiencies of the two 1-D transmission functions $t_1(x)$ and $t_2(y)$, respectively. One-dimensional diffraction efficiency is defined by Dammann [4.12] as

$$\eta_1 = \sum_{n=-N}^{n=N} I_n \tag{4.27}$$

where diffraction order n produced by the grating has an intensity of $I_n = |A_n|^2$, with A_n being its amplitude. The diffraction order intensities I_n are normalised such that

$$\sum_{n=-\infty}^{n=+\infty} I_n = 1 \tag{4.28}$$

Eq. (4.27) defines diffraction efficiency for an odd number ($2N$+1) of diffraction orders but this definition is easily adapted to gratings that produce any periodic array of diffraction orders.

Referring to §4.2.1, the intensities of the five central diffraction orders from the Dammann grating whose unit cell is defined by transition points $x_t = \pm\{0.019, 0.368\}\Delta x$,

are $I_0 = 0.1568$, $I_{\pm1} = 0.1550$ and $I_{\pm2} = 0.1539$. Thus the diffraction efficiency of this one-dimensional grating is equal to

$$\eta = I_0 + 2\sum_{n=1}^{2} I_n = I_0 + 2(I_{\pm1} + I_{\pm2}) = 0.7747$$

In other words (in the one-dimensional case) 77.47% of the beam intensity transmitted through the grating is diffracted into the five central diffraction orders and the remaining power is distributed amongst higher-order parasitic orders. Similarly the other Dammann gratings solution mentioned in §4.2.1 with transition points at $x_t = \pm\{0.132, 0.481\}\Delta x$ produces an array of five diffraction order with the same intensities and therefore that grating also registers a diffraction efficiency value of $\eta = 77.47\%$.

Although diffraction efficiency is defined with a maximum value of unity, investigations have shown that upper bounds can be placed on realisable efficiencies. For example Krackhardt *et al* [4.19] have shown that for array generators (beam-splitters) required to produce arrays of more than five equi-intense diffraction orders, the upper bound on diffraction efficiency for (0, π) binary phase gratings (such as Dammann gratings) phase gratings ranges between 83% and 84%; for (0, non-π) binary phase gratings the upper bound is 87-88%; and for continuous, non-binary phase gratings the upper bound is 97-99%. Whereas continuous phase gratings have effectively a very large number of degrees of freedom, binary phase gratings have considerably less thereby reducing the maximum diffraction efficiency they can achieve.

An equally important criterion for beam-splitting applications is that the intensity be distributed evenly between the signal orders. For example, when used for laser welding a beam-splitter must deliver equal power to each weld spot. Figure 4-8 shows the far field intensity produced by two different continuous-phase gratings that are designed to generate arrays of six equi-intense signal orders. Both solutions yield similarly high diffraction efficiencies (~93%), however the diffraction order intensities I_n in Figure 4-8(*a*) vary greatly from peak to peak compared to those in Figure 4-8(*b*). By definition diffraction efficiency only provides a measure of the fraction of incident power that is diffracted into the array of signal orders. It gives no indication of the power distribution between those beams so by itself is an insufficient measure of performance [4.19].

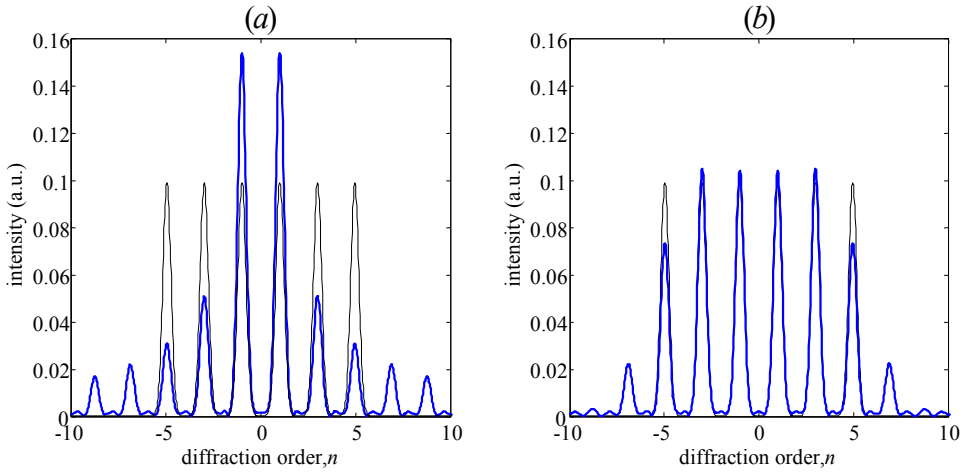


Figure 4-8. Far field intensity patterns from two Fourier phase gratings designed to produce a target array (plotted in black) of six equally intense diffraction orders. Clearly power is distributed very unevenly between the six orders in (*a*) compared to the pattern in (*b*) yet the diffraction efficiency evaluated for both intensities yield almost identical values of ~93%.

A measure of how power is distributed amongst the array of signal orders is given by the uniformity, U of the diffraction order intensities generated by the grating. Several definitions of beam uniformity are found in literature. For example Barnes [4.15] defines beam uniformity as

$$U = (I_{max} - I_{min})/(I_{max} + I_{max}) \tag{4.29}$$

where I_{max} and I_{min} denote the intensities of the strongest and weakest spot, respectively. The definition used by Krackhardt *et al* [4.19] specifies beam uniformity in decibels as

$$U = 10 \log_{10}\left(\frac{I_{max} + I_{min}}{I_{max} - I_{min}}\right) \tag{4.30}$$

With this definition a value of $U > 40$dB indicates a solution with perfect uniformity. A useful definition, based on the latter, is as follows

$$U = 1 - \left(\frac{I_{max} + I_{min}}{I_{max} - I_{min}}\right) \tag{4.31}$$

where $U = 1$ indicates perfect uniformity.

We define mean power difference (*MPD*) to measure the non-uniformity between the intensity of N signal orders as follows. The sum of intensity differences between each order and the strongest order is calculated and divided by N to give

$$MPD = \frac{1}{N}\sum_{i=1}^{N} |I_{max} - I_i| \tag{4.32}$$

with $0 \le MPD < 1$. Note that the subscript 'i' on intensities I_i, refer to a beam number and not the diffraction order. If power is distributed equally amongst the N diffraction

orders $MPD = 0$. Conversely we define mean power uniformity as $MPU = (1 - MPD)$ so that a perfectly uniform array of N diffraction orders yields a value of $MPU = 1$. For the far field intensities shown in Figure 4-8 the mean power difference between the six diffraction orders was calculated to be ~74% and ~21%, respectively. Conversely the mean power uniformity for each is ~26% and ~79%, thus indicating the superior quality of the second solution over the first.

The standard deviation (SD) of the diffraction order intensities I_0, $I_{\pm 1}$, …, $I_{\pm N}$, is another commonly used way to estimate beam uniformity. The definition of standard deviation of intensities that we have used is

$$\sigma = \sqrt{\frac{1}{2N+1} \sum_{n=-N}^{n=+N} (I_n - \langle I \rangle)^2} \tag{4.33}$$

where $\langle I \rangle$, the mean value of the intensities of the $2N+1$ diffraction orders is given by

$$\langle I \rangle = \frac{1}{2N+1} \sum_{n=-N}^{n=+N} I_n \tag{4.34}$$

The standard deviation provides a measure of error in the energy distribution between the signal orders. In other words it is a measure of beam non-uniformity, and a small value indicating a more uniform intensity distribution than a large value.

As was noted in [4.19], when searching for phase grating solutions one usually encounters a trade-off between multiple criteria (e.g. diffraction efficiency and beam uniformity). Consequently one must account for trade-offs by defining a design metric that combines two or more merit functions into a single figure of merit. For example Jacobsson *et al* [4.20] combined efficiency, cross correlation and average deviation to quantify the degree to which the intensity distribution from a kinoform replicated a desired intensity distribution. Since we wish to find solutions that exhibit high diffraction efficiency and a high degree of beam uniformity we define the weighted diffraction efficiency as

$$\eta_{weight} = \eta \cdot (1 - MPD) = \eta \cdot MPU \tag{4.35}$$

where a value of unity indicates a solution with unit diffraction efficiency and perfect beam uniformity. The weighted diffraction efficiency for the intensity in Figure 4-8(*a*) is 0.93(0.26) = 0.24, or 24%, while the value for the intensity in Figure 4-8(*b*) is 0.93(0.79) = 0.735, or 73.5%. Similarly one could combine diffraction efficiency with standard deviation to give a measure of weighted diffraction efficiency as $\eta_{weight} = \eta \cdot (1 - \sigma)$.

Another option is to define a design metric as the weighted sum of an error metric ε_1 and a diffraction efficiency metric ε_2 [4.21] as

$$\varepsilon = \varepsilon_1 + \varepsilon_2 = \alpha(I_{max} - I_{min})^2 + \sum_{n=-N}^{n=+N} |\frac{\beta}{2N+1} - I_n|^2 \tag{4.36}$$

where the weighting between ε_1 and ε_2 is determined by the constant scaling factors α and β that can be used to adjust the relative importance of each term thereby balancing the trade-off.

4.2.3 Quasi-Optical Design of Phase Gratings

When designing a phase grating one assumes planar illumination of the grating. In other words the only phase function at the grating plane is the phase modulation introduced by the grating itself. A lens or mirror is required to collimate the illuminating beam to provide a planar phase front at the grating. The distance at which the diffraction pattern from the grating is formed follows the criterion for far-field image formation, which states that if allowed to propagate freely the far-field image will form only after a distance of $z >> a^2/\lambda$, where a is the largest dimension of the diffracting object – in this case the grating diameter [4.22]. For instance, the far field image from a 100 mm diameter grating illuminated by a source radiating at a frequency of 100 GHz (λ ~3 mm) will form at a propagation distance of approximately 3.3 metres. Clearly such an arrangement is unsuitable for tabletop experiments because of the large separation between the grating and the plane of observation. Also the size of the observation plane would need to be very large to accommodate the array of diffraction orders at this distance. A more compact arrangement is to place a second lens/mirror after the grating such that the grating plane coincides with the common focal plane of the two lenses/mirrors. If the focal lengths of the two lenses/mirrors are equal the system is referred to as a 4-*f* Fourier optical system, or Gaussian beam telescope. When the grating is omitted from the system the input beam is formed with unit magnification. The effect of the second lens is to Fourier Transform the field at the grating plane onto the output plane (the back focal plane L_2/M_2). Thus, Fourier optics is commonly used in the design and analysis of phase gratings.

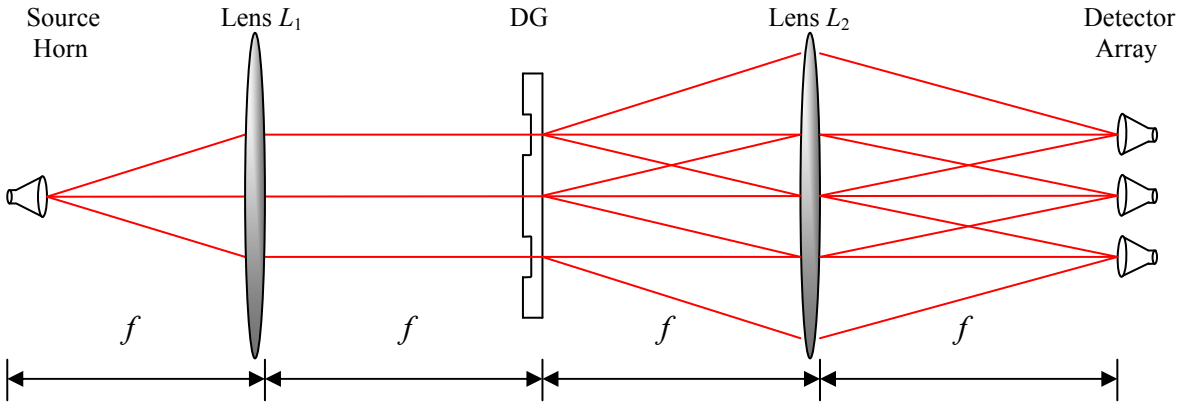


Figure 4-9. Optical configuration with two lenses of equal focal length f for beam multiplexing with a Dammann grating. The grating produces an array of images of the source horn for coupling to an array of horns (on the right).

A typical 4-f Fourier imaging system is shown in Figure 4-9. The system is fed by a single coherent source on the left and propagation occurs from left to right. In the far-infrared and terahertz wavebands sources of coherent radiation are typically horn antennas, which produce quasi-Gaussian illumination patterns rather than uniform illumination of a grating. When fed by a conical corrugated horn antenna the phase grating generates an array of images of the feed horn at the output plane. A horn antenna has a waist position (where the phase front is planar) a distance Δz behind its aperture so the field transmitted from the source can be modelled to a good approximation by a simple Gaussian field distribution

$$|E_S(x, y)| \propto Gauss(x, y; W_S) = e^{-\frac{x^2+y^2}{W_S^2}} \tag{4.37}$$

where W_S is the Gaussian beam radius at the source waist position. The source horn is located such that its waist position is at the focal plane of lens L_1 of the 4-f system The first focusing element, lens L_1, quasi-collimates the source field, thereby illuminating the grating plane with the Fourier Transform of $Gauss(x, y; W_S)$, which of course, assuming ideal collimation by L_1 and correct alignment, is approximated by another Gaussian beam with waist radius W_G

$$b(x, y) = |E_G(x, y)| \propto Gauss(x, y; W_G) = e^{-\frac{x^2+y^2}{W_G^2}} \tag{4.38}$$

where $W_G = \lambda f_1/\pi W_S$. It is vital that the grating be situated at the common focal plane of the two lenses, where the Gaussian beam has a waist position, to permit the assumption that the field incident on the grating has (effectively) no curvature so that the only phase

contribution at the grating plane is due solely to the phase modulation of the grating itself.

The diffraction order spacing $\Delta x'$ (the inter-beam spacing in the output plane) is related to the grating period Δx by

$$\Delta x' = \lambda f_2 / \Delta x \tag{4.39}$$

where f_2 is the focal length of the second (collecting/focusing) lens or mirror. The ratio of output beam spacing $\Delta x'$ to beam radius W_F is determined by how many grating cells are illuminated, since by the convolution theorem if the grating function, $t(x, y)$ is illuminated by the incident field $b(x, y)$ (the collimated horn field) then in the Fourier (far field) plane

$$E(u, v) = \Im\{b(x, y) \cdot t_G(x, y)\} = \Im\{b(x, y)\} \otimes \Im\{t_G(x, y)\} \tag{4.40}$$

Thus

$$E(u, v) = B(u, v) \otimes [T(u, v) \cdot A(u, v)] \tag{4.41}$$

so that each element of the array of diffraction orders is smoothed by the Fourier transform image $B(u, v)$ of the incident field $b(x, y)$. For an incident field $b(x, y)$ with a Gaussian amplitude profile $B(u, v)$ is also Gaussian, i.e.

$$|B(u, v)| = |E(x', y')| \propto \text{Gauss}(x', y'; W_F) = e^{-\frac{x'^2 + y'^2}{W_F^2}} \tag{4.42}$$

where W_F, the Gaussian beam radius at the Fourier plane, is given by

$$W_F = \frac{\lambda f_2}{\pi W_G} = W_S \cdot \frac{f_2}{f_1} \tag{4.43}$$

which implies that the incident Gaussian beam radius (W_G) to cell length (Δx) ratio at the grating is inversely proportional to the Gaussian beam radius (W_F) to inter-beam spacing ($\Delta x'$) in the output plane array of images, i.e.

$$\frac{W_G}{\Delta x} = \frac{1}{\pi} \frac{\Delta x'}{W_F} \tag{4.44}$$

Notice that for a closely packed array of output beams, for example with $W_F/\Delta x' = 0.3$, Eq. (4.44) implies that $W_G/\Delta x \approx 1$, i.e. the radius of the Gaussian beam incident on the grating is on the order of the grating period (cell size) and therefore only a small number of cells are illuminated. In other words, a closely packed beam array can be generated with a grating that has a small number of repeat cells [4.18]. Such arrays of closely packed Gaussian beams find use, for example for feeding a compact array of detector horns with an array of quasi-optically coupled local oscillator (LO) beams fed by a single LO source beam.

4.2.4 Crossed linear phase gratings for 2-D dispersion

Generalising from the one-dimensional beam arrays discussed above, Dammann gratings are used to produce regular two-dimensional arrays of equi-intense output beams that are separable in the transverse directions x and y, i.e. rectangular spot-arrays. The grating phase function $\phi(x, y)$ needed to produce a 2-D spot array is derived from the two corresponding 1-D transmission functions $t_G(x)$ and $t_G(y)$, so that

$$t_G(x, y) = e^{-i\phi(x, y)} = e^{-i\phi_1(x)} \cdot e^{-i\phi_2(y)} = t_1(x) \cdot t_2(y) = e^{-i[\phi_1(x) + \phi_2(y)]} \qquad (4.45)$$

where the one-dimensional phase functions $\phi_1(x)$ and $\phi_2(y)$ generate 1-D arrays of M and N beams, respectively. Equivalently a 2-D grating transmission function $t_G(x, y)$ can be derived by convolving the 2-D unit cell transmission function $t_{cell}(x, y) = t_{cell}(x) \cdot t_{cell}(y)$ by a finite array function $a(x, y)$ of delta functions at intervals of Δx and Δy as

$$t_G(x, y) = a(x, y) \otimes t_{cell}(x, y) \qquad (4.46)$$

This construction was used to produce the transmission function of the 5×5 Dammann grating shown in Figure 4-10 with 4×4 unit cells where the unit cell is defined by the transmission points $x_t = \pm\{0.132, 0.481\}\Delta x$.

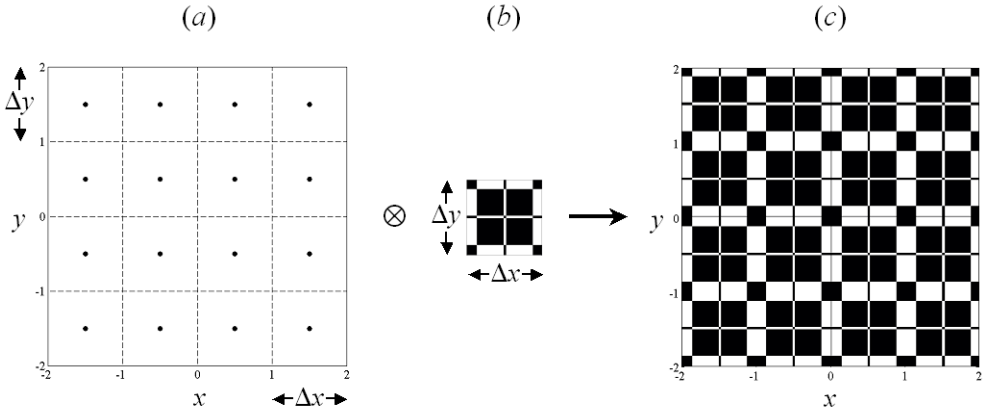


Figure 4-10. The convolution of (*a*) a 2-D array of regularly spaced delta functions $a(x, y)$ – where the positions of the delta functions are shown as black dots – with (*b*) the 2-D unit cell transmission function $t_{cell}(x, y)$ for a 5×5 Dammann grating yields (*c*) the 2-D grating transmission function $t_G(x, y)$ in which the unit cell is repeated at the positions of each delta function.

For completeness the Fourier transform of the 5×5 grating is shown in Figure 4-11. The interference term $A(u, v)$ (a periodic array of sinc functions) is multiplied by the diffraction envelope $T(u, v)$ due to the unit cell to yield a square 5×5 array of sinc functions of uniform intensity. The one-dimensional diffraction efficiency for this

solution, assuming uniform illumination of an infinite number of cells has a quoted value in the literature of 77.5% [4.11] and so a two-dimensional efficiency of ~60%.

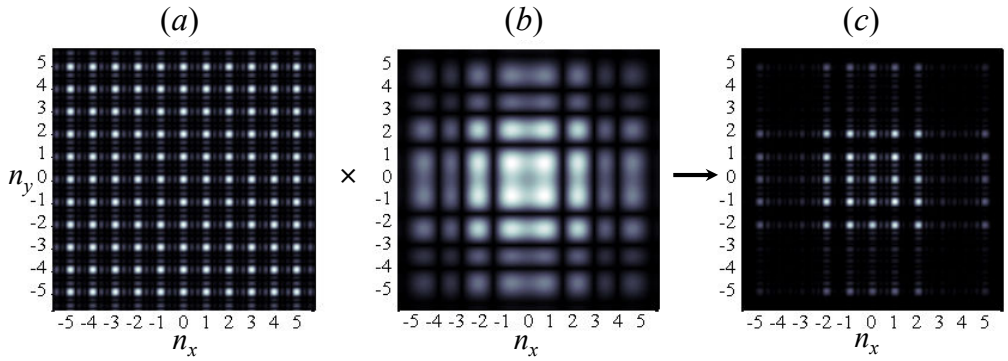


Figure 4-11. (*a*) The interference term $A(u, v)$ when multiplied by (*b*) the diffraction envelope $T_{cell}(u, v)$, (the Fourier transform of the unit cell transmission function $t_{cell}(x, y)$ in Figure 4-10(*b*)) yields (*c*) the Fourier transform of the grating transmission function $t_G(x, y)$ of Figure 4-10(*c*) with most of the power contained in the central 5×5 diffraction orders.

Figure 4-12 shows a profile view of an example 5×5 DG phase profile with unit grating period. The 2-D phase function $\phi(x, y)$ of such a two-dimensional dispersion grating has a checkerboard like design. While the manufacture at visible wavelengths is straight forward, at longer (e.g. mm, sub-mm) wavelengths the vertical phase jumps in discrete-level phase gratings makes exact milling of such structures with sharp concave corners impossible due to the relatively large diameter of the endmill used in the manufacture of these components (see §4.7.2).

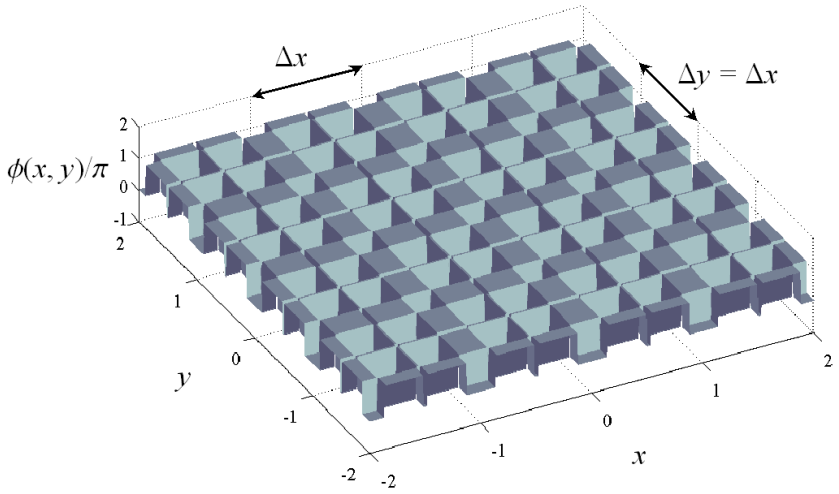


Figure 4-12. Profile view of the 2-D binary phase function for 5×5 DG derived from the with transition points $x_t = \{0.132, 0.481\}\Delta x$. The grating has 4 cells in each direction.

An alternative approach for realising a 2-D grating transmission function $t_G(x, y)$ involves using a *stacked* design. In this arrangement (also referred to as a crossed grating) two-dimensional dispersion is achieved by overlaying in orthogonal directions two separate gratings that impose individually transmission functions $t_1(x)$ and $t_2(y)$. This allows fabrication of a 2-D grating as a combination of two linear gratings[†] [4.23]. The advantages of this method are that firstly machining is made simpler and secondly each linear grating can be characterized independently of the other and so can be combined with other plates to generate different beam array patterns. For example Figure 4-13 shows two linear Dammann gratings, one to generate a 3×1 array of Gaussian beams and the other a 1×4 array, that when stacked orthogonally (such that the grooves of one grating run perpendicular to the grooves of the other) generate a two-dimensional 3×4 spot-array pattern of equi-intense beams. Note however that the diffraction efficiency of the 3×1 linear array is not so high and this is evidenced by the presence of 2nd-order off-axis diffraction orders with non-negligible intensities.

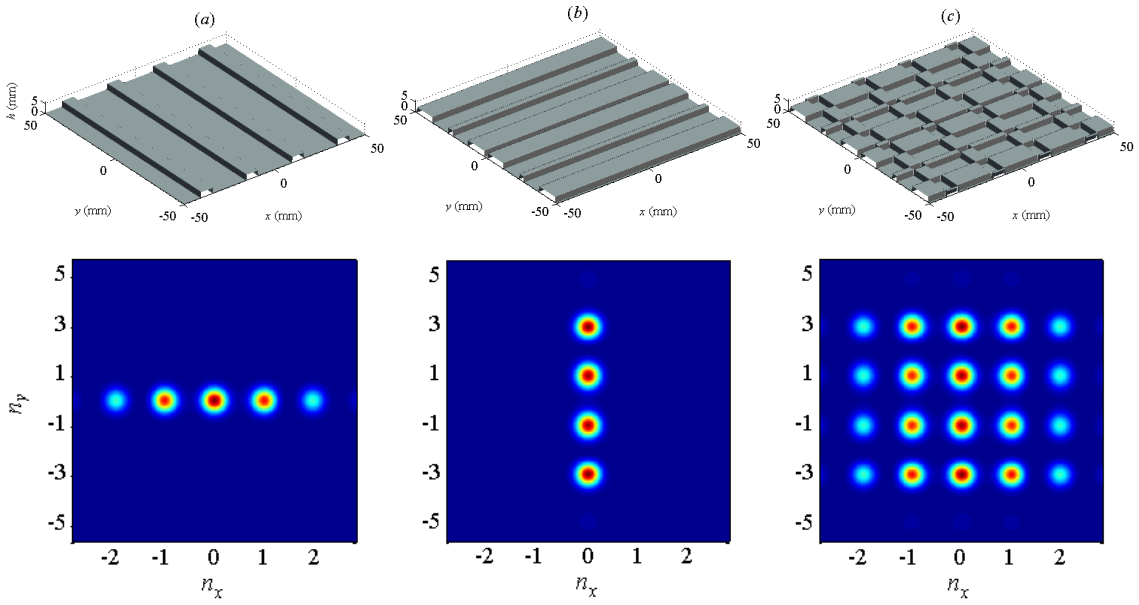


Figure 4-13. The three 2-D binary phase gratings in the upper plots, when illuminated with a collimated Gaussian beam, produce the Fourier plane arrays of Gaussian beams seen below. The two linear 1-D DG's in (*a*) and (*b*) generate arrays of 3 and 4 Gaussian beams. When stacked orthogonally, they produce the equivalent two-dimensional phase grating shown in (*c*) and generate a 3×4 spot-array beam pattern with equal intensities. Notice that the grating in (*a*) has four unit cells, while that in (*b*) has just two. This is to ensure equal diffraction order spacing in *x* and *y*.

[†] The term *linear grating* refers to a two-dimensional grating with periodic grooves in just one direction: *x* or *y*, which therefore generates a one-dimensional line array of diffraction orders in the same direction.

The most tightly packed array of output beams is produced by overlaying two linear gratings such that their grooves cross at an angle of 60° (rather than 90° as above). Killat [4.14] demonstrated this at visible wavelengths with two binary-level phase gratings to generate an array of 35 diffraction orders in a 5×7 formation. Figure 4-14 shows the 35-beam array configuration produced by that particular crossed grating as well as another example of a crossed grating to produce a tightly packed array of 9 beams by overlaying at 60° two identical phase gratings that individually produce linear arrays of three signal orders.

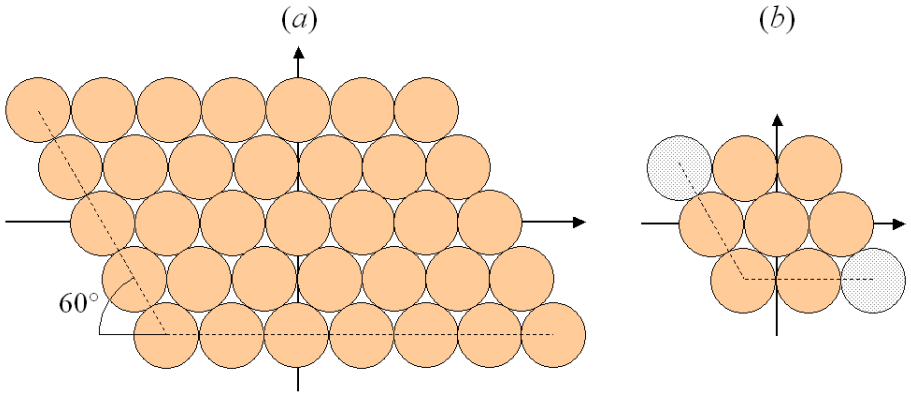


Figure 4-14. Closely packed beam array patterns generated by crossing two linear phase gratings at 60°. The 35-beam array in (*a*) is produced by crossing two linear phase gratings that each produce seven and five beams, respectively. The 9-beam array in (*b*) is produced by crossing two linear phase gratings that each generates an array of 3 beams. This in turn provides a good starting point for finding a solution to generate a hexagonal 7-beam array (as used on the heterodyne array receiver DesertSTAR [4.28]) by eliminating the two beams furthest from the central beam (coloured grey).

4.3 Phase Grating Design: Multivariable Optimisation

For the design of complex beam-splitting gratings to produce more than two/three equi-intense diffraction orders, the form that the grating unit cell takes becomes more elaborate as the number of diffraction order intensities to be controlled increases and so calculations needed to determine the unit cell become very difficult. Optical design and analysis makes prevalent use of optimisation techniques to improve the performance of optical systems [4.29]. Likewise phase gratings such as Dammann gratings, as well as more complicated multilevel phase gratings (discussed in Chapter 5), are usually designed using non-linear multivariable optimisation techniques.

The problem of finding an appropriate phase-only modulation to fulfil the particular criteria of a specific beam-shaping or beam-splitting problem can be regarded as an instance of the phase retrieval problem. Methods to solve the phase retrieval problem can be grouped into two classes: unidirectional and bi-directional techniques. Here we concentrate on unidirectional techniques and we will return to bi-directional algorithms in Chapter 5.

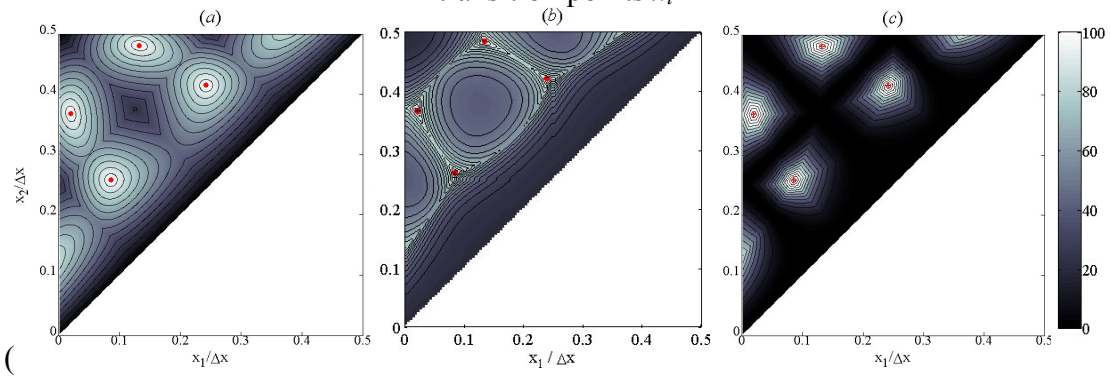
When designing a diffractive grating one must "encode" the unit cell of the grating. The encoding scheme for Dammann gratings assumes two phase levels and varies feature sizes by changing the positions of a discrete number of transition points within the unit cell. The transition points are the free parameters of the system, which act as design variables that can be varied to yield different far field diffraction envelopes and therefore different diffraction efficiencies and standard deviations between diffraction order intensities. Optimisation employs some iterative technique to methodically change transition point locations and evaluate grating performance upon subsequent changes.

For optimisation one must define an objective, or cost function $f(x)$, which is a function of the M parameters $\{x_1, x_2, \ldots x_M\}$ of the system. The only constraint on the form and content of the objective function is that the fitness value returned by it is in some manner proportional to the "desirability" of a given trial solution (the set of parameters) that is input to the objective function. The objective function is the only link between the physical problem being optimised and the optimising routine. In the case of a Dammann grating with reflection symmetry (see §4.5.1) that is required to generate an array of $2N+1$ equi-intense spots each solution is characterised by a set of N independent transition points. The goal of Dammann grating design is to find a binary phase-only element that generates an array of equi-intense diffraction orders. Thus an appropriate cost function to optimise would be some measure of beam uniformity (standard deviation or uniformity). After the cost function is chosen an optimisation routine is selected and used to search for a solution (a set of transition points) that yields the minimum value of that cost function.

4.3.1 Optimisation with Direct Search Methods

Optimisation routines can be divided into classical gradient-based algorithms and direct search algorithms. The former rely on the objective function being differentiable. When this is not the case direct search methods are used.

The simplest of these "generate-and-test" methods is a *brute force search*, which involves simply evaluating the objective function at all grid points in a bounded region and storing the current best point (solution). Applying the brute force search method to the problem of finding a Dammann grating solution to generate $(2N+1) = 5$ equi-intense diffraction orders is trivial since each solution is parameterised by just $N = 2$ transition points $x_t = \{x_1, x_2\}\Delta x$. The standard deviation σ, mean power uniformity and uniformity U of diffraction order intensities were evaluated for all (allowed) permutations of transition points x_t



(Figure 4-15). Transition point locations are subject to the constraint that $x_i \leq x_{i+1}$ so valid solution vectors are restricted to the upper portion of the 2-D solution space (above the line $x_2 = x_1$). Although the surface defined by each objective function is different the maxima/minima occur at the same positions.

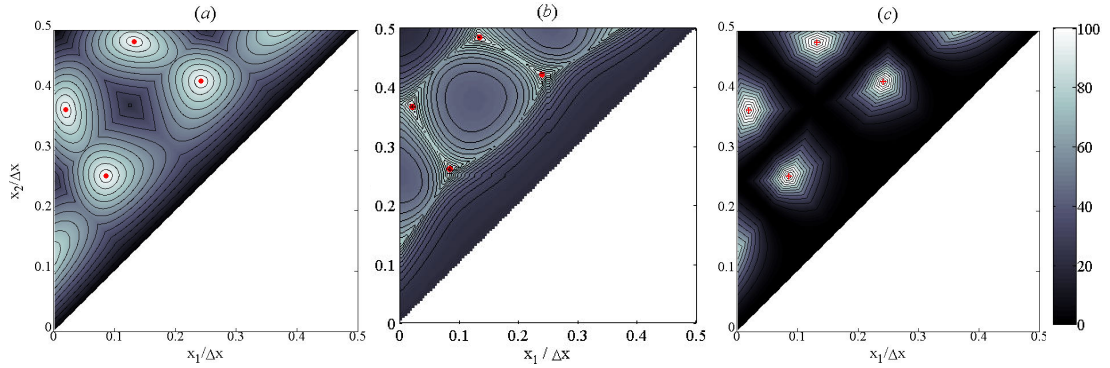


Figure 4-15. Results of brute force search for Dammann grating solutions to generate five equi-intense diffraction orders. The objective functions shown indicate the degree of uniformity between the five central diffraction order intensities for each solution point, or vector, $\{x_1, x_2\}\Delta x$ that characterises a single symmetric Dammann grating. The three objective functions are (*a*) standard deviation, σ, between diffraction order intensities, (*b*) mean power uniformity (MPU) and (*c*) uniformity U as defined by Eq. (4.31). Note that for visualisation $(1-\sigma)$ is plotted in (*a*) so maxima (high-valued regions) indicate solutions with high uniformity. The red markers indicate the four different solutions obtainable with a symmetric Dammann grating characterised by two transition points.

Note that in

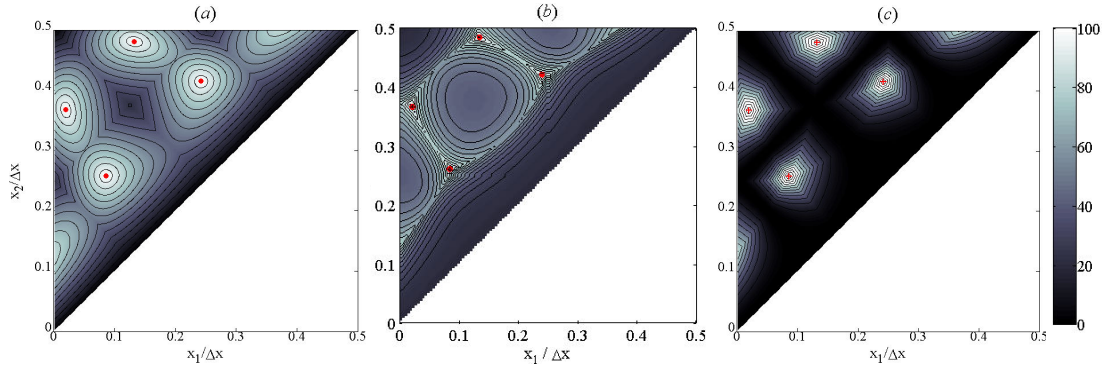
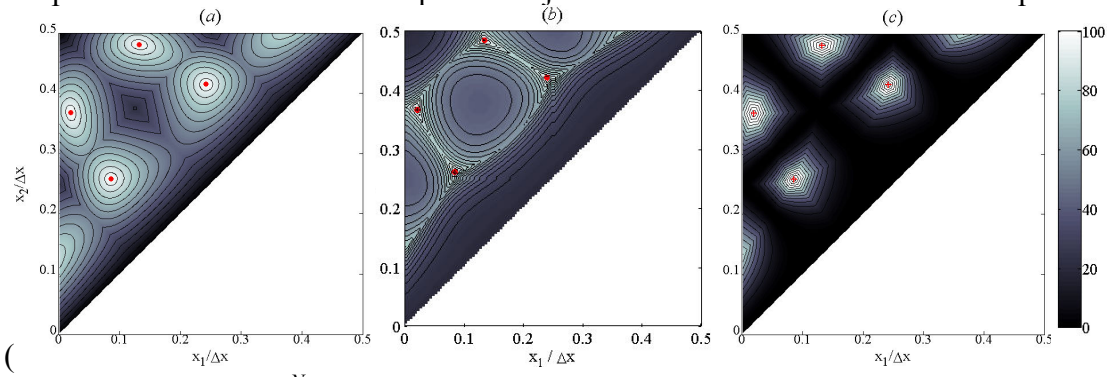


Figure 4-15(*a*) the objective function used is the standard deviation of the magnitudes, rather than intensities, of the diffraction orders, i.e.

$$\sigma = \sqrt{\frac{1}{2N+1}\sum_{n=-N}^{n=+N} |A_n - \langle A \rangle|^2}$$

As stated by Jahns *et al* [4.9], since a Dammann grating has N independent transition points and because of the quadratic relationship between diffraction order intensities and amplitudes ($I_n = |A_n|^2$) more than one solution exists. For Dammann gratings (with 0 and π phases) the diffraction orders are all real-valued and allowed to have positive or negative signs. Furthermore for binary gratings $A_{+n} = A_{-n}$, so there are 2^N possible solutions. For example the objective functions for $N = 2$ transition points



(Figure 4-15) contains $2^N = 4$ maxima. These four solution points, or vectors, correspond to the four known solutions for a symmetric Dammann grating to generate five equi-intense diffraction orders. A function, or solution space, with more than one solution is called multi-modal (whereas a function with one solution is called uni-modal). The goal of multi-modal function optimisation is to find the global maximum/minimum from amongst all other local maxima/minima.

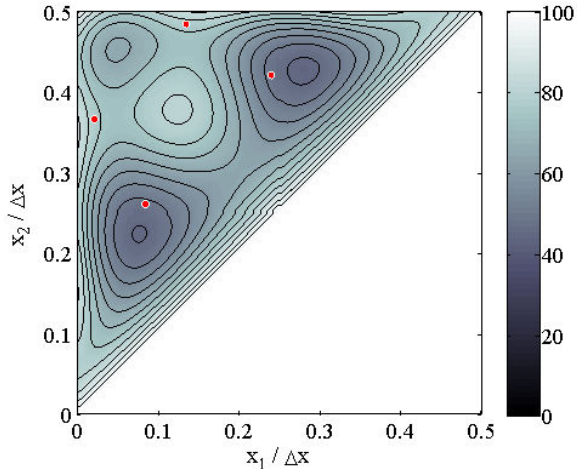


Figure 4-16. Diffraction efficiency η evaluated for all trial solution vectors $x_t = \{x_1,x_2\}\Delta x$. The four red markers are the four solution vectors identified as producing the most uniformly intense array of equi-intense diffraction orders. The two solution vectors closest to the left and upper borders yield higher diffraction efficiencies than the other two vectors and are therefore the global solutions to this problem.

The other requirement of phase grating design is that diffraction efficiency must be maximised. This criterion can be used to differentiate between multiple solutions that appear to be equally good solutions. For example in the 5-beam Dammann grating example, if we now calculate the diffraction efficiency associated with the four solution vectors that were identified in

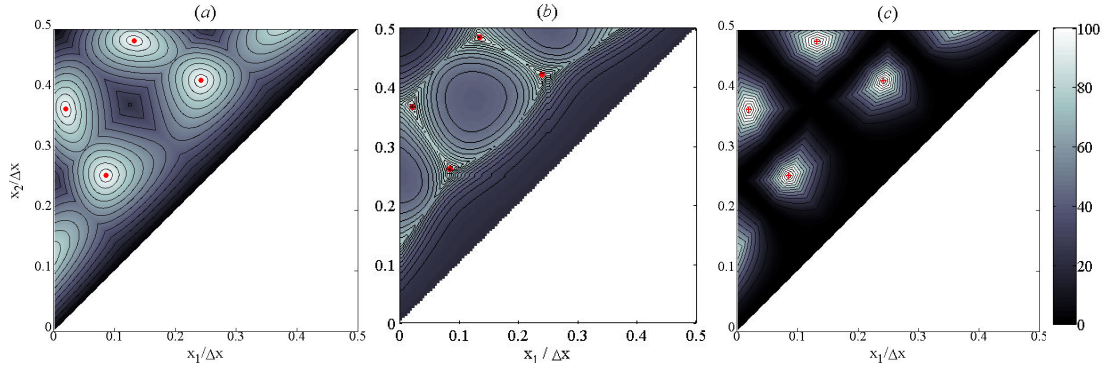


Figure 4-15 the solution with the highest diffraction efficiency is then the global maximum. Figure 4-16 shows the diffraction efficiency calculated for all vectors $\{x_1, x_2\}$ for the 5-beam Dammann grating problem. The four solution vectors resulting in maximum beam uniformity are superimposed and show that the two vectors nearest the line $x_2 = x_1$ produce an array of diffraction orders with much lower diffraction efficiency (48.3%) than the two vectors nearest the $x_1 = 0$ and $x_2 = 0.5\Delta x$ (77.5%). Thus the latter are referred to as the global solutions for the problem of designing a symmetric Dammann grating to produce an array of five equi-intense signal orders.

Notice in Figure 4-16 that the maximum value of diffraction efficiency occurs for vector solutions close to the line $x_2 = x_1$. Solution vectors on the line $x_2 = x_1$ correspond to a unit cell with no transition points and all of the transmitted power will

be diffracted into a single on-axis peak and therefore unit diffraction efficiency is achieved. Clearly however such solutions are not good beam-splitting solutions since we also require that diffraction order intensity be evenly distributed, which clearly it is not for solutions where x_2 has a value close to x_1. The point to note is that diffraction efficiency cannot be used as the sole measure of beam-splitting quality, but it is useful for differentiating between solutions that generate diffraction order arrays with equal beam uniformity.

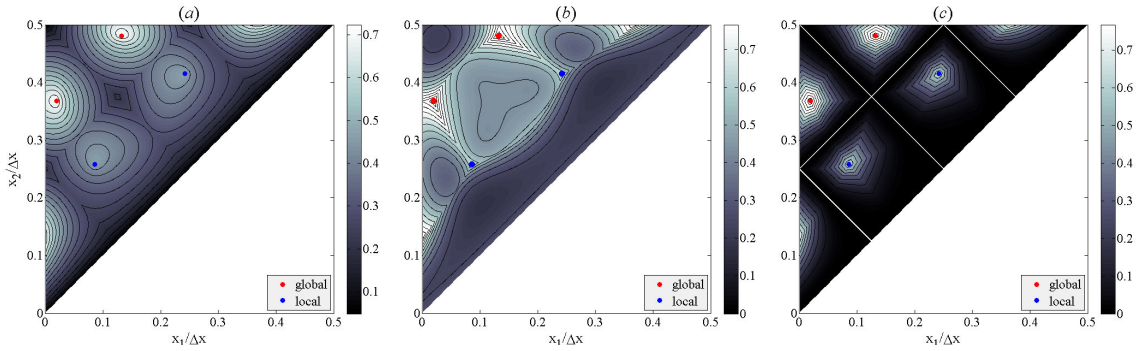


Figure 4-17. Objective functions for the 5-beam Dammann grating solution vectors $x_t = \{x_1, x_2\}\Delta x$. Each plot corresponds to a different objective function $f(x_t)$ that weights the diffraction efficiency η by one of the three measures of beam uniformity. (*a*) $f(x_t) = \eta(1-\sigma)$, (*b*) $f(x_t) = \eta \cdot MPU$ and (*c*) $f(x_t) = \eta \cdot U$ the objective function used is diffraction efficiency, η weighted by standard deviation, σ, i.e. $f(x_t) = \eta(1-\sigma)$. The white lines in (*c*) represent boundaries between the four basins of attraction (see text below) that lead to the four local solutions. The red and blue markers indicate the global maxima and local maxima, respectively.

If numerical optimisation is used to solve the 5-beam Dammann grating problem with an objective function measuring beam uniformity it is possible that the solution found is one of the two with low diffraction efficiency. Thus the two criteria (high beam uniformity and high diffraction efficiency) can be combined into a single objective function to ensure the optimisation routine seeks out solutions that simultaneously satisfy both criteria. For example combining diffraction efficiency with standard deviation to results in the objective function $f(x_t) = \eta\,(1-\sigma)$, as shown in Figure 4-17(*a*). Or combining η with *MPU*, gives an objective function $f(x_t) = (\eta \cdot MPU)$, as shown in Figure 4-17(*b*).

The problem with a brute force search is that if nothing is known about the objective function it is difficult to decide on grid sampling. If the objective function is sampled too coarsely the optimum point may be missed. However a high sample rate results in execution time exploding, since sampling each of *D* dimensions (parameters)

with N points results in N^D grid points. Thus brute force is only used for objective functions with a small number of parameters. More sophisticated algorithms are required for objective functions with higher dimensionality.

4.3.2 Deterministic Algorithms

Direct search optimisation routines can be grouped into deterministic or nondeterministic algorithms. Deterministic algorithms are those in which the solution that the routine converges on is determined by the starting point that the search begins at. They are known also as local optimisers because due to the "greedy" selection used to choose optimisation paths they only find the best solution in their immediate locality. A *basin of attraction* [4.30] refers to a group of vectors that when used as starting points by a deterministic search algorithm all result in algorithm locating the same local optimum solution. For example the white lines in Figure 4-17(*c*) represent the boundaries between basins of attraction of the four solutions for the 5-beam DG problem for the objective function $f(x_t) = \eta U$.

Although deterministic algorithms are guaranteed to converge on a local solution (generally, in a short time) there is no guarantee that the solution found will be the global solution. Examples of deterministic search algorithms include the Hooke-Jeeves method [4.31] and "simplex" method of Nelder and Mead [4.32]. The latter optimises a function of n variables by employing a (n+1)-dimensional polyhedron, or simplex, in which each of the n+1 vertices corresponds to a unique trial vector. Four operators (reflection, expansion, contraction and shrinkage) are applied to adapt the shape of the simplex to the local landscape of the function in order to locate a minimum. The reflection operator reflects the highest-value vertex through the centroid of the other vertices. An improved version of the Nelder-Mead (NM) method [4.33] uses objective function values at the vertices to make an informed choice about the point through which reflections occur. The advantage of the NM method over the Hooke-Jeeves method, which uses only decreasing step sizes during a search, is that the simplex can expand as well as contract thus allowing the step size to adapt to the local topography of the objective function.

The NM method available in MATLAB, which is based on Ref. 4.34, was used to find the four solutions for the 5-beam Dammann grating problem. Figure 4-18 shows the paths taken by the NM algorithm from four different starting points. Each starting

point was chosen specifically to be in a different basin of attraction so as to illustrate how each one leads to a different local solution. The random walk method is not technically deterministic since because it accepts random search directions there is a finite possibility that the solution vector found will be in a different basin of attraction than the search began with.

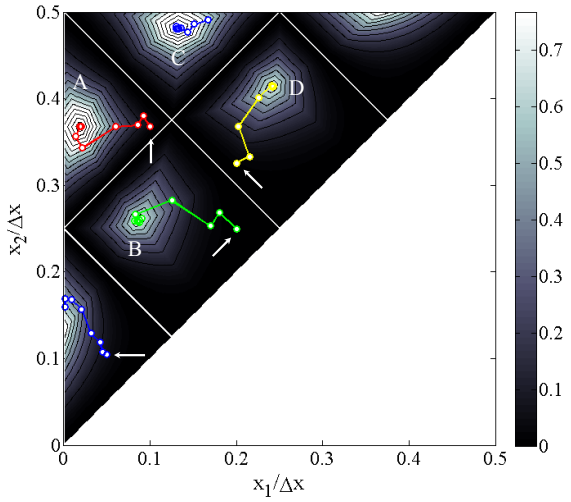


Figure 4-18. Nelder-Mead algorithm applied to four starting points (white arrows) in the 5-beam Dammann grating problem. The objective function was $f(x_t) = \eta \cdot U$. Each starting point is in a different basin of attraction and therefore leads to a different solution (labelled A, B, C and D).

The starting point problem (referring to the tendency of "greedy" optimisers to find only local optimum solutions) implies that in order to locate the global optimum with a deterministic algorithm, the objective function must be sampled in the vicinity of the global optimum solution. A simple means of increasing the chances of finding the global optimum with a local optimiser is to use a multi-start technique. As the name suggests, this involves starting a local optimisation routine from many different starting points. The simplest implementation is to choose starting points at random and for a large enough number of randomly chosen starting points we will necessarily find the global optimum [4.10]. The problem is that without having knowledge of the objective function, it is difficult to know how many starting points are required, since many of the starting points may be in the same basin of attraction and therefore lead to the same local minimum. A multi-start algorithm can be modified by applying a clustering algorithm to identify starting points that belong to the same basin of attraction or cluster. Local optimisation is then applied to a single point in each cluster. Because of high computational requirements clustering algorithms are usually limited in applicability to problems with a small number of parameters.

A multi-start NM algorithm was used to search for Dammann grating solutions to produce an array of seven equi-intense diffraction orders. Solutions are characterised by the positions of $N = 3$ transition points, $x_t = \{x_1, x_2, x_3\}$, thus we expect there to be $2^N = 8$ solutions. For this problem, 1000 optimisation searches (with up to 500 iterations) were performed. Each search was initiated with a randomly chosen starting point and the objective function chosen was beam uniformity U. The weighted diffraction efficiency $\eta_{weight} = \eta \cdot U$ was calculated for each of the 1000 endpoints and are plotted in Figure 4-19. We note that ~70% of all searches converged on one of the eight known solutions (as calculated by Dammann and Klotz[4.12] as well as by Heanue [4.35]) to this problem. The remaining searches converged on local optima in the 3-D solution space with much lower values of beam uniformity. An important consideration in grating design is that the grating should be easily machined thus solutions in which transition points are closely spaced should not be considered. In Figure 4-19 the solutions are colour coded according to minimum transition point separation: solutions marked in blue have a minimum transition point spacing, $\min\{x_{i+1}-x_i\} \leq 0.05\Delta x$ and solutions marked in red have $\min\{x_{i+1}-x_i\} > 0.05\Delta x$. Notice that the eight known solutions (with nearly perfect beam uniformity) all have transition point spacing greater than $0.05\Delta x$.

Another means of optimising symmetric Dammann gratings is to use a multidimensional error feedback algorithm [4.36]. Because this method does not define an objective function it does not introduce additional local minima and thereby avoids entirely the starting point problem.

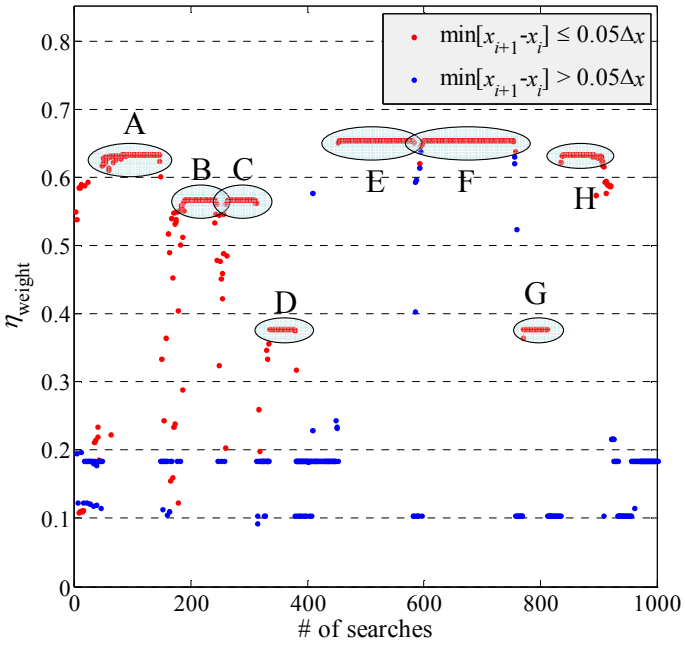


Figure 4-19. Weighted diffraction efficiency of Dammann grating solutions to produce seven equi-intense diffraction orders. The solutions were found using a multi-start search with the Nelder-Mead algorithm. Solutions are sorted by location of the first transition point x_1. The eight known solutions are encircled and labelled A to H. Solutions are divided into those with a minimum transition point separation above (red) and below (blue) $0.05\Delta x$, where Δx is the grating period

4.3.3 Nondeterministic Algorithms

Conventional optimization techniques, i.e. those based on greedy, local optimisation routines, are poorly suited to problems involving high dimensional, multimodal objective functions [4.37]. Nondeterministic, or stochastic, algorithms sample the objective function more thoroughly so as to avoid converging on local optima. Simulated annealing (SA) [4.38] is analogous to annealing in metals. It modifies a random walk search by allowing the routine to accept some poorer short-term results which gives the algorithm a chance to move from one basin of attraction to another. Although originally proposed as a combinatorial optimiser (for objective functions defined by discrete parameters) SA has since been modified to allow for optimisation of functions of continuous variables [4.39]. SA was employed, in combination with damped least squares algorithm, by Turunen *et al* [4.40] in the design of beam-splitting Dammann gratings for arrays of up to 53 equi-intense signal orders. It has also been used to design computer-generated holograms [4.21,4.41] as well as continuous-phase gratings (kinoforms) [4.42]. Stochastic methods were also used in the design of multilevel phase gratings by Barnes *et al* [4.15].

The other type of nondeterministic direct search algorithms are population-based optimisers and are collectively referred to as evolutionary algorithms (EA). They include *differential evolution* (DE), *evolution strategies* (ES) and *genetic algorithms* (GA) and are based on the Darwinian notion of natural selection and evolution. An initial population of trial starting vectors (or base points) are randomly chosen and the algorithm attempts to "evolve" the population in parallel towards an optimal solution under the "selective pressure" of the objective function. As with multi-point algorithms, an EA tackles the starting point problem by creating an initial population of randomly chosen points/vectors. But whereas multi-start algorithm optimises base points independently of each other, members of an EA population interact with each other (through breeding, mutation and selection operations) to generate a new generation of vectors. EA's have been successfully applied to a wide variety of optical design problems. For example GA have been used for the synthesis of shaped beam antenna patterns for linear arrays [4.37], the design of refractive beam shaping elements [4.43], lenses to achieve specific resolution and distortion requirements [4.44, 4.45], as well as lightguides for a clinical diagnostic instrument [4.46]. DE algorithms have been used for automated mirror design [4.47].

Although stochastic algorithms such as SA and GA provide the best solutions for phase grating design [4.8] this performance comes at a heavy computational cost in terms of execution time because of the large number of function evaluations needed. GA requires large population sizes to ensure greater diversity so that the solution space is adequately sampled, while SA must reduce the probability of accepting poor solutions very slowly to avoid deterministic behaviour. For example SA is characterised by the need to perform about three orders-of-magnitude times the number of function evaluations as is typically required of local optimisation algorithms [4.39]. Next we examine how symmetries can be used to reduce computational complexity in grating design.

4.4 Symmetry Considerations in Phase Grating Design

In this section we extend the discussion on the design criteria for millimetre-wave gratings presented in [4.25], including symmetry consideration in the discussion and report on results in the literature for phase gratings at visible wavelengths. The number

of transition points required to generate a specific beam pattern using a binary-phase grating grows with the number of signal orders, N. Typically each specified diffraction order requires one independent parameter in the design. Thus a $1\times N$ spot array requires on the order of N parameters to describe the grating solution if we require that all diffraction order intensities be adjusted independently. Dammann gratings are computed by nonlinear optimisation techniques and it has been shown by Jahns *et al* [4.9] that the computational complexity (the time taken to find a reasonable solution) grows exponentially with N. If the surface is limited to a periodic regular binary-phase structure the intensities of each positive-negative pair of diffraction orders ($\pm n$) are equal so only approximately $N/2$ transition points are needed. This is one consideration that reduces the complexity of the problem by reducing the number of phase transitions. Another means of substantially reducing the complexity of phase grating design is to incorporate, where applicable, reflection and/or translational symmetry into the grating design [4.48].

4.4.1 Reflection symmetry

The primary objective in designing a regular rectangular spot array is to ensure that all the generated orders have equal intensity. Although Dammann gratings do not in fact require special symmetry considerations to produce symmetric diffraction patterns, for simplicity we choose to illustrate the inclusion of reflection symmetry in grating design as it applies to binary phase grating design.

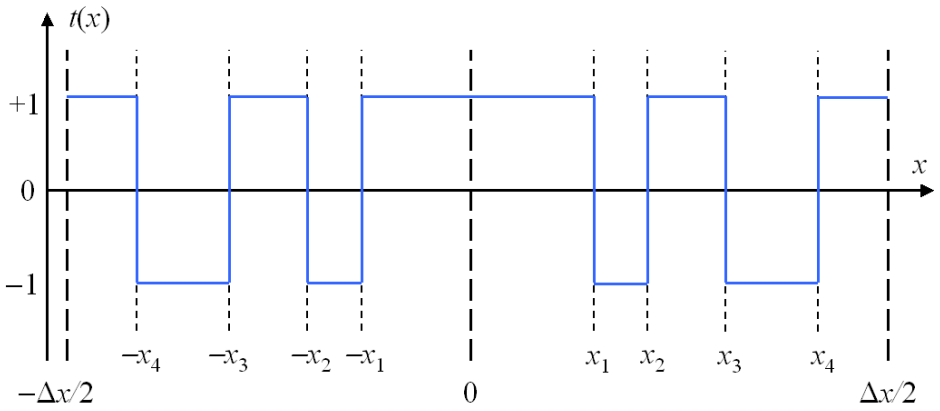


Figure 4-20. A one-dimensional binary transmission function $t(x)$ incorporating reflection symmetry, with period Δx and $M = 4$ transition points.

The reflection of $t(x)$ about the midpoint of the basis cell guarantees that each pair of order $\pm n$ (positive-negative order pair) have matching intensities since the Fourier Transform of any even function is itself even [4.49]. Figure 4-20 shows a single period (the unit cell) of a binary grating with reflection symmetry imposed about the cell midpoint (here taken to be at $x = 0$). The diffraction pattern (envelope) from a single period of a one-dimensional phase-only transmission function $t(x)$ is

$$T(u) = \int_{-\Delta x/2}^{\Delta x/2} t(x) \cdot e^{-i2\pi(ux)}\, dx = \int_{-\Delta x/2}^{\Delta x/2} e^{-i\phi(x)} \cdot e^{-i2\pi(ux)}\, dx \tag{4.47}$$

where the phase modulation of the unit cell is $\phi(x)$. Since we are dealing with a phase-only modulation we integrate across the entire unit cell (from $x = -\Delta x/2$ to $+\Delta x/2$). Imposing reflection symmetry about the centre of the unit cell implies a grating period transmission function $t(x) = t(-x)$ with transition points $x_t = \{\pm x_1, \pm x_2, \pm x_3, \ldots \pm x_m\}$. In other words only information (the position of transition points) on a single half of the unit/basis cell is required. The previous integral can be expressed as

$$T(u) = \int_{-\Delta x/2}^{0} e^{-i\phi(x)} \cdot e^{-i2\pi(ux)}\, dx + \int_{0}^{\Delta x/2} e^{-i\phi(x)} \cdot e^{-i2\pi(ux)}\, dx \tag{4.48}$$

$$T(u) = 2\int_{0}^{\Delta x/2} e^{-i\phi(x)} \cdot \cos(2\pi ux)\, dx \tag{4.49}$$

since $e^{-i\phi(x)}$ is an even function of x. The alternating phase term $e^{-i\phi(x)}$ is dependent only on the transition points x_t and so the above can be written as

$$T(u) = 2\sum_{m=0}^{M} (-1)^m \cdot \int_{x_m}^{x_{m+1}} \cos(2\pi ux)\, dx \tag{4.50}$$

where for $m = 0$, $x_0 = 0$, $x_{M+1} = \Delta x/2$, and where the term $(-1)^m$ term takes into account the phase change that occurs at each transition point. The above equation implies that the diffraction pattern (T as a function of u) is given by

$$T(0) = 2\sum_{m=0}^{M} (-1)^m \cdot (x_{m+1} - x_m) \tag{4.51}$$

$$T(u) = \frac{1}{2\pi u} \sum_{m=0}^{M} (-1)^m \cdot \left[\sin(2\pi u x_{m+1}) - \sin(2\pi u x_m)\right] \tag{4.52}$$

Thus the zeroth-order ($u = 0$) beam has a functional dependence different from the remaining, off-axis ($u \neq 0$) output beams. The position of the diffraction orders n are of course determined by the grating equation (array function), which for a unit cell width

of Δx is $n\lambda = \Delta x \sin\theta_n$. Therefore in terms of spatial frequency coordinates u, the diffraction orders are located at discrete frequencies $u_n = \sin\theta_n/\lambda = n/\Delta x$, as expected.

The inclusion of reflection symmetry reduces the number of parameters needed to characterise the grating design and so reduces the complexity of the optimisation process used to find solutions. For Dammann gratings, where the solution is parameterised by a set of transition points and a phase difference of $\Delta\phi = \pi$, the result is that only half of the transition points of the unit cell are independent and so only half need to be optimised to give the required output pattern. To produce an array of $N = 2M+1$ diffraction orders of equal intensity requires a grating period with the same number of phase levels. This translates to $2M$ transition points per grating period, but if reflection symmetry (about grating cell midpoint) is imposed the number of independent positions for the transition points between $x = 0$ and $+\Delta x/2$ (the cell edge) is reduced to M [4.11]. For example to produce an array of 5 equally intense output beams, only 2 independent positions for transition points (x_1 and x_2) are required to characterise the basis cell. Numerous Dammann grating solutions with reflection symmetry to produce odd-numbered beam arrays have been reported in references [4.48, 4.12, 4.51, 4.35].

The intensity of the n^{th} diffraction order is $I_n = A_n^2$, where A_n is its amplitude. Because of this quadratic relationship between intensity and amplitude when searching for a maximum diffraction efficiency η there exists more than one solution to the problem of producing an array of $2M+1$ signal orders, since the diffraction orders are real-valued and are permitted a positive or negative sign (or may even be complex, in the more general case). Theoretically, mathematical considerations predict that for M parameters, the number of possible solutions, S_M is 2^M [4.35]. According to Dammann however, the number of "essentially different" solutions is $2^{(M-1)}$ [4.12]. The difference in number of solutions is due to the fact that many of the solutions are equivalent to another solution. Consider for example the problem of generating an array of five equi-intense diffraction orders. In this case $M = 2$ so there are $2^2 = 4$ possible solutions. Two of these solutions are [4.12,4.35]

$$S_1 = \pm\{0.019, 0.368\}\Delta x\ ,\ \ S_2 = \pm\{0.132, 0.481\}\Delta x$$

The transition points of these two solutions are related through $S_1 = \Delta x/2 - S_1$ (the order of transition points within each set being irrelevant). In effect, the transition points of S_2 are those of S_1 translated by half a grating period, $\Delta x/2$. This point is made clearer if we consider a grating composed of several cells (right hand side of Figure 4-21). We see

that the grating structure derived from solution S_2 is the same as that derived from solution S_1 after translation by $\Delta x/2$. Thus $2^M = 2(2^{M-1})$ or 2^{M-1} are equally valid answers to the number of possible solutions available, depending on whether or not such equivalent, or degenerate, solutions are included in the tally.

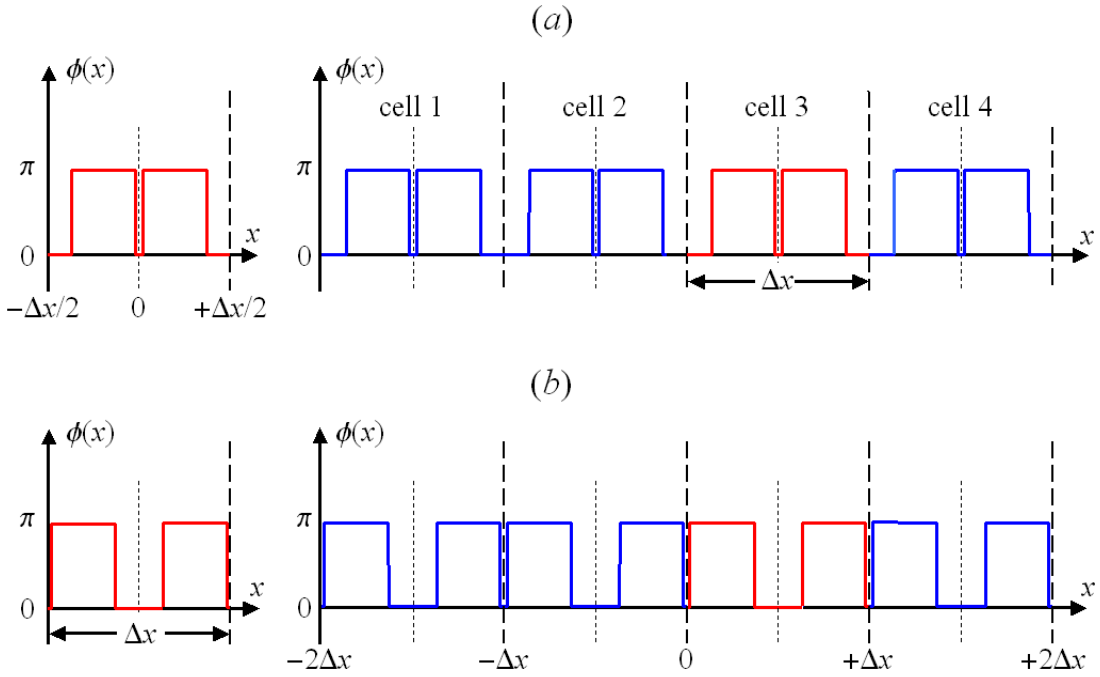


Figure 4-21. Equivalent one-dimensional DG solutions to generate five equi-intense diffraction orders. **Left:** unit cell. **Right:** a 4-cell grating structure. The solutions in (*a*) and (*b*) correspond to solutions S_1 and S_2 respectively (see preceding text). Since an even number of cells is used, the unit cell is positioned with its leftmost edge at the grating centre ($x = 0$). Clearly, however, if the unit cell in (*a*) is centred, i.e. shifted by $\pm\Delta x/2$ from its current position, the resulting grating structure is that shown in (*b*). Therefore the two basis cells shown are just shifted versions of each other. Hence solutions S_1 and S_2 are equivalent.

The far field diffraction patterns produced by the two 4-cell phase gratings of Figure 4-21 are shown in Figure 4-22. The intensities of the 5 central orders from each grating are (theoretically) identical. Since the grating consists of four cells there are two faint secondary maxima between adjacent principal maxima. The only significant difference between the two diffraction patterns in Figure 4-22 is the relative intensity distribution between secondary maxima, which is due to the difference in $T(u)$ that exists between diffraction orders (see Figure 4-7). If the grating consists of a large number of cells the influence of the diffraction envelope on the difference in secondary maxima intensity becomes less pronounced. The effective equivalence of solutions S_1 and S_2 is also evidenced by the fact that the quoted diffraction efficiencies for both are 77.5%.

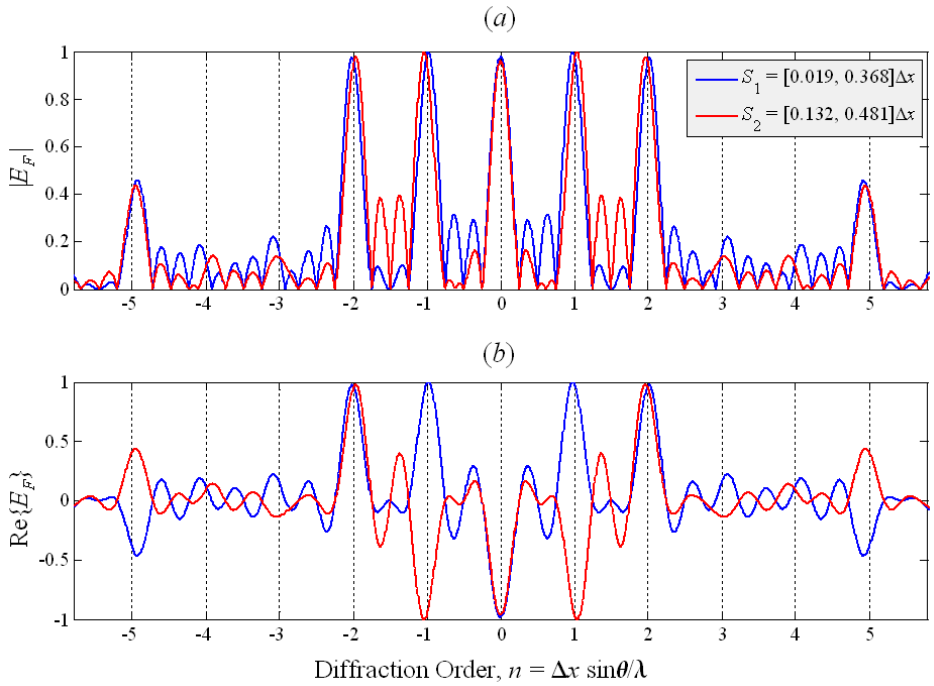


Figure 4-22. (*a*) Amplitude and (*b*) real-valued far field diffraction pattern produced by the 4-cell DG's in Figure 4-21 when uniformly illuminated. The plots in (*a*) are the same except for distribution of power amongst secondary maxima. The plots in (*b*) show that polarity of diffraction orders are different for each solution: even-numbered diffraction orders (n = 0, 2) have the same polarity; odd-numbered orders (n = 1, 3, 5) are opposite in polarity. Since diffraction efficiency calculalations involve only diffraction order intensities, polarities of orders are not recognised. Thus in terms of diffraction efficiency the solutions S_1 and S_2 are identical.

Although multiple solutions may exist for a given problem typically the different solutions will yield different efficiencies, due to the amount of power in higher diffraction efficiencies beyond orders ±M. For example the other two solutions for M = 2 are

$$S_3 = \pm\{0.086,\ 0.258\}\Delta x\ ,\quad S_4 = \pm\{0.242,\ 0.415\}\Delta x$$

and the quoted diffraction efficiencies in the literature for these two solutions are only 48.3% [4.12,4.35]. Although these two solutions also produce an array of five equi-intense diffraction orders, they also generate diffraction orders at n = ±3 that are nearly twice as intense as the orders $|n| \le 2$, hence the low diffraction efficiency.

Although Dammann [4.11,4.12] did employ reflection symmetry when searching for solutions to generate odd-numbered spot arrays it was noted in [4.48] that because binary phase gratings automatically generate equally intense positive- and negative-order diffraction spots ($I_{-n} = I_{+n}$) reflection symmetry is not actually required in binary-

level phase gratings. Indeed binary-level solutions without reflection symmetry have been found to yield higher efficiencies, as reported by Killat *et al* [4.14]. For example using Dammann's method to solve the problem of generating 7 and 9 equi-intense diffraction orders yields maximum diffraction efficiencies of 65.7% and 66.3%, respectively [4.12], whereas solutions obtained by Killat *et al* to the same problems yielded efficiencies of 84.5% and 80.2%. A Dammann grating defined with reflection symmetry requires $M = n_{max}$ transition points to generate $N = 2n_{max}+1$ diffraction orders (n_{max} being the highest-order spot). However without reflection symmetry the number of transition points depends on whether n_{max} is even or odd [4.14] as follows

$$M = \begin{cases} (n_{max}-1) + 1 = n_{max} & n_{max} \text{ odd} \\ n_{max}+1 & n_{max} \text{ even} \end{cases} \tag{4.53}$$

Besides removing reflection symmetry from the unit cell the other difference in the binary grating solutions to generate odd-numbered diffraction order arrays that were reported in [4.14] is the use of a non-π phase difference between the two phase levels in the binary grating.

Since multi-level and continuous-level (Fourier) phase gratings do not automatically generate equally intense positive/negative diffraction order pairs, reflection symmetry is needed in the design of all non-binary phase gratings. It will be seen in Chapter 5 that for multilevel gratings $M = (n_{max}+1)$ transition points are needed to produce an array of $2n_{max}+1$ diffraction orders. When using Gaussian Beam Mode Analysis for grating design reflection symmetry is achieved by restricting the choice of mode set to one containing only even-numbered modes (as described in §2.8).

4.4.2 Translation symmetry for even-numbered beam arrays

In general we would expect a symmetric grating to produce an on-axis maximum and pairs of diffraction orders on either side with equal intensity. This implies that for a grating to generate a symmetric array with an even number of equi-intense spots, the diffraction pattern must contain alternating high-intensity odd-numbered and suppressed even-numbered diffraction orders (including the on-axis zeroth order). Normally, in binary phase gratings, as well as multilevel gratings exhibiting reflection symmetry, the central diffraction order ($n = 0$) has a functional dependence on the grating function different from that of the other orders (viz. Eq. 4.51), which results in greater sensitivity of the zeroth order to errors in phase depth [4.9]. By designing a grating such that all

even-numbered orders n = {0, ±2, ±4…} are suppressed and only odd-numbered orders n = {±1, ±3, ±5…} remain, this sensitivity is no longer such an issue [4.48]. Thus a grating designed to generate an even-numbered spot array is a more stable design than one to generate an odd-numbered spot array.

The earliest attempts to find binary phase grating solutions that could produce even-numbered beam arrays (apart from the trivial case of splitting a single beam into two beams) appear to be due to *Killat et al* in 1982 [4.14], in which solutions for array sizes up to N = 28 were reported. However there were several problems with the approach used. Firstly, the diffraction orders were not equally spaced, since even values of N were achieved by searching for solutions in which only the zeroth diffraction order was suppressed. For example an array of six beams consisted of diffraction orders n = {±1, ±2, ±3}, so the two central beams (n = ±1) were separated by a distance twice that of any other pair of neighbouring beams. Secondly, the solutions obtained were (like Dammann gratings) sensitive to errors in phase depth since the n = 0 diffraction order could only be fully suppressed when the phase difference $\Delta\phi$ was set exactly equal to π. The zeroth-order diffraction spot sensitivity was so great in fact that the solutions that generated even-numbered arrays were subsequently used as the basis of solutions to generate odd-numbered arrays by simply altering the value of $\Delta\phi$ appropriately.

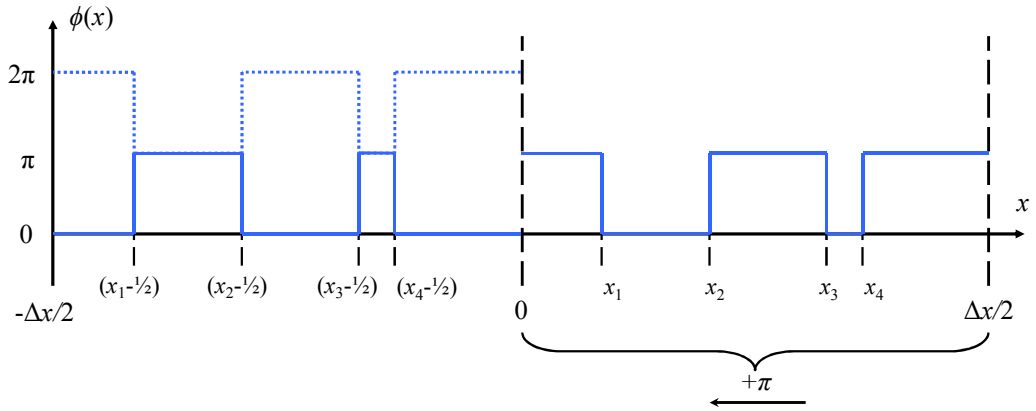


Figure 4-23. The binary phase profile $\phi(x)$ with skew symmetry is defined by Morrison [4.48] by translating the phase from the first half ($0 \le x \le \Delta x/2$) of the unit cell into the second half ($-\Delta x/2 \le x \le 0$) with a π phase shift between the two halves of the unit cell. Thus only half of the transition points are independent parameters.

Later investigations [4.48] showed that even-numbered spot arrays could be readily produced with a phase grating whose unit cell is derived by translating one half of the unit cell into the second half, and with a π phase shift added between the two

halves, as illustrated in Figure 4-23. The result of this translational, or skew, symmetry is that all even-numbered (including $n = 0$) diffraction orders are suppressed so that the output array contains an even number of beams. The advantage for system design is that elimination of the zero-order spot removes the uniformity problem mentioned above.

New conditions are imposed on the unit cell transmission function $t(x)$ as follows. A translational shift relates the second half of the cell period ($-\Delta x/2 < x < 0$) to the first half ($0 < x < +\Delta x/2$) so that

$$t(x) = -t(x - ½), \qquad 0 \le x \le +\Delta x/2 \tag{4.54}$$

and the negative sign on $t(x–½)$ corresponds to the π phase shift between the two halves. Similarly the unit cell phase function $\phi(x)$ is defined as

$$\phi(x) = \pi + \phi(x - ½), \quad 0 \le x \le +\Delta x/2 \tag{4.55}$$

Note that if the negative sign in Eq.4.54 is omitted no phase shift exists between the two halves of the unit cell and the phase structure in the half-cell is simply replicated at twice the frequency (in other words the grating period Δx is halved). The resulting diffraction envelope is

$$T(u) = \int_0^{+\Delta x/2} t(x) \,.\, e^{-i2\pi(ux)}[1 + e^{i\pi u\Delta x}e^{i\pi}]dx = \int_0^{+\Delta x/2} t(x) \,.\, e^{-i2\pi(ux)}[1 - e^{i\pi u\Delta x}]dx \tag{4.56}$$

and since maxima for the array function occur at discrete spatial frequencies, $u = n/\Delta x$, where n is integer-valued, this implies that

$$T(u) = \begin{cases} 0 & n \text{ even} \\ 2\int_0^{\Delta x} t(x) \,.\, e^{-i2\pi(ux)}dx & n \text{ odd} \end{cases} \tag{4.57}$$

The transition points for two one-dimensional Dammann grating solutions exhibiting translational symmetry to generate even-numbered arrays of equi-intense diffraction orders are reproduced in Table 4-2. The two solutions are designed to generate arrays of four and eight beams, respectively and were reported in [4.51] and [4.48]. Whereas a DG with reflection symmetry requires M transition points per unit cell to generate $2M+1$ bright diffraction orders, apparently no such exact relationship exists for a DG exhibiting translational symmetry. The best that one can say is that a $N\times1$ array with N bright orders and $N–1$ suppressed orders requires approximately $N/2$ independent parameters [4.48]. Clearly, of course a similar number of transition points is required since although $N–1$ orders are suppressed, the same number of transmission points is

required for their suppression. For the two solutions in Table 4-2 the number of transition points M is related to the total number of output beams N_{total} equal to N_{even}, the number of (suppressed) even-numbered orders, plus N_{odd}, the number of odd-numbered orders through

$$N_{total} = N_{odd} + N_{even} = 2^M - 1$$

However since this relation is based solely on data for only two solutions it may simply be coincidence and may not hold for other values of M.

N_{total}	N_{odd}	N_{even}	M	$x_1/\Delta x$	$x_2/\Delta x$	$x_3/\Delta x$	$x_4/\Delta x$	$\eta(\%)$
7	4	3	3	0.025	0.250	0.470	–	70.7
15	8	7	4	0.1812	0.2956	0.3282	0.4392	75.9

Table 4-2. Unit cell transition points for one-dimensional Dammann grating solutions incorporating translational symmetry. The M transition points generate N_{odd} numbered diffraction orders and N_{even} = (N_{odd} – 1) suppressed orders.

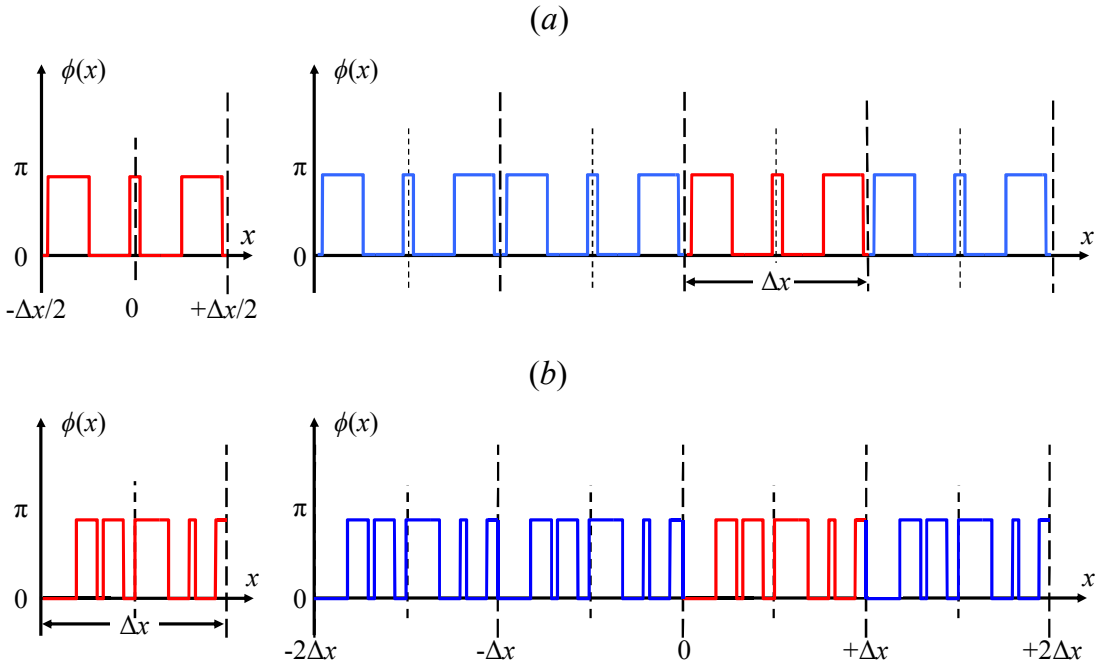


Figure 4-24. Dammann gratings with translational symmetry to produce even-numbered spot arrays. The unit cell (left) is used to form a 4-cell grating. The Dammann grating in (*a*), which is characterised by three transition points x_t = {0.025, 0.25, 0.47}Δx, is designed to generate four equi-intense diffraction orders with a quoted diffraction efficiency of 70.7%. The grating in (*b*) is, which is characterised by four transition points x_t = {0.025, 0.25, 0.47}Δx, is designed to generate eight equi-intense diffraction orders with a quoted diffraction efficiency of 75.9%.

The phase profiles of the 4-beam and 8-beam solution sets (Table 4-2) are shown in Figure 4-24. Notice that for the 8-beam solution there exist phase steps at the midpoint ($x = 0$) and endpoints ($x = \pm\Delta x/2$). These points are not counted as transition points since they only arise due to the π phase shift between the two halves of the unit cell. These additional phase steps do not occur in a unit cell derived from an odd number of transition points M, such as the 4-beam solution.

Figure 4-25 shows a plot of the far field intensity generated by a 4×1 DG whose unit cell is defined by the transition points given in Table 4-2 superimposed on the output generated by a 7×1 DG with reflection symmetry. The two gratings have the same dimensions (four cells of cell length Δx) to ensure overlapping diffraction orders and are uniformly illuminated (indicated by the presence of secondary maxima). Besides reasonable uniformity amongst the four central diffraction orders ($n = \pm 1, \pm 3$) in the 4×1 grating output pattern, all even-numbered orders, $n = \{0, \pm 2, \pm 4\ldots\}$, are completely suppressed.

The far field diffraction pattern of a diffraction grating has, as well as a set of principal maxima corresponding to the diffraction orders, a set of

$$N_{secondary} = N_{cell} - 2 \tag{4.58}$$

secondary maxima between adjacent principle maxima, where N_{cell} is the number of grating cells, or periods. Since both the 4×1 and 7×1 gratings are defined with four repeat cells ($N_{cell} = 4$) we would expect to observe be $N_{secondary} = 2$ secondary maxima in the far field diffraction pattern from both gratings. While the diffraction pattern from the 7×1 grating contains the expected number of secondary maxima, the diffraction pattern from the 4×1 grating contains not two, but six secondary maxima (Figure 4-25), suggesting that the grating contains eight instead of four grating periods. Therefore, as far as secondary maxima are concerned, the unit cell of the 4×1 acts as two cells (of width $\Delta x/2$), suggesting an alternative interpretation of how the phase grating produces the even-numbered array. This is of course consistent with the alternative interpretation that the grating consists of 8 repeated cells, with each cell shifted out of phase by π radians relative to its neighbouring cells. Thus we conclude that any grating can be made to have the skew symmetry (translational symmetry with a π phase shift) simply by adding a π phase shift between adjacent cells.

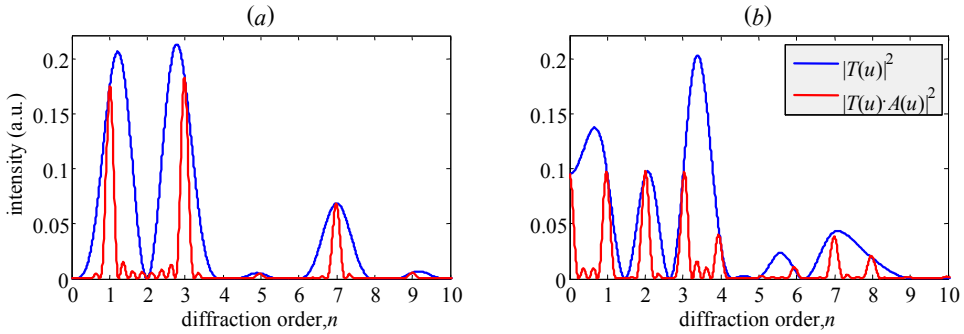


Figure 4-25. The far field intensity from two Dammann gratings designed to produce (*a*) 4 equi-intense diffraction orders and (*b*) 7 equi-intense diffraction orders. Each plot shows the intensity of the diffraction envelope $T(u)$ from the grating unit cell and the diffraction pattern from a grating with a number of unit cells, $T(u)\cdot A(u)$. The array of four diffraction orders in (*a*) is produced by the DG defined by the solution in Table 4-2. The array of seven diffraction orders in (*b*) are produced by a DG with reflection symmetry and transition points at $x_t = \pm\{0.123, 0.344, 0.395\}\Delta x$.

Recall that constructive interference from an array of N equally spaced point sources occur at points when the phase difference $\delta = k\Delta$ between adjacent sources satisfies the condition $\delta = 2n\pi$, for $n = 0, \pm1, \pm2,\ldots$, where $\Delta = d\sin\theta$ is the difference in path length taken by two adjacent point sources (separated by a distance d). This leads to the grating equation for normal incidence from a diffraction grating: $n\lambda = d\sin\theta$. Now if every second point source has a π phase shift added to it constructive interference occurs at angles satisfying the condition $\delta = (2n\pi + \pi)$. This means that diffraction orders now occur at angles satisfying the equation $(n+\frac{1}{2})\lambda = d\sin\theta$. Notice that the separation of diffraction orders is the same as for a regular diffraction grating (without the π phase shift) so effectively the diffraction orders are just shifted off-axis.

The operation of a phase grating defined with skew symmetry discussed above can be elegantly explained in terms of Fourier transforms as follows. Recall that the Fourier transform of a periodic grating is equal to the product of the diffraction envelope $T(u)$ with the interference term $A(u)$. Normally the periodicity of a grating is represented by an array function $a(x)$ that takes the form of a periodic array of Dirac delta functions, which for an infinite array can be written as $a(x) = \text{comb}(x/\Delta x)$, the Fourier transform of which gives $A(u) = \text{comb}(u/\Delta u)$, where $\Delta u = 1/\Delta x$. However with alternative cells now phase-shifted by π, every second delta spike in $a(x)$ has a value of -1, instead of $+1$. Thus the array function is now given by the convolution of a comb function of period $2\Delta x$ with an odd impulse pair as

$$a(x) = \delta\delta\left(\frac{x}{\Delta x/2}\right) \otimes \mathrm{comb}\left(\frac{x}{2\Delta x}\right)$$

as shown in Figure 4-26.

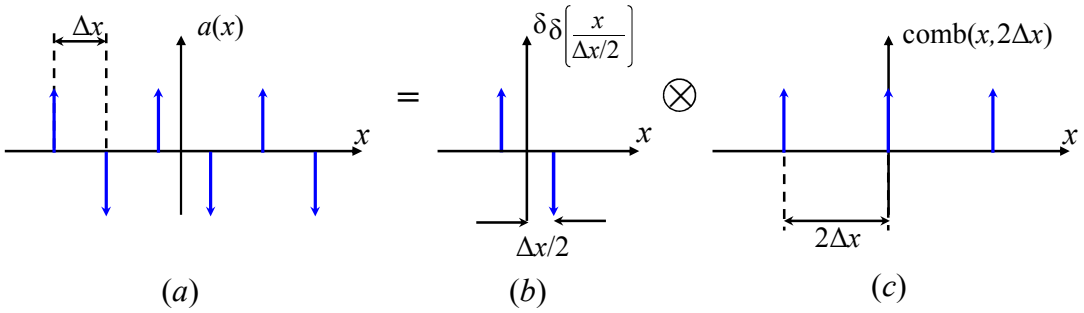


Figure 4-26. (*a*) The array function $a(x)$ can be expressed as the convolution of (*b*) an odd impulse pair with (*c*) a comb function with twice the periodicity ($2\Delta x$) of $a(x)$.

Using the convolution theorem the interference term $A(u)$ is thus given by

$$A(u) = i\,\sin\left(2\pi(\Delta x/2)u\right) \cdot \mathrm{comb}\left(\frac{u}{\Delta u'}\right)$$

where $\Delta u' = 1/2\Delta x = \Delta u/2$. Thus the periodicity of $\mathrm{comb}(u/\Delta u')$ is half that of $\mathrm{comb}(u/\Delta u)$ and its maxima occur at $u_n = (n/2)\Delta u$. However the sine function means that all peaks corresponding to even n are zero, therefore the maxima of $A(u)$ occur at $u_n = (n+½)\Delta u$, with a spacing equal to the original period Δu. While $A(u)$ is an imaginary-valued function, its magnitude is given by

$$|A(u)| = \mathrm{comb}\left(\frac{u-(\Delta u/2)}{\Delta u}\right)$$

a comb function whose central peak is shifted by $(\Delta u/2)$ from the origin (Figure 4-27). Of course if the array of delta functions is finite, the output will consist of an array of sinc functions instead.

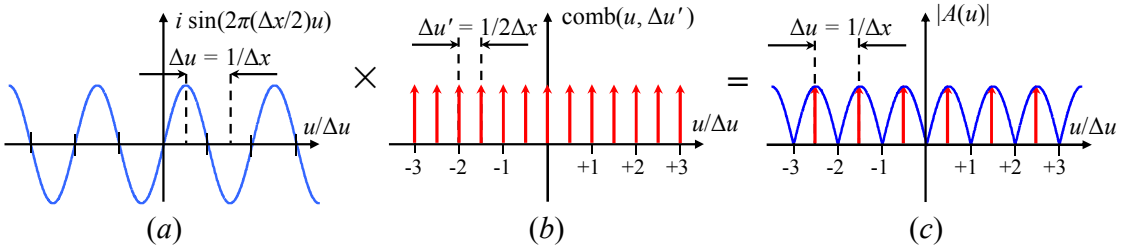


Figure 4-27. The interference term $A(u)$ from the array function $a(x)$ in Figure 4-26 is given by the product of (*a*) an imaginary-valued sine function (the Fourier transform of an odd impulse pair) and (*b*) a comb function with peaks at $u_n = n\Delta u'$. The result is that the magnitude of $A(u)$ is (*c*) a comb function (solid red lines) with delta functions located at $u_n = (n+½)\Delta u$.

4.5 Gaussian Beam Mode Analysis of Phase Gratings

So far we have used Fourier analysis to analytically simulate the operation of periodic phase gratings. The Array Theorem was invoked to calculate the far field diffraction pattern by treating the grating as the convolution of an array function $a(x, y)$ – representing the periodicity of the grating – with the transmission function $t_{cell}(x, y)$ of the periodically repeated unit cell of the grating. Such a treatment, while convenient for analysis of gratings under ideal conditions, is limited in applicability. To analyse grating performance under non-ideal conditions requires a numerical approach in which the grating function $t(x, y)$ is represented as a discretely sampled array. Typically numerical scalar wave diffraction, particularly at visible wavelengths, is computed with a discrete or fast Fourier transform. For quasi-optical systems however, a Gaussian Beam mode analysis (GBMA), i.e. decomposition of the given field into a summation of independently propagating Gaussian beam modes, is a computationally effective alternative that accounts for the long wavelength diffractive effects of such systems. In this section we outline how GBMA can be applied to the analysis of phase gratings, in particular Dammann gratings.

When performing GBMA of a phase grating the field $E_0(x, y)$ transmitted (or reflected) from the grating (located at z_0) is represented as a summation of a set of appropriately weighted Gaussian-Hermite modes, $\psi_{mn}(x, y; z_0)$ as

$$E_0(x, y) = \sum_{m=0}^{m_{\max}} \sum_{n=0}^{n_{\max}} A_{mn} \cdot \psi_{mn}(x, y; z_0) \tag{4.59}$$

where the contribution from mode $\psi_{mn}(x, y; z_0)$ is determined by the value of its corresponding mode coefficient A_{mn}. The diffracted wavefront at propagation distance z ($> z_0$) is then given by a summation of the weighted propagated modes $\Psi_{mn}(x, y; z)$ as

$$E_z(x, y) = \sum_{m=0}^{m_{\max}} \sum_{n=0}^{n_{\max}} A_{mn} \cdot \Psi_{mn}(x, y; z) \tag{4.60}$$

For a grating field $E_0(x, y)$ that is separable in terms of x and y into two one-dimensional terms $E_0(x)$ and $E_0(y)$, e.g. a Dammann grating, we need only analyse the system in terms of x (or y), i.e.

$$E_0(x) = \sum_{m=0}^{m_{max}} A_m \cdot \psi_m(x; z_0) \tag{4.61}$$

and

$$E_z(x) = \sum_{m=0}^{m_{max}} A_m \cdot \Psi_m(x; z) \tag{4.62}$$

In a quasi-optical system ideally phase grating illumination is provided by a quasi-collimated Gaussian beam with an amplitude distribution $Gauss(x,W_G)$, where W_G is the $1/e$ Gaussian beam radius. Thus the grating plane should be defined at a waist position of the incident Gaussian beam thus the grating position is defined as $z_0 = 0$, to ensure illumination with uniform phase.

Numerically the one-dimensional grating field $E_0(x)$ is treated as a discretely sampled array defined at transverse coordinates x, with a constant sample interval dx. With the grating located at $z_0 = 0$ a set of discretely-sampled normalised Gaussian-Hermite modes $h_m(x;W_0)$, representing the continuous modes $\psi_m(x;W_0)$, are created according to Eq. (2.2). The value of mode coefficient A_m that determines the contribution from mode $h_m(x;W_0)$ is given by the one-dimensional form of Eq. (2.11) as

$$A_m = \int_{-\infty}^{+\infty} E_0(x) \cdot h_m^*(x;W_0)\, dx = \int_{-\infty}^{+\infty} E_0(x) \cdot h_m(x;W_0)\, dx \tag{4.63}$$

since the modes $h_m(x;W_0)$ are real-valued at $z_0 = 0$. If the discretely-sampled grating field $E_0(x)$ is represented by row vector $\mathbf{E_G}$ whose i^{th} entry corresponds to value of $E_0(x)$ at $x = x_i$, so that

$$E_0(x_i) = \sum_{m=0}^{m_{max}} A_m \cdot h_m(x_i\,;\, W_0)_{z=0} \tag{4.64}$$

and $\mathbf{A}$ is a row vector whose m^{th} entry is the value of mode coefficient A_m of mode m and $\mathbf{H}$ is a two-dimensional matrix with $(m_{max}+1)$ rows and i_{max} columns such that entry in position (m, i) corresponds to the value of the m^{th} mode at position x_i, i.e.

$$\mathbf{H}_{m,i} = h_m(x_i\,;\, W_0)_{z=0} \tag{4.65}$$

then we can write

$$\mathbf{E_G} = \mathbf{A} \cdot \mathbf{H} \tag{4.66}$$

The mode coefficients are then

$$\mathbf{A} = \mathbf{H}^{-1} \cdot \mathbf{E_G} \tag{4.67}$$

where $\mathbf{H}^{-1}$ is the pseudoinverse of the matrix $\mathbf{H}$. Alternatively we can also solve for values of A_m with a least squares fitting routine.

4.5.1 The far field diffraction from a phase grating

After the transmitted/reflected field from a phase grating has been decomposed into a set of Gaussian beam modes, i.e. after mode coefficients A_m have been calculated, the diffraction pattern on an observation plane a propagation distance z from the grating is calculated by propagating the modes the appropriate distance and then recalculating the weighted sum of propagated modes. At propagation distance z the Gaussian-Hermite mode of order m is given by the one-dimensional form of Eq. (2.2) as

$$\psi_m(x; z) = h_m(x; z) \exp\left[-ik\left(z + \frac{x^2}{2R}\right) + i\phi_m(W; R)\right] \tag{4.68}$$

Since we are concerned primarily with the far field diffraction pattern produced by a phase grating the far field approximation can be employed so the modes simplify to

$$\psi_m(x; z) = h_m(x; z)\, i^{m+\frac{1}{2}} \tag{4.69}$$

where the phase curvature term is suppressed. To model propagation from a grating through a single focusing element, such as in a 4-f system, the ABCD matrix method (see §2.5) is used to keep track of the mode parameters W, R and ϕ_m from the grating plane to the back focal plane of the lens. The diffracted wavefront $E_F(x; z)$ is now calculated with matrix multiplication using

$$\mathbf{E_F} = \mathbf{A} \cdot \mathbf{P} \tag{4.70}$$

where the rectangular matrix **P** is the far field equivalent of matrix **H** in Eq. (4.66).

4.5.2 Choosing an appropriate mode set

Before propagation can be performed an appropriately scaled mode set must be chosen to analyse the phase-modulated beam transmitted from a grating. The defining equations of the normalised Gaussian-Hermite function imply that a mode set is characterised by just two parameters: the mode set size (the number of modes) and the fundamental beam waist radius W_0. Thus when decomposing a given wavefront into a set of Gaussian beam modes the mode set can be tailored to achieve accurate reconstruction of the field by experimenting with different combinations of values for W_0 and maximum mode index, m_{max}. The fundamental mode waist radius W_0 controls the scale, i.e. the width, of the Gaussian beam modes. In some situations, such as modelling the field from a feed horn, it is convenient to scale the mode set such that the fundamental mode, $m = 0$ contains most of the power. However we have seen in

Chapter 2 that for the case of a uniformly illuminated aperture (a top-hat function) accurate reconstruction is achieved by scaling the mode set such that the highest-order mode fits just inside the aperture, i.e. by choosing a value of W_0 so that the effective width of the highest-order mode equals the aperture width. Whatever the problem, the goal is the same: to describe accurately the given field using the minimum number of modes possible to ensure computationally efficient analysis.

As with the top-hat example in §2.4, accurate reconstruction of a discrete-level phase grating requires many modes to account for the high spatial frequency content due to the sharp phase steps (at locations of transition points). Assuming a mode set with (m_{max}+1) modes with mode indices m = {0, 1, 2, …m_{max}}, a suitable value for the mode waist parameter W_0 must now be chosen to minimise the maximum mode order m_{max}. Two approaches for selecting an appropriate value of W_0 were examined.

Fitting the mode set to the grating aperture

The first method for choosing W_0 is based on the method used for reconstructing a top-hat function. This involves choosing a value of W_0 such that the effective extent, or length L, of the highest-order mode, m_{max} equals the width of the top-hat. For analysis of a grating this means finding a value of W_0 so that the width of the highest-order mode equals the grating width, D. The width of a Gaussian-Hermite mode of order m is given approximately by

$$L_m \approx 2W_0\sqrt{m} \tag{4.71}$$

and we require that the width of mode m_{max} equals the grating width D, i.e.

$$D = L_{m_{max}} \tag{4.72}$$

which yields a fundamental beam waist radius of

$$W_0 \approx \frac{D}{2\sqrt{m_{max}}} \tag{4.73}$$

Note that Eq. (4.71) is only an approximation for the mode length. A simple optimisation routine can be used to find a value of W_0 that more closely satisfies Eq. (4.72). In fact if mode length is defined as the distance between the two outermost zero crossings a more exact expression is given by $L_m = W_0(2\sqrt{m}-1)$, but Eq. (4.71) makes for simpler analysis later on.

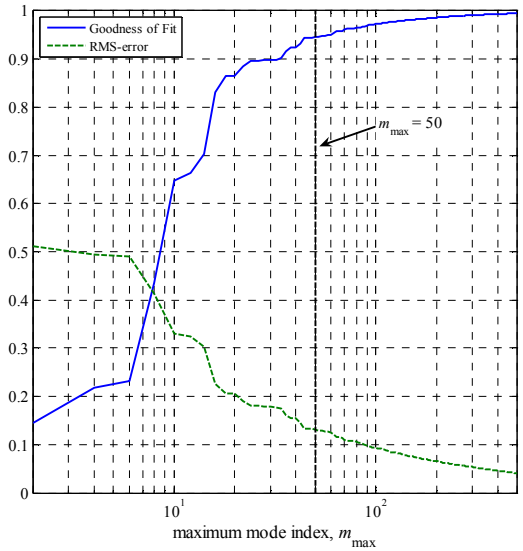


Figure 4-28. Root-mean squared (RMS) error and intensity correlation metrics for a Gaussian beam mode expansion of the field $E_G(x)$ from a Gaussian-illuminated Dammann grating to produce three equi-intense diffraction orders. As the maximum mode index m_{max} increases the reconstruction quality improves.

For example consider the Dammann grating solution to produce three equi-intense diffraction orders (referred to hereafter as a 3-beam DG). The grating is to consist of four unit cells ($D = 4\Delta x$) and is illuminated with a quasi-collimated Gaussian beam (of waist radius $W_G = \Delta x$). Eq. (4.73) gives $W_0 = 2\Delta x/\sqrt{m_{max}}$, which ensures that the outer zero crossings of mode m_{max} coincide with the grating endpoints (at $x = \pm D/2$). A series of beam mode expansions were performed for the 3-beam DG wavefront $E_G(x)$. In each expansion a different number of modes was used and the reconstruction quality estimated (by calculating the RMS-error and intensity correlation between the expected and reconstructed wavefront amplitude distributions $|E_G(x)|$ and $|E_G'(x)|$, respectively) – see Figure 4-28. Beyond a certain point ($\sim m_{max} = 50$) increasing the number of modes yields little improvement in reconstruction quality. In other words only a small number of modes are needed to accurately reconstruct the phase-modulated field.

Figure 4-29 show the amplitude and phase profiles for trial reconstructions of the grating field $E_G(x)$ with mode sets defined by $m_{max} = 20$, 50 and 300. When $m_{max} = 20$ the reconstruction fails to accurately reconstruct the phase modulation at the grating, missing the two outermost phase steps. Furthermore the reconstructed grating amplitude $|E_G'(x)|$ does not match the ideal incident Gaussian beam profile very well. The reconstruction with $m_{max} = 50$ yield much better results with all phase steps being accounted for. As the number of modes increases the reconstruction becomes increasingly accurate, however the magnitude of the high-frequency ringing in the

amplitude distribution in the vicinity of the discontinuous phase jumps never completely disappears but approaches a finite limit, similar to the Gibbs phenomenon observed in Fourier analysis.

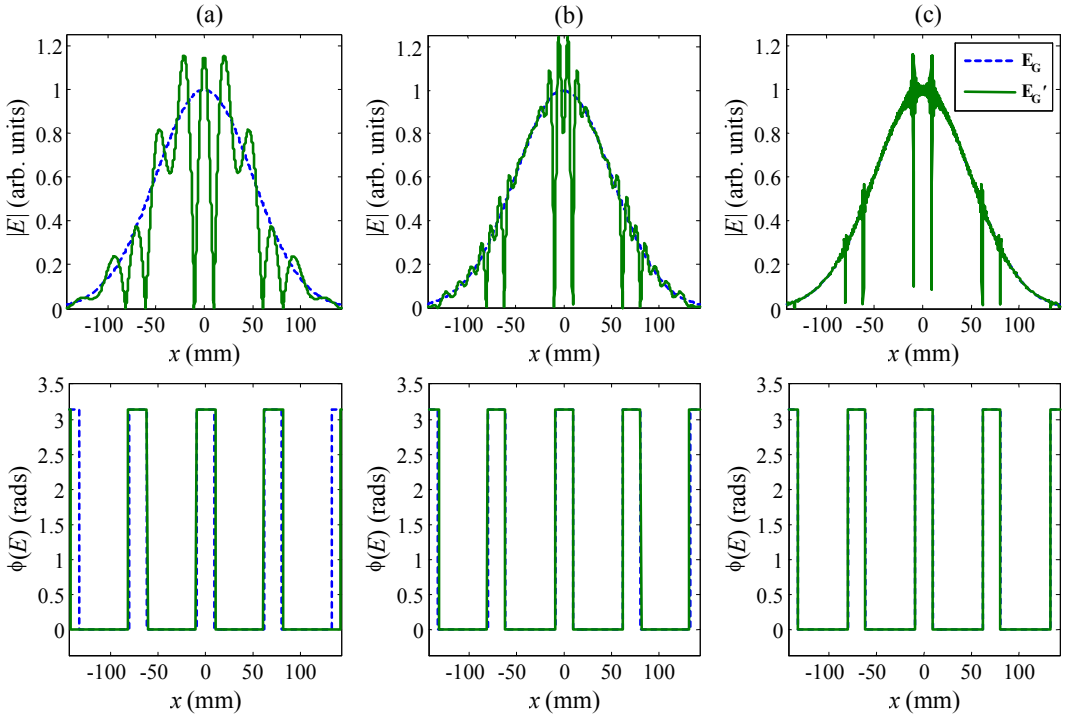


Figure 4-29. **Top:** Grating amplitude. **Bottom:** Grating Phase. Gaussian beam mode reconstructions of a Gaussian illuminated Dammann grating with four unit cells (each with two transition points) with mode sets defined by a highest-order mode index of (*a*) m_{max} = 20, (*b*) m_{max} = 50 and (*c*) m_{max} = 300.

Figure 4-30 shows a bar chart of the amplitude and phase values of the mode coefficients A_m obtained from the reconstruction with m_{max} = 50. Note that all odd-numbered modes have zero amplitude, i.e. they do not contribute to the reconstructed grating field. This is as expected since the 3-beam DG is symmetric about x = 0 and so can only support symmetric (even-numbered) modes. Another thing to note from Figure 4-3 is that the phases of A_m are restricted to values of zero or π only, reflecting the binary nature of the grating phase. Because the grating phase is restricted to values of only 0 and π, the field at the grating is in fact real-valued so modes with a phase of 0 correspond to positive-valued regions of the E-field and modes with a phase of π correspond to negative-valued regions of the E-field. In other words the mode coefficients A_m are themselves real-valued.

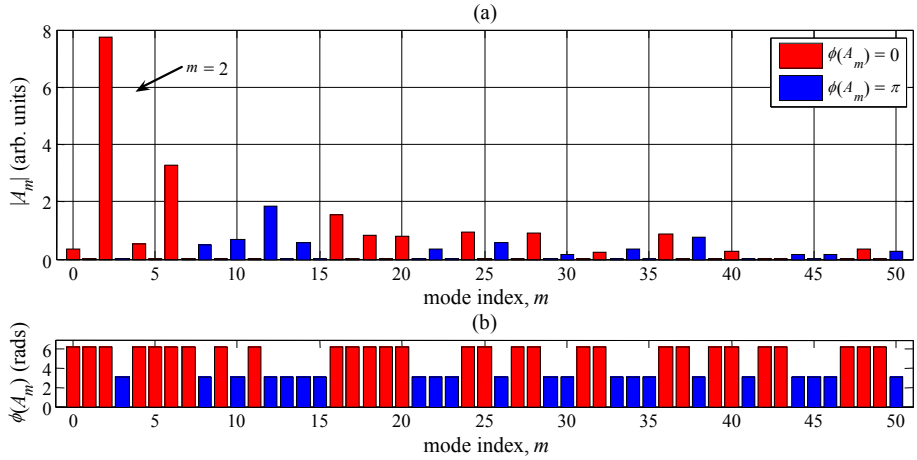


Figure 4-30. (*a*) Amplitude and (*b*) phase of mode coefficients A_m for the GBM-reconstruction of the Gaussian illuminated 3-beam DG with $m_{max} = 50$. Note that mode $m = 2$ is the dominant mode. Note also that the symmetric nature of the grating is reflected in the fact that only symmetric (even-numbered) modes contain power, while all asymmetric (odd-numbered) modes are suppressed. Another characteristic property is that the phases of the mode set are restricted to values of 0 (red bars) and π (blue bars) only, due to the fact that the grating phase profile is itself binary. For clarity 0-valued phases are plotted with equivalent values of 2π.

Notice in Figure 4-30 that the most intense mode is the second-order mode, which is not surprising since the intensity profile of a second-order Gaussian-Hermite (with three equally spaced peaks) is similar to the expected far field diffraction pattern from a 3-beam DG. By itself however the second-order mode cannot reproduce the diffraction pattern. From Figure 4-31(*a*), the central order is both narrower and of lower intensity than the surrounding ($n = \pm 1$) diffraction orders – a feature characteristic of Gaussian-Hermite modes that becomes much more noticeable for higher order modes (the two outer intensity peaks are broader and more intense than the inner $m-1$ intensity peaks). Also, since a second-order mode has only two zero-crossings it cannot reproduce all phase transitions in the grating. Clearly several higher order modes are needed as well.

Figure 4-31 shows a number of reconstructions of the far field amplitude with progressively larger mode subsets from the mode-set with $m_{max} = 50$ (whose coefficients are shown in Figure 4-30). Figure 4-31(*a*) shows a reconstruction with just the second-order mode. The reconstruction in Figure 4-31(*b*) uses only the 3 most intense modes; (*c*) the 8 most intense modes and (*d*) the 12 most intense modes. The last (generated with 12 modes) is in good agreement with that using all 51 modes. This further illustrates the efficiency of Gaussian beam mode expansion for modelling propagation in quasi-optical systems, even with complex components like Dammann gratings.

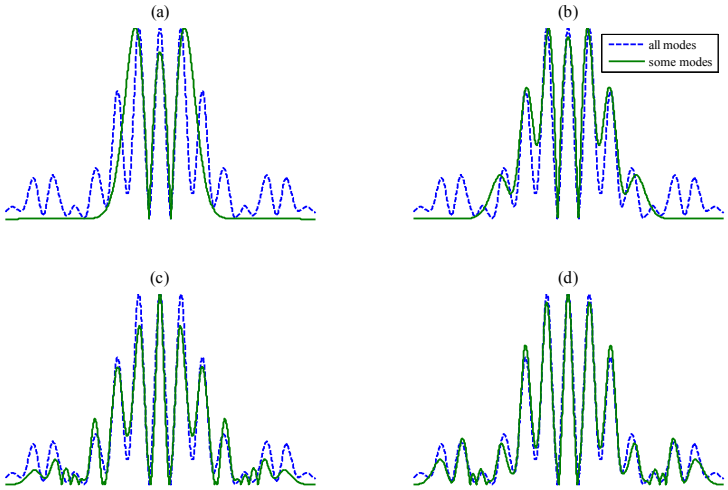


Figure 4-31. Far-field mode reconstruction with mode subsets containing a small number of modes from the set of modes with m_{max} = 50, whose coefficients are plotted in Figure 4-30. The blue curves represent the reconstruction using all (m_{max}+1) = 51 modes, while the green curves represent reconstructions using only (*a*) 1, (*b*) 3, (*c*) 8 and (*d*) 12 most intense, or dominant, mode(s).

Figure 4-32 shows the estimated reconstruction quality (in terms of RMS-error and intensity correlation between the GBM-reconstructed intensity and the true far field intensity from a Gaussian-illuminated 3-beam Dammann grating) for a series of beam mode expansions as a function of mode subset size. The number of beam mode expansions performed is 26 since only the even-numbered modes are involved.

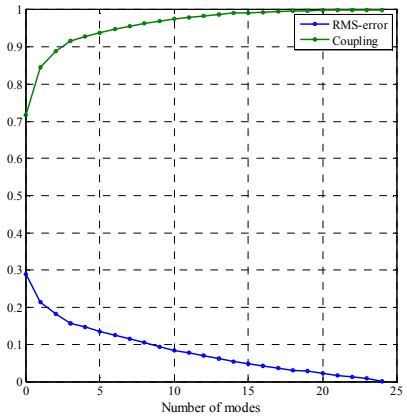


Figure 4-32. Reconstruction quality (RMS-error and intensity correlation) for a series of trial beam mode expansion as a function of mode set size. The mode set with a single term contains only m = 2 mode. Reasonably good reconstruction quality is achieved with as few as 12 modes.

Maximising the power in a specific mode

In the previous analysis of the 3-beam DG the second-order mode $\psi_2(x)$ was the most intense mode. The intensity profile of $\psi_2(x)$ is similar in shape to the expected far field

diffraction pattern from the 3-beam grating, i.e. an array of 3 intensity maxima with approximately equal spacing and equal magnitude. As such the 3-beam DG can be thought of as a mode-selective device that effectively transforms the incident Gaussian beam, approximated by the fundamental mode h_0, into the second-order mode h_2. This suggests a more intuitive approach for choosing an appropriate value for W_0 by considering the far field diffraction pattern that the grating is designed to generate. When illuminated with a single incident Gaussian beam a beam-splitting element ideally generates an array of equally spaced and equally intense Gaussian beams in the far field of the grating. Thus we could choose a mode set in which the mode most resembling the expected far field diffraction pattern from a grating is maximised.

In the analysis of the 3×3 DG it was merely coincidental that the dominant mode was one that matched the form of the far field intensity, since W_0 was chosen merely to fulfil the criterion that the size of the highest-order mode fit matched the size of the grating aperture. If we view any one-dimensional beam-splitting phase grating as a mode-switching device then it makes sense to optimise W_0 such that power in the mode that most closely resembles the far field array of N diffraction orders is maximised. The intensity profile of a Gaussian-Hermite mode of order m has m+1 intensity peaks, so for a grating designed to generate an array of N equally spaced, and equally intense diffraction orders, W_0 should be selected to maximises power in mode $m = N–1$. Note that to "maximise" the power in the m^{th} mode, it is meant that mode h_m contain more power than it otherwise would given a different value of W_0. It is not implied that mode h_m should be the most intense mode since such a situation may yield poor reconstruction quality. For instance, while the far field diffraction pattern shape may be dominated by the profile of the m^{th} mode, the reconstructed intensity at the grating plane must resemble as closely as possible the incident Gaussian wavefront, whose intensity is most closely approximated by the fundamental mode. Therefore the fundamental mode, or similarly low-order modes, must contain a substantial fraction of power. If this consideration is not taken into account and W_0 is optimised such that the mode $m = (N–1)$ becomes dominant and the fundamental mode coefficient amplitude $|A_0|$ has a low value a large number of high-frequency modes will be required to accurately reproduce the Gaussian intensity at the grating plane, which will inevitably introduce unwanted high-spatial frequency ringing in the approximated wavefront.

Figure 4-33(a) shows the amplitude of the fundamental and second-order mode coefficients $|A_0|$ and $|A_2|$ as a function of W_0 for reconstruction of the 3-beam DG field. Choosing the value of W_0 that produces a maximum value of $|A_2|$, i.e. $W_0 \approx 0.1D$, was then used to reconstruct the grating field and its far field diffraction pattern, the amplitude of which is shown in Figure 4-33(b). This choice of value for W_0 results in the second-order mode (with three intensity peaks) matching most closely the far field diffraction pattern.

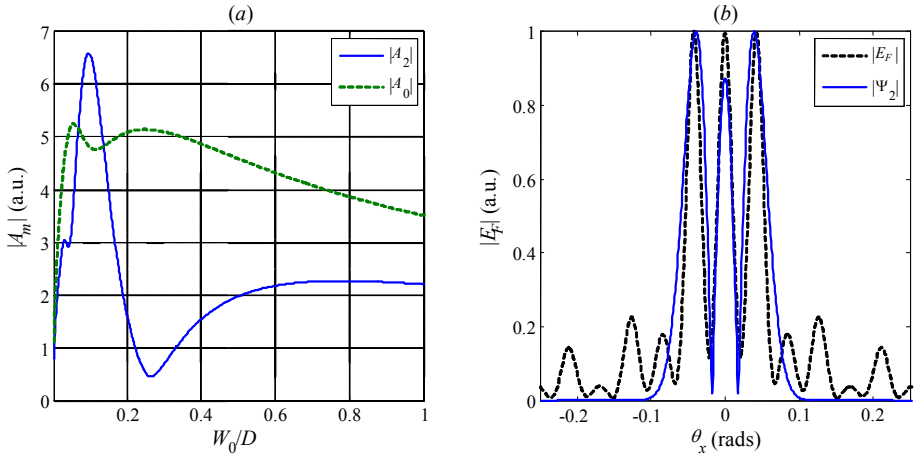


Figure 4-33. (a) Mode coefficient amplitudes $|A_m|$ of modes m = 0 and 2 as a function of W_0 for the 3-beam DG example. $|A_2|$ has a maximum when $W_0 \approx 0.1D$. (b) GBM-propagated far-field pattern for a mode set defined by $W_0 = 0.1D$ and $m_{max} = 50$, superimposed on the far-field pattern of the second-order mode, $\psi_2(x)$.

Limiting the spatial frequency content of the mode set

The previous approach for selecting a value of W_0 is only appropriate (a) when one knows the expected output from the element and (b) when that expected output has the symmetric (or asymmetric) form of a Gaussian-Hermite mode. Generally speaking this is not the case and a more general means of choosing W_0 is required. The problem with selecting a value of W_0 by imposing the condition that the highest-order mode size be equal to the size of the grating aperture is that it makes the number of modes an arbitrary decision. As we have seen a mode set with too few modes may result in unexpected loss of features (such as phase transitions), while a mode set with too many modes is computationally inefficient.

We saw for the 3-beam DG example that high spatial frequencies associated with the sharp edges at transition points require the inclusion of high-order modes to achieve good quality reconstruction. However in a real optical system, where the diffraction pattern is formed on the focal plane of a lens/mirror the large high-order

modes transmitted from the grating plane will be truncated at the edges of the focusing element [4.18]. In other words high spatial frequency features at the grating plane may not contribute to the image formed on the observation plane. Furthermore, high spatial frequencies at the grating plane correspond to high off-axis features in the Fourier plane. Since we are only interested in the diffraction orders formed within a narrow cone in the far field, the absence (or presence) of very high spatial frequencies at the grating has little impact on the diffraction pattern in the far field region that we are interested in.

If we relax the demands placed on the accuracy of the grating field reconstruction by placing a lower limit on the spatial frequency content of the grating field a mode set can be found that reconstructs the general form of the grating field, without necessarily reproducing the sharp edges at transition points. This leads to a mode set with substantially fewer modes, thus increasing computational efficiency. Also, restricting the spatial frequency content of the grating allows one to determine not only the beam mode radius W_0, but also the number of modes needed.

The minimum feature size δ of any discrete-level phase grating is equal to the minimum distance between two transition points and is the minimum spatial frequency of the grating field. Hence by choosing a mode set with a minimum spatial frequency equal to δ a GBM expansion of the grating field will include features at least as small as the minimum feature size of the grating. A Gaussian-Hermite mode of order m contains $m/2$ full quasi-sinusoidal periods of approximately equal size Λ_m across its width L_m, therefore

$$L_m \approx \frac{m}{2}\Lambda_m \tag{4.74}$$

and substituting Eq. (4.71) for L_m above gives the spatial frequency of mode m as

$$\Lambda_m \approx \frac{4W_0}{\sqrt{m}} \tag{4.75}$$

Now we impose the condition that the highest spatial frequency, $\Lambda_{max} = 4W_0/\sqrt{m_{max}}$, equals the minimum feature size δ. But we also require that the highest-order mode does not exceed the width of the grating aperture D, so we require that

$$D = 2W_0\sqrt{m_{max}} \tag{4.76}$$

and

$$\delta = \Lambda_{max} = \frac{4W_0}{\sqrt{m_{max}}} \tag{4.77}$$

are simultaneously satisfied. Solving for W_0 and m_{max} yields

$$m_{\max} = 2\frac{D}{\delta} \qquad \text{and} \qquad W_0 = \sqrt{\frac{D\,\delta}{8}} \tag{4.78}$$

Thus given a grating, or any other aperture, of width D and minimum feature size δ a mode set defined by Eq. (4.78) will result in a reconstruction accurate to within δ.

For example consider the 5×5 Dammann grating (5-beam DG) solution, the phase modulation of which is characterised by a unit cell with transition points at $x = \pm\{0.02,\ 0.481\}\Delta x$. The minimum feature size for this grating is $\delta = 2(0.02)\Delta x$. For a grating with four unit cells $D = 4\Delta x$, from Eq. (4.78) the grating can be reconstructed with a mode set defined by the parameters $m_{max} = 2(4/0.04) = 200$ and $W_0 = \sqrt{4\times0.04/8}\ \Delta x \approx 0.14\Delta x$. Figure 4-34 shows the ideal and GBM-reconstructed real-valued fields $E_G(x)$ and $E_G'(x)$ for the Gaussian-illuminated 5-beam DG. Since the unit cell is symmetric about $x = 0$, only symmetric (even-numbered) modes contribute in the GBM summation, so only $m_{max}/2 = 100$ modes are needed.

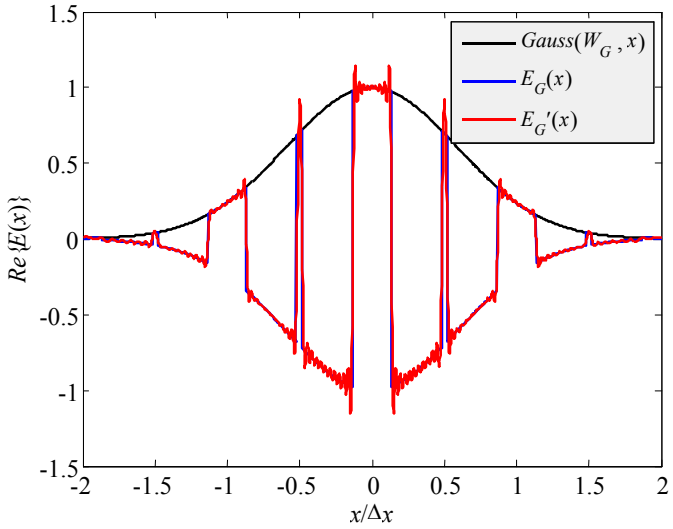


Figure 4-34. GBMA of a Gaussian-illuminated 4-cell Dammann grating to produce 5 diffraction orders. Shown are the real-valued parts of the ideal grating field $E_G(x)$ and the GBM-reconstructed field $E_G'(x)$. The mode set is defined by $m_{max} = 200$ and $W_0 = \sqrt{0.02}\Delta x$, which allow the reconstruction to reproduce the smallest feature in the original grating field.

Figure 4-35 shows the field and phase at the fourth cell (i.e. over $x = +\Delta x$ to $+2\Delta x$) of the GBM-expanded fields superimposed on the original fields for both the Gaussian-illuminated and uniformly-illuminated grating fields. The particular choice of m_{max} and W_0 results in the smallest feature of the grating field at the midpoint of the unit cell (the phase step near $x = 1.5\Delta x$) being reproduced as a single "ripple" instead of the ideal

sharp rise and fall of the original grating profile. Notice that the phase of $E_G'(x)$ for the Gaussian-illuminated grating contains additional phase jumps near the end of the grating (towards $x = 2.0\Delta x$), whereas the phase of $E_G'(x)$ for the uniformly illuminated grating contains the correct number of discontinuities. The additional phase jumps in $E_G'(x)$ under Gaussian illumination occur near the edges of the grating where the illuminating Gaussian amplitude drops below some minimum value. This behaviour was observed when performing modal expansions on other fields in regions where the intensity drops below a minimum value and presumably occurs because the phase becomes undefined in regions of low intensity. In this example the illuminating Gaussian beam had a radius of $W_G = 0.837\Delta x$, therefore at $|x| > 2W_G \approx 1.67\Delta x$ the intensity of the grating field falls below e^{-4}. Since the illuminating intensity at the edges of the grating falls to such a low level in these regions the fact that the GBM approximated field does not replicate the phase exactly in these regions is irrelevant, since the low intensity means that the phase contributes very little compared to the phase in well illuminated regions of the grating.

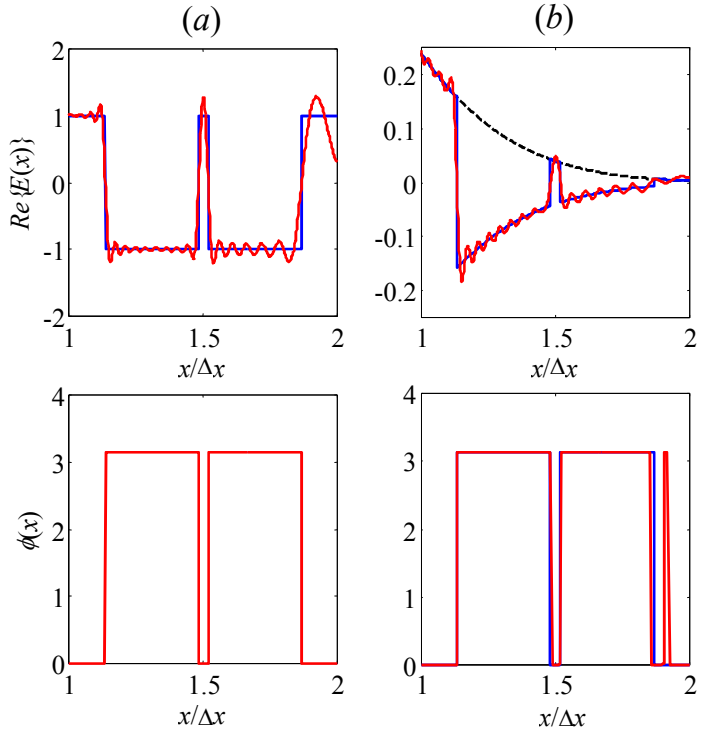


Figure 4-35. Close-ups of the rightmost cell in the 4-cell 5-beam DG. Shown are the real-valued parts (upper) and phase (lower) of the GBM-reconstructed grating under (*a*) uniform illumination and (*b*) Gaussian illumination. The small groove at $x = 1.5\Delta x$ is the smallest feature in the grating and was used to determine the highest-order mode m_{max} and the value of W_0. Additional phase jumps occur in the GBM-approximated grating field for Gaussian illumination near the end of the grating where the intensity of the illuminating Gaussian has fallen below some threshold and the phase has become undefined.

We now return to the 3-beam DG, the transition points of which are $x = \pm 0.132\Delta x$. The minimum feature size at the grating is $\delta = 2(0.132\Delta x) \approx 0.25\Delta x$. For a grating with four unit cells ($D = 4\Delta x$) the mode-set needed to reproduce feature size δ is defined by $m_{max} = 2(4\Delta x/0.25\Delta x) = 32$ and $W_0 = \sqrt{4\times 0.25/8}\Delta x \approx 0.36\Delta x$. Since the unit cell is symmetric about $x = 0$ only symmetric modes are required, so in fact only 16 modes are needed. Figure 4-36 shows the real-valued GBM-approximated field superimposed on the actual grating field, as well as the reconstructed and original phase of the 3-beam DG. For this example we assume illumination of the grating with a Gaussian beam of radius $W_G = \Delta x$, so the incident amplitude does not fall below e^{-4} at any point on the grating. The larger value of W_G ensures that no additional phase jumps are introduced into the GBM-approximated phase. In other words the phase modulation across the entire grating plane contributes to the field transmitted from the grating.

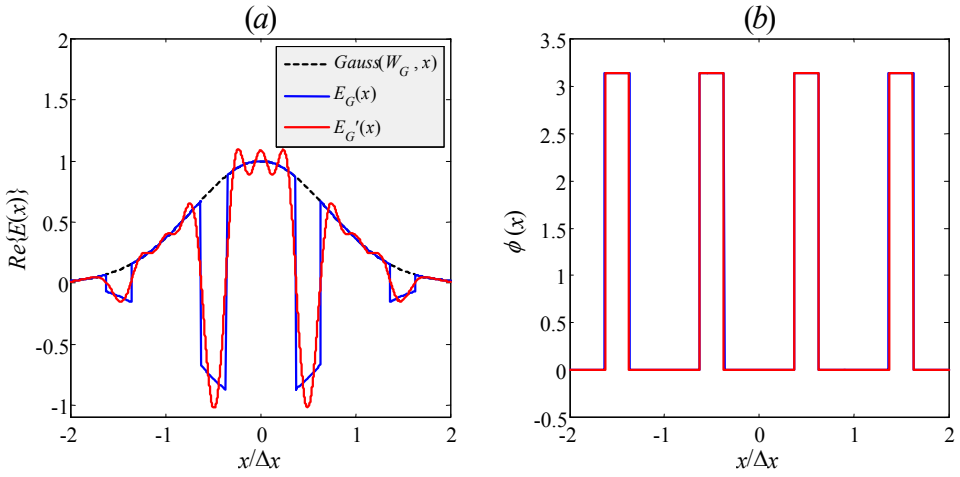


Figure 4-36. GBMA of the Gaussian-illuminated 3-beam Dammann grating. The mode set is defined by $W_0 \approx 0.36\Delta x$ and $m_{mmax} = 32$. The minimum feature size δ is equal to the width of each of the four π-valued regions in (*b*). With this choice of mode set parameters the smallest grating features (the negative-valued regions in (*a*)) are reproduced as single troughs in $E_G'(x)$.

In the last two examples the mode set parameters m_{max} and W_0 were defined by setting the highest spatial frequency of the mode set equal to the minimum feature size of the original grating field. If computational resources permit, a more accurate reconstruction can be achieved by setting the highest spatial frequency of the mode set to some fraction of the smallest feature size δ. Previously we had

$$m_{max} = 2\frac{L_m}{\Lambda_{min}} \quad \text{and} \quad W_0 = \sqrt{\frac{L_m\Lambda_{min}}{8}}$$

where $\Lambda_{min} = \delta$. However if we now require that the highest spatial frequency is given by $\Lambda_{min} = \delta/n$ then

$$m_{max} = 2n\frac{D}{\delta} \quad \text{and} \quad W_0 = \sqrt{\frac{D\delta}{8n}} \tag{4.79}$$

Thus the accuracy of the Gaussian beam mode decomposed field can be controlled by simply varying the value of n, i.e. by increasing the highest spatial frequency of the mode set.

4.6 Practical Considerations in Phase Grating Design

In this section we examine firstly how the required phase modulation needed to produce the desired far field diffraction pattern from a particular phase grating is achieved, before examining how the phase modulation due to a phase grating changes over a finite bandwidth.

4.6.1 DPE fabrication: inducing the phase modulation

The phase transformation imparted by a diffractive phase element (DPE) must be induced in a refractive material or in reflection. One method is to construct the DPE in segmented form, such that neighbouring segments are composed of materials with different refractive indices so that an electromagnetic beam transmitted through the device propagates at different speeds thus distorting the incident wavefront. For example a Dammann grating would require two different materials to produce a binary phase modulation. In practise this of course could be quite difficult to realise. Another method is to construct the device from a material whose refractive index varies with position across the surface thereby allowing the element to be constructed from a sheet of uniform thickness. A common technique at visible wavelengths is the controlled bleaching of photographic emulsions, which induces refractive index change owing to a chemical reaction at the exposed parts of the emulsion [4.9].

The most straightforward and commonly used technique, however is to construct a DPE by varying the thickness of a single slab of dielectric material with uniform refractive index. The variation in depth across the slab, combined with a refractive index difference between the material and surrounding medium (typically air, $n_0 = 0$) causes different parts of the transmitted wavefront to travel different optical path

lengths, which corresponds to a phase delay. At visible wavelengths techniques such as plasma etching and reactive ion etching are used to realise a surface relief groove pattern by etching the relief pattern into a substrate such as quartz glass (e.g. SiO_2, with a refractive index of ~1.5), whereby the etch depth determines the phase depth. Fabrication of visible-wavelength phase gratings with these methods are described in [4.9]. At millimetre wavelengths these groove patterns are typically milled into a sheet of dielectric using a CNC milling machine [4.23]. A reflective device can be realised by cutting a groove pattern into the surface of a sheet of metal (cutting grade aluminium was used at Maynooth, but other groups have used copper).

Rapid prototyping is a recently available option for constructing reflective surface relief patterns [4.52]. In this process the surface profile is built up in plastic using a direct-write laser process and a reflective coating is then applied to the surface. The advantage of rapid prototyping is that much higher surface accuracy and highly reflective surfaces can be obtained. However it cannot be used to construct transmission devices since the density of the material is not constant throughout and would result in an effective non-uniform refractive index.

Realising a transmission DPE

The surface profile is specified by a two-dimensional height function $h(x, y)$. We wish to find a height function $h(x)$ that produces a one-dimensional phase modulation $\phi(x)$.

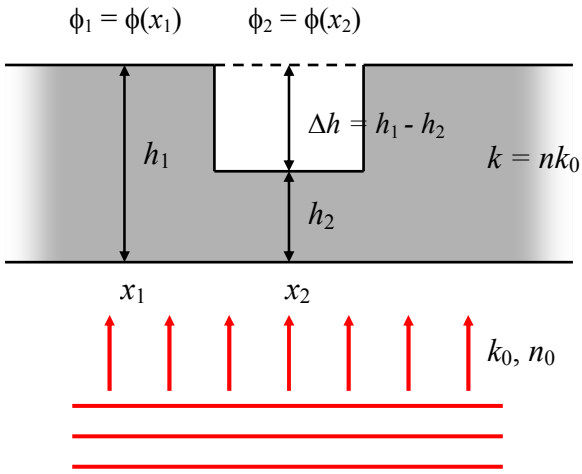


Figure 4-37. Cross-sectional view through a slab of dielectric (with refractive index n) situated in air with a single groove of depth Δh that produces a phase difference, $\Delta\phi = \phi_1 - \phi_2$ due to the difference in height between the recessed surface and top surface.

Consider a plane wave incident on the binary-level DPE shown in Figure 4-37, made from a piece of dielectric with refractive index n. Between points at x_1 and x_2 on the incident wavefront there exist a phase difference $\Delta\phi = \phi_2 - \phi_1$. The phase lag at any point x between the bottom and top of the slab is given by summing the optical path length through the refractive material and through the surrounding medium. For example at points x_1 and x_2

$$\phi(x_1) = kh_1 \qquad \text{and} \qquad \phi(x_2) = kh_2 + k_0\Delta h \tag{4.80}$$

where $\Delta h = h_1 - h_2$ and the wavenumbers in air and within the dielectric, respectively are

$$k_0 = 2\pi/\lambda,\ k = n2\pi/\lambda = nk_0 \tag{4.81}$$

The phase difference, $\Delta\phi = \phi(x_1) - \phi(x_2)$ is therefore

$$\Delta\phi = kh_1 - (kh_2 + k_0\Delta h) = k_0(n-1)\Delta h \tag{4.82}$$

Solving for Δh above yields the depth of the groove required. This allows us to calculate the step height(s) of a discrete-level phase grating as

$$\Delta h = \frac{\Delta\phi}{k_0(n-1)} \tag{4.83}$$

or the height function corresponding to a smoothly varying phase function $\phi(x)$ as

$$h(x) = \frac{\phi(x)}{k_0(n-1)} \tag{4.84}$$

Here it is assumed that the refractive index n of the dielectric substrate does not vary appreciably with wavelength. Although in reality refractive index is a function of wavelength the particular plastic materials used for the gratings described in this thesis have refractive indices that are practically wavelength-independent, i.e. non-dispersive, over the frequency range that is of interest to us: high-density polyethylene (HDPE) and poly tetrafluorothylene (Teflon) with refractive indices of 1.525 and 1.38, respectively, that remain approximately constant between 100 GHz (3 mm) and 4.5 THz (75 μm) [4.53,4.54].

Realising a reflection DPE

Reflection gratings offer lower losses compared to transmission through an absorptive dielectric and additionally avoid standing wave effects. In a reflection DPE a phase modulation is imposed by varying the distance that various points on the incident wavefront must travel on reflection to the output plane. Points in the incident wavefront

with longer distances to travel lag behind those with shorter distances to cover thus inducing the required relative phase shifts.

The height function $h(x)$ of a reflection DPE is calculated as follows. The portion of a wavefront reflected at a point P on the surface accumulates a phase lag of

$$\phi_P = k_0 r_1 + k_0 r_2 \tag{4.85}$$

relative to the wavefront reflected at a point Q. Since the incident and reflected angles are equal in magnitude, distances r_1 and r_2 are equal therefore the accumulated phase is

$$\phi_P = 2k_0 r \tag{4.86}$$

where $r = r_1 = r_2$. If point P is at a depth h_P relative to the height of point Q, the added phase is given by

$$\phi_P = \frac{2k_0 h_P}{\cos(\alpha)} \tag{4.87}$$

where α is the angle of throw between incident and reflected beams. The height $h(x)$ needed to induce a phase modulation $\phi(x)$ for all x in a reflective DPE is thus

$$h(x) = \frac{\phi(x)}{2k_0 \cos(\alpha)} \tag{4.88}$$

When illuminated with a beam at an oblique angle of incidence ($\alpha_{inc} > 0$) the profile of the incident beam becomes elongated along the surface. To compensate for this projection effect, the grating surface profile must be elongated in the same direction, by simply stretching the x-coordinates (i.e. along the surface) by the same amount. The grating surface $h(x)$ is now defined in terms of transverse coordinates

$$x' = \frac{x}{\cos(\alpha_{inc})} \tag{4.89}$$

Alternatively a reflective DPE can be modified so that either the incident or reflected beam is normal to the surface, i.e. $\alpha_{inc} = 0$, or $\alpha_{ref} = 0$. In the latter case we do not need to correct for projection effects. To achieve such a configuration requires the introduction of a blazed phase component into the gratings phase modulation.

4.6.2 Bandwidth of a Diffractive Phase Element

Quasi-optical devices can, broadly speaking, be separated into two categories: frequency-independent components and frequency-selective components. Although DPE's fall into the latter category, since maximum diffraction efficiency occurs at a single design wavelength λ_0, they generally operate within certain tolerance limits over

a finite bandwidth about λ_0. To evaluate the useful bandwidth of a particular DPE we must determine the effect that a change in wavelength has on the phase modulation imparted by a DPE.

The only wavelength dependent characteristic exhibited by a periodic amplitude-modulated multi-slit diffraction grating is the angular separation of its far field diffraction orders, while the amplitude of those orders depend on the ratio of the slit width to slit separation. However periodic phase-modulating diffractive gratings have a more complicated frequency response. As well as the wavelength-dependence of the angular separation of diffraction orders, the gratings phase modulation $\phi(x)$ is itself wavelength-dependent and therefore affects the intensities of the diffraction orders in a wavelength dependent manner.

Phase modulation imparted by a DPE at non-design wavelengths

The height function of a transmission grating is given by Eq. (4.84). Away from the design wavelength ($\lambda \neq \lambda_0$) the phase modulation experienced by an incident wavefront is given by solving for phase in Eq. (4.84) and substituting k (= $2\pi/\lambda$) for k_0, which gives

$$\phi(x) = h(x) \cdot k(n - 1) \tag{4.90}$$

where we assume again that refractive index n is not a strong function of λ. In terms of the design phase modulation, $\phi_0(x)$ this is

$$\phi(x) = \phi_0(x) \cdot \frac{k}{k_0} = \phi_0(x) \cdot \frac{\lambda_0}{\lambda} = \phi_0(x) \cdot \frac{\nu}{\nu_0} \tag{4.91}$$

Thus the phase modulation imparted by a DPE at frequency $\nu \neq \nu_0$ is equal to the design phase modulation $\phi_0(x)$ scaled by the ratio of ν to ν_0. The wavelength dependence of Dammann gratings is examined in §4.7.1.

4.7 Experimental Testing and Verification of Dammann Phase Grating Designs

This section describes the results of experimental measurements that were made of two Dammann phase gratings using the intensity measurement system TOAST described in Chapter 3. As well as verifying the grating theory developed previously in this chapter, these measurements also provided an opportunity to compare experimentally obtained beam pattern measurements with predictions made using numerical simulations developed in MODAL, particularly in terms of the aberrations and distortions introduced by any optics in the test system.

4.7.1 Transmission Dammann Grating (3×3 spot array)

The first grating is a transmission Dammann grating designed to generate a two-dimensional 3×3 array of equally intense diffraction orders, hereafter referred to as the 3×3 DG.

Design and Fabrication

The grating surface profile of the 3×3 DG was derived from a one-dimensional solution by Dammann [4.11,4.12] to generate an array of 3 equally intense diffraction orders. The unit cell for this particular solution is characterised by a phase difference, $\Delta\phi = \pi$ radians between neighbouring recessed and raised regions, a unit cell with reflection symmetry and a single independent transition point at $x_1 = 0.132\Delta x$. The expected one-dimensional (two-dimensional) diffraction efficiency for this solution is $\eta_1 = 66.4\%$ ($\eta_2 \sim 44.1\%$). The reason for low efficiency is that power is not well suppressed in the unwanted diffraction orders $n = \pm 2$.

The phase grating was designed to operate in transmission at a centre frequency of 100 GHz (a wavelength of 3 mm). The material used to induce the required phase modulation is high density polyethylene (HDPE), which at 100 GHz has a refractive index, n of 1.525. The two-dimensional phase modulation $\phi(x, y)$ was converted into a dielectric height function $h(x, y)$ using Eq. (4.84). Since $\phi(x, y)$ is a binary function then so too is the surface height function $h(x, y)$ and the difference in height between the milled and unmilled parts of the surface (i.e. the depth of the milled grooves) is given

by Eq. (4.83). For a phase difference of $\Delta\phi = \pi$ radians the required groove depth is then $\Delta h = \lambda_0/[2(n-1)]$, which at the design wavelength $\lambda_0 \approx 3$mm corresponds to a groove depth of approximately 3 mm. The two-dimensional surface profile was milled into one side of a 10mm thick HDPE disc. With four unit cells of periodicity $\Delta x = 27\text{mm} \approx 9\lambda_0$ in both x and y directions the grating dimensions are 108mm × 108mm.

Figure 4-38. The square 3×3 Dammann grating milled from a 10mm thick HDPE disc. The grating is mounted in a Perspex holder, which is supported by an aluminium frame.

Test Arrangement 1

The first measurement of the 3×3 DG was obtained using an inline 4-*f* Fourier optics system i.e. two lenses with the same focal length (HDPE plano-convex lenses of focal length 230 mm [3.2]) that are separated by the sum of their focal lengths and with the input and output planes also coinciding with the corresponding focal planes of the lenses. The grating is placed at the intermediate focal plane (Fourier plane). The measured output plane intensity (Figure 4-39) clearly shows a 3×3 array of Gaussian beams, surrounded by weaker, poorly suppressed second-order beams. In the *x*-direction the angular direction of diffraction order *m* is given by the grating equation as

$$\theta_m = \sin^{-1}\left(\frac{m\lambda}{\Delta x}\right) \tag{4.92}$$

The off-axis position of diffraction order *m* on the output plane (at the focal length f_2 from the focusing lens L_2) is thus $x_m = f_2 \tan(\theta_m)$. In the paraxial approximation $\sin(\theta_m) \approx \tan(\theta_m) \approx \theta_m$ and therefore diffraction order *m* located at

$$x_m \approx f_2\left(\frac{m\lambda}{\Delta x}\right) \tag{4.93}$$

The 3×3 DG unit cell width is $\Delta x \approx 9\lambda_0$ and the 230 mm focal length of the lenses means that the first- and second-order diffraction spots should, at 100 GHz, be located at $|x_{\pm 1}| \approx 25.8$mm and $|x_{\pm 2}| \approx 52.5$mm. The expected diffraction order positions are indicated in Figure 4-39 with black crosses.

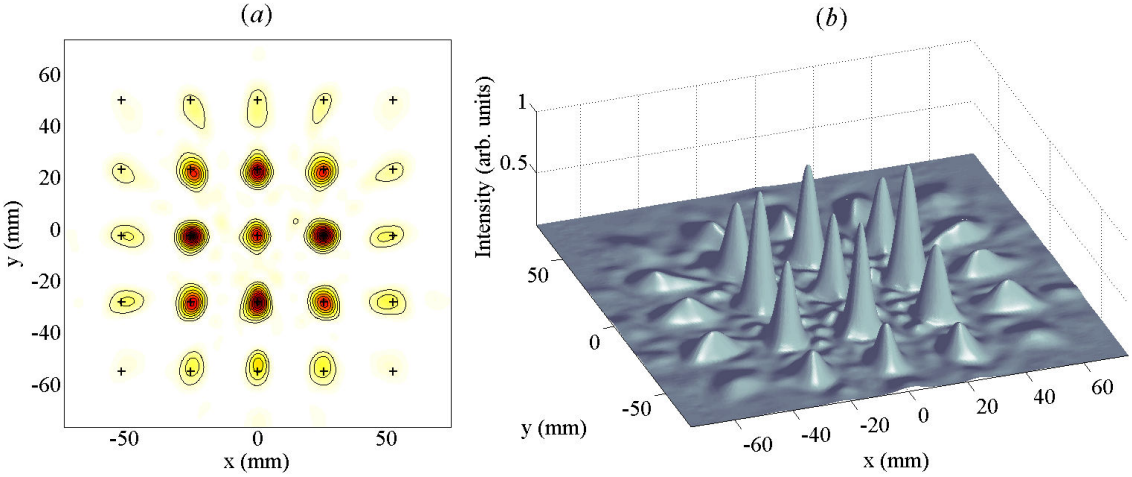


Figure 4-39. (*a*) Contour plot and (*b*) 3-D plot of measured intensity at the output focal plane of lens L_2. The minimum contour has a value of e^{-2} of the maximum measured intensity thus indicating the radius of the Gaussian beams of maximum intensity.

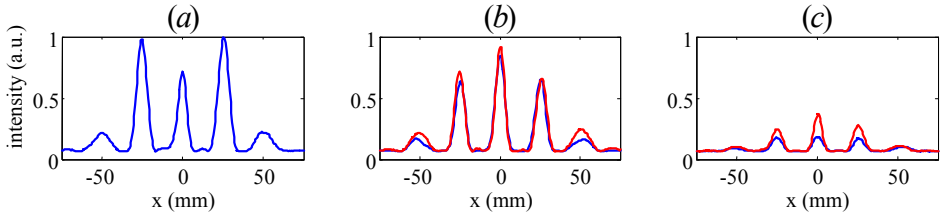


Figure 4-40 Horizontal cuts through measured output plane intensity. Each plot corresponds to a horizontal cut through a different row of beams, with y-direction diffraction order (*a*) $n = 0$, (*b*) $n = \pm 1$ and (*c*) $n = \pm 2$.

Note that the intensity is not evenly distributed between all nine diffraction orders within the 3×3 spot array. The on-axis zeroth-order and four corner beams are less intense than the remaining beams. This may be caused by one of four factors: the source might not have been correctly calibrated so the wavelength of the illuminating radiation may not have been equal to the design wavelength λ_0; the recessed parts of the grating surface may have been cut to an incorrect depth; or the actual transition point locations may not be at their correct positions (i.e. the grooves may have incorrect widths).

Test Arrangement 2

Next the 3×3 DG was measured using a 4-*f* arrangement with the two of the off-axis parabolic mirrors as collimating and focusing elements (see Figure 4-41). The mirrors each have a 350mm focal length and a 90° angle of throw.

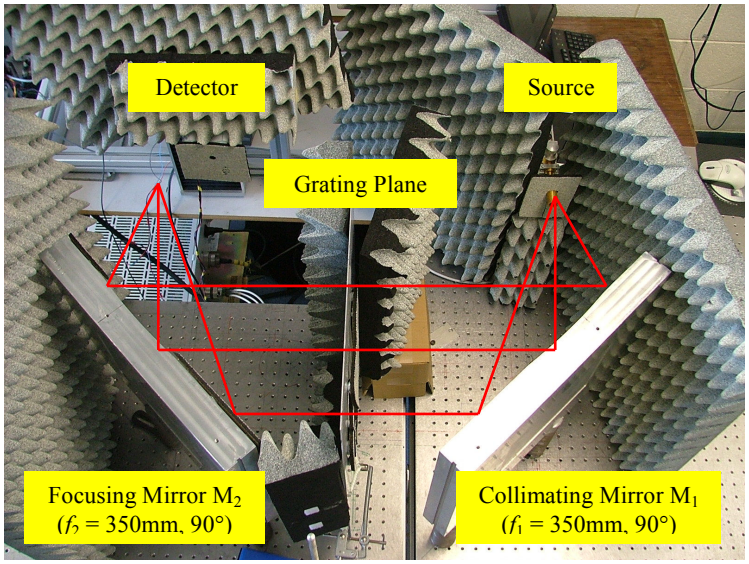


Figure 4-41. The 4-*f* grating test system with two off-axis parabolic mirrors of focal length 350mm and 90° angle of throw. The 3×3 DG is located at the intermediate focal plane of M_1 and M_2. Red lines indicate the path taken by a beam from the source to the detector plane when the grating is not present.

If the Gaussian beam incident on the grating is too small, not enough periods (cells) are illuminated and the far field diffraction pattern resembles more closely the diffraction envelope $T(u)$ of a single cell, rather than that of the multi-celled grating. If, however the grating is over-illuminated (with an incident Gaussian that is much larger than the grating) the grating modulates the phase at the centre of the beam. If not blocked at the grating plane, the beam spill-over at the grating will propagate through the rest of the system (i.e. focused by mirror M_2 onto the output plane) and will interfere with the far field diffraction pattern from the grating. The 350 mm focal length of the parabolic mirrors means that the radius of the Gaussian beam incident on the grating is approximately 71mm (when the source is tuned to 100 GHz) so the grating is over-illuminated. Figure 4-42 shows two different output plane intensity measurements. The image in Figure 4-42 (*a*) was obtained when the beam spill-over was not terminated but allowed to propagate past the grating plane. The image in Figure 4-42(*b*) was measured after an Ecosorb collar was used at the grating to block the outer part of the incident beam. The on-axis zeroth-order diffraction spot is clearly much more intense when the Ecosorb collar is not used, as shown by intensity cuts in Figure 4-43.

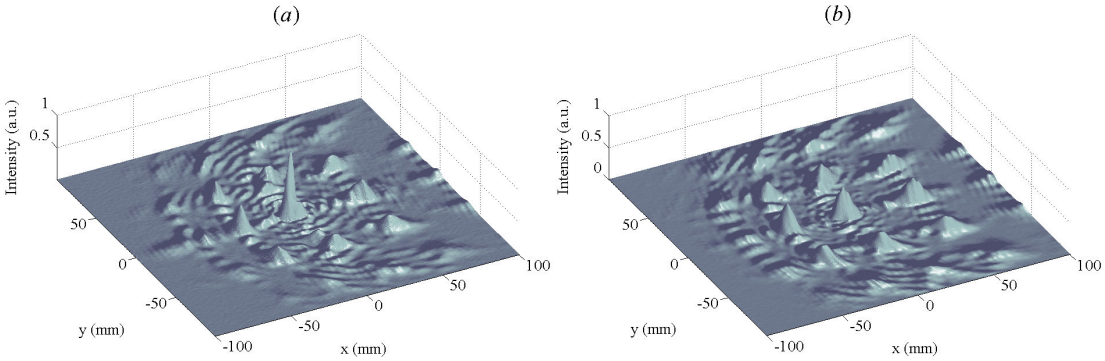


Figure 4-42. Surface plots of output plane intensity images made at 100 GHz (*a*) without and (*b*) with the Ecosorb collar at the grating plane.

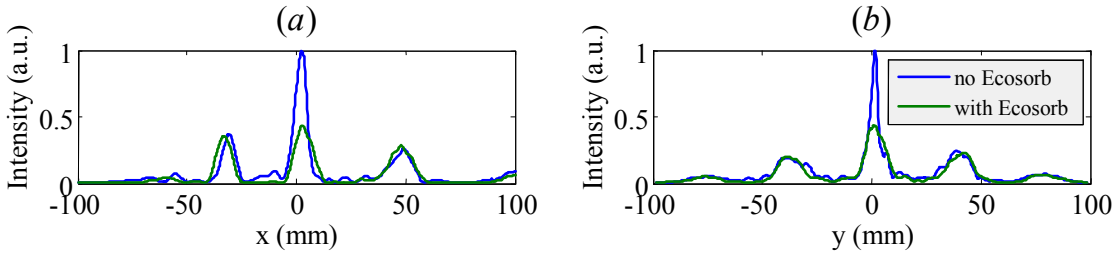


Figure 4-43. Horizontal and vertical cuts through the centre of the intensity measurements in Figure 4-42. The intensity in the on-axis zeroth-order beam is ~2.3 times more intense when the incident beam beyond the edges of the grating is not blocked with Ecosorb.

The appearance of the very intense on-axis peak is explained as follows. The incident beam at the grating plane is the Fourier transform of the source beam (i.e. formed by the source horn antenna). Since the grating filters only the centre of the incident beam, the rest of beam can be considered as the Fourier transform of the source beam with a hole at its centre, which after Fourier transformation by mirror M_2 forms a band-pass filtered image of the source beam (i.e. source horn) on the output plane. We expect this image to be dominated by a relatively intense on-axis single beam of the same width as the beams formed by the grating. At the output plane the filtered image of the source beam interferes with the diffraction pattern from the grating. Since the zeroth-order diffraction spot (produced by the grating) and the filtered image of the source beam are both focused onto the optical axis they interfere to produce an on-axis beam. However if significant power spills past the grating, the source image dominates and the on-axis maximum is much more intense than all other diffraction orders. If the Ecosorb collar is used it acts as an aperture that truncates any part of the beam beyond the grating edges (spillover) and therefore the far field array of diffraction orders is convolved with a sinc function, which results in the appearance of secondary orders between the primary diffraction orders.

A simulation of the 4-*f* system (test arrangement 2 with off-axis paraboloidal mirrors) was developed in MODAL, a screen-shot from which is shown in Figure 4-44. In this model propagation of the wavefront from the grating to the output plane was calculated using the scalar diffraction (Fresnel integrals) option in MODAL.

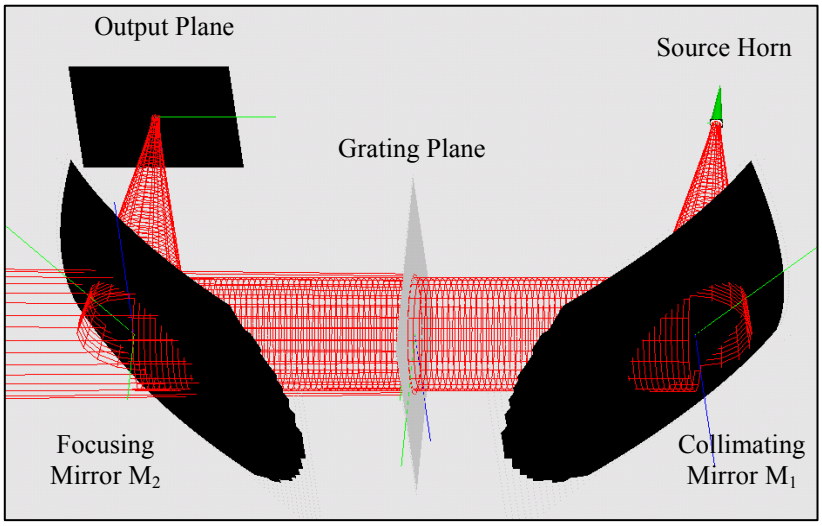


Figure 4-44. Screen shot from MODAL showing the 4-*f* set-up with the off-axis paraboloidal mirrors (with 90° angle of throw) used to test the 3×3 Dammann grating.

The simulated output plane field amplitude is almost identical in form to the amplitude measured at the detector plane (Figure 4-45). Note that by 'measured amplitude' we mean the square root of the measured intensity.

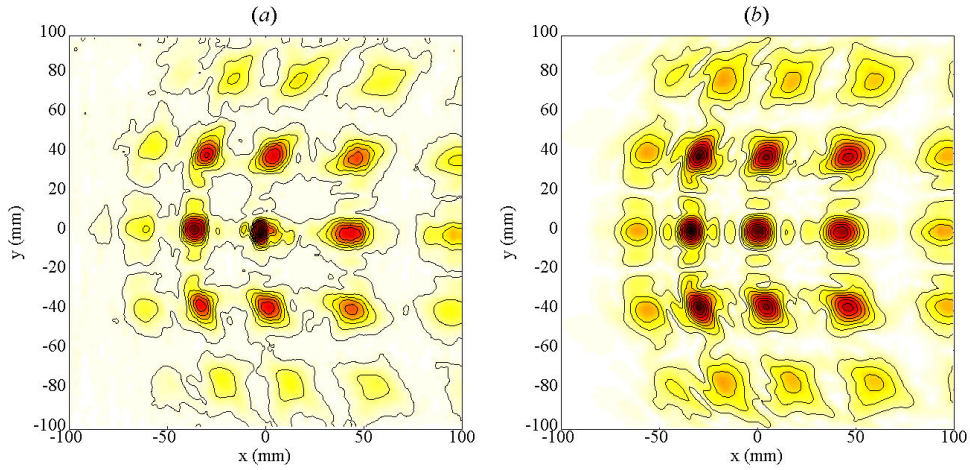


Figure 4-45. Contour plots of (*a*) the measured amplitude and (*b*) field amplitude simulated with MODAL at the output plane of the 4-*f* system (test arrangement 2 with the parabolic mirrors). Contours are in steps of 10% of maximum amplitude from 10% to 100%.

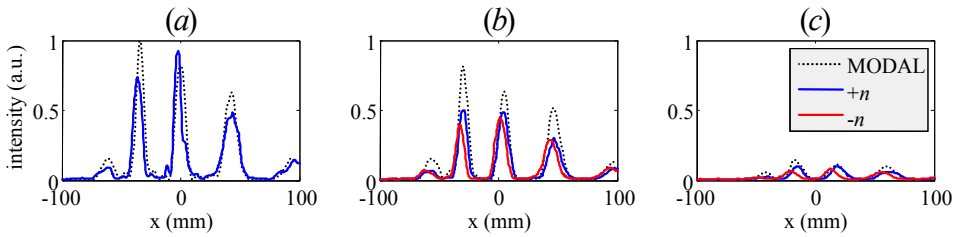


Figure 4-46. Horizontal cuts through MODAL predicted (dotted black curves) and measured (solid curves) output plane intensity images. Each plot corresponds to the intensity through a single row of observed diffraction spots. The plot in (*a*) corresponds to the on-axis row, i.e. with diffraction order $n = 0$, (*b*) to the positive and negative first-order ($n = \pm 1$) rows and (*c*) to the positive and negative second-order ($n = \pm 2$) rows.

The measured and simulated images obtained with the current arrangement are clearly of lower quality than was obtained with the previous in-line set-up (test arrangement 1 with the two plano-convex lenses). The two main sources of aberration are distortion and field curvature. Severe distortional aberrations are introduced because of the high angle of throw of the off-axis focusing mirror M_2. Because of distortion all of the off-axis diffraction orders are shifted right of their expected positions (as shown in Figure 4-45, with the magnitude of deviations increasing from left to right across the output plane).

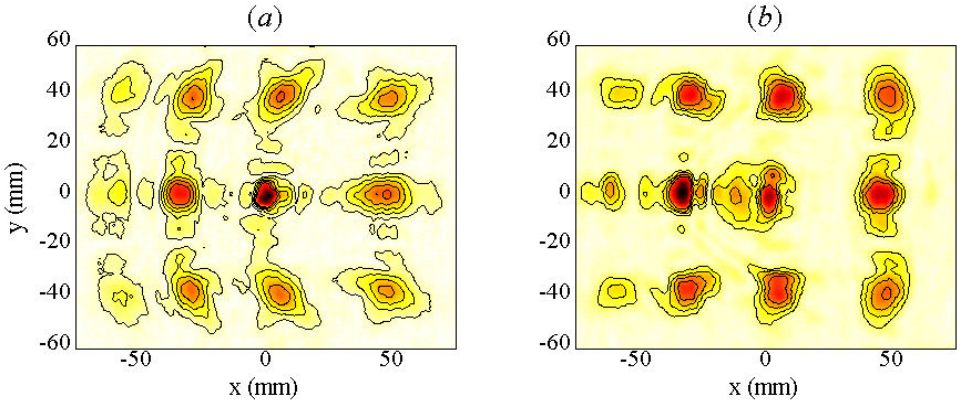


Figure 4-47. Measured output plane intensity from the 4-*f* set-up (test arrangement 2 with off-axis parabolic mirrors) to test the 3×3 DG with the source horn antenna set at (*a*) $z_S = f_1$ and (*b*) $z_S = f_1 + 60$mm from off-axis parabolic mirror M_1 of focal length $f_1 = 350$mm.

The other main source of aberration is field curvature, which is indicated by the fact that only the central diffraction order appears to be in focus while the other diffraction orders are less intense (see Figure 4-46) and the contours in Figure 4-45 are elongated and not circular, indicating that they are not in focus at the detector plane. A measurement to verify that field curvature was indeed an issue was performed: the

source horn antenna was set at a distance $z_S = f_1 + 60\text{mm}$ from mirror M_1, which, since mirrors M_1 and M_2 have equal focal lengths, means that the image formed by mirror M_2 will be focused at a 60mm in front of the detector plane (i.e. closer to M_2). The measurements obtained for source distance $z_S = f_1$ and $z_S = f_1+60\text{mm}$ are shown in Figure 4-47. The first-order spots are more intense and are in focus in the second image compared to the first.

Bandwidth Characteristics of Dammann gratings (3×3 DG)

A series of measurements of the 3×3 Dammann grating was made at eight frequencies spanning the W-band (between 75 and 110 GHz in steps of approximately 5 GHz). The exact frequencies used along with the expected first- and second- order diffraction order positions from the 3×3 DG at each frequency are given in Table 4-3.

ν (GHz)	λ (mm)	$\lvert x_{\pm1}\rvert$ (mm)	$\lvert x_{\pm2}\rvert$ (mm)
74.67	4.01	52.63	109.02
79.87	3.75	49.13	101.31
84.77	3.54	46.24	95.01
90.36	3.32	43.34	88.74
95.54	3.14	40.94	83.64
99.524	3.01	39.29	80.12
105.246	2.85	37.13	75.55
110.008	2.73	35.51	72.14

Table 4-3. Source frequencies ν and equivalent wavelengths λ used in measurements of the 3×3 DG and the expected output plane coordinates of first and second order diffraction spots assuming a grating period of 27 mm and focal length $f_2 = 350\text{mm}$.

As explained in Chapter 3 two Gunn diode sources were used to cover the required frequency range. The first six measurements were made with one source (whose frequency range is 75 to 100 GHz) and the remaining two measurements with the other (whose frequency range is 100 to 110 GHz). False-colour plots of the measured intensity and contour plots of measured amplitude at each of these frequencies are shown in Figure 4-48 and Figure 4-49, respectively. The false-colour plots give a good sense of overall intensity distribution, while the contour plots are included to highlight low-level features.

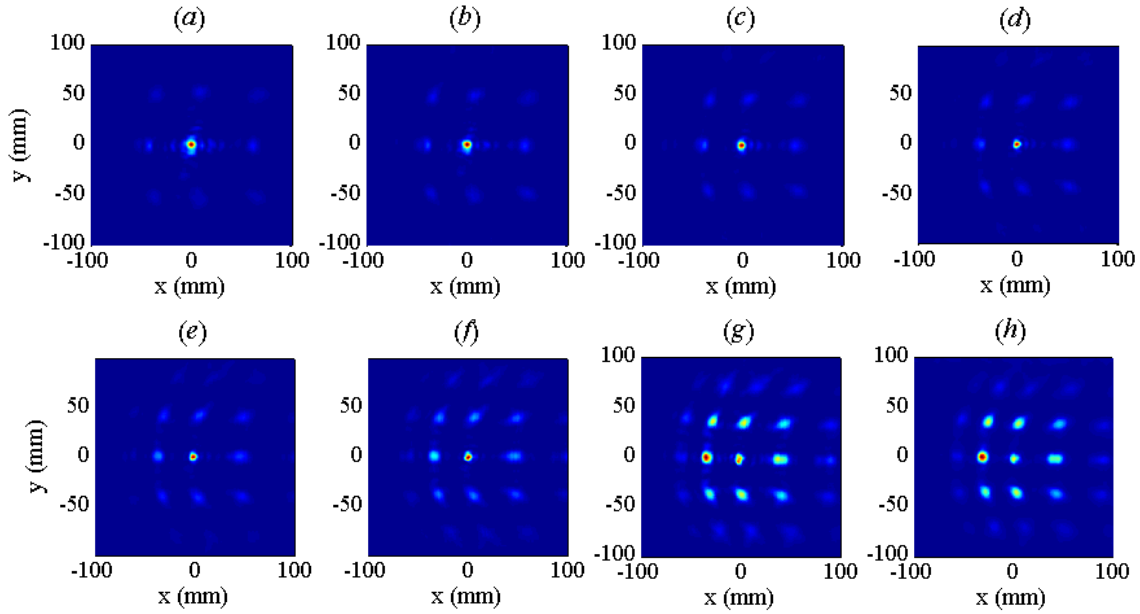


Figure 4-48. Measured output plane intensity patterns from the 3×3 DG at frequencies of (*a*) ~75 GHz to (*h*) ~110 GHz in steps of 5 GHz.

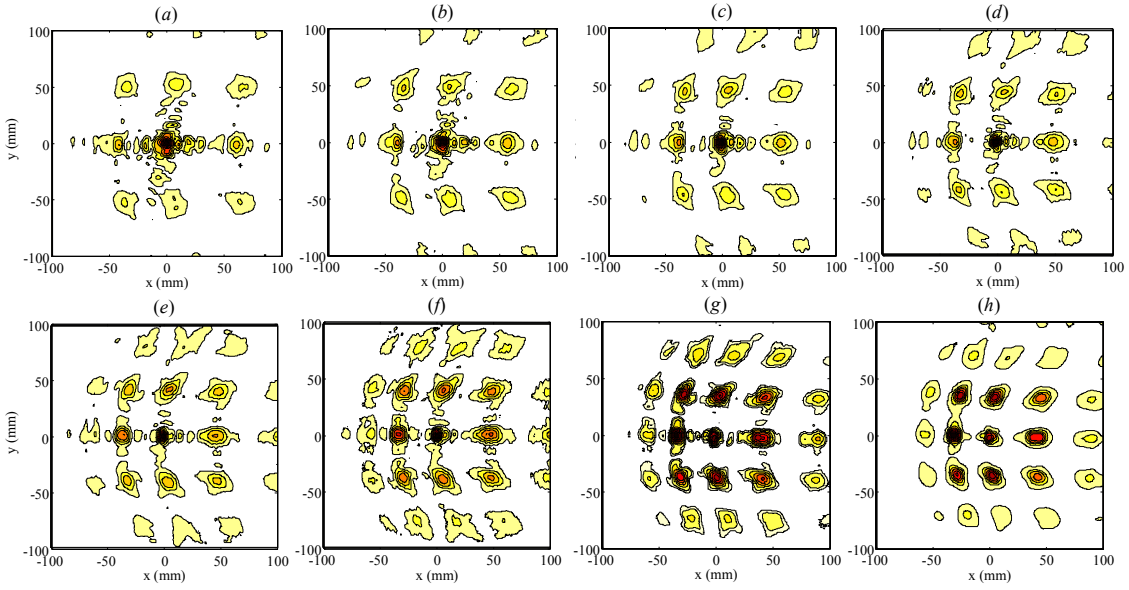


Figure 4-49. Contour plots of measured field amplitude patterns at the output plane at frequencies of (*a*) ~75 GHz to (*h*) ~110 GHz in steps of 5 GHz. Contours are at 10% intervals between the minimum and maximum amplitude values of each image.

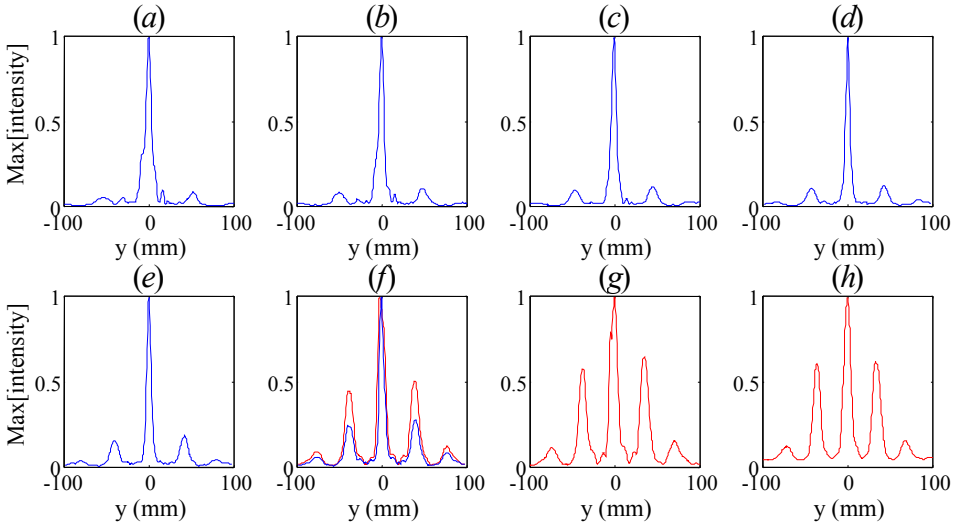


Figure 4-50. Cuts across central row with spot of maximum intensity for frequencies 75 (*a*) to 110 GHz (*h*) in 5 GHz intervals. Blue curves show results obtained using the first source whose tunable frequency range is 75-100 GHz and without Ecosorb surrounding the grating. Red curves show results obtained with the second source (100-110 GHz) with Ecosorb surrounding the grating. A measurement made at 100 GHz with both sources, is shown in (*f*). Note that the last plot, corresponding to 110 GHz scan, has very low contrast. The low power level available from the source at this frequency means that images obtained at this frequency have a relatively low signal-to-noise ratio (SNR) and the filtering used to extract a useful image results in low contrast.

The 75-100 GHz measurements were made without the Ecosorb collar surrounding the grating, which of course caused undesirable effects with respect to the intensity of the on-axis maximum (Figure 4-50). Ignoring the fact that the central on-axis diffraction spot is extremely intense, the relative intensity of the first-order diffraction spots increases with frequency as we approach the design frequency. The maximum first-order intensity occurs for the 105 and 110 GHz cases. If we take the refractive index of HDPE to be 1.54 then for a groove depth of 3 mm a phase pattern with steps of π radians occurs at a frequency of 108 GHz – thus in agreement with measurements.

Modelling the frequency response of the 3×3 DG

To complement the experimental measurements, the frequency response of the grating was modelled as a one-dimensional function using FFT. Grating operation over a 150 GHz bandwidth about a design frequency, $\nu_0 = 100$ GHz is illustrated in Figure 4-51. At each frequency the phase modulation $\phi(x)$ is calculated, the phase-modulated incident plane wavefront propagated to the far field (using FFT) and the one-dimensional diffraction efficiency, η_1 was calculated.

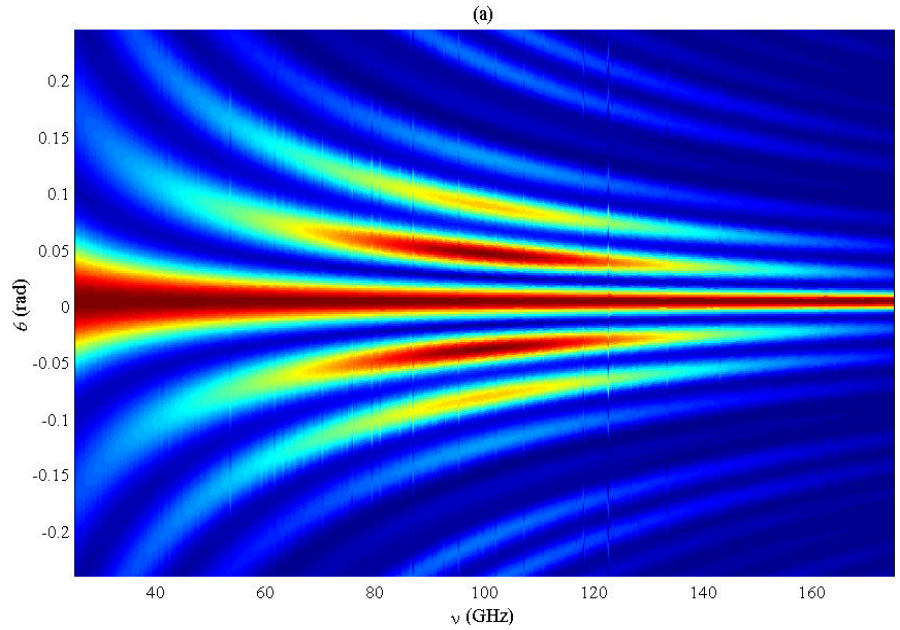


Figure 4-51. Simulated one-dimensional far-field amplitude pattern from a 3-spot DG (designed to yield maximum diffraction efficiency at a design frequency of ν_0 = 100 GHz) evaluated at frequencies between 25 and 175 GHz (on the *x*-axis). The grating width is $4W_x \approx 281$mm for $W_x \approx 71$mm at ν_0. The colour scheme used displays lower intensity values as dark blue and higher intensity values as red.

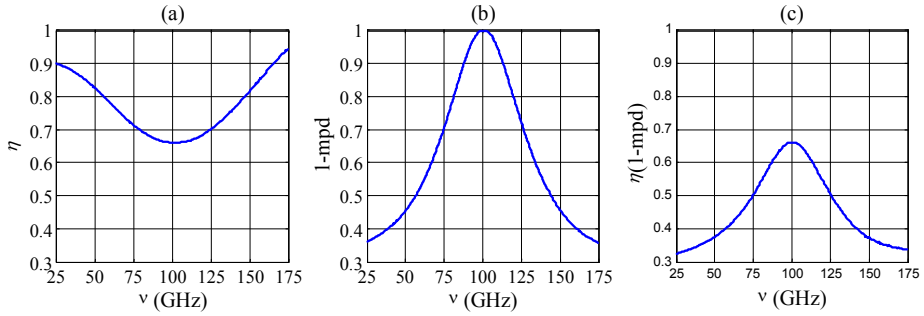


Figure 4-52. Performance of the 3-spot DG in terms of one-dimensional (*a*) grating diffraction efficiency η_1, (*b*) mean power uniformity (1–*mpd*) between the 3 equi-intense diffraction orders and (*c*) weighted efficiency $\eta_{\text{mpd}} = \eta_1(1-mpd)$

Besides the usual frequency-dependence of the angular spacing between diffraction orders, the frequency response (Figure 4-51) appears to be reasonably symmetric about the design frequency, ν_0 = 100 GHz. This behaviour is reflected in the plots of grating efficiency and mean-power uniformity (Figure 4-52) and is explained as follows.

At the design frequency ν_0 the two phase levels $\phi_1 = 0$ and $\phi_2 = \pi$ produce a phase difference of $\Delta\phi_0 = \pi$ which corresponds to a delay of half a wavelength. At a lower frequency, $\nu^- = (\nu_0 - \Delta\nu)$, from equation (4.91) the phase difference is $\Delta\phi^- = -\pi(\Delta\nu/\nu_0)$. Since only two phase levels are involved the value of $\Delta\phi$ at a higher frequency

$\nu^+ = (\nu_0 + \Delta\nu)$ is $\Delta\phi^+ = +\pi(\Delta\nu/\nu_0)$, i.e. the same in absolute value as at the lower frequency ν^-. Thus for a binary-level DPE, except for a change in sign, the phase difference $\Delta\phi$ imparted on an incident wavefront is symmetric about the design frequency ν_0.

The useful bandwidth of the device can be gauged by examining the range of frequencies over which the desired diffraction pattern is produced. For this modelled grating good uniformity between the three central diffraction orders is maintained reasonably well between 80 GHz and 120 GHz, i.e. a bandwidth of ±20% about the centre frequency. Unfortunately the symmetric nature of the frequency response could not be demonstrated through experimental measurements because the design frequency of the grating is 108 GHz and the upper source frequency available to us was 110 GHz.

Truncation analysis of the 3×3 DG with GBMA

Now we investigate what effects, if any, truncation by the collecting element (lens/mirror) has on the experimental measurements of the 3×3 Dammann grating. Truncation analysis (described in Chapter 2) is particularly important for phase grating design because the phase-modulated wavefront scatters to larger off-axis distances the further it travels from the grating. So whereas a beam with a simple Gaussian-profile might pass largely unobstructed through a lens/mirror, the same optical element may cause problems when used to focus the diffraction pattern generated by a phase grating.

First we present truncation analysis of the 3×3 Dammann grating with the experimental set-up using two HDPE lenses (each with 230 mm focal length and 220 mm diameter). The mode parameters that define the set of Gaussian beam modes used to decompose the field transmitted through the grating were chosen by setting the highest spatial frequency of the mode set equal to the minimum feature size of the grating, which gives a mode set with highest-order mode, $m_{max} = 30$ and a fundamental Gaussian beam waist radius at the grating of $W_0 = 0.3636\Delta x$, where $\Delta x = 27$mm for the 3×3 DG. Since the grating has the same profile in both the x and y directions initially we need only calculate mode coefficients A_m of a 1-D grating field $E_G(x)$. The modes are then propagated to the truncating aperture of lens L_2 at $z = f_2$. Since the aperture at L_2 is circular, truncation and subsequent propagation to the output plane must be treated two-dimensionally. The 2-D field incident on the lens, $E_{in}(x, y)$ is created and multiplied

by a binary mask representing the circular aperture (of radius a) of the lens to give the field $E_{out}(x, y)$ transmitted through the lens.

Next the truncated mode coefficients B_{mn} are calculated by performing the relevant overlap integral. The symmetric nature of the grating and the aperture means that only symmetric (even-numbered) modes contain power, so only 16×16 modes are needed to describe the wavefront at lens L_2 and beyond. The small number of modes involved means that the pseudoinverse approach can be used to calculate the scattered mode coefficients B_{mn}. Figure 4-53 shows the magnitude of the symmetric (even-numbered) mode coefficients before and after truncation, i.e. $|A_{mn}|$ and $|B_{mn}|$. The only difference between the plots is that slightly more power exists in some higher-order modes after truncation. Thus we should expect that truncation by this particular lens will have little effect on the output plane intensity $|E_F|^2$.

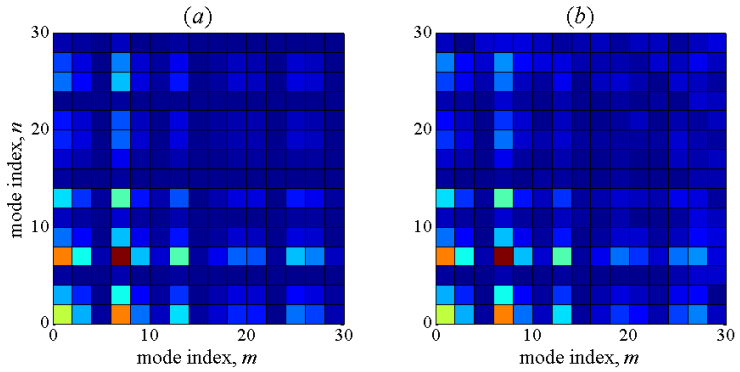


Figure 4-53. Mode coefficient magnitudes for the 3×3 Dammann grating (*a*) before and (*b*) after truncation by a circular aperture of radius a = 110mm representing the aperture of the 230mm focal length lens L_2. Only symmetric (even-numbered) mode coefficients are displayed displayed.

Figure 4-54 shows log-scaled plots of the beam intensity before and after truncation at lens L_2. Since we are only interested in the field within the aperture of the lens/mirror, to maximise resolution at the lens plane, the field $E_{in}(x, y)$ is calculated on a rectangular array whose height and width is equal to the aperture diameter ($2a$). Thus even before truncating with a circular binary mask, $E_{in}(x, y)$ is already a truncated version of the field from the grating. To calculate the truncated mode coefficients B_{mn} with a scattering matrix S_{mn} requires a slight modification. By the time the Gaussian beam modes have reached the truncating aperture, they will probably have spread into an area that is larger than the size of the aperture. Thus the modes at the lens plane must now be defined on a rectangular array that is large enough to contain the highest-order mode. Since the width of a Gaussian-Hermite mode of order m is approximately given by $2W_0\sqrt{m}$ following

on from arguments in Chapter 2 the modes were defined over a plane of width $2W_0\sqrt{m_{max}}$ and height $2W_0\sqrt{n_{max}}$.

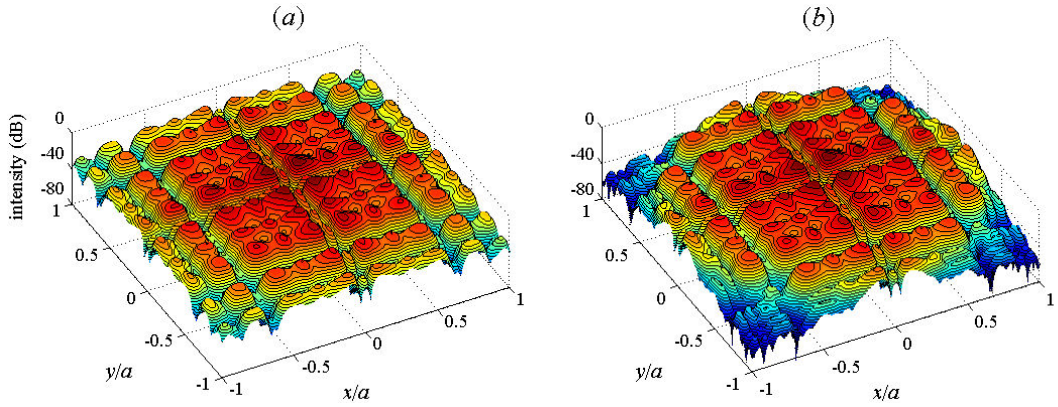


Figure 4-54. Log-scaled plots of the intensity of the GBM-approximated field from the four-cell 3×3 Dammann grating at lens L_2 (*a*) before and (*b*) after truncation with a circular aperture of radius a = 110mm.

The output plane intensity patterns with and without truncation effects are shown in Figure 4-55 & Figure 4-56. The difference in intensity between the central array of nine diffraction orders with and without truncation effects included is minimal. The only noticeable difference when truncation occurs is that the intensity level between diffraction orders is higher; the intensity of the spots outside the central block of nine diffraction orders is lower; and the diffraction orders further off-axis have a less circular profile than those closer to the optical axis (similar to what was observed in experiments measurements – see Figure 4-39).

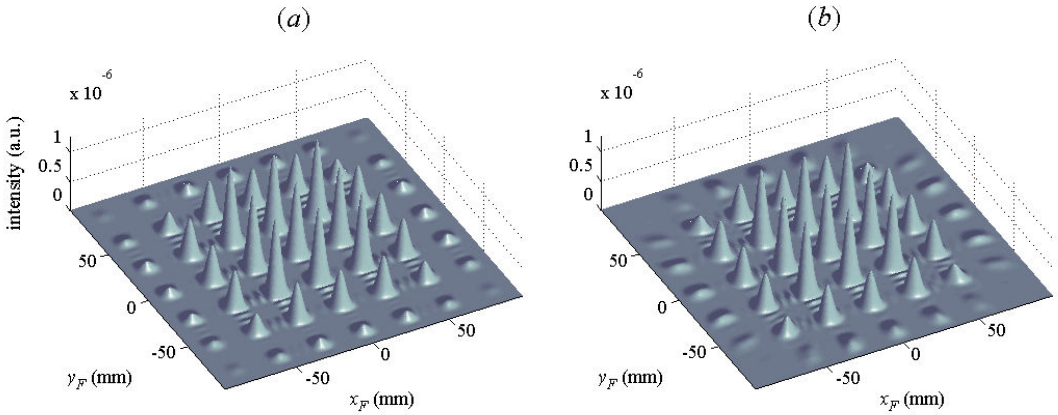


Figure 4-55. Linear plots of simulated output plane intensity distribution (*a*) without and (*b*) with truncation effects included at lens L_2 (f_2 = 230mm and a = 110mm).

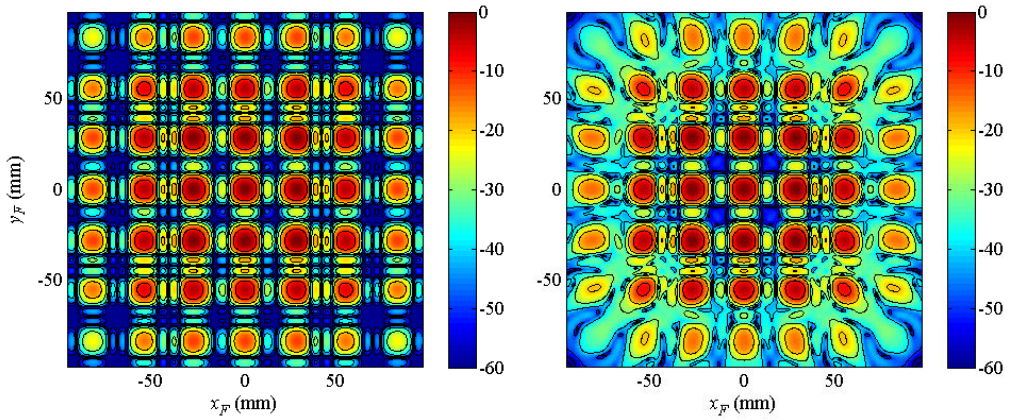


Figure 4-56. Log-scale plots of simulated output plane intensity distribution (*a*) without and (*b*) with truncation effects included at lens L_2 (f_2 = 230mm and a = 110mm).

Next truncation analysis was performed for the second experimental arrangement: the 4-*f* system with two 350mm focal length mirrors. These mirrors were designed to have a radius of a = 142.37mm. The mode-set was chosen so that it could reproduce details in the grating plane half the size of smallest feature, which required a mode-set with m_{max} = 60 and W_0 = 6.942mm. The mode coefficients magnitudes before and after truncation, i.e. $|A_{mn}|$ and $|B_{mn}|$ are shown in Figure 4-57. As with the 230 mm focal length lens, there is very little scattering of power between modes, so again we expect that truncation will have little impact on the output plane intensity.

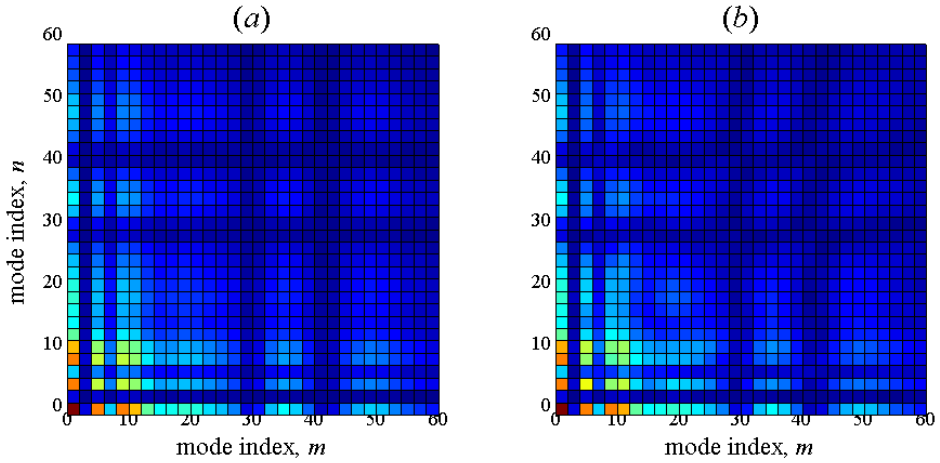


Figure 4-57. Magnitude of (*a*) input mode coefficient $|A_{mn}|$ and output (scattered) mode coefficients, $|B_{mn}|$ from the 3×3 DG for truncation with a circular aperture (of radius, a = 142.37mm) representing the aperture of mirror M_2 (focal length f_2 = 350mm) used in test arrangement 2 (off-axis parabolic mirrors).

Figure 4-58 shows the simulated output plane intensity patterns with and without truncation effects at M_2 included. Figure 4-59 shows x and y cuts through the intensity at the centre of the output plane. In this case a slight decrease in intensity of the central

beams is observed. However the decrease is the same for all nine beams so array intensity uniformity is unaffected. Note that neither aberrational losses, nor transmission losses in lenses are accounted for in our GBMA model of truncation, which limits its usefulness to on-axis systems.

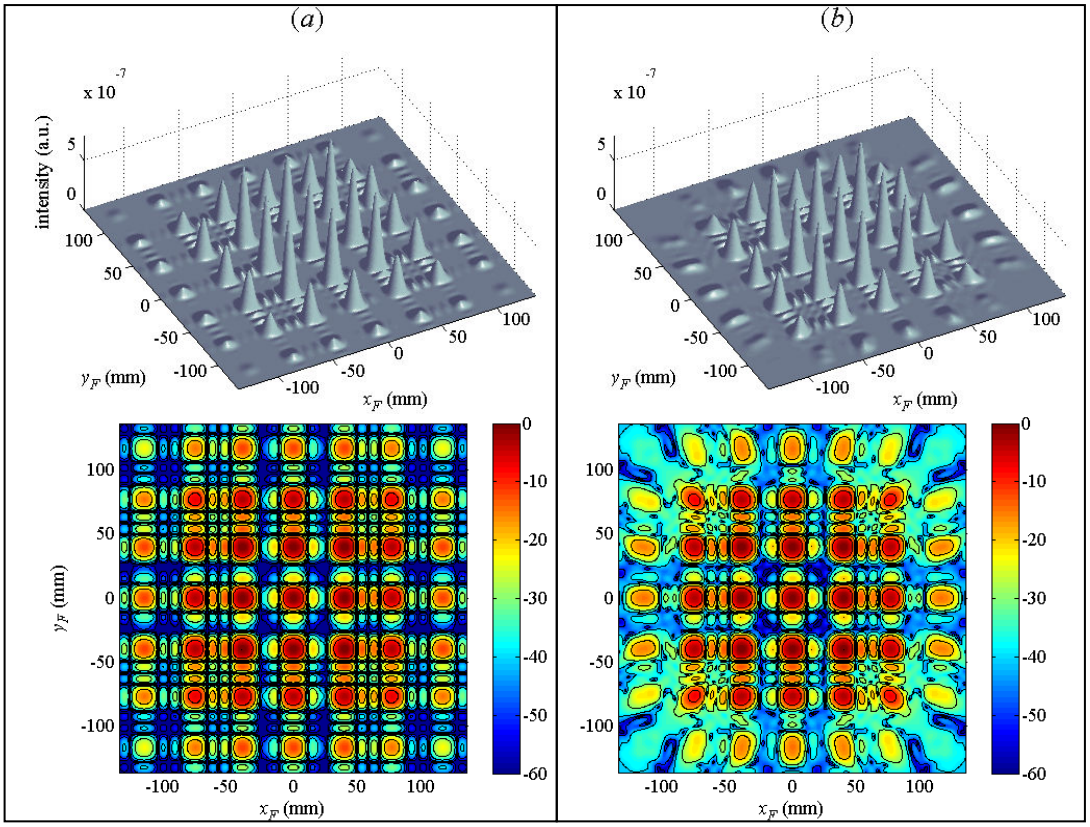


Figure 4-58. Linear (top) and log-scale (bottom) plots of simulated output plane intensity from the 3×3 DG (*a*) without and (*b*) with truncation at mirror M_2. (f_2 = 350mm, a = 142.37mm)

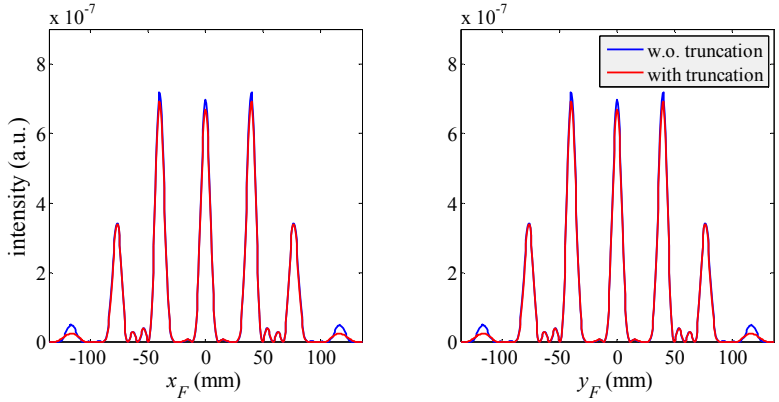


Figure 4-59. X- and Y- cuts through centre of output plane intensity patterns from the 3×3 DG with and without truncation by the 350mm focal length mirror M_2. As well as the reduced intensity in the off-axis diffraction orders $|m|$, $|n|$ = 3, the intensity in the three central diffraction orders is also lower. However since all three central orders decrease by the same amount in intensity beam uniformity is unaffected.

4.7.2 Transmission Dammann Grating (5×5 spot array)

The second Dammann grating was designed to generate a square array of 25 beams in a 5×5 configuration and is hereafter referred to as the 5×5 DG. It was designed by another member of the THz Optics group during a previous research programme [4.18] and originally tested with a 4-*f* Fourier optics arrangement (with two 150 mm focal length parabolic mirrors) and measured with the scanning tool GHOST. The higher resolution and faster scan times offered by the new scanning device TOAST (see Chapter 3) and the greater range of optical components (the newly designed paraboloidal and ellipsoidal mirrors) now at our disposal permitted retesting of this grating using a number of optical arrangements with different combinations of mirrors.

Design and Fabrication

The surface profile of the 5×5 DG was derived from a solution by Dammann [4.11], which specifies a binary one-dimensional unit-cell phase function $\phi(x)$ characterised by the transition points, $x_t = \pm\{0.132, 0.481\}\Delta x$ and a π phase difference between the two phase levels. The grating has a cell period of $\Delta x = \Delta y = 32$mm and the unit-cell is repeated four times in both the x and y directions, so the square grating is 128 mm wide.

The surface relief profile was cut into a quartz plate of thickness ~6.2mm using the technique of diamond turning. Whereas the phase modulation for the previously described 3×3 DG was realised by milling the 2-D surface profile entirely onto one side of a HDPE disc, for this grating the phase function was imposed by machining the surface profile of a linear 1-D grating (that generates a linear 1-D array of 5 equi-intense diffraction orders) onto each side of the quartz plate with the direction of dispersion at the two sides orthogonal to each other to provide the required 2-D phase modulation. With this arrangement the surface on each side of the plate consists only of either horizontally or vertically aligned linear grooves and so can be machined with greater surface accuracy.

Test Arrangement 1

The first measurements of the 5×5 DG were made using the 4*f* Fourier optics system that was used to measure the 3×3 DG with the two parabolic mirrors (of focal length f = 350mm and 90° angle of throw). The output plane intensity was measured at a frequency of ~100 GHz (its design frequency) and is shown in Figure 4-60. As with the

3×3 DG, the output plane pattern obtained with this set-up is extremely degraded due to severe distortional aberrations introduced by mirror M_2. A simulation of the test arrangement was developed in MODAL and the output plane intensity pattern predicted (using scalar diffraction) by this model (Figure 4-61) is in very close agreement with the experimental measurements (Figure 4-60).

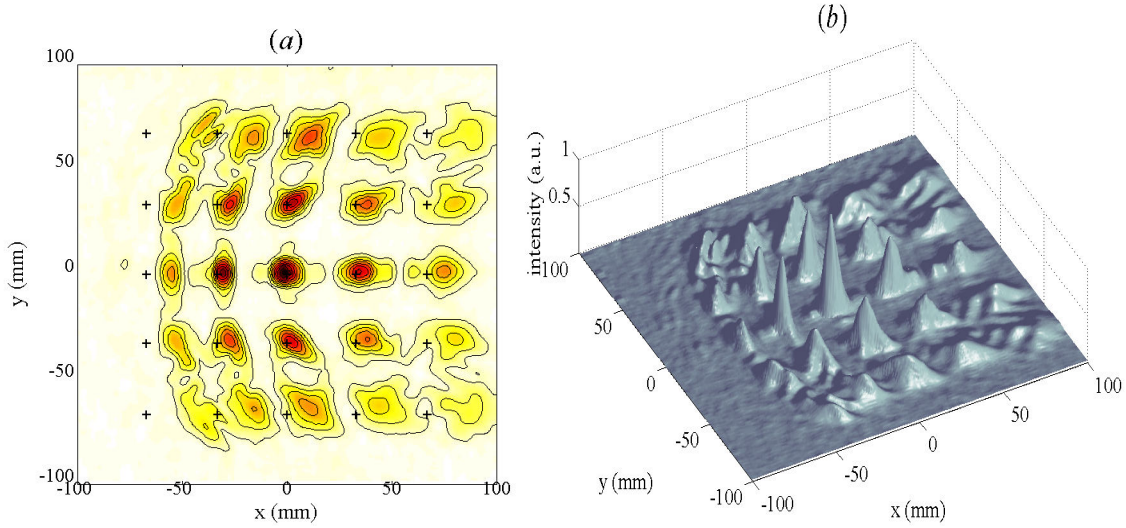


Figure 4-60. Output plane patterns measured at a frequency of 99.7 GHz with the 4-*f* Fourier optics arrangement using two off-axis parabolic mirrors. The two plots show (*a*) amplitude in contours and (*b*) a 3-D plot of intensity. The high angle of throw of parabolic mirror M_2 causes extreme distortion.

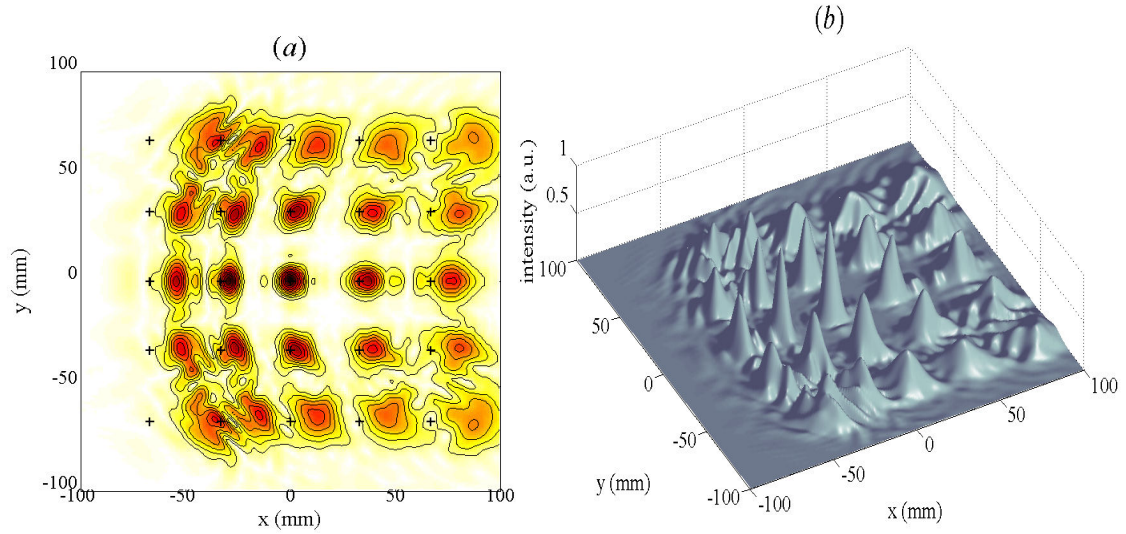


Figure 4-61. MODAL simulated output plane intensity with the 4-*f* Fourier optics arrangement comprising two parabolic mirrors (f = 350, angle of throw = 90°) at a frequency of 99.9GHz (λ = 3mm). Plot (*a*) shows a contour plot of amplitude and (*b*) shows a 3-D plot of intensity.

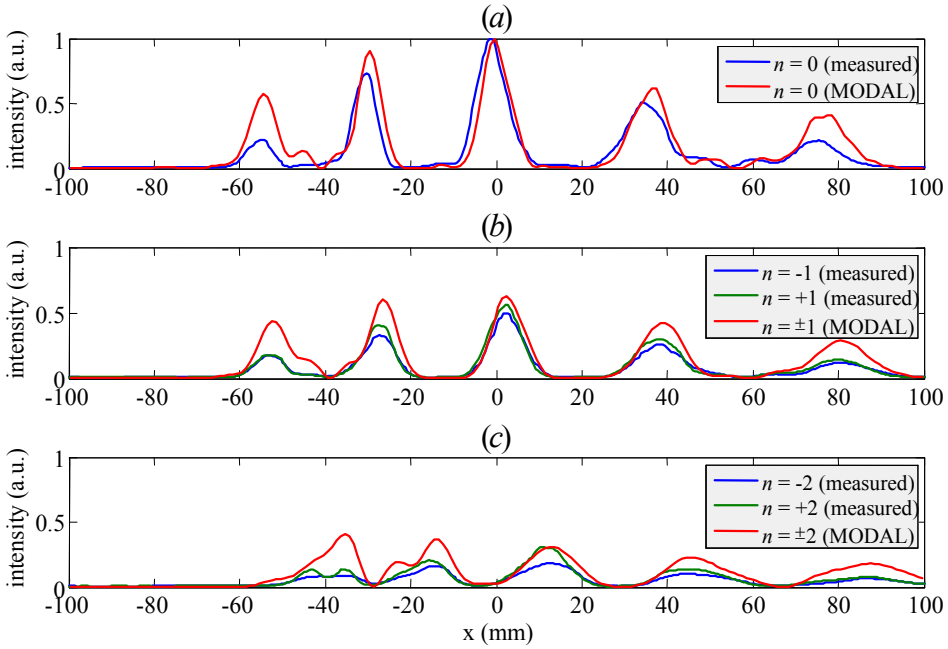


Figure 4-62. Cuts through the centre of the five horizontal rows of diffraction orders in the measured and MODAL simulated (using MODAL) intensity patterns from the 5×5 DG at the output plane of the 4-*f* system at ~100 GHz.

With reference to the test arrangement shown in Figure 4-44, the effects of distortion in a 4-*f* system with single beam propagation can be removed using a compensating set-up (Figure 4-63) in which the beam path through the system is S-shaped. This arrangement was simulated in MODAL for the 5×5 DG but with no improvement (in terms of a reduced amount of distortion). This is not unexpected since several widely spread beams are now involved instead of simply an on-axis beam.

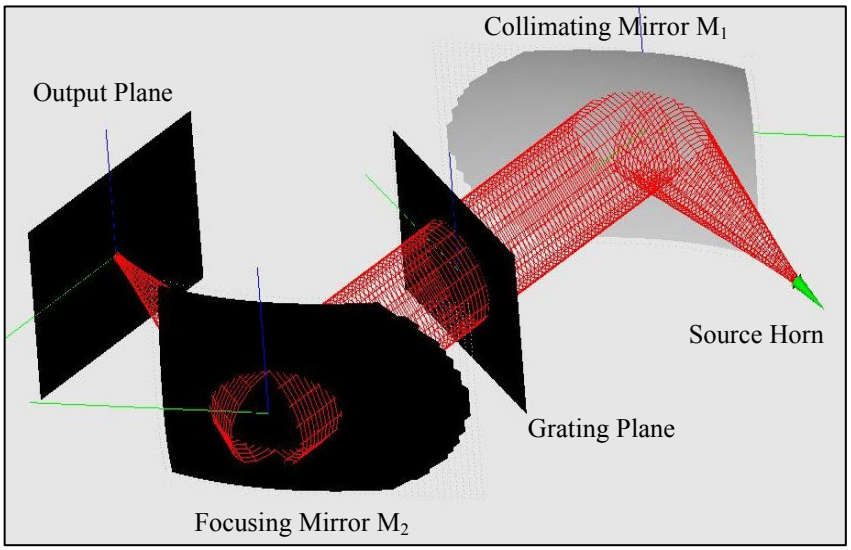


Figure 4-63. Screen shot from MODAL showing a compensated 4-*f* Fourier optics system with the two off-axis paraboloidal mirrors M_1 and M_2 arranged so that the beam path (indicated by red mesh) is S-shaped.

Test Arrangement 2

Another set of measurements was made using a compensated 4-*f* Fourier optics set-up with two off-axis ellipsoidal mirrors with 500 mm focal lengths and 45° angle of throw (see Figure 4-64). The smaller angle of throw means that the amount of distortion that was seen when an off-axis parabolic mirror (with a 90° angle of throw) was used to focus the diffraction orders onto the output plane should be reduced considerably. The longer focal length of mirror M_1 means that the grating is now illuminated with a Gaussian beam of radius W_G = 101 mm. Thus truncation will be more of an issue than it was for the off-axis parabolic mirrors with focal lengths of 350 mm.

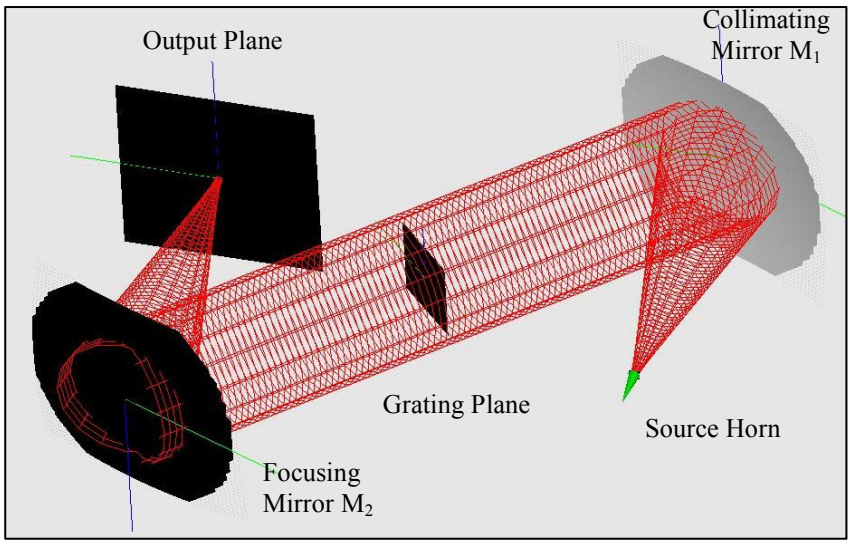


Figure 4-64. Screen shot from MODAL showing a compensating 4-*f* Fourier optics system comprising two ellipsoidal mirrors M_1 and M_2 (each with 500 mm focal length and 45° angle of throw). The longer focal lengths means that the grating is relatively small compared to the illuminating beam size.

Figure 4-65 shows the affect of collimating the source beam with ideal focusing elements of various focal lengths. Cleary, as shown in Figure 4-65(*a*), a collimating mirror M_1 with a longer focal length produces a larger Gaussian beam radius W_G at the grating plane. However, the greater the value of W_G, the more truncated the Gaussian beam for a fixed grating width. The effect at the output plane, as shown in Figure 4-65(*b*) is that image formation appears more like that associated with uniform (top-hat) illumination than with Gaussian illumination with the appearance of secondary maxima between primary maxima (the diffraction orders), the relative intensities of which increase as the incident Gaussian beam width is made larger. Furthermore, a wider incident Gaussian beam produces a narrower output plane Gaussian beam. This

particular grating (of width L = 128 mm) is best illuminated by a Gaussian with a waist radius of $W_G = 0.32 \times L$. With the particular source horn antenna used this requires a lens or mirror of focal length 200 mm. The original measurements made of the 5×5 DG used 150mm focal length mirrors, which provided a Gaussian beam waist of radius $W_G = L_x/4$. However such a small Gaussian beam incident on the grating will result in wide output plane Gaussian beams that overlap slightly with non-negligible intensity between the diffraction order peaks.

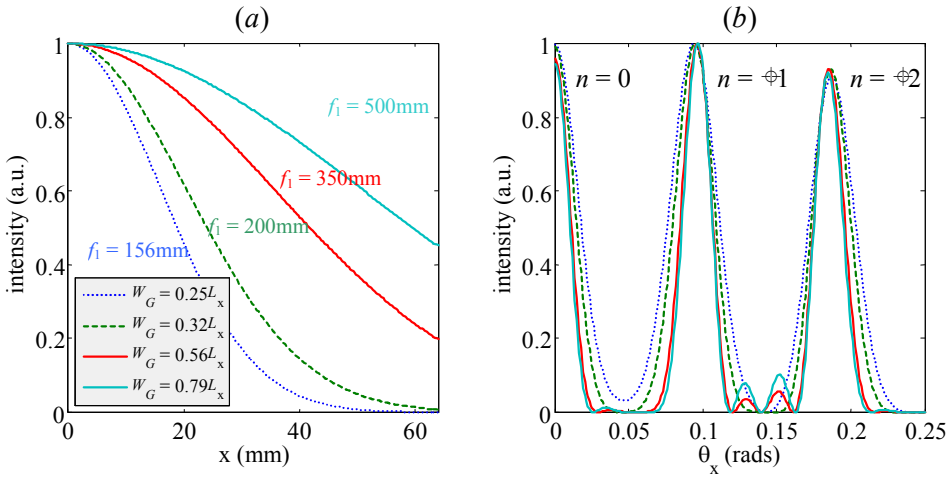


Figure 4-65. 5-beam Dammann grating illumination with a collimated Gaussian beam provided by an idealised optical element of various focal lengths (*a*). A shorter the focal length produces a smaller value of Gaussian beam waist W_G. The far field intensity is shown in (*b*) for a number of different sized Gaussian beams incident on a 1-D 5×5 DG.

The experimentally obtained output plane intensity pattern from the 5×5 DG is shown in Figure 4-66. Sheets of Eccosorb were used to absorb any overspill of the illuminating beam from interfering with the on-axis diffraction order. The system was also simulated with the MODAL software and the calculated intensity at the output plane is shown in Figure 4-67. The most noticeable difference between the simulated and measured beam patterns is in the degree of uniformity in intensities of the various diffraction orders, which are less uniform in the measured than in the simulated image. The diffraction envelope is much more uniform in the simulated images than in the measured images, which shows a significant decrease in power with increased distance in the *x*-direction from the on-axis (zeroth-order) diffraction order. Again the uneven intensity distribution observed in the experimental measurement may be due to the source not being correctly calibrated and so the emitted radiation not being equal to the design frequency.

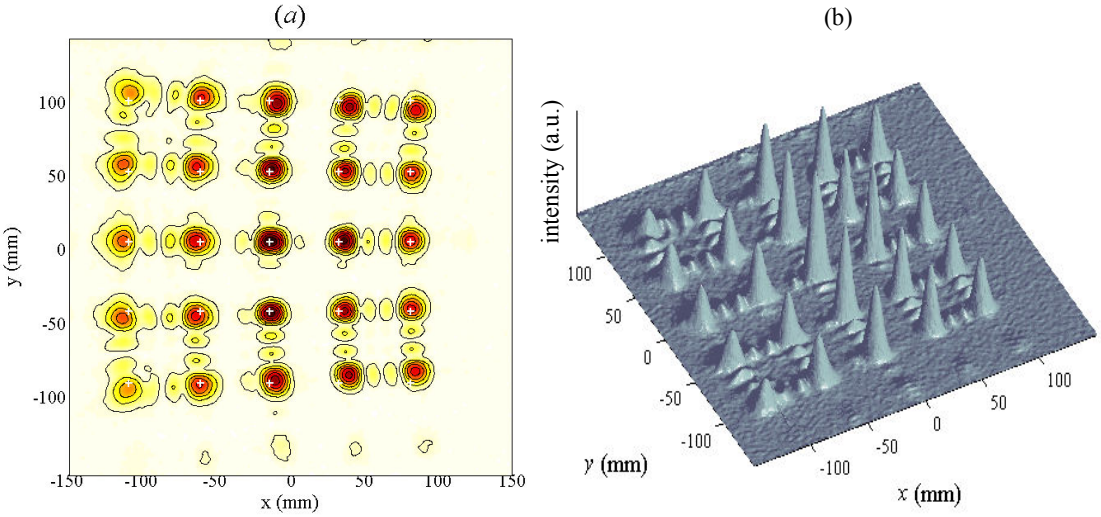


Figure 4-66. Intensity measurements of diffraction patterns generated by 5×5 DG obtained using the compensated 4-*f* Fourier optics arrangement illustrated in Figure 4-64. The two secondary maxima between adjacent primary maxima (diffraction orders) are due to the relatively large size of the incident Gaussian beam illuminating the grating. Some distortion is introduced by the ellipsoidal mirror, which is manifested as a decrease in vertical beam spacing from left-to-right. As with the parabolic mirrors the diffraction orders that are further off-axis appear out-of-focus compared to the on-axis beams, indicating that the focal plane of the ellipsoidal mirror is curved. To ensure maximum coupling to the beam array a set of horns would need to be arranged such that their phase centres lie on the curved focal plane.

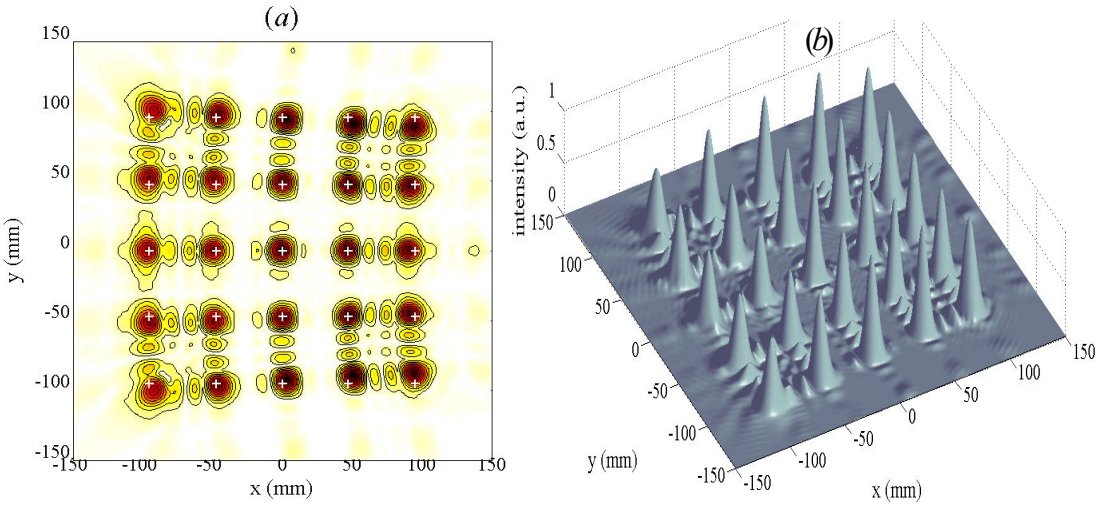


Figure 4-67. The output plane diffraction pattern calculated using a simulation developed in MODAL showing (*a*) amplitude and (*b*) intensity profiles at a frequency of 99.9 GHz (λ = 3mm).

Frequency Response of the 5×5 DG

Measurements were made at eight frequencies between 75 and 110 GHz using the current set-up to evaluate the frequency response of the 5×5 DG. The scanned area was reduced to a narrow vertical strip (15 mm wide and 300 mm long) centred on the on-axis diffraction order. Contour plots of the measured intensity are shown in Figure 4-68.

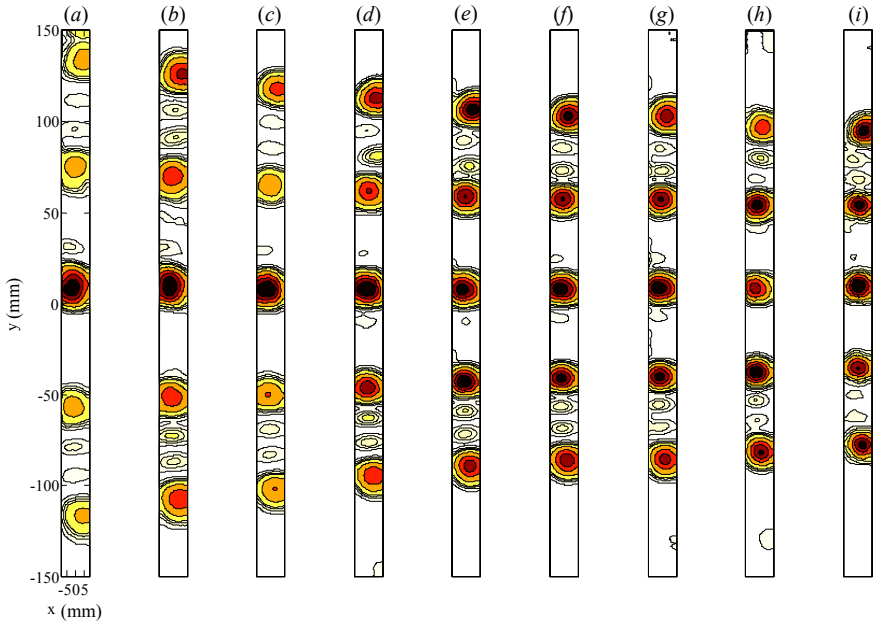


Figure 4-68. Contour plots of (15 mm × 300 mm) scans centred on the central diffraction order at the output plane measured at frequencies 75 GHz (*a*) to 110 GHz (*h*), in steps of ~5 GHz.

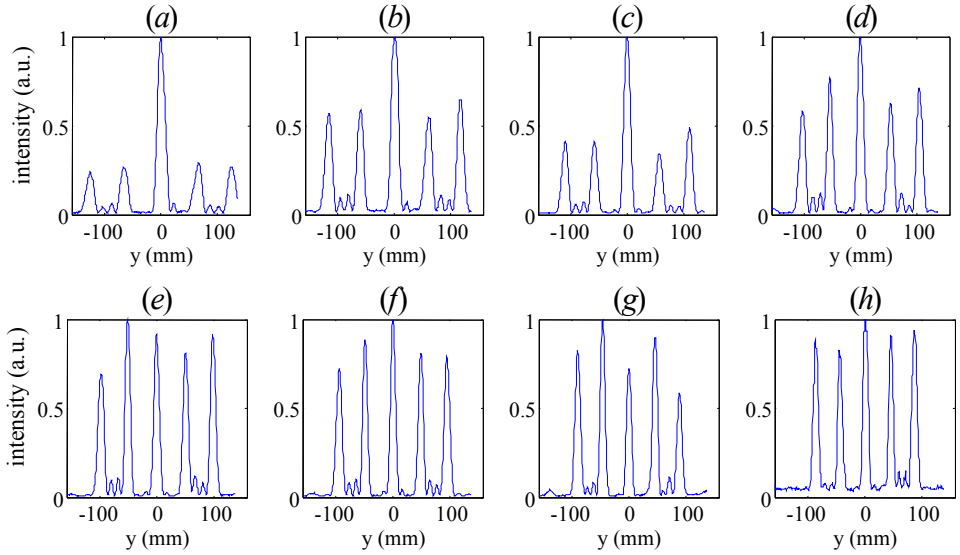


Figure 4-69. Plots of the maximum intensity in each row of the narrow 2-D scans taken at frequencies of 75 GHz (*a*) to 110 GHz (*h*) in steps of approximately 5 GHz.

From the one-dimensional cuts shown in Figure 4-69 the centre of each of the five diffraction orders was located for each test frequency. The diffraction order positions are shown in Figure 4-70 against expected positions (as predicted by the grating equation). There is some deviation (on the order of a few percent) between expected and measured first- and second-order positions, especially at lower frequencies, which is due to distortion but in general there is quite good agreement. Again, the possibility that the source spanning 75-100 GHz is not correctly calibrated must also be taken into account when analysing the intensity patterns measured in this frequency range.

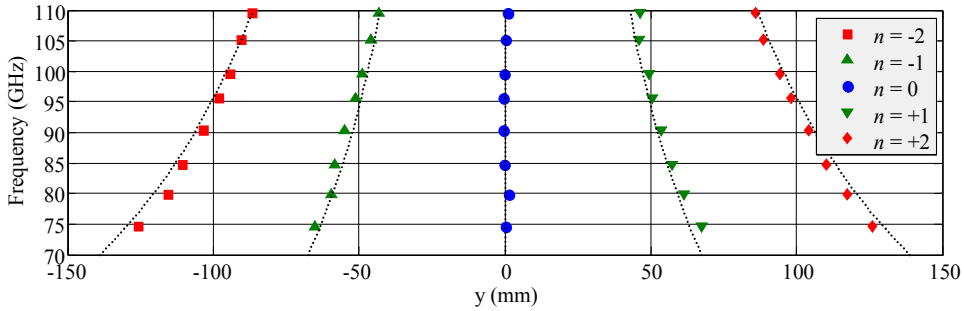


Figure 4-70. Measured and predicted positions of diffraction orders $n = -2$ to $n = +2$ produced by the 5×5 Dammann grating at frequencies between ~75 GHz and ~110 GHz. Measured positions are indicated by the various markers while expected positions (from the grating equation) are shown as dotted black lines.

Note that even at the design frequency of 100 GHz (see Figure 4-69(*f*)) the five diffraction orders have unequal intensity, with reduced intensity in higher-order beams, which was also observed in the full two-dimensional scan (Figure 4-66). In the original measurements of this grating [4.18] the decrease in intensity with increasing off-axis position was also observed and at the time attributed to the loss of high spatial frequency components because of truncation at the rims of the focusing mirror M_2. In simulations developed in MODAL the mirrors are accurately defined to represent the real optics used in the laboratory and although a drop in intensity was observed in the calculated output plane intensity, it is not as extreme as that in the experimentally obtained images. This indicates that the culprit is not the optics but the grating itself. One factor that may be responsible for (or at least contribute to) the observed decrease in beam intensity is the surface accuracy of the grating itself (i.e. manufacture tolerances). These are now investigated.

Limited grating surface accuracy

The 5×5 DG was mechanically machined from a quartz plate by diamond turning. While this process can produce components with higher surface accuracy than can be achieved using an endmill on a CNC milling machine, the two methods are similar in that both are a multi-stage process whereby stages of machining are carried out using a series of cutting passes of increasing depth with increasingly smaller cutting tools. Because cutting is done with finite sized cutting tools the surface will never match exactly the target surface and this can adversely affect the far field diffraction pattern.

The effect of limited surface accuracy in the 5×5 DG was modelled by replacing each perpendicular concave corner of the one-dimensional grating height function with a rounded corner (a segment of a circle whose radius, $R_{\min}$ is equal to the radius of the cutting tool we wish to simulate). After modifying the height function, the equivalent phase modulation is calculated and the far field diffraction intensity calculated.

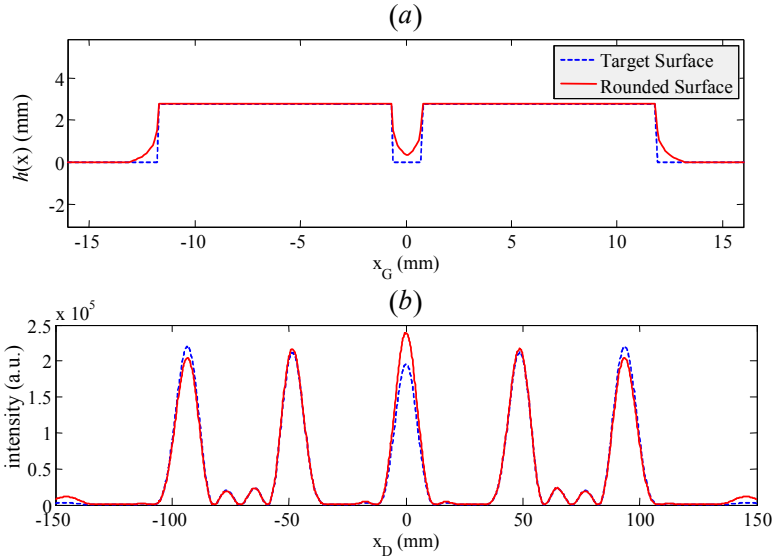


Figure 4-71. (*a*) 1-D height functions of the ideal binary surface (blue) and with rounded corners (red) to simulate the affect of using a cutting tool of minimum radius R_{min} = 2mm. (*b*) Far field intensity from a 4-cell grating. The rounded corners introduces non-uniformity into the beam pattern, reducing the weighted diffraction efficiency, $\eta_{weight} = \eta(1-mpd) = \eta\sigma^2$.

Figure 4-71 shows the 1-D unit-cell height function $h(x)$ with and without rounded corners and the resulting far field diffraction patterns from a grating with four unit cells when illuminated by a Gaussian beam of radius W_G equal to that with the 4-*f* Fourier optics set-up shown in Figure 4-64. The resulting far field intensity shows a more intense zeroth-order beam and less intense second-order beams, qualitatively in agreement with experimental observations. The simulation was repeated for values of

R_{min} between 0 mm and 2 mm and the weighted diffraction efficiency calculated each time (Figure 4-72). As $R_{\min}$ increases, the surface departs from the ideal grating surface and efficiency drops quite dramatically. However, we note that the minimum radius of the actual grating is much less than 2mm and so surface accuracy error is unlikely to be the main contributor to the non-uniform beam intensity observed in measurements.

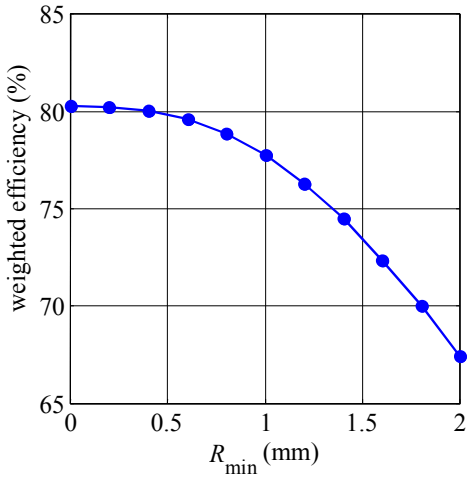


Figure 4-72. The smaller the minimum cutting tool radius $R_{\min}$ the closer the surface matches the target surface and the greater the weighted diffraction efficiency, η_{weight}. The value $R_{\min} = 0$, corresponds to the ideal surface profile and so corresponds to optimum diffraction efficiency.

Another concern for phase gratings is grating depth accuracy. For odd-numbered Dammann gratings the zeroth-order diffraction spot has a dependence on the groove depth different from the other diffraction orders, such that if the phase difference $\Delta\phi \neq \pi$ then the zeroth-order diffraction spot becomes brighter than the other diffraction orders. Jahns *et al* [4.9] has shown that even for gratings that produce small beam arrays the grating depth must be accurate to within 1%. Thus a grating thickness of 3mm must be accurate to ~0.03mm, which is well within the 0.001mm tolerance of the milling machine in the department workshop.

Test Arrangement 3

Operation of the 5×5 DG with a third Fourier optics arrangement was evaluated using a simulation developed in MODAL, but no experimental measurement of this set-up was made. The source beam is collimated with a parabolic mirror and an ellipsoidal mirror focuses the wavefront from the grating onto its focal plane (Figure 4-73). The shorter 350mm focal length of the off-axis parabolic mirror provides a more suitably sized Gaussian beam for grating illumination, while the 45° angle of throw of the off-axis

ellipsoidal mirror and the longer focal length means that the output plane intensity is less prone to distortion (than with the 90° angle of throw paraboloidal mirror). The simulated output plane intensity pattern calculated using scalar diffraction is shown in Figure 4-74.

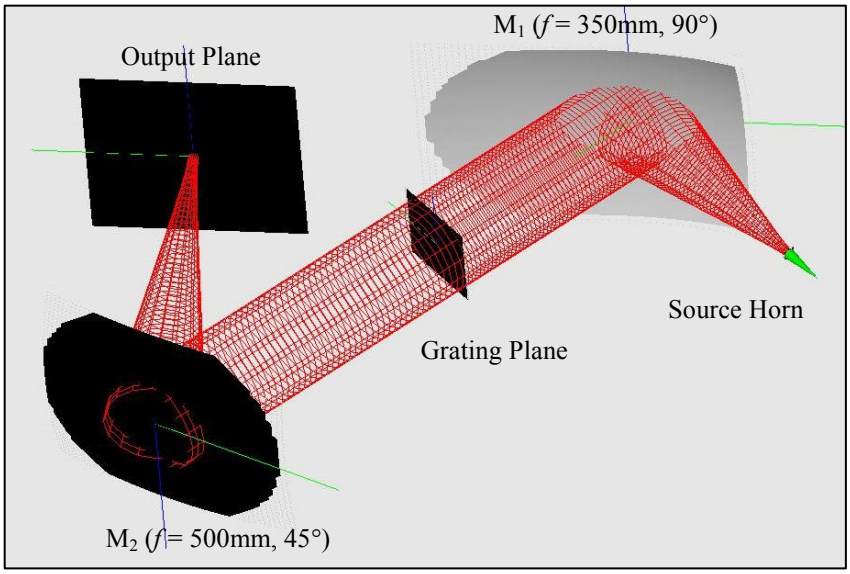


Figure 4-73. The Fourier optics arrangement where the source beam is collimated by off-axis parabolic mirror M_1 (f_1 = 350 mm, 90° angle of throw) and the grating diffraction pattern is focused onto the output plane by off-axis ellipsoidal mirror M_2 (f_2 = 500 mm, 45° angle of throw)

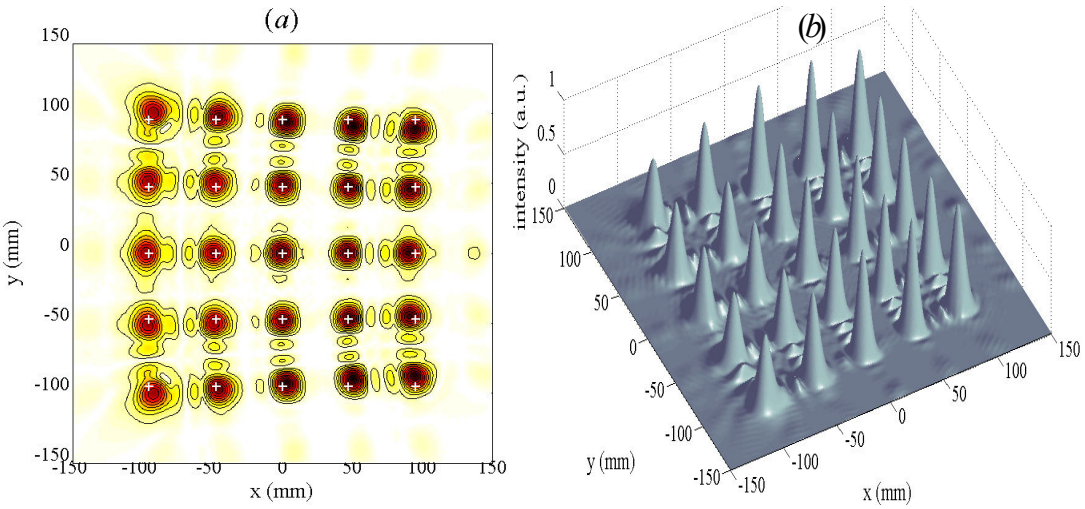


Figure 4-74. Simulated output plane (*a*) amplitude and (*b*) intensity distributions from the Fourier optics arrangement shown in Figure 4-73 for a source frequency of 99.9 GHz (λ = 3mm).

Two differences between the output plane intensity obtained with this arrangement and the previous (with two ellipsoidal mirrors) are observed. Firstly, less power is diverted from primary to secondary maxima (Figure 4-74) because the shorter focal length f_1 of mirror M_1 provides a smaller incident Gaussian beam radius at the grating. Secondly,

the Gaussian beams on the output plane are larger. Previously focal lengths f_1 and f_2 were equal, thus giving unit magnification of input-to-output beam size, so that Gaussian beam radii, W_D and W_S at the detector and source planes, respectively, are equal (assuming negligible truncation at the mirrors). However if $f_1 \neq f_2$ the Gaussian beams at the output plane have a radius of

$$W_D = W_S \cdot \frac{f_2}{f_1} \tag{4.94}$$

which for f_1 = 350mm and f_2 = 500mm gives a larger value of $W_D \cong 1.43 \times W_S$ = 6.74mm (see Figure 4-75).

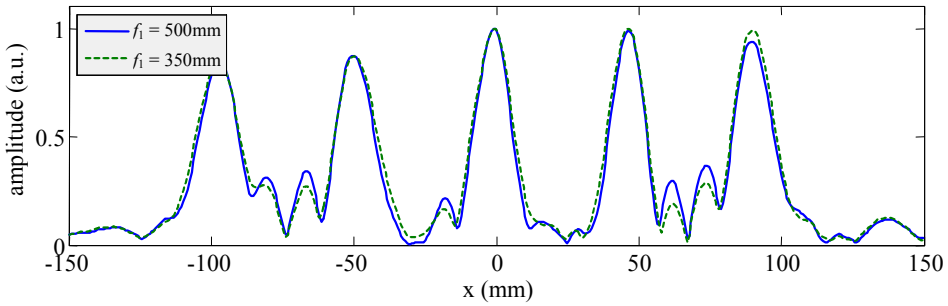


Figure 4-75. Cuts through the centre at $(x, y) = (0, 0)$ of the MODAL simulated output plane amplitude for two different Fourier optics arrangements. In both system the focusing mirror M_2 has a focal length f_2 = 500mm, but the focal lengths of collimating mirror M_1 are f_1 = 350mm (dashed green curve), and f_1 = 500mm (solid blue curve). For f_1 = 350mm the grating is illuminated with a smaller Gaussian beam and so less power is truncated at the edges of the grating. Therefore the secondary maxima (between the diffraction orders) are less intense.

Truncation analysis of the 5×5 DG with GBMA

The phase profile of the 5×5 DG has much smaller features than that of the 3×3 DG so the transmitted field will spread out into a larger area and so is potentially at greater risk of suffering from truncation effects at the collecting lens/mirror.

The set of Gaussian Beam Modes used to analyse the field transmitted from the grating had a highest-order mode index of m_{max} = 219 and a scaling factor W_0 = $0.1352\Delta x$. Only symmetric modes contribute to the grating field so the 2-D mode set requires (110 × 110) modes. The large number of modes needed to describe the grating field meant that because of computational limitations[♣] SVD could not be used to calculate the scattered mode coefficients B_{mn} after truncation by the lens. Instead mode coefficients were calculated by numerical integration.

[♣] The maximum matrix size permitted in MATLAB was exceeded when calculating the pseudoinverse.

We begin by considering truncation in a 4-*f* system with two 150mm focal length parabolic mirrors, for which the 5×5 DG was originally designed to be imaged. When the system is fed with a corrugated conical horn at 100 GHz a mirror of focal length 150mm provides illumination of the grating with a Gaussian beam whose waist radius is W_G = 30.80mm, thus adequately illuminating all four cells of the grating and producing a closely spaced array of Gaussian beams at the output plane.

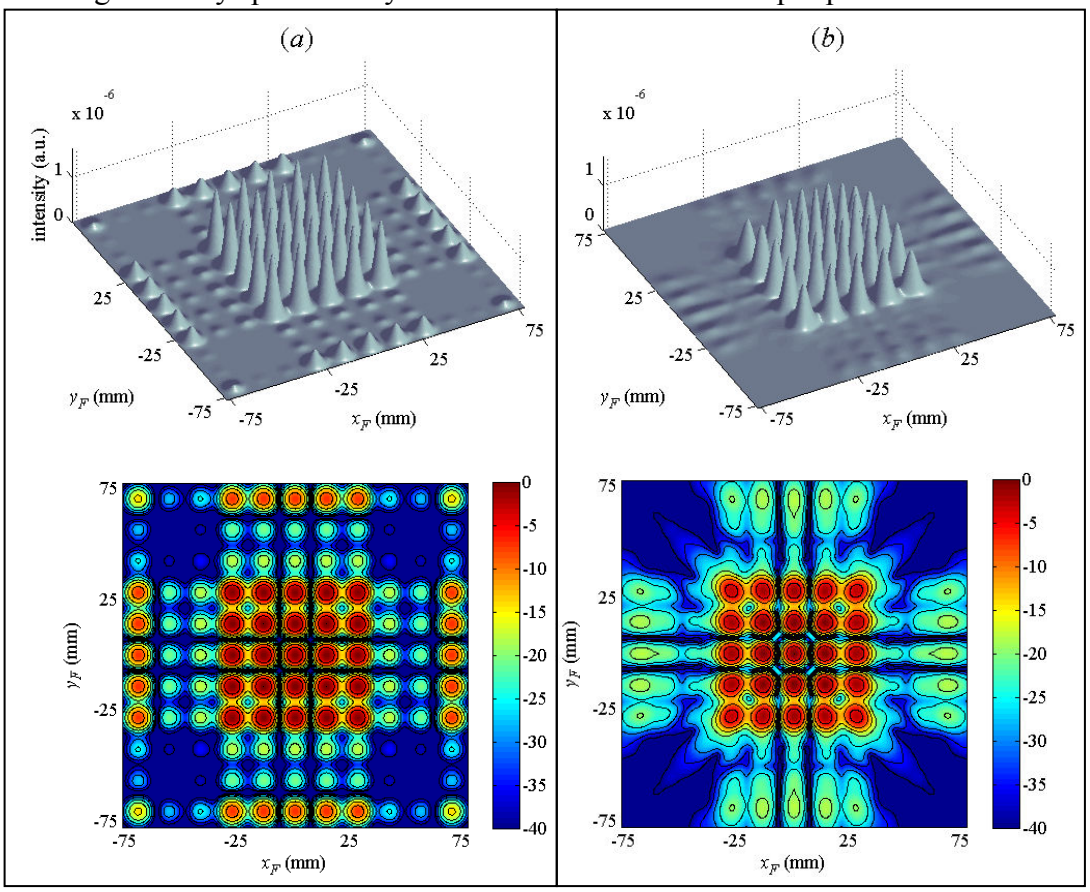


Figure 4-76. Output plane intensity from the 5×5 DG when modelled as part of a 4-*f* set-up with two 150mm focal length mirrors of radius a = 135mm (*a*) without and (*b*) with truncation effects at lens L_2.

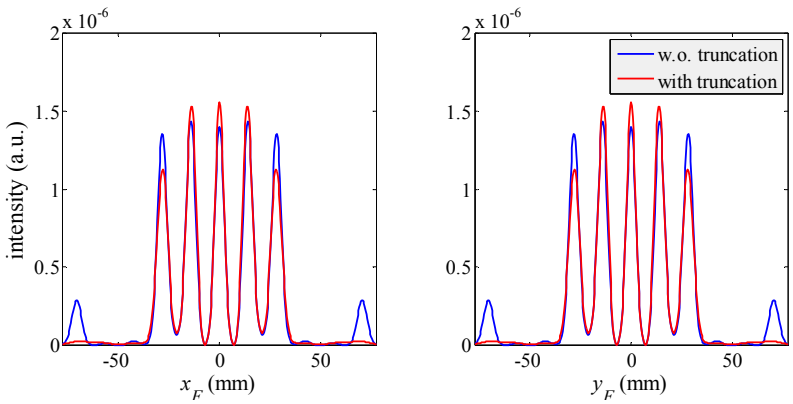


Figure 4-77. X- and Y-cuts of intensity from the 5×5 DG through the optical axis of the output plane (x_F, y_F) = (0,0) for the 4-f set-up with two 150mm focal length mirrors.

Figure 4-76(*a*) shows that the set-up with two 150mm focal length mirrors produces a closely spaced array of 25 Gaussian beams without any power being sent into unwanted secondary maxima between the principle diffraction orders. However Figure 4-76(*b*) and Figure 4-77 shows that truncation by mirror M_2 (treated as a circular aperture with radius, a = 135mm) results in much reduced power in the second-order beams as well as a relative increase in power in the zeroth- and first-order beams, resulting in a significant non-uniformity in the square array of 25 beams as a whole.

Figure 4-78 shows the intensity and phase of the mode coefficients before and after truncation by mirror M_2. There is a large difference in power distribution as well as phase values between the input mode coefficients A_{mn} and output (scattered) B_{mn}. While the gross distribution of power in $|A_{mn}|$ and $|B_{mn}|$ is similar, power is distributed more smoothly between modes in $|B_{mn}|$ than it is in $|A_{mn}|$. Furthermore, while the phase values of A_{mn} are restricted to values of 0 and π (due to the binary surface of the phase grating itself), the phase values of B_{mn} take on a continuum of values in the range [−π, +π).

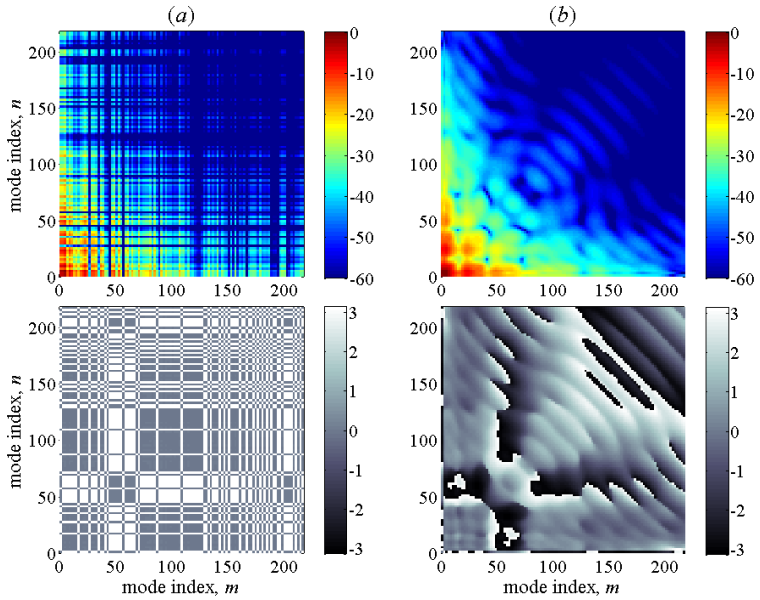


Figure 4-78. Intensity (top) and phase (bottom) of mode coefficients (*a*) A_{mn} and (*b*) B_{mn} of the GBM expanded 5×5 D×G field. B_{mn} are the mode coefficients after truncation with a circular aperture of radius a = 335mm at a distance f_2 = 150mm from the grating. Only symmetric modes are shown.

Next we modelled the test arrangement with two 350 mm focal length off-axis parabolic mirrors. The aperture of mirror M_2 is treated as a circular aperture of radius a = 142.37mm (= 2W, where W = 71.187mm is the radius of a Gaussian beam at a distance of 350mm from one of the corrugated conical horns when fed by a 100 GHz

source.) Figure 4-79 shows the magnitudes of the mode coefficient magnitudes A_{mn} and B_{mn}. The plot of $|B_{mn}|$ is essentially a smoothened version of $|A_{mn}|$.

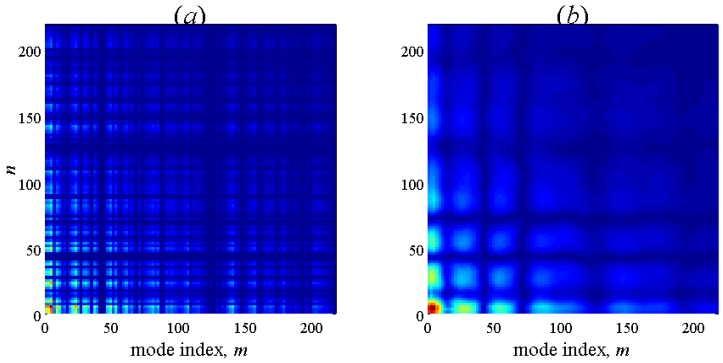


Figure 4-79. Magnitude of (*a*) input mode coefficients A_{mn} and (*b*) output (scattered) mode coefficients B_{mn} of the 5×5 DG field for a 4-*f* set-up with two 350mm focal length mirrors.

Figure 4-80 shows the output plane intensity with and without truncation at M_2. When truncation is included power in diffraction orders $|m|, |n| \geq 3$ is reduced significantly but there is only a slight decrease in power in the central 25 beam array. However the 2nd-order beams are less symmetric in profile than those near the optical axis.

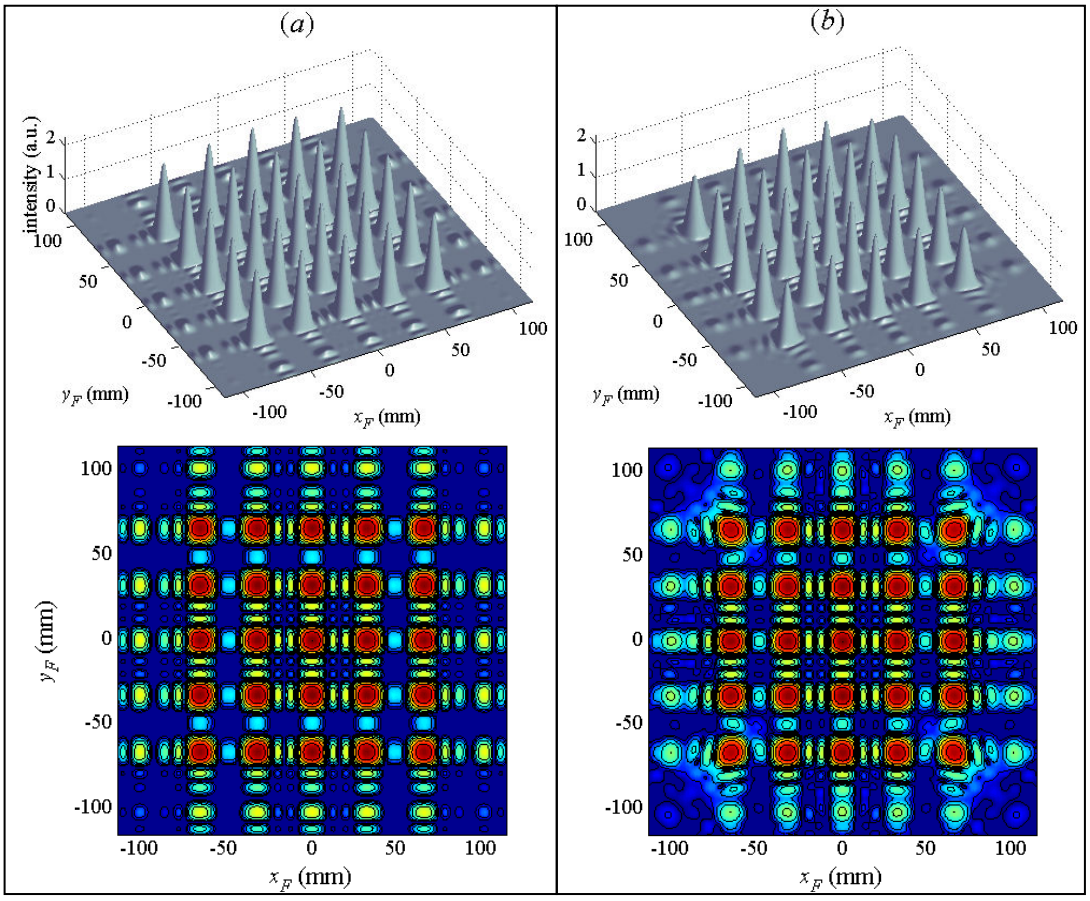


Figure 4-80. Linear (top) and log-scaled (bottom) plots of the output plane intensity from GBMA of the 5×5 Dammann grating (*a*) without and (*b*) with truncation by 350mm focal length mirror M_2.

Finally, we performed truncation analysis of the third test arrangement: a Fourier optics arrangement with a 350mm focal length mirror, M_1 and a 500mm focal length mirror, M_2. The radius of the truncating aperture for M_2 is a = 202.95 (= $2W$, where W = 101.473mm from one of the corrugated conical horns when fed by a 100 GHz source). The reflective surface of the 500mm focal length mirrors were cut into rectangular blocks with dimensions of 404×335mm, or $2a \times 1.65a$. Thus the truncating aperture is not fully circular but is cut off at the top and bottom by ~34.5mm. Figure 4-81 shows the beam intensity at a distance f_2 from the grating after truncation by the non-circular aperture of mirror M_2. Note that aberrational effects have not been included.

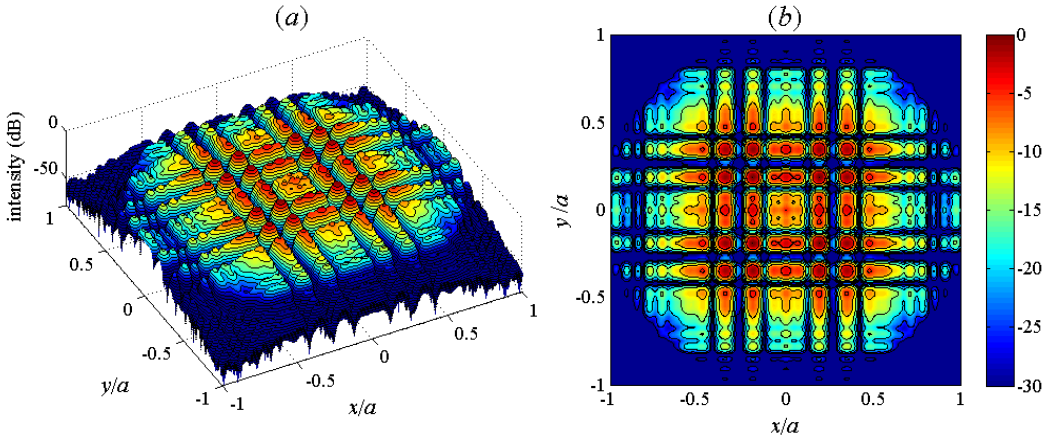


Figure 4-81. Log-scaled plots of the intensity from the 5×5 DG at the plane of the 500mm mirror M_2 whose aperture is treated as a truncated circular aperture. In other words as well as the power being set to zero outside a radius $r \geq a$ we also truncate power in the regions at the top and bottom where $|y| \geq 0.825a$.

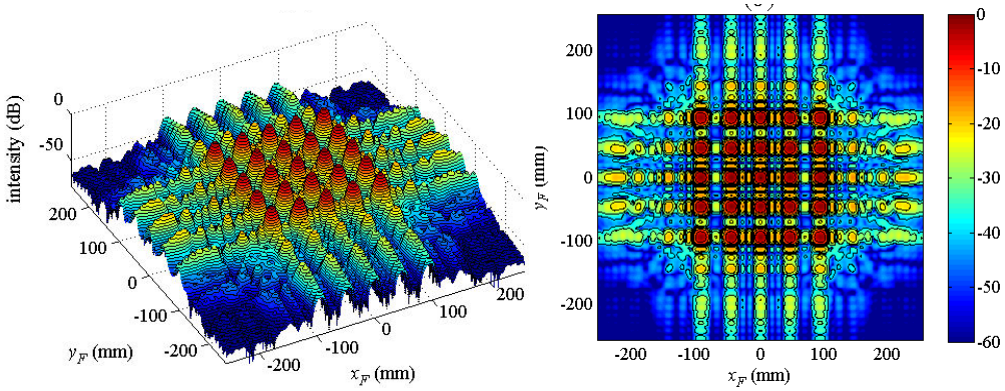


Figure 4-82. Log-scaled plots of output plane intensity from the 5×5 DG after truncation with the 500mm focal length mirror M_2 with a non-circular aperture. Because the truncating aperture is narrower in the y-direction than it is in the x-direction the diffraction orders above and below the central block of 25 beams are less intense than those left and right.

Because the width of the aperture is greater than its height we can expect that the diffraction orders in the y-direction will be reduced in power relative to those in the x-

direction. Figure 4-82 shows a log-scale plot of output plane intensity after truncation and indeed the diffraction orders nearest the top and bottom edges of the image are weaker than those nearest the left and right edges, however the fact that the aperture is non-circular has little impact on the intensity of the central block of 25 diffraction orders (see Figure 4-83 and Figure 4-84).

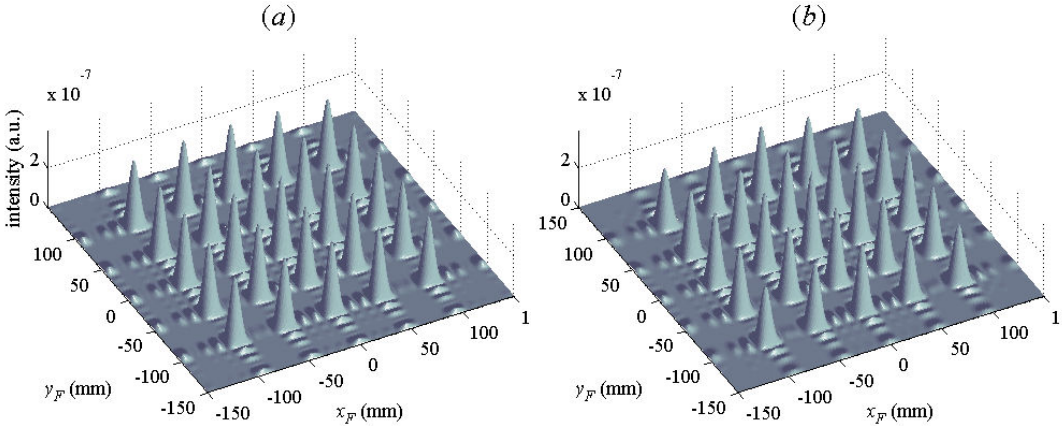


Figure 4-83. Output plane intensity from the 5×5 Dammann grating, operated in a 4-*f* set-up with mirrors M_1 and M_2 of focal lengths 350mm and 500mm (*a*) without and (*b*) with truncation effects from mirror M_2 included.

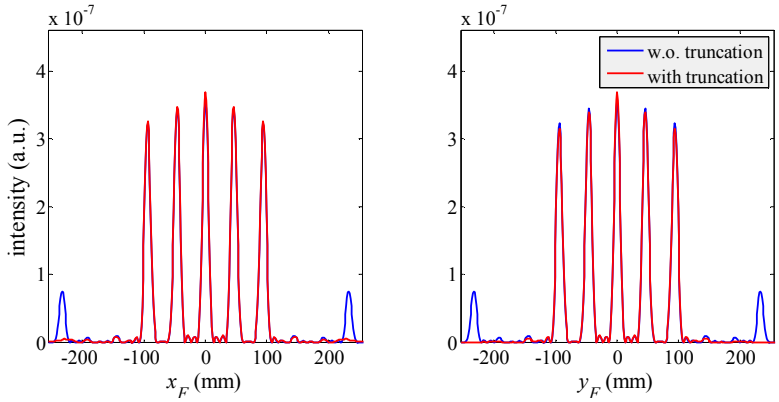


Figure 4-84. X- and Y-cuts through output plane intensity from the 5×5 DG with and without truncation effects from the 500mm focal length mirror M_2. The diffraction orders in (*b*) the *y*-direction are slightly less intense than those in (*a*) the *x*-direction, however overall the effect of truncation on the central diffraction orders is negligible and a significant reduction in beam intensity is only observed in higher diffraction orders, namely $|m|, |n| \geq 5$.

4.8 Chapter Conclusions

This chapter began by introducing the concept of diffractive phase elements, or phase gratings, in particular those needed to produce regularly spaced far field beam arrays, and which can be used as optical multiplexers. The theory of a specific type of binary-level phase grating – the Dammann grating – was developed before discussing methodologies used to design and quantify the performance of these passive devices, as well as how symmetry considerations in the design process can reduce computational overhead (by reducing the number of free parameters involved). The application of modal (GBMA) techniques to the analysis of Dammann gratings was next illustrated and a procedure for mode-set scaling that ensures accurate reconstruction of the wavefront emitted from a phase grating was presented. Although Gaussian beam mode analysis was used to accurately predict phase grating operation, as was seen in Chapter 3, when used for accurate analysis of wavefronts with complicated profiles computational overhead increases to the point that it becomes less efficient.

The construction and experimental measurements of two specific examples of Dammann gratings were presented. The effective bandwidth of the second of these gratings was determined by analysis of experimental measurements. The various test arrangements with which these gratings were measured used real, non-ideal focusing components (mirrors and lenses) which meant that expected output from an ideal treatments of phase gratings (as components in an in-line system with illumination and imaging by ideal lenses) did not match experimental results, due to truncation and/or amplitude distortions (introduced by the off-axis mirrors employed). GBMA was used to include truncation effects at the finite apertures of the various lenses and mirrors used in the different test arrangements. However, our propagation model did not include the facility to model off-axis reflectors. To include these effects the optical simulation software package MODAL was employed to simulate the various test arrangements. The numerical results it produced compared extremely well with experimentally obtained results, thus verifying the proper operation of the two example gratings as well as illustrating the importance of MODAL as an accurate and efficient tool for the design and analysis of complicated multi-element optical systems.

Chapter 5.

Design, Analysis and Experimental Investigation of Fourier Phase Gratings

5.1 Introduction

The limited number of degrees of freedom available in the design of binary phase elements, means that devices such as the Fresnel phase plate (a binary phase-only version of the Fresnel zone plane) and the Dammann grating are relatively inefficient. To achieve higher efficiency requires more free parameters to describe the grating profile. In the context of phase gratings increased efficiency can be achieved by increasing the number of phase levels in the phase modulation. Discrete- or multi-level phase gratings are more efficient elements and are discussed in §5.2. More efficient still are continuous-level, or Fourier, phase gratings with which the majority of this chapter is concerned.

Fourier phase gratings are phase-only implementations of the kinoform. Although initially designed for use at visible wavelengths, limitations in fabrication technology in that part of the electromagnetic spectrum prevented exact realisation of smoothly varying surface relief profiles. Therefore continuous phase functions were approximated by digitised phase functions, thus giving rise to discrete- or multi-level phase gratings mentioned above. Due to advances in fabrication technology in the intervening years, today fabrication of components at visible wavelengths can include many more phase-levels than was previously possible, so that today approximate digital solutions that deviate very little from the continuous solution can be realised. Besides advances in fabrication techniques, increased computational power means that today solutions to high multi-dimensional problems can be solved with greater ease. The distinct advantage that one has when designing Fourier phase gratings for use in the submm and THz wavebands is that the relatively long wavelengths involved mean that typically these devices can be easily fabricated using a CNC milling machine. Because of the high surface accuracy-to-wavelength ratio afforded by this process, continuous phase gratings can be realised without the need to resort to digitised approximations.

The large number of free parameters available in Fourier phase grating design requires more efficient design methods than those described in Chapter 4. Iterative methods that are relatively simple to implement and yield good solutions to phase grating design problems on short time scales (requiring much fewer iterations than, say genetic algorithms) are presented in §5.3, including a novel approach that is described in terms of Gaussian beam modes. These methods were used in the design of example

Fourier phase grating, two versions of which (one in reflection, the other in transmission) were fabricated, experimentally tested and analysed using MODAL.

5.1.1 The heterodyne array receiver CHAMP

An interesting example to illustrate the evolution of strategies for local oscillator (L.O.) beam power distribution by increasingly efficient means is the design of the Carbon Heterodyne Array of the MPIfR (CHAMP) for the Heinrich-Hertz-Telescope (HHT) designed to operate across a 40 GHz band with a centre frequency of 480 GHz. Originally it was envisaged that L.O. beam power distribution would be performed using a set of dielectric power-splitting foils and rotating grids to adjust L.O. distribution to individual mixer elements [5.1]. Later [5.28] a more elegant approach incorporating phase gratings was used to distribute L.O. beam power equally to each of the 16 receiver elements, via two L.O. chains that feed two interleaved 8-pixel sub-arrays (each measuring one of two orthogonal polarisation components) arranged in a 2-4-2 configuration (see Figure 5-1).

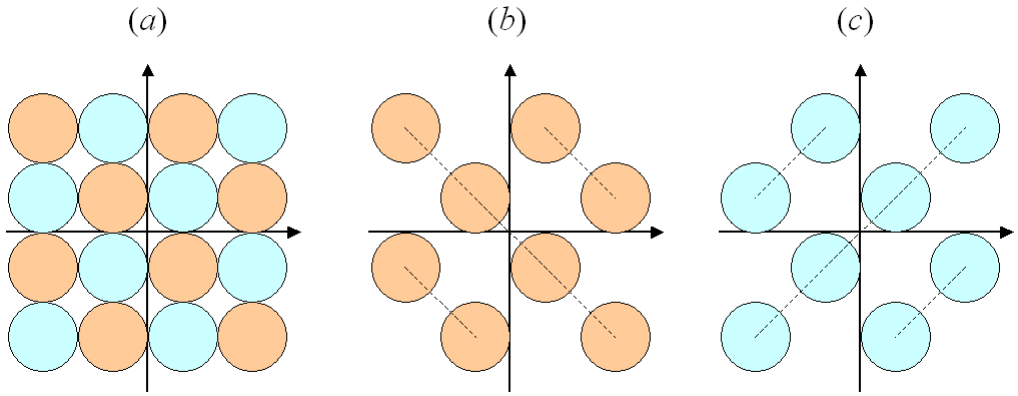


Figure 5-1. Geometry of the (*a*) 4×4 16-element pixel array configuration for CHAMP array receiver (colours indicating polarisations), and (*b*) 8-pixel 2-4-2 sub-array for a single polarisation.

In each sub-array the L.O. beam was distributed using a crossed grating arrangement consisting of two layered one-dimensional binary-level phase gratings that individually generate 3- and 4-beam arrays that when stacked in orthogonal directions produce a 3×4 beam array, with a diffraction efficiency of 72%. The four corner beams would then be eliminated with an absorber (Figure 5-2). Therefore a third of the L.O. power (in the four blocked corner beams) is effectively lost resulting in a much-reduced efficiency of 48%.

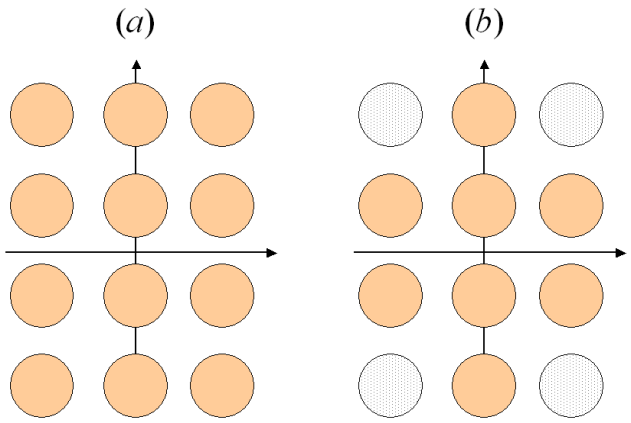


Figure 5-2. Geometry of (*a*) 4×3 beam array produced by crossing 4- and 3-beam linear phase gratings and (*b*) with corner beams blocked to yield the 2-4-2 beam pattern of one of the 8-pixel sub-arrays.

An alternative approach to this problem is to search for a grating solution to generate the 2-4-2 beam array pattern directly. Since this particular nontrivial beam pattern is not separable into two 1-D solutions, a fully two-dimensional treatment of the problem is required. In [5.3] an initial trial solution to generate a 4×3 beam array was derived from crossing four-level 3- and 4-beam linear phase gratings – the increased number of phase levels resulting in a slightly improved efficiency of 74%. The resulting 2-D transmission function $t_{4\times3}(x, y)$ was then used as the starting point in an optimisation scheme that had as its goal the 2-4-2 beam array far field diffraction pattern. The resulting solution yielded a 15% improvement in efficiency.

A third reported attempt [5.4] at the same problem involved optimising a small number (13×13) of Fourier coefficients in a Fourier summation of sine functions only (since the beam pattern is symmetric and therefore does not require cosine functions) that yielded a diffraction efficiency of ~84%. Finally this problem was tackled using a Fourier Transform implemented phase retrieval algorithm [5.5]. By forcing the algorithm to search for only symmetric solutions (by beginning with a symmetric estimate for the far-field phase distribution) a good solution was found on a short time scale (30 iterations). Interestingly the resulting grating phase function $\phi(x, y)$ appears quite similar in structure to that found with the above-mentioned Fourier coefficient optimisation technique.

5.2 Multi-Level Phase Gratings

Despite their popularity Dammann gratings have some serious drawbacks. The main simplification of restricting the transmission function $t(x, y)$ to one that is separable into one-dimensional functions $t(x)$ and $t(y)$ prevents the generation of arbitrary two-dimensional beam patterns that are non-separable in x and y. Even simple geometric arrangements such as a circle or a cross of beams, as well as more complicated beam patterns such as irregularly spaced beam arrays and randomised arrays are thus not possible with DG's. For beam patterns that DG's are capable of generating (rectangular arrays of regularly spaced beams) most solutions result in far field diffraction patterns with relatively low diffraction efficiencies. The main reason for the poor performance of DG's is due to the limited number of optimisable free parameters afforded by periodic binary phase gratings. More fundamental however, is the fact that it is impossible to completely suppress higher order diffraction spots through use of a binary-level grating because of the high spatial frequencies associated with the sharp edges of such a grating. At the same time as mixer development advances to ever-higher frequencies, L.O. power is difficult to generate resulting in much lower levels so diffraction efficiency becomes an important issue in multiplexing applications. As noted by Dammann [5.9,5.10], efficiency losses due to light being diffracted to higher unwanted diffraction orders can be reduced by using phase gratings with multiple phase levels.

Multi-level phase gratings provide a means of increasing diffraction efficiency for multiplexing L.O. power into several beams by increasing the degrees of freedom available (phase levels). Obviously increasing the number of phase levels eventually converges into a grating design with smoothly varying phase structure – a Fourier grating phase. Although in theory it is possible to design a so-called kinoform [5.11] structure with a continuous phase profile, in the past it was not possible to manufacture such devices at visible wavelengths (for which these devices were originally developed) their manufacture proving too challenging with the fabrication techniques [5.6]. Fabrication of visible wavelength diffraction gratings relied on lithographic methods developed for the semiconductor industry. These methods produce a surface profile with discrete heights, corresponding to a finite number of quantised phase levels. Due to quantisation imposed by the fabrications process it was therefore necessary to convert a smoothly varying solution into one with into a number of discrete phase levels. Thus multilevel gratings were a good way of approximating Fourier grating designs.

Subsequent optimisation of the discrete solution is then performed to reduce the negative impact of digitisation [5.9] and restore uniform intensity to the array of output beams, which can be done using an iterative quantisation method [5.12]. Morrison states [5.8] that quantisation and subsequent optimisation can produce results that incur a net loss of only ~1% in diffraction efficiency.

At visible wavelengths the finite number of phase levels involved (often 2^m) is determined by m, the number of lithographic fabrication steps. A single binary mask produces a two-level (binary) surface, a second mask transforms that binary surface into a 4-level grating, etc. Typical quantisation schemes assume phase step sizes of $\Delta\phi = 2\pi/2^m$ and Table 5-1 lists example phase level sets for 2-, 4- and 8-level gratings. At millimetre wavelengths gratings are typically fabricated by directly milling the multi-level surface into a transmission/reflection substrate so no such restrictions on the numbers of phase levels exist.

# fabrication steps m	# phase levels $(= 2^m)$	step size, $\Delta\phi$ $(= 2\pi/2^m)$	Phase Levels $\{\phi_m\} = \{0,1,2\ldots m\}\Delta\phi$
1	2	π	$\{0, 1\}\ \pi$
2	4	$\pi/2$	$\{0, 1, 2, 3\}\ \pi/2$
3	8	$\pi/4$	$\{0, 1, 2, 3, 4, 5, 6, 7\}\ \pi/4$

Table 5-1. The phase levels and number of fabrications steps required for multilevel phase gratings

When searching for solutions to DG's the only free parameters available for optimisation are the transition points (since typically the single phase level difference $\Delta\phi$ is restricted to π). With multi-level phase gratings finding a good solution requires optimisation of both the transition points and the relative phase values between neighbouring transition points. Morrison states [5.8] that computationally it is more efficient to hold the phase levels steady and allow only transition point positions to vary.

Whereas binary level phase gratings are guaranteed to produce positive-negative pairs of diffraction orders with equal intensity[†] this is not the case with multi-level gratings. Reflection symmetry, as described in §4.4.1, must be applied about the midpoint of the gratings periodic cell to ensure matching order-pair intensities. Two reflection symmetry configurations are possible. In the first case a phase offset of π

[†] a property not recognised by Dammann, but later exploited by Killat to find binary-level phase grating solutions with much higher diffraction efficiencies than previously reported.

radians is imposed upon one half of the grating period with respect to the other, so that a phase jump occurs at the midpoint of the unit cell (x = 0), which results in the suppression of the on-axis zeroth order beam as expected[‡]. The other standard configuration is one without a central phase offset and consequently a diffraction pattern that does contain a central zeroth order diffraction spot. The latter approach was the scheme used in the design of the first multilevel phase gratings, as described by Walker and Jahns [5.13]. Their gratings were 4-level, 8-level and 16-level devices designed to produce square 5×5 equi-intense beam arrays. The maximum one-dimensional efficiency achieved with a 4-level design was 79.9% (approximately 3% higher than was achieved with binary level gratings [5.6, 5.7]). The maximum efficiency increased by approximately 5% (to 84.7%) when L = 8 phase levels were used and by a further 5% with L = 16 phase levels. In all three cases the number of independent parameters (transition points) is independent of the number of phase levels L and is solely a function of array size N. For odd $N = 2n_{max}+1$, where n_{max} is the highest diffraction order, the number of degrees of freedom needed is $F = n_{max}+1$, to specify the zeroth to the n_{max} order, since because reflection symmetry is imposed the negative orders do not require additional parameters.

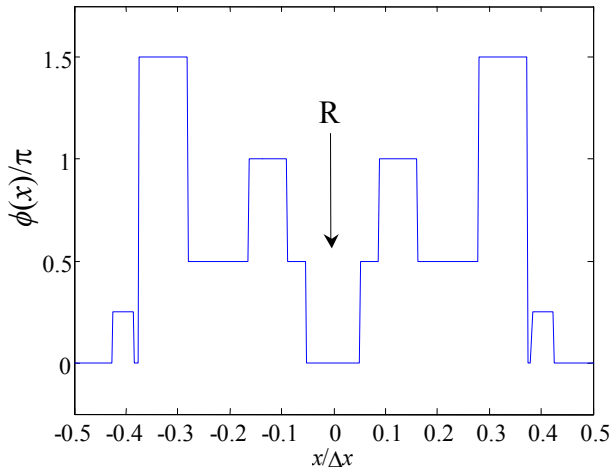


Figure 5-3. Phase profile of the 8-level unit cell to generate 13 equi-intense diffraction orders. Reflection symmetry (R) about the cell midpoint (x = 0) ensures equally intense positive-negative pairs of diffraction orders.

Figure 5-3 shows the unit cell of a 8-level phase grating designed to generate an array of N = 13 equi-intense output beams, i.e. a maximum diffraction order of n_{max}= 6. Because

[‡] Similar to an odd-numbered one-dimensional Gaussian-Hermite Beam Mode that has an on-axis null associated with a π phase difference at x = 0.

of reflection symmetry about the period midpoint (x = 0) the number of independent transition points needed to characterise this solution is F = 7 (only half the total number in the full period). The set of independent transition points for this particular solution is given in [5.8] as

$$x_t = \pm\{0.0523, 0.0887, 0.1642, 0.2804, 0.3743, 0.3836, 0.4253\}\Delta x \qquad (5.1)$$

with associated phase levels (in steps of $\Delta\phi = \pi/4$) of

$$\phi_t = \{0, 2, 4, 2, 6, 0, 1, 0\}\Delta\phi \qquad (5.2)$$

and a quoted diffraction efficiency of 82.8% (compared to 78.0% for a binary phase grating with a π phase difference without reflection symmetry [5.8] and 77.2% for a binary phase grating with a non-π phase difference and also without reflection symmetry [5.7]).

According to Morrison [5.8] the unit cell of a phase grating to generate an array with an even number of equi-intense spots requires both a translational shift and a phase shift about the cell midpoint (x = 0) as well as reflection symmetry about the centre of each half period, i.e. reflection about the half period midpoints ($x = \pm\Delta x/4$). The π phase shift ensures an even-numbered array, while reflection symmetry guarantees equally intense positive-negative diffraction order pairs.

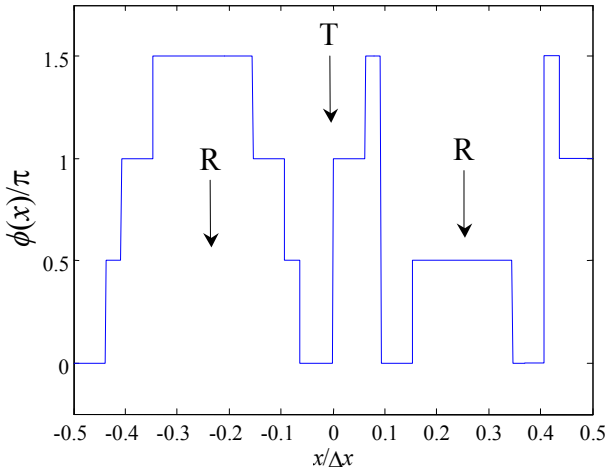


Figure 5-4. The unit-cell of a 4-level phase grating designed to generate an array of 8 equi-intense beams. The unit cell is defined with translational symmetry (T) about the period midpoint (x = 0), and with reflection symmetry (R) about each half period $x = \pm 0.25\Delta x$.

Figure 5-4 shows the basis cell for a 4-level phase grating designed to generate the even-numbered output array of 8 bright equi-intense beams shown in Figure 5-5. The translational shift with a π phase offset at the period midpoint (x = 0) guarantees the even-numbered array, while reflection symmetry about the midpoint of the half-periods

(at $x = \pm\Delta x/4$) midpoints ensures that matching order pairs have equal intensity. The π phase offset, added modulo 2π to the negative x-axis half-period, eliminates the zero-order diffraction beam. Because both translational and reflection symmetry are used the number of independent transition points needed to characterise the solution is reduced even further than might be expected. The solution is parameterised by the transition points in just the first positive quarter-period $[0, \Delta x/4]$,

$$x_t = \{0.0632, 0.0931, 0.1548\}\Delta x \tag{5.3}$$

The phase levels (in steps of $\Delta\phi = \pi/2$) for the first quarter-period are

$$\phi_t = \{2, 3, 0, 1\}\Delta\phi \tag{5.4}$$

The remaining transition points in the positive half-period $[0, \Delta x/2]$ are inferred by reflecting the three existing transition points about $x = \Delta x/4$ and the transition points in the negative half-period $[-\Delta x/2, 0]$ by translating those in the first half. The theoretical diffraction efficiency for this solution is said to be 85.3% (compared to binary level solutions with efficiencies of 75.9% [5.8] and 78.9% [5.7]).

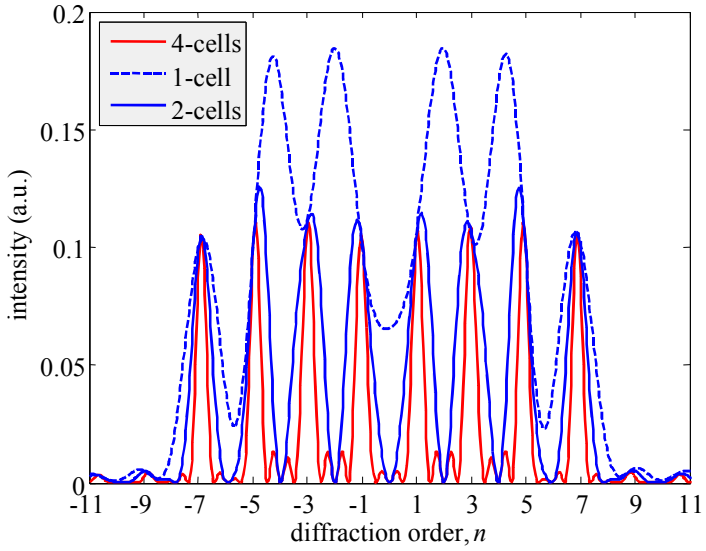


Figure 5-5. The far field intensity from a uniformly illuminated 4-level phase grating to produce 8 equi-intense diffraction orders. The unit cell is defined as one half of the cell shown in Figure 5-4. Plotted are the far field diffraction patterns from a single unit cell (dashed blue curve), two unit cells (which is a single unit cell if the π phase shift is included in the definition of the grating period) and a periodic grating with four unit cells (reflected by the presence of two secondary maxima between adjacent principal maxima).

The two preceding examples show that when searching for solutions to generate output beam array patterns the incorporation of symmetries can reduce significantly the computational complexity of the problem. In summary: the use of reflection *or*

translational symmetry reduces the number of independent transition points needed to characterise a solution by a half, so an N-odd numbered beam array requires approximately $N/2$ independent parameters. While for even-numbered arrays in which both translational *and* reflection symmetry are imposed only a quarter of the transition points are independent parameters and require optimisation. Thus an even-numbered array consisting of N bright and $N-1$ suppressed orders also requires approximately $N/2$ independent parameters to specify the design.

Note that Morrison includes the π phase offset about the period midpoint of the unit cell for a phase grating to produce an even-numbered spot array. However recall from §4.4.2 that this is equivalent to defining a unit cell of half the size and including the π phase shift not in the definition of the unit cell itself, but rather to alternating cells within the grating as a whole. Thus the phase profile that is shown in Figure 5-4 can also be thought of as two unit cells with one shifted by π radians relative to the other. In this case the concept of translation symmetry within the unit cell itself is meaningless and only reflection symmetry is required at the points indicated by 'R' in Figure 5-4.

5.2.1 GBMA of Multilevel Phase Gratings

We now apply Gaussian beam mode analysis (GBMA) to model the 4-level phase grating designed to produce an array of eight equi-intense diffraction orders. In this example the phase grating is modelled with two unit cells (each consisting of two identical half-cells with a π phase shift between them) which gives a grating length of $D = 2\Delta x$. Illumination is provided by a Gaussian beam with a waist radius of $W_G = D/4 = \Delta x/2$, to ensure adequate illumination of the two cells (amplitude falls off to e^{-4} at the edges of the grating).

First we perform GBMA of the grating with a mode set whose parameters are defined by setting the maximum spatial frequency Λ_{max} of the mode set equal to the minimum feature size, δ. The unit cell of this particular grating has a total of thirteen transition points and the separation between the two closest transition points, i.e. the minimum feature size, is $\delta = 0.0299\Delta x$. So we require that $\Lambda_{max} \approx 0.03\Delta x$, which, from Eq. (4.78), yields mode set parameters $W_0 = \sqrt{(2\Delta x)(0.03\Delta x)/8} \approx 0.086\Delta x$ and highest-order mode $m_{max} = 2(2\Delta x/0.03\Delta x) \approx 133$. Because of the π phase shift between the two halves of the unit cell only asymmetric (odd-numbered) modes contribute to the

summation of the grating field approximation $E_G'(x)$ so only 67 modes are required, as shown in Figure 5-6 for Gaussian illumination.

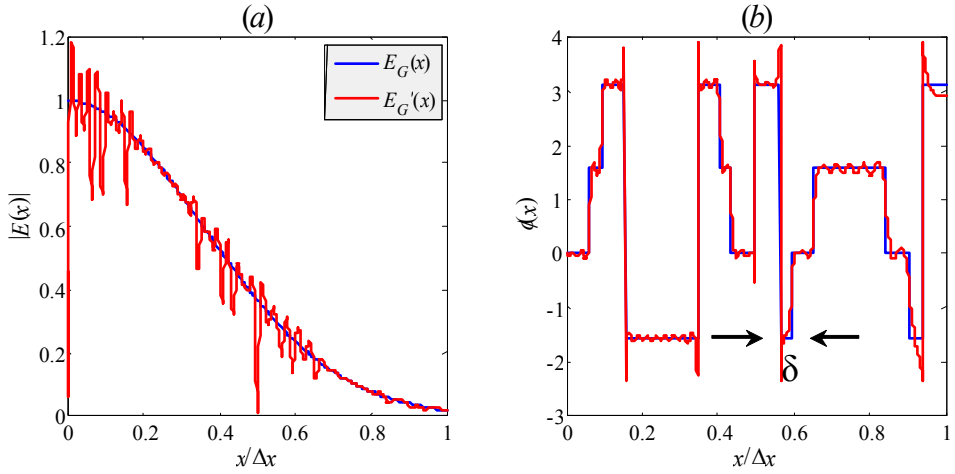


Figure 5-6. Gaussian beam mode analysis of a 4-level phase grating designed to produce an array of 8 equi-intense diffraction orders. (*a*) Amplitude and (*b*) phase distributions of the original field $E_G(x)$ and the GBM-approximated field $E_G'(x)$. Only one half of the grating profile (x > 0) is shown. The Gaussian beam mode parameters were chosen by setting the maximum spatial frequency of the mode set equal to the minimum feature size, δ, which gave $m_{max} = 133$ and $W_0 \approx 0.086\Delta x$.

A different set of Gaussian beam modes was chosen to improve reconstruction quality of the grating field. The mode-set parameters were chosen by setting the maximum spatial frequency to δ/5, which, from Eq. (4.79), yields $m_{max} = 20/0.03 \approx 666$ and $W_0 \approx 0.038\Delta x$. Again only half the number of modes is needed since the symmetric (even-numbered) modes do not contribute. The reconstructed grating amplitude and phase are shown in Figure 5-7. This particular mode-set allows the grooves in the grating phase profile $\phi(x)$ to be reproduced with much greater accuracy than above.

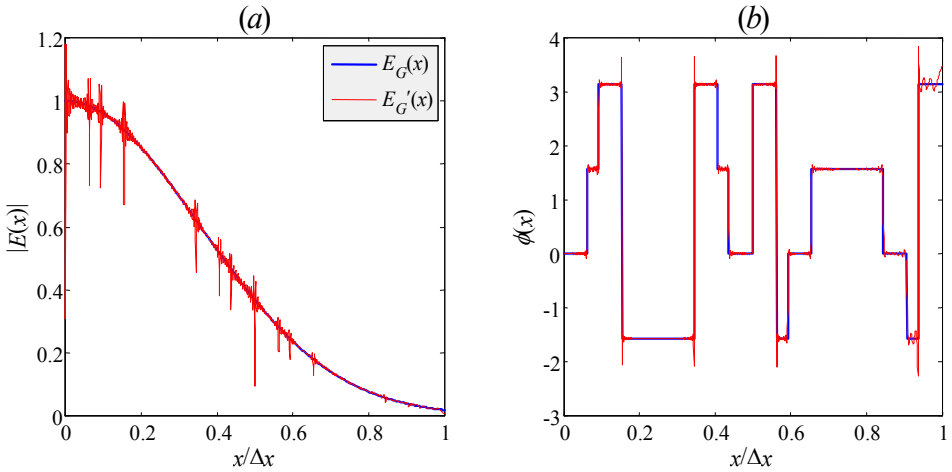


Figure 5-7. GBMA of 4-level phase grating to produce 8 equi-intense diffraction orders. (*a*) Amplitude and (*b*) phase of original and GBM-approximated fields, $E_G(x)$ and $E_G'(x)$.

5.3 Phase retrieval for phase grating design

The design variables used to encode the unit cell of a Dammann grating are the locations of transition points. A more common encoding technique for grating design involves representing the unit cell as a rectilinear array of pixels. In so-called array encoding the unit cell consists of a fixed array of uniformly sized pixels and it is the phase values of each pixel that are treated as the free parameters, rather than transition points. Grating design then involves finding an appropriate set of values for the phase associated with each pixel in the array that produces the desired far field diffraction intensity distribution. This encoding technique is clearly suited to the design of multilevel and continuous-level phase gratings since it gives one greater design freedom. Also, because pixels are uniformly spaced calculating the Fourier transform of the unit cell is straightforward using a discrete Fourier transform (DFT), typically performed using a fast Fourier transform (FFT).

A straightforward search method suited to array encoding is now examined with application to binary-level phase gratings. The proposed method is based on *direct binary search* (DBS), an iterative algorithm originally developed for the design of computer-generated holograms [5.16]. When applied to find a binary-valued transmittance function for a digital hologram DBS manipulates directly the binary transmittance values (0 and 1) of the hologram pixels in order to generate the required wavefront at some distance beyond the hologram. The algorithm is easily adapted to search for binary phase solutions by assigning values of +1 or −1 to individual pixels. The algorithm begins by generating a random binary transmittance $t(x)$ with a uniform distribution of +1's and −1's and computing the objective function for this initial trial solution. The unit cell is then scanned point-by-point. At each point x_i the transmittance $t(x_i)$ is inverted (multiplied by −1) and the objective function evaluated and compared with the previous figure of merit. If the inversion results in an improved solution (a higher objective function value) then the change to $t(x)$ is accepted, otherwise $t(x_i)$ is restored to its original value. An iteration is said to be complete when inversion of every addressable point in the unit cell has been considered. The algorithm continues until no inversions are accepted during an entire iteration. The main problem with DBS is its slow execution time when a large number of unit cell pixels are used. Another problem with DBS, as noted by Seldowitz *et al* [5.16], is that is a local optimiser, but it can be

turned into a global optimisation technique by using simulated annealing to probabilistically accept point inversions that in the short term yield worse results.

Our deterministic DBS algorithm was used to search for solutions to the problem of generating an array of five equi-intense diffraction orders with a binary phase grating, the results of which are shown in Figure 5-8. Although the diffraction efficiency for the best Dammann grating solution is higher than the solution found using our DBS (77% compared to 71%) the beam uniformity is much higher (almost unity) for our solution and in most applications the improved uniformity would outweigh slightly reduced efficiency[5.14]. The problem with the final solution is that on the last set of iterations the unit cell was sampled with 1000 points and this resulted in the grating structure becoming extremely complicated by the introduction of narrowly separated phase jumps that would be difficult to manufacture. Of course this problem could be overcome by including in the algorithm a mechanism for avoiding or removing separations between phase jumps below some user-defined minimum threshold.

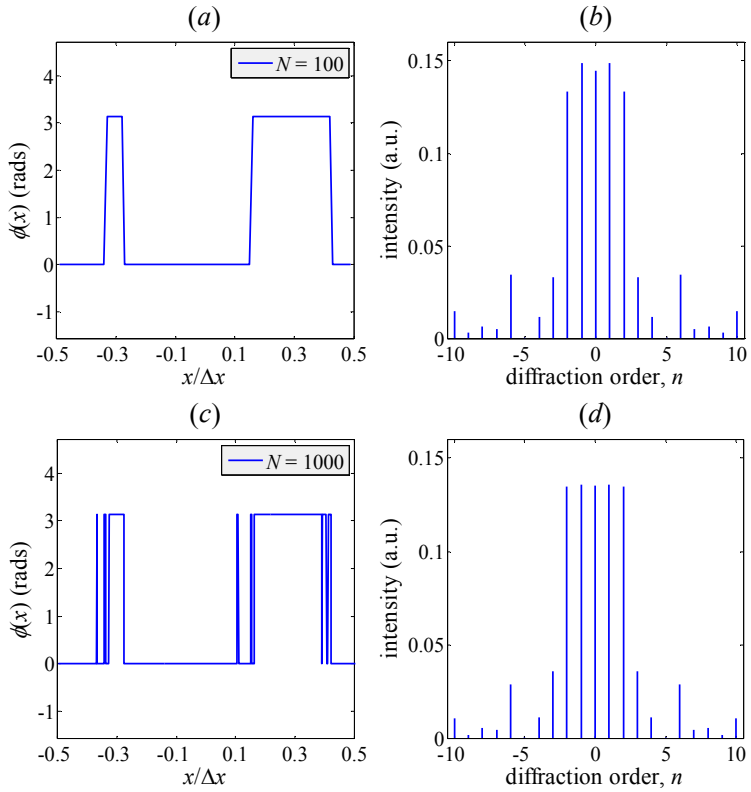


Figure 5-8. Results of applying DBS to find a binary phase grating to generate five equi-intense diffraction orders. **Left:** Phase distribution $\phi(x)$ of the grating unit cell when the unit cell is sampled with (*a*) N = 100 points and (*c*) N = 1000 points. **Right:** Diffraction order intensities from infinite gratings composed with unit cells shown on the left. The diffraction efficiencies, η of the two solutions are 70.9% and 67.6%, respectively. In other words the simpler structure yields higher efficiency. However the solution with more phase jumps has 99.7% uniformity, compared to 94.6% for the other.

5.3.2 Bidirectional algorithms (Iterative Phase Retrieval)

The increased design freedom offered by array encoding means that direct search techniques may struggle to yield solutions on adequately short timescales since the phase of each pixel must be set independently of the others. Fortunately alternatives to the direct searches mentioned above exist.

The problem of finding a solution for a Fourier phase grating can be viewed as an instance of *the phase retrieval problem*. The phase retrieval problem, or image reconstruction problem [5.15], is defined as follows: given the intensity distributions of two complex-valued wavefronts (that are defined at the object plane and the image plane) one is required to find the phase distributions of the wavefronts at both these planes. Phase retrieval is encountered in many areas, for example in determining the phase distribution associated with the measured intensity from an astronomical source. In the context of phase grating design, the two intensity distributions are the intensity of the incident beam at the grating plane (usually a laser beam, or similar with a Gaussian amplitude profile) and the far-field diffraction intensity pattern that the grating is required to generate. As with Dammann gratings, the problem is simplified by assuming that the illuminating wavefront has a uniform phase distribution and as such the only phase term present at the grating is the phase modulation from the grating.

Direct search methods are also referred to as unidirectional algorithms, to distinguish them from another class of algorithms called bi-directional, or inverse, techniques. Unidirectional algorithms proceed by calculating a system transform in the forward direction only and evaluating the merit function. The basis for improving the design is by observing the impact that a change in variables has on the merit function. Bidirectional algorithms on the other hand involve a forward system transform as well as an inverse transform. Thus the use of a bidirectional algorithm requires not only an understanding of the system transform in the forward direction but also an understanding of its inverse, i.e. an understanding of how variations in the response from the grating (e.g. the far field wavefront) affects the grating [5.17]. Therefore bidirectional algorithms are applicable only to optimisation problems in which the system function can be easily inverted. Under the assumption that scalar diffraction is valid the optical system transform can be modelled with near-field (Fresnel) diffraction or far-field (Fourier) diffraction, both of which are easily inverted so bidirectional algorithms are well suited to the design of phase gratings.

Bidirectional algorithms are an extension of *direct inversion*, a method often used to design optical elements for information processing, such as matched spatial filters for pattern recognition [5.14]. For the design of the unit cell of a phase grating direct inversion is applied as follows. Since the relative phases of the signal orders is (usually) irrelevant we can assign random phase values $\phi_{m,n}$ to the diffraction orders with indices (m, n). Thus the complex amplitude of the diffraction order array may be expressed as

$$F_{m,n}(u, v) = A_{m,n}(u, v) \exp(i\phi_{m,n}(u, v))$$

where the intensity of the diffraction order array is

$$I_{m,n}(u, v) = |F_{m,n}(u, v)|^2$$

The inverse Fourier transform of $F_{m,n}(u, v)$ gives the grating unit cell field $f_{m,n}(x, y)$. The requirement of a phase-only unit cell means that the amplitude $|f_{m,n}(x, y)|$ must now be set to unity. After this amplitude truncation, the phase-modulated grating is then Fourier transformed. Ideally the resulting spot pattern intensity $|F_{m,n}(u, v)|^2$ should equal $I_{m,n}(u, v)$ but the result of mapping $|f_{m,n}(x, y)|$ to unity is the introduction of significant noise and nonuniformity in the Fourier plane intensity. This is because the intensity constraint imposed at the grating plane is a dramatic truncation that results in the loss of a lot of information in the unit cell. The reason for assigning random phases $\phi_{m,n}(u, v)$ to the signal orders is to distribute power in the grating plane. If a uniform phase was assigned to the diffraction orders, the inverse transform amplitude would consist mainly of an on-axis component and little else. Although random phases do help to distribute power at the grating plane, amplitude truncation remains a problem. The reconstructed far field intensity would better match the target intensity $I_{m,n}(u, v)$ if $|f_{m,n}(x, y)|$ were more nearly uniform since then less information would be lost during amplitude truncation. By varying the phase values $\phi_{m,n}(u, v)$ assigned to the initial Fourier plane complex amplitude we can expect that some set of phases will result in more uniform $|f_{m,n}(x, y)|$ and so minimise the impact of amplitude truncation. The aim is thus to determine the set of phase values that will require minimal amplitude modulation at the grating plane. This is the subject of phase retrieval, whereby one seeks to retrieve the phase of an image that has been lost because one has only access to intensity measurements.

A discrete, or fast, Fourier transform operates on amplitude and phase values defined on a sampled grid. If the amplitude and phase values are known in one domain, then they can be calculated in the other using a DFT or FFT. Consider a unit cell

sampled on a grid with $M \times N$ points. If all MN unit cell amplitude and phase values are defined then one can calculate all MN far field amplitudes and phase values. Gerchberg and Saxton [5.18] showed however that if the amplitudes are known in both domains, but the phases are unknown we have 2(MN) known values and 2(MN) unknown values and the Fourier transform relationship can be used to generate 2(MN) equations in 2(MN) unknowns. Thus the unknown phase values can be retrieved when the amplitude in both domains is known. Solving the 2(MN) equations is an involved process but bidirectional algorithms offer another way of solving the unknown phase values in a fraction of the time that would be required to invert 2(MN) equations.

The iterative bi-directional algorithm used by the author of this thesis to solve phase retrieval in the context of phase grating design is based on the algorithm of Gerchberg and Saxton [5.18]. Typically the Gerchberg-Saxton algorithm (GSA) is implemented using the fast Fourier transform (FFT) to propagate back and forth between the grating plane and the Fourier plane, where the required intensity output plane pattern from the grating is formed.

A bidirectional algorithm begins by generating a first "estimate" of the transmission function of unit cell, typically by assigning randomly generated values to the pixels in the unit cell. A system transform (FFT) is applied to the unit cell and performance constraints (usually intensity requirements) are imposed on the field in the output (Fourier) plane. An objective/merit function is evaluated at this point. Next an inverse transform is applied with the resulting field being the updated estimate of the unit cell transmission function. Finally amplitude truncation is performed, thus fulfilling the coordinate domain constraints. As this cycle is repeated the unconstrained phase values are driven towards values that most nearly fit the system constraints so that the system converges to some optimum estimate of unit cell phase values. It has been shown by Gerchberg and Saxton [5.18,5.19] that the unknown grating and Fourier plane phase values can be reconstructed in just a few iterations. Note that whereas the objective function is used to guide unidirectional algorithms, in bidirectional algorithms the merit function has no influence on the path taken and serves only as an indicator of the performance of the current design and algorithm progress towards the desired target intensity. Typically the bidirectional algorithm of Gerchberg and Saxton is implemented with an objective function that measures the (mean squared) error between target and trial far field intensities. It was shown by Gerchberg and Saxton [5.19] that the error

must decrease or at least remain constant with increasing number of iterations, hence the Gerchberg-Saxton algorithm is also known as the error-reduction algorithm [5.20]. The system transform typically used with the Gerchberg-Saxton algorithm is the Fourier transform, and so the algorithm is often also referred to as the Iterative Fourier Transform algorithm (IFTA). The steps involved in IFTA are illustrated in Figure 5-9.

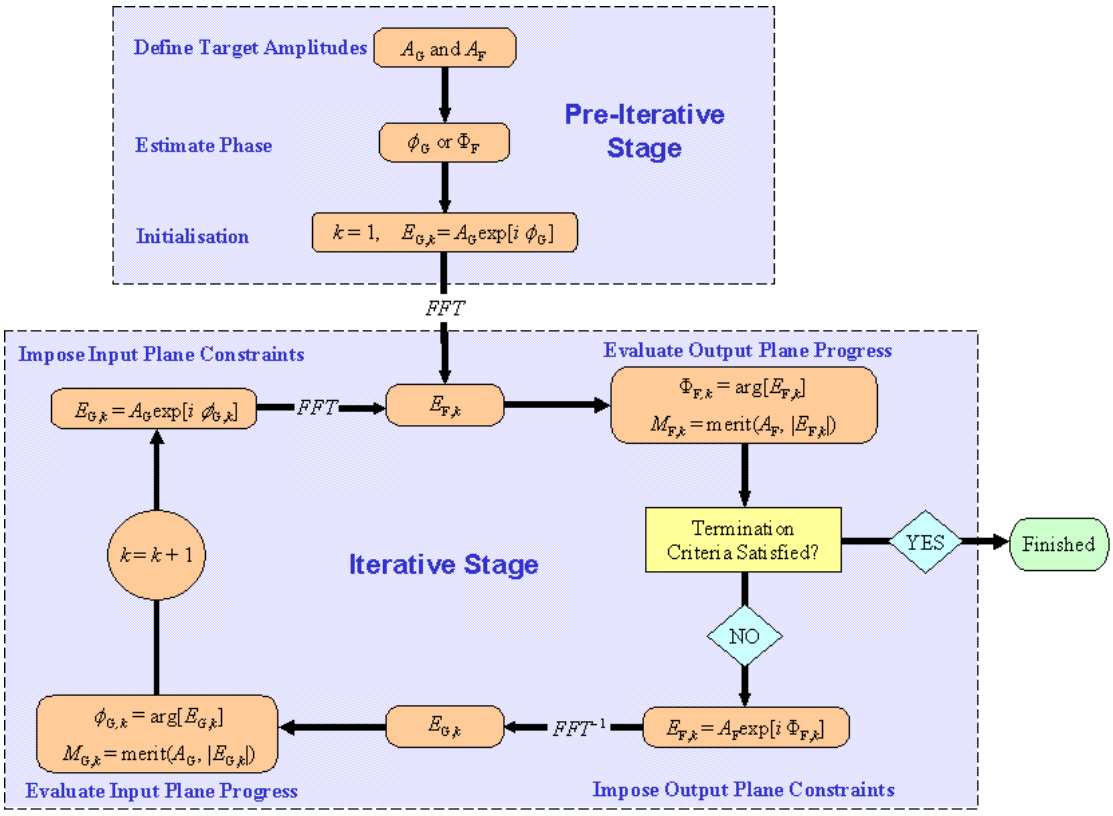


Figure 5-9 The basic steps of the iterative transform algorithm with propagation between input and output planes implemented in terms of Fourier Transforms.

Another encoding scheme that can be used for grating design is to use modal analysis. This involves expanding the grating and diffraction plane fields into a summation of appropriately weighted plane-wave or Gaussian beam modes (GBM). With this scheme the expansion coefficients that determine the relative contribution of each plane-wave/GBM component to the grating field are treated as the free parameters of the system. Unidirectional search methods can then be employed to find phase-only solutions to produce the target far field intensity distribution. One example of optimisation in terms of Fourier coefficients is reported in [5.13], in which a small number of coefficients were optimised to find a grating solution to generate a two-dimensional 2-3-2 array of far field diffraction orders. Only a few Fourier coefficients were used in the optimisation which yielded a phase grating with a smoothly varying

profile. An equivalent array encoding scheme would necessitate a much larger number of variables representing the phase at each pixel in the grating unit cell. We present a novel implementation of a bi-directional algorithm in which propagation is performed using Gaussian Beam Mode Analysis (GBMA). Because we have chosen to implement this type of phase retrieval algorithm in terms of both Fourier transforms and Gaussian beam modes, in this thesis we refer to it as an Iterative Phase Retrieval Algorithm (IPRA) so as not to imply the use of any specific propagation technique (Fourier transform or otherwise).

The advantage of inverse methods over direct methods is that they are simple to implement and converge to solutions very quickly even for systems with a large number of variables. The fast convergence is due to the fact that on each step the diffractive element is modified globally rather than pixel-by-pixel. This parallelism, as well as the use of FFT to evaluate propagation integrals means that inverse methods are usually far superior to direct methods in terms of computational efficiency [5.21] and are particularly suited to the design of continuous-level gratings which involve a large number of design variables. The drawback is that the basic bidirectional algorithm is deterministic, i.e. the quality of the solution is determined by the starting point (the first estimate of a solution). If the solution space is complex with many local minima, the bidirectional algorithm has a much greater chance of getting stuck, or stagnating, at one of these sub-optimal solutions. In practice the behaviour one observes is that the error between target and trial intensities decreases quickly during early iterations but decreases extremely slowly thereafter, requiring an impractically large number of iterations for convergence. Attempts to increase the speed of convergence of bidirectional algorithms have resulted in the development of a family of input-output algorithms [5.22]. These algorithms are based on a principle similar to that of negative feedback [5.20] and, as noted by Mait [5.17], essentially transform the bidirectional algorithm into a unidirectional algorithm. Since different starting points are likely to yield different solutions, the usual approach is to execute the algorithm using multiple randomly chosen starting points, i.e. to incorporate the algorithm into a multi-start algorithm. Alternatively one may exploit design freedoms [5.23] to overcome the stagnation problem. Application of complex-wave amplitude freedom for example implies that the element need only generate the desired intensity distribution within a bounded region, the signal window, in the output plane; it is not necessary to constrain its performance elsewhere [5.17]. Although careful treatments to establish the existence

of and uniqueness of [5.24] a solution that satisfies the constraints of a design problem are possible they are not necessary. For practical grating design any solution that satisfies the constraints within prescribed limits is acceptable.

5.4 Design, Analysis and Measurement of Fourier Phase Gratings

We first describe the design, manufacture and testing of a reflection Fourier phase grating that was designed to produce a linear 3-spot array using a FFT version of an iterative phase retrieval algorithm (IPRA) based on the Gerchberg-Saxton algorithm. The other two gratings presented are transmission and reflection implementations designed to produce a sparse array of eight Gaussian beams arranged in a circular formation. This design was arrived at using a Gaussian Beam Mode version of the IPRA. Phase unwrapping was then applied to the grating design to reduce manufacture error that might otherwise be incurred during the machining of the grating. Two gratings (one in reflection, the other in transmission) were made from the unwrapped phase design. Measurements were made of the two gratings and were found to compare extremely well with numerical simulations developed in MODAL (the in-house Maynooth CAD software package described in Chapter 2).

5.4.1 Reflection 3-beam Blazed Fourier Grating

Next we consider the manufacture and testing of a reflection Fourier phase grating. The two Dammann gratings examined in Chapter 4 were transmission devices that allowed for in-line axial arrangements. However they are subject to standing wave effects [5.26]. Reflection gratings are free of absorptive losses and standing wave effects that occur with transmission devices. It was therefore decided to investigate a simple reflection Fourier phase grating to gain experience with this type of device. The function of this grating is to generate a linear array of three equally intense off-axis diffraction orders. In the design stage the grating was treated as an ideal transmission device, but was then fabricated as a reflection component instead because the small feature sizes could be more easily machined by cutting the complex profile into a block of aluminium than, for example, Teflon or HDPE.

Design

The goal was to find a phase modulation that can split a single incident beam into an off-axis row of three diffraction orders. The problem can be treated as a one-dimensional phase retrieval problem and a solution was found using a one-dimensional FFT version of the iterative phase retrieval algorithm (IPRA) described previously. The target intensities at the object (grating) and image (Fourier) planes were defined as a single Gaussian beam and an off-axis array of equally intense Gaussian beams, respectively. The three Fourier plane Gaussians correspond to diffraction orders $n = -1$, 0, +1 with the zeroth-order beam set at an angle of $\alpha = 38°$ to the grating normal and the angular separation between zeroth- and first-order beams set to 3.58°. The IPRA algorithm was initialised with an initial guess of the Fourier plane phase distribution that consisted of plane phase fronts across each of the three Gaussian beams as this would be the most desirable outcome for coupling to an array of feed horns. The grating plane target intensity is a single Gaussian beam with a waist radius at the grating of W_G = 89.08mm. To minimise edge diffraction (truncation losses) the grating length was set to $4 \times W_G$ = 360mm.

The solution phase modulation $\phi(x)$ found by the IPRA was transformed into a reflection height function $h(x)$ using Eq. (4.88) with the angle between incident and reflected beams set to zero, i.e.

$$h(x) = \frac{\phi(x)}{2k_0} \tag{5.5}$$

A 50mm long segment ($1/7^{th}$ of the total grating length) of $h(x)$ is shown in Figure 5-10 and appears to consist entirely of an approximately periodic saw-tooth function.

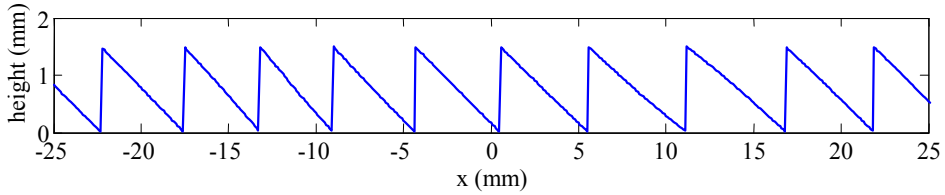


Figure 5-10. A 50mm long segment of the central portion of the surface profile of the Fourier phase grating designed to generate an array of three equally spaced and equi-intense diffraction orders, that propagate at an angle $\alpha = +38°$ relative to the grating normal. Because the grating is designed for reflection the maximum peak-to-trough depth $\sim\lambda_0/2$, which corresponds to a phase jump magnitude of 2π.

An alternative approach to solve this particular phase retrieval problem would have been to have the algorithm search for a beam-splitting (multiplexing) phase

function $\phi_{B\text{-}S}(x)$ that produces an on-axis array of three diffraction orders (i.e. three images of the incident beam) and then modify that solution by adding a suitable blazed phase term $\phi_{blazed}(x)$ to direct the beam array off-axis, thus yielding a total grating phase function of

$$\phi(x) = \phi_{blazed}(x) + \phi_{B\text{-}S}(x) \tag{5.6}$$

The phase term of a blazed grating that directs light into its first diffraction order at an angle α is given by

$$\phi_{blazed}(x) = k_0\, x\, \sin(\alpha), \tag{5.7}$$

which when wrapped about the range $\pm\pi$ radians is a periodic saw-tooth function. Thus the dominant structure observed in the solution arrived at by the IPRA is that of the blazed grating term that is needed to direct the beam array centred at the required offset angle $\alpha = 38°$. The beam-splitting phase function $\phi_{B\text{-}S}(x)$ can be extracted from $\phi(x)$ by subtracting the appropriate blazed phase term $\phi_{blazed}(x)$ as follows

$$\phi_{B\text{-}S}(x) = unwrap\{\phi_G(x) - k_0\, x\, \sin(\alpha)\} \tag{5.8}$$

where the unwrapping operator $unwrap\{\ \}$ is used to limit the range of phase values in $\phi_{B\text{-}S}(x)$ to $\pm\pi$ radians.

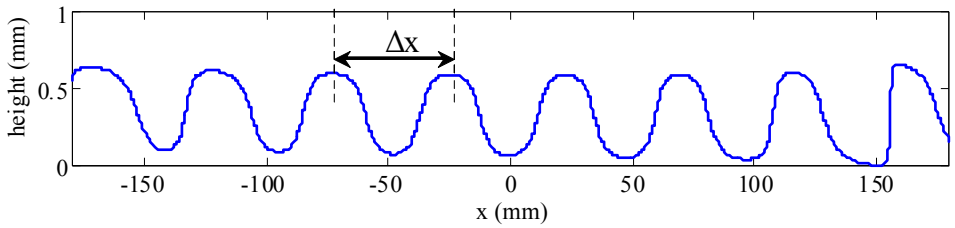


Figure 5-11. The height function $h_{B\text{-}S}(x)$ of the beam-splitting phase function $\phi_{BS}(x)$ responsible for splitting the single incident Gaussian beam into three far-field diffraction orders. The function has approximately 7.5 unit cells, with a period Δx = ~48mm. The maximum peak-to-trough depth is ~$\lambda/4$, corresponding to a maximum phase difference of π radians.

Figure 5-11 shows the height function $h_{B\text{-}S}(x)$ corresponding to the beam-splitting phase function $\phi_{B\text{-}S}(x)$ that was extracted from $\phi_{B\text{-}S}(x)$ using Eq. (5.8). The phase profile is periodic with approximately 7.5 repeated cells approximately 48mm wide. This is as expected since if we had required that beam array be produced on-axis (with the zeroth-order spot at 0°) the specific angular separation of $\theta_{\pm1}$ = ~3.58° (between the zeroth- and first-order beams) would correspond to an in-line diffraction grating with a period of

$$\Delta x = \frac{(\pm1)\lambda_0}{\sin(\theta_{\pm1})} \cong 48\text{mm}$$

for the design wavelength λ_0 = 3mm. Given that the grating length is 360mm such an in-line grating would therefore contain $L_x/\Delta x = 7.5$ repeated periods, or cells.

The image formation from this blazed, beam-splitting multiplexing phase grating is explained in terms of Fourier optics. Assuming, for now, that the Gaussian illuminated grating consists only of a blazed phase term, the transmitted wavefront at the grating is

$$E_{blazed}(x) = Gauss(x;\, W_G) \cdot \exp[i\phi_{blazed}(x)] = Gauss(x;\, W_G) \cdot e^{i\, k0x\, \sin(\alpha)} \qquad (5.9)$$

where $Gauss(x;\, W_G)$ is the incident Gaussian beam at the grating of width W_G. The wavefront in the Fourier plane is then

$$E_{blazed}(u) = Gauss(u;\, W_F) \otimes \Im\left\{e^{i\, k0x\, \sin(\alpha)}\right\} \qquad (5.10)$$

which is a single Gaussian beam (of width W_F) convolved with a sinc function shifted to $u = u_{off}$, i.e. a shifted Gaussian beam (assuming negligible truncation)

$$E_{blazed}(u) = Gauss(u - u_{off};\, W_F) \qquad (5.11)$$

If the beam-splitting multiplexing component $\phi_{B\text{-}S}(x)$ is now included at the grating the transmitted wavefront is

$$E(x) = Gauss(x;\, W_G) \cdot \exp[i\phi_{blazed}(x)] \cdot \exp[i\phi_{B\text{-}S}(x)] = E_{blazed}(x) \cdot \exp[i\phi_{B\text{-}S}(x)] \qquad (5.12)$$

The Fourier transformation of which is

$$E(u) = \Im\left\{E_{blazed}(x)\right\} \otimes \Im\left\{\exp[i\phi_{B\text{-}S}(x)]\right\} \qquad (5.13)$$

$$E(u) = Gauss(u - u_{off};\, W_F) \otimes \Im\left\{\exp[i\phi_{B\text{-}S}(x)]\right\} \qquad (5.14)$$

from Eq. (5.11), where the Fourier transform of the beam-splitting term produces an on-axis array of three diffraction orders (Delta functions). When convolved with the shifted Gaussian beam $Gauss(u - u_{off};\, W_F)$ the Fourier plane wavefront is an array of Gaussian beams shifted to position $u = u_{off}$. Figure 5-12 illustrates image formation in terms of the convolution of the Fourier transforms of a blazed and a beam-splitting phase grating.

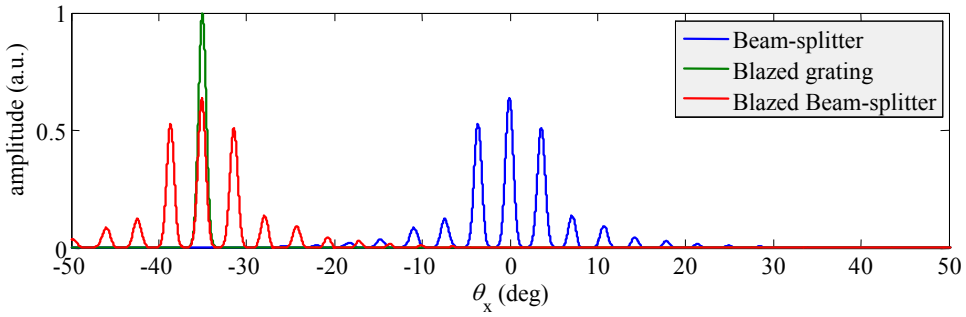


Figure 5-12. The Fourier transform of the beam-splitting phase grating is an on-axis array of three diffraction orders (blue), while the FT of the blazed grating is a single off-axis Gaussian at an angle α = 38° (green). Convolving these two FT's produces the off-axis array of three diffraction orders (red).

Fabrication

The surface relief profile was milled into a rectangular sheet of cutting-grade aluminium with dimensions of 360mm × 203mm, on a CNC milling machine in the department workshop (Figure 5-13).

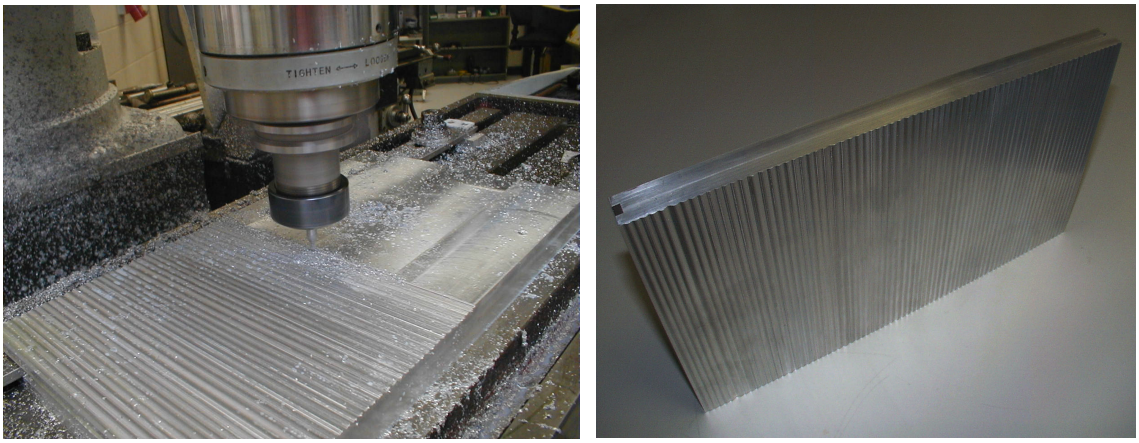


Figure 5-13. **Left:** The grating during machining on a CNC milling machine in the department workshop. **Right:** The finished reflective Fourier phase grating.

Although the one-dimensional height function $h(x)$ is shown above with a jagged saw-tooth profile, the finished grating surface is much smoother with rounded peaks and troughs and its overall appearance is more similar to a sine-wave function, in fact. Two factors that contributed to the surface being less than ideal are under sampling of $h(x)$ and limited surface accuracy afforded by the milling process. At the design stage the phase modulation $\phi(x)$ was sampled every 0.1mm (a rate of 30 points per wavelength). With the particular grating dimensions (360mm × 203mm) this requires the two-dimensional surface to be represented by 3600 × 2030 sample points. To reduce computational overhead on the software controlling the CNC milling machine $\phi(x)$ was resampled at a much lower resolution. Because of the high spatial frequency of the blazed grating term indiscriminate under sampling resulted in the loss of critical data points from $\phi(x)$. Figure 5-14(a) shows $h(x)$ sampled at the sample rates dx = 0.1mm, 0.5mm and 1.0mm and the corresponding Fourier plane amplitude distributions in each case. For larger values of dx the blazed (sloped) surfaces in $h(x)$ are retained but because some of the data points associated with peaks and troughs are now omitted 1) the vertical phase jumps are replaced by blazed surfaces and 2) the peak-to-trough height is reduced.

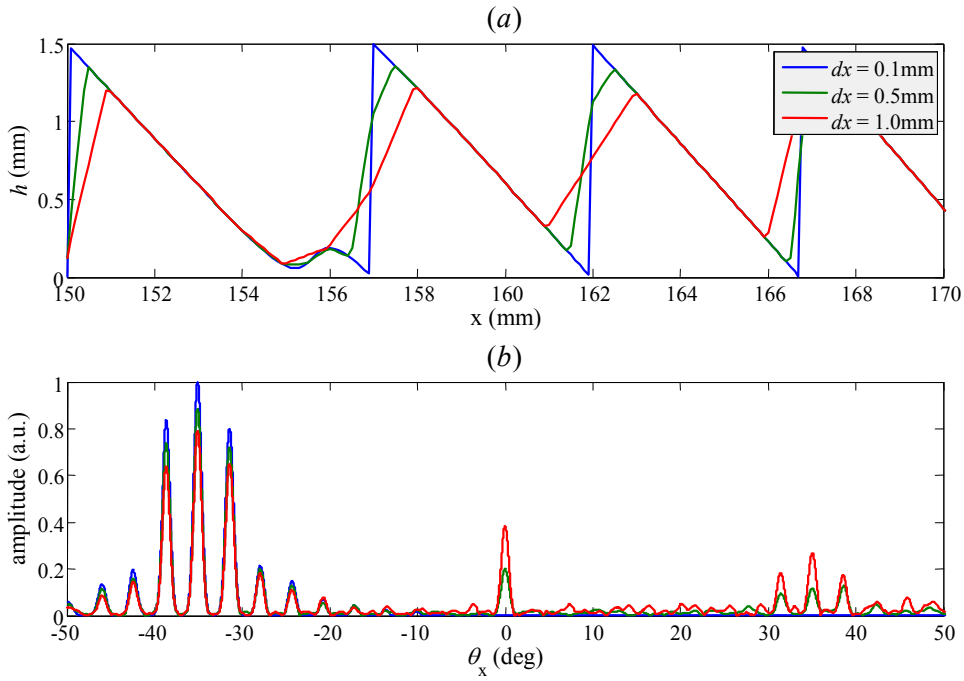


Figure 5-14. When the grating surface height $h(x)$ in (*a*) is resampled at a lower sample rate dx the result is that in the Fourier plane some of the transmitted power is diverted from the target array of three diffraction orders (centred on –38°) and redirected into an on-axis beam and another array of diffraction orders (centred on +38°).

The other surface degrading factor is the limited precision to which the surface can be machined because of the finite size of the cutting tools used to mill the surface. Because a blazed grating directs all light into the first diffraction order its period is given by

$$\Lambda = \frac{\lambda_0}{\sin(\alpha)} \tag{5.15}$$

which for the offset angle $\alpha = 38°$ gives a period of $\Lambda = 4.91$mm. However the smallest cutting tool used had a radius of 1 mm so clearly such small surface features cannot be machined to exact specifications.

Measurements

The first arrangement used to measure the reflection blazed beam-splitter is shown in Figure 5-15. Although the direction of propagation of the reflected wavefront from the grating is high at 38° it is still shallow enough that if the collimating lens is placed at f_1 = 250mm in front of the grating, a sizeable portion of the reflected wavefront will be obscured by the lens. Therefore the lens was set at $2\times f_1$ = 500mm in front of the grating.

Since this distance is much less than the confocal length (which is several metres) the illuminating Gaussian beam is still collimated at the grating plane as required.

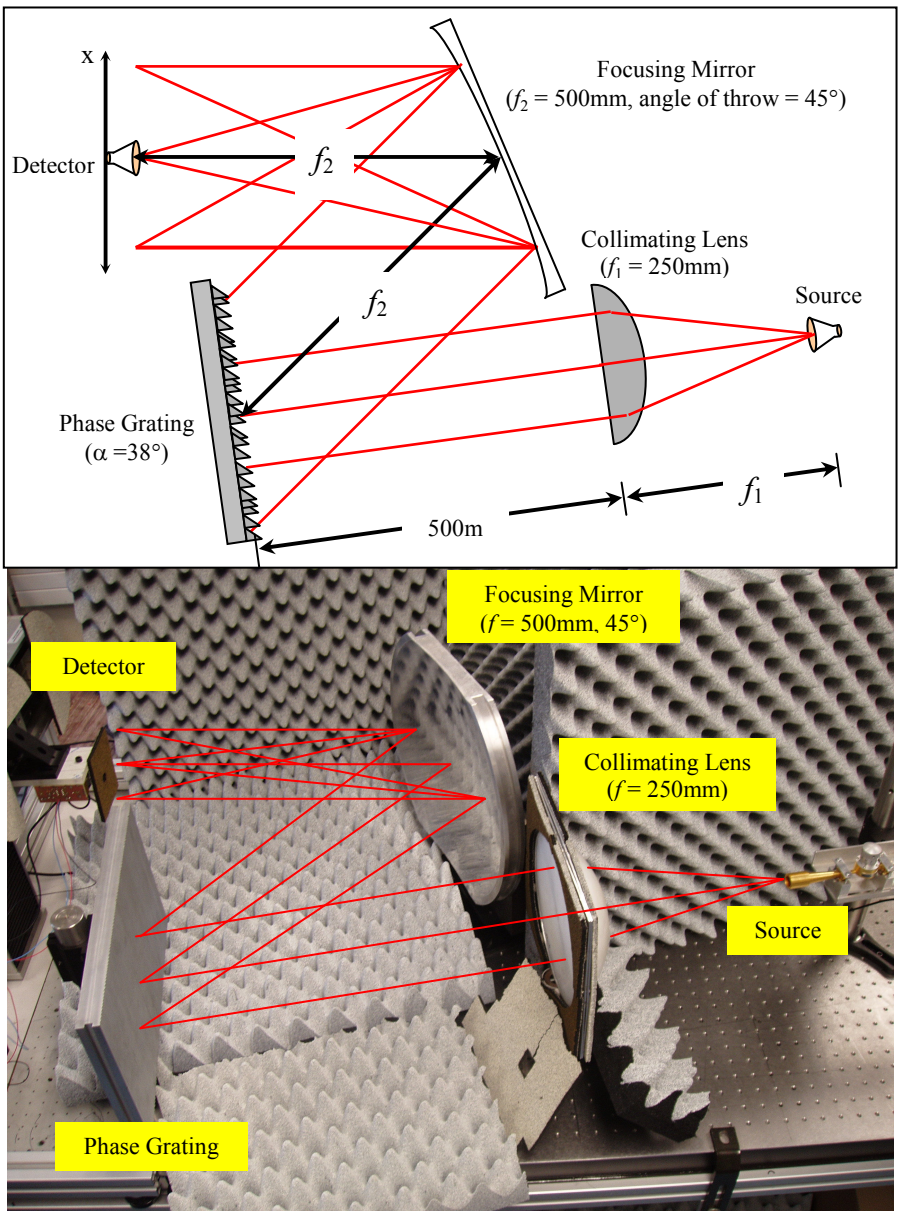


Figure 5-15. Schematic and photograph of a Fourier optics arrangement for testing the blazed beam-splitting phase grating. The source beam is collimated by a HDPE plano-convex lens to provide a collimated Gaussian beam at the phase grating. The mirror then focuses the far field diffraction order array onto the output (detector) plane.

The problem with this set-up is that the incident Gaussian beam (provided by the lens) is very small compared to the grating width. Thus only the central part of the grating is involved in modulating the incident wavefront. Ideally all of the grating cells should contribute to the process. The lens was thus replaced with a 500 mm focal length off-axis ellipsoidal mirror (Figure 5-16). Now the Gaussian beam incident on the grating

has a radius of approximately 110mm, which is sufficient to illuminate the entire grating.

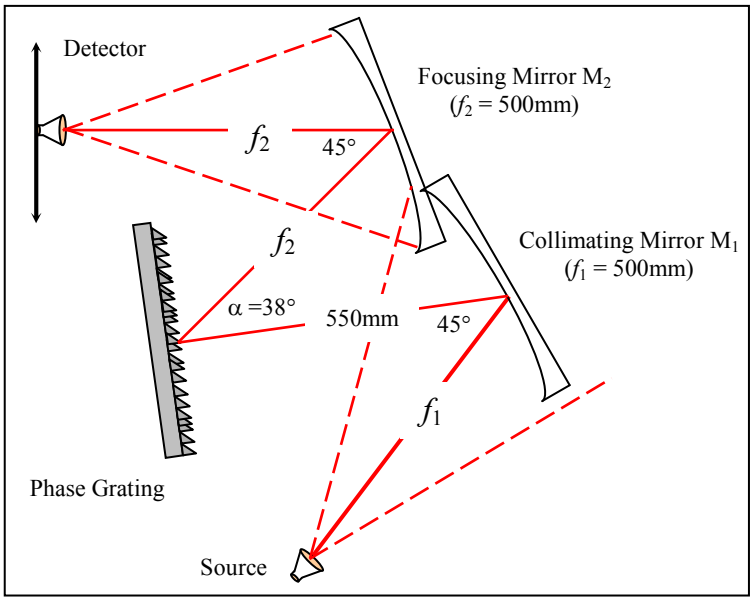


Figure 5-16. Schematic of the 4-*f* Fourier optics set-up used to test the blazed beam-splitting phase grating. The longer focal length of mirror M_1 means the grating is illuminated with a Gaussian beam wide enough to ensure adequate coverage of the entire grating. However note that some of the power from the source is collected directly by mirror M_2, which may interfere with the far field image measured at the detector plane.

Because the actual manufactured grating surface was more sine-wave like than saw-tooth (because of the finite size of the cutting tool) it was difficult to determine by visual inspection on which side of the grating normal the array of diffraction orders would propagate. Two trial measurements were performed to establish the direction of propagation. The grating was mounted using the 4-*f* Fourier optics set-up (Figure 5-16) and the intensity measured at the output plane. The grating was then rotated 180° about the optical axis and a second set of measurements made. Naturally, we expect to observe the array of three Gaussian beams in just one of the measured images and little or no power in the other. However three diffraction orders are clearly observed in both measurements (Figure 5-17). Thus the grating in fact generates two arrays at +38° and –38° to the grating normal.

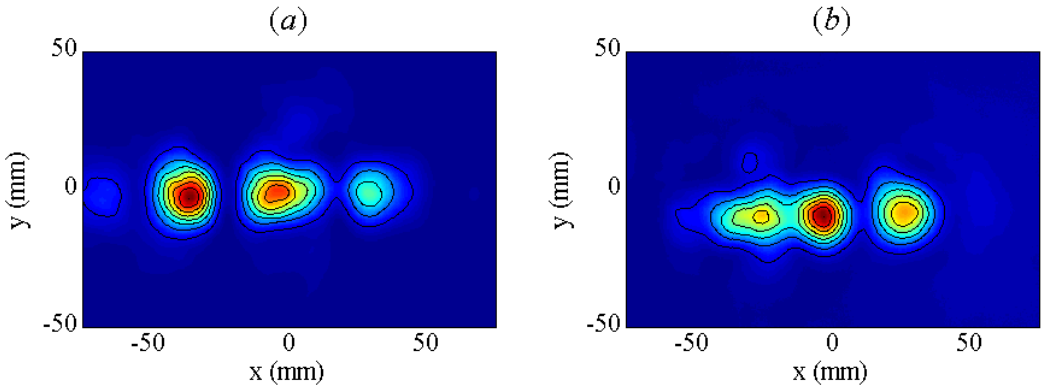


Figure 5-17. Contour plots of the output plane intensity measured with the phase grating arranged such that the centre of focusing mirror M_2 is at an angle of (*a*) −38° and (*b*) +38° to the grating normal.

The presence of the second set of diffraction orders was already predicted in Figure 5-14(*b*) and can be explained by taking into account the surface error due to the milling process (the finite size of the cutting tool). We can simulate the effect of this in an alternative approach to previously as follows. The height function $h(x)$ is transformed into a smoothened height $h'(x)$, representing the machined surface, by use of a smoothing operator defined by

$$h'(x) = smooth\{h(x); s(x)\} = \beta\,[h(x) \otimes s(x)] \tag{5.16}$$

where $s(x)$ is a discretely sampled 'smoothing' function, β is a scaling factor needed to ensure that the range of height values $h'(x)$ matches those of $h(x)$ and ⊗ represents convolution. After $h'(x)$ has been calculated, the equivalent phase front $\phi'(x)$ is then given by

$$\phi'(x) = 2\,k_0\,h'(x) \tag{5.17}$$

Figure 5-18 shows the results of smoothing $h(x)$ using Eq. (5.16) where a sine-wave function $\sin(\pi x/n\Delta x)$ with n sample points and sampling interval Δx was used for $s(x)$. As n increases $h'(x)$ becomes smoother and the saw-tooth features take on a more sine-wave like appearance (similar to what occurred when $h(x)$ was undersampled). Also note that as n increases the peaks-to-trough height of some of the grooves declines, so the surface undulations become shallower.

The effect that the smoothing operator has on the Fourier plane intensity is that as the convolving function $s(x)$ is made larger (by increasing samples n) more power is diverted from the target array of diffraction orders into an on-axis array of beams and an array on the opposite of the Fourier plane at −38°, as observed in measured data.

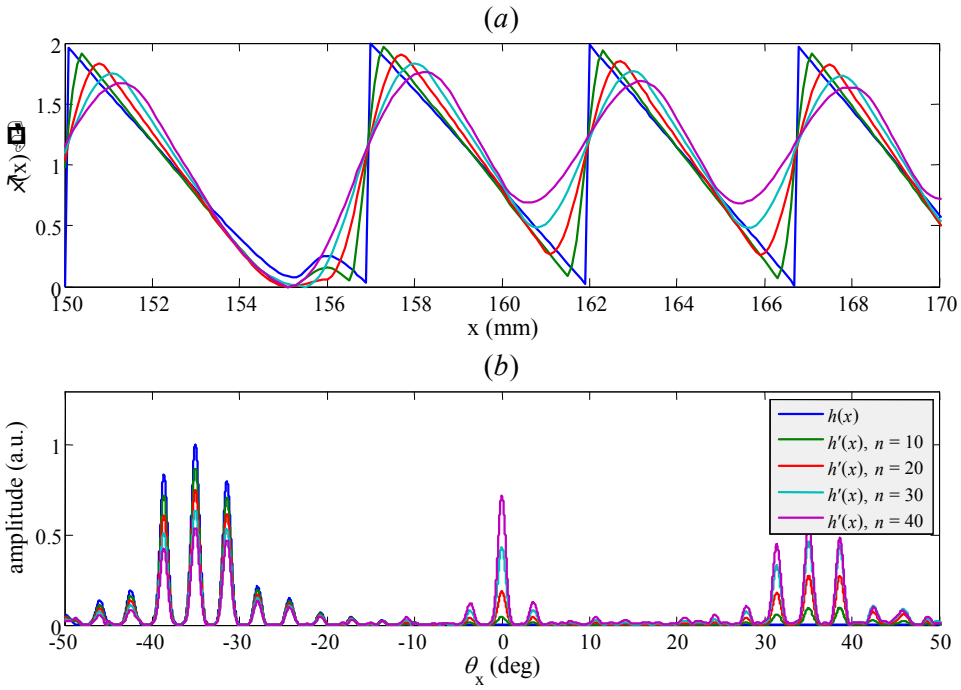


Figure 5-18. Smoothing the height function $h(x)$ is accompanied by the appearance of three diffraction orders on the opposite side of the normal and centred on $-\alpha = +38°$. As the size of the smoothing function, defined by its number of samples n, increases more power is directed into the extra set of diffraction orders as well as the on-axis beam..

This behaviour is explained qualitatively as follows. The mirrored set of diffraction orders appears because as the saw-tooth structure is smoothed it becomes more sine-wave like and since the Fourier transform of a sine-wave is an odd-impulse pair [5.32], i.e. a pair of Delta functions, the array of diffraction orders produced by the beam-splitting component of the phase grating is now convolved with two Gaussian beams at angular positions of ±38°. Furthermore, in this simulation the peak-to-trough depth of $h'(x)$ is less than that of the ideal height function $h(x)$. Thus the phase modulation $\phi'(x)$ is also shallower than $\phi(x)$, which results in the appearance of an on-axis diffraction order, the intensity of which increases as the $\phi'(x)$ is made shallower (as the surface is made smoother). The mechanism that produces this on-axis beam is the same affect employed in the design of beam samplers [5.33]. Of course the experimental arrangement used did not permit verification of the presence of the on-axis diffraction order since it would propagate back in the direction of the illuminating beam and be imaged at the source horn (and also potentially give rise to standing waves). An alternative arrangement incorporating polarising grids and a Faraday rotator would allow operation of the reflection grating in normal incidence.

Frequency Response

The consequence of combining a blazed grating with a multiplexer is that not only is the angular separation of the diffraction orders (the beams within the array) dependent on wavelength but so too is the position of the array as a whole: the blazing effect. At wavelength λ the angular position of the first-order diffraction beam from a blazed grating of period Λ is given by

$$\alpha = \sin^{-1}\left(\frac{\lambda}{\Lambda}\right) \tag{5.18}$$

Since the beam array (produced by the multiplexing part of the grating) is convolved with this diffraction order, the centre (zeroth-order spot) of the array of beams is also given by Eq. (5.18).

The source was tuned to 101 GHz and a set of intensity measurements made of the pattern within a 160mm × 30mm region centred roughly on the zeroth-order spot of the three-beam diffraction order array (labelled as position $x \approx 0$). Thirteen more measurements were made of the same region in the output plane at successively lower frequencies at intervals of approximately 2 GHz. The 2-D intensity images obtained are shown in Figure 5-19 along with horizontal cuts through the centre ($y = 0$) of each scan. As the source frequency is reduced the spot array moves across the output plane in the negative x-direction, until at 77 GHz the 3-beam array has moved entirely out of view and all that can be seen are several higher-order ($n = +2$ and $+3$) diffraction spots. Note that negative values of x correspond to larger distances from the grating – that is larger off-axis angles with respect to the grating normal.

The positions, x_n of peak intensity of the diffraction orders visible in each scan were identified and plotted against frequency in Figure 5-20(*a*). At each frequency the zeroth-order spot position, x_0 corresponds to the expected position x_{blazed} of the first-order diffraction beam produced by the blazed grating term given by

$$x_{blazed} = f_2 \sin^{-1}\left(\frac{\lambda}{\Lambda}\right) \tag{5.19}$$

Figure 5-20(*b*) shows a plot of $(x_n - x_{blazed})$ against frequency to illustrate the wavelength dependent nature of diffraction order spacing from the beam-splitting component of the grating, typical of any periodic grating.

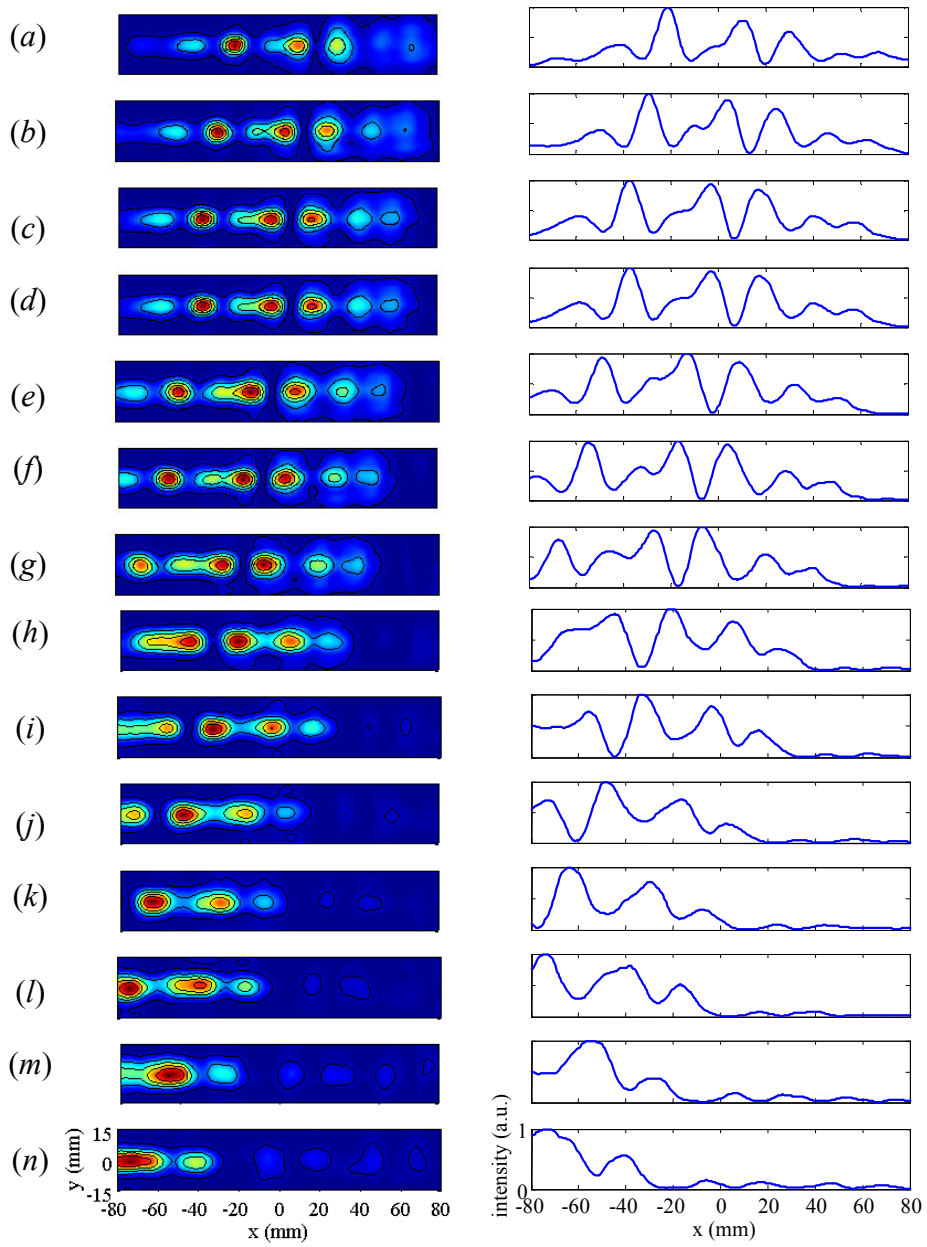


Figure 5-19. **Left:** Contour plots of output plane field amplitude measured at frequencies of ~101 GHz (*a*) to ~75 GHz (*n*) at approximately 2 GHz intervals. **Right:** Horizontal cuts through the centre (at y = 0mm) of the 2-D intensity maps on the left. The exact source frequencies were (*a*) 100.9, (*b*) 98.07, (*c*) 96.61, (*d*) 95.45, (*e*) 94.36, (*f*) 93.23, (*g*) 91.26, (*h*) 88.03, (*i*) 85.65, (*j*) 83.30, (*k*) 80.75, (*l*) 79.05, (*m*) 77.06 and (*n*) 74.92 GHz. Note that x = 0 approximately coincides with the position of the centre of the array at the design frequency (i.e. 38° off-axis with respect to the grating normal).

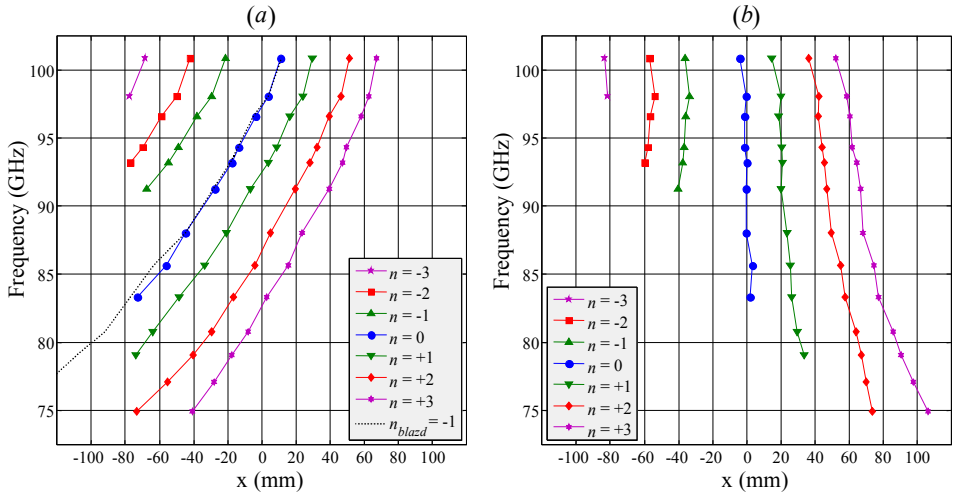


Figure 5-20. (*a*) Measured beam centre positions x_n as estimated from the measured intensity patterns in Figure 5-19 and (*b*) beam centre positions (x_n–x_0) relative to the centre of the beam array, where x_0 is the predicted position of the array centre. The second plot emphasises that the diffraction order spacing of the beam array produced by this blazed multiplexer has the usual frequency dependence of a periodic grating.

Notice that although the task of the IPRA was to find a phase modulation to generate an array of three equally intense diffraction orders, the actual beam uniformity is poor since the zeroth-order spot is more intense than the two neighbouring first-order beams, as predicted by simulations. However as mentioned above a shallower grating depth results in a more intense on-axis beam so conversely a deeper grating profile will produces a less intense on-axis beam. It follows that by appropriate scaling of the beam-splitting (multiplexing) phase term (in isolation of the blazed phase component) the intensity of the zeroth-order beam can be reduced and a solution found that distributes power more evenly between the three diffraction orders. Figure 5-21 shows the result of redefining the grating phase term as

$$\phi(x) = \phi_{blazed}(x) + \left[1.125\ \phi_{BS}(x)\right] \tag{5.20}$$

which results in an array of three diffraction orders with roughly equal intensity.

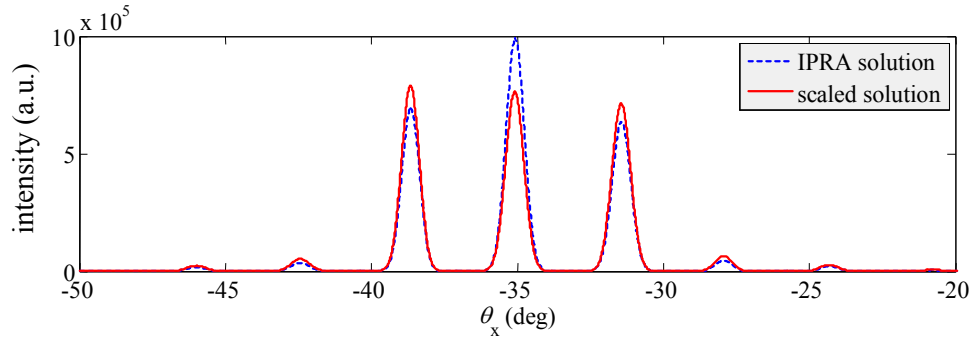


Figure 5-21. Fourier plane intensity from a grating derived from the solution provided by the IPRA (dotted blue curve) and from a deeper version of the same grating (solid red curve).

However, the actual measurements of the grating demonstrate an even poorer performance than predicted through realistic simulations of the actual manufactured grating. In particular, the beams are weaker and less uniform than predicted. This is most likely due to the non-ideal illumination of the grating in test arrangement 2 (using two off-axis mirrors) due to truncation of the incident beam on the grating. In general one conclusion that can be drawn is that blazed phase gratings at these wavelengths are difficult to utilise as the compact nature of the optics given the long wavelength can give rise to vignetting effects. Therefore it was decided to pursue only on-axis designs in further investigations of Fourier phase gratings.

5.4.2 Two-dimensional Fourier phase grating(s) designed using the Gaussian Beam Mode Iterative Phase Retrieval Algorithm

In terms of designing an example of a 2-D Fourier phase grating for careful testing and design verification it was decided to choose a grating that would produce a sparse array of images of a source beam which does not possess the rectangular array configuration of a Dammann grating. The example chosen was that of a circular array of image beams with no on-axis beam. The grating design was achieved using a Gaussian beam mode version of the two-dimensional iterative phase retrieval algorithm (IPRA). The function of the grating is to split the incident wavefront (a collimated Gaussian beam) into a circular array of eight Gaussian beams in the far field.

Grating Design

The grating design consisted of the following steps:

1) Define target amplitudes A_G and A_F
2) Initialise image phase ϕ_F
3) Define an appropriate set of Gaussian Beam Modes
4) Begin Iterative Phase Retrieval

In step 1 we must define the specific problem, in other words the target intensities at the grating and Fourier planes. It is our intention that the grating be tested with a Fourier optics arrangement in which the incident beam is collimated by an off-axis parabolic mirror of focal length 350mm, which produces a Gaussian beam at its focal plane with a waist radius W_G = 71mm. Thus the target amplitude at the grating is a single Gaussian

beam with the same radius. The target image plane intensity distribution is a circular array of eight Gaussian beams. All of the previous gratings had a zeroth-order diffraction spot but here the aim is to completely suppress the power in this beam. The target signal intensity is shown in Figure 5-22.

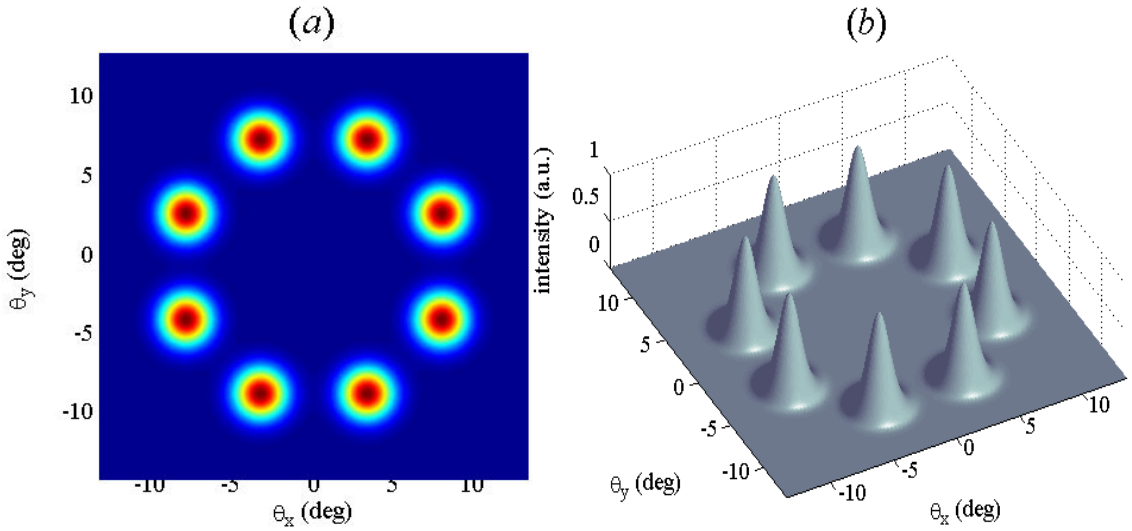


Figure 5-22. Plots of the far-field target (*a*) amplitude and (*b*) intensity for the sparse array of eight Gaussian beams in a circular arrangement.

Step 2 involves defining an initial phase at either the object or image plane with which to start the algorithm. Ideally the initial phase should be a close estimate of the solution phase that will satisfy the intensity constraints at both planes. For this particular problem the far field intensity consists of well-separated regions of high intensity, between which the intensity goes to zero (i.e. a high contrast sparse pattern). Thus uniform phase fronts were assigned to each beam individually, since a solution of this type would be desirable for coupling an array of Gaussian beams to a set of feed horns. Since the optimal arrangement of phases on each beam is unknown, an arbitrarily set of phases were chosen such that the phase difference between beam n and n+1 (i.e. neighbouring beams) is equal to $+\pi/4$ radians. It was found that this initial phase distribution was a very poor estimate of the far field phase distribution for the solution found. Despite the poor choice of initial signal phase the algorithm was still able to find an acceptable and useful solution to the problem.

In Step 3 we define the basis set of Gaussian beam modes needed for propagating between the grating and image planes during the iterative stage of the IPRA. Because propagation between the grating and image (far field) plane is performed using GBM's the image plane is not strictly located at infinity (as it would be

in a FFT version of the algorithm) but at a propagation distance that is so large that it can be considered to be in the far field of the grating plane. The Gaussian beam mode basis set is characterised by the following parameters and properties

- Scaling factor (i.e. beam width W)
- Mode set size (number of modes)
- Symmetry parameter (even, odd, or none)

We used Gaussian-Hermite modes defined (as in Chapter 2) by $\psi_{mn}(x, y) = \psi_m(x)\psi_n(y)$, where $\psi_m(x)$ is a one-dimensional Hermite mode of index m. Thus the mode-set size is given by $(m_{max}+1)\times(n_{max}+1)$, where m_{max} and n_{max} are the maximum mode indices for the x and y 1-D modes, respectively. The scaling factor is determined by the waist radius of the fundamental mode, which may require different values in each direction (i.e. $W_{0,x} \neq W_{0,y}$) depending on the particular problem. The symmetry parameter defines which modes are included in the mode-set. If a symmetry parameter is not specified then the mode-set consists of all modes with indices $m = [0, m_{max}.]$. A symmetric mode-set is defined as one consisting of only even-numbered (symmetric) modes, whereas an asymmetric mode-set contains only odd-numbered (asymmetric) modes. Imposing a symmetry constraint on the mode-set necessarily restricts the solutions that can be attained. Clearly, a symmetric mode-set can only produce solutions whose fields are also symmetric. It also means that redundant modes are automatically omitted from consideration and this can reduce considerably computational overhead (memory and execution time).

When analysing an existing phase grating solution, such as a Dammann grating, in terms of Gaussian beam modes the task of finding a suitable mode-set to describe the phase grating is relatively trivial. We simply experiment with different values of the three mode-set parameters and pick those that give the best reconstruction of the original grating field in terms of a least squares fit to the field (for example). However choosing a suitable mode-set for the Fourier grating phase retrieval problem is more difficult since we do not know the phase distributions at the grating or image planes and so do not know the spatial frequency content of the field. In this situation the best that one can do is choose a set of parameters that yield accurate reconstruction of the target amplitudes (particularly at the image plane with a best guess for the phase solution).

A routine was written to allow the user to experiment with different value of the three mode-set characteristics alluded to above. For the current problem it was found

that both the grating and far field target amplitudes (with uniform phase for the image plane field) could be accurately reconstructed (with unit intensity correlation at both planes) using a mode-set with a maximum mode number of $m_{\max} = n_{\max} = 34$ with optimum mode-set scaling achieved with a fundamental beam mode radius of $W_{0,x} = W_{0,y} = 27.23$ mm. Thus the mode-set contains (34+1)×(34+1) = 1225 modes. This, coupled with the sampling requirement at each plane, leads to extremely large matrices for storing the modes and their pseudoinverses at each plane. If the mode set is restricted to symmetric modes only the number of 2-D modes is reduced 18×18 = 324, which reduces the memory requirements considerably. A symmetric mode-set was chosen because the target intensities $|A_G|$ and $|A_F|$ are themselves symmetric about the x- and y-axes. However, this of course will only give rise to solutions with symmetric phase distributions at both planes.

In two-dimensional GBMA, mode coefficients are more efficiently calculated using the pseudoinverse SVD approach rather than by integrating over individual modes (see Chapter 2). Thus as well as the mode arrays $\mathbf{\Psi}_G$ and $\mathbf{\Psi}_F$ (in which each mode is represented as a column array) that contain mode values at each plane we also require two arrays $\mathbf{\Psi}_G^+$ and $\mathbf{\Psi}_F^+$, to store the pseudoinverses of $\mathbf{\Psi}_G$ and $\mathbf{\Psi}_F$, respectively. The evaluation of the pseudoinverse of a large array is extremely time consuming and therefore $\mathbf{\Psi}_G^+$ and $\mathbf{\Psi}_F^+$ were calculated before beginning the iterative stage of the IPRA.

Analysis of the grating solution

The GBM-IPRA was run for 1000 iterations. The grating and far field intensities at the end of the algorithm are shown in Figure 5-23 and Figure 5-24. The circular array is clearly formed, however none of the Gaussian beams have the desired symmetric profile, and instead appear to be slightly squeezed in different directions. The surface plot in Figure 5-23(*b*) shows that there are intensity nulls at different points around the edge of each of the beams. Since only symmetric modes were used the solution is itself symmetric, thus the beams labelled 'A' and 'B' in the third quadrant in Figure 5-23(*a*) have equal counterparts in the other three quadrants upon reflection about the x- and y-axes. Figure 5-24 shows the intensity of the GBM-approximated grating plane wavefront after target signal intensity replacement (with the circular beam array). Thus if the grating were illuminated with the intensity shown in Figure 5-24 the far field

intensity would be exactly equal to the target intensity (a circular array of eight beams). Since however we require a phase-only solution this intensity is replaced with the target grating intensity (a single Gaussian beam), which yields the actual far field intensity seen (Figure 5-23).

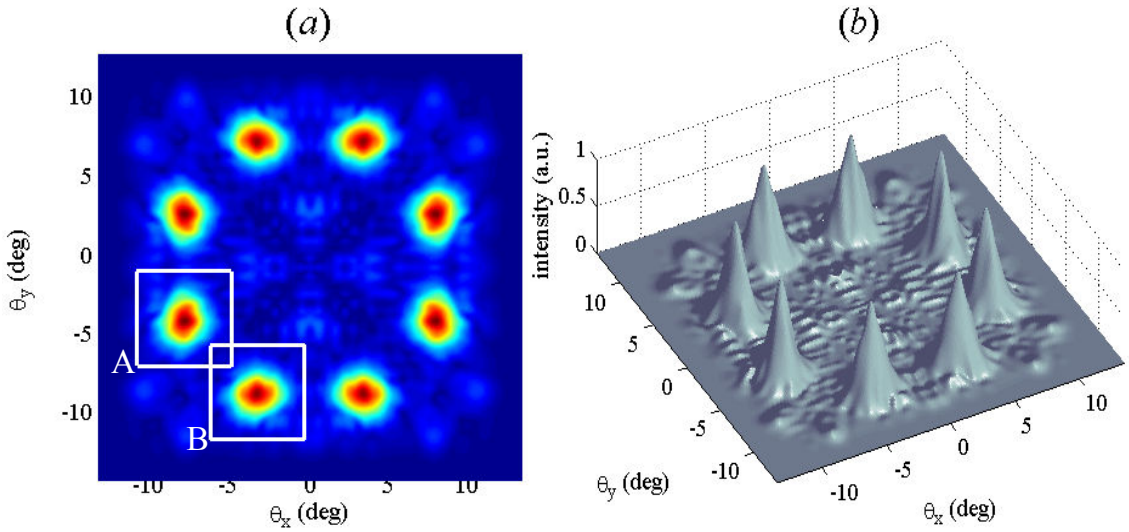


Figure 5-23. Far field solution found by GBM version of IPRA showing (*a*) linear-scale false-coloured plot of amplitude and (*b*) 3-D plot of intensity (*b*).

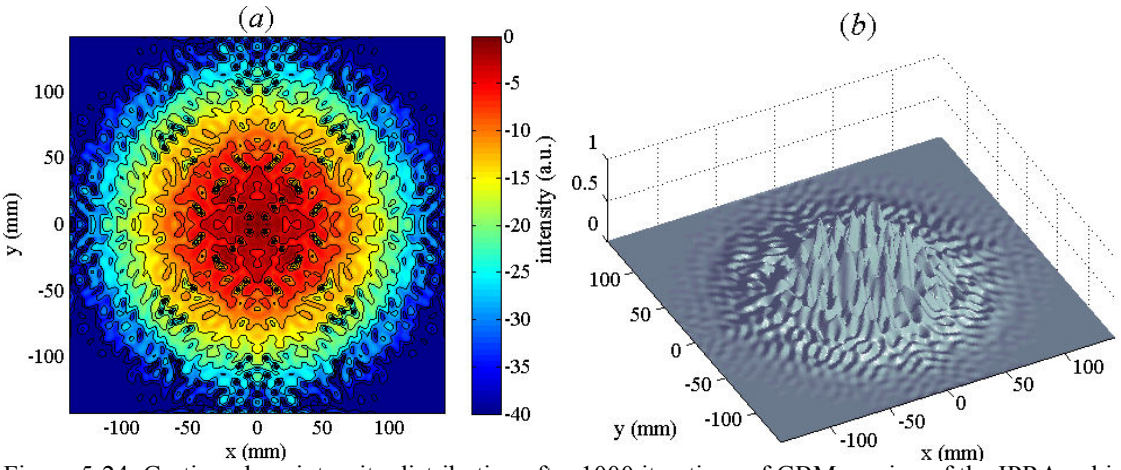


Figure 5-24. Grating plane intensity distribution after 1000 iterations of GBM version of the IPRA, which yields the far-field intensity pattern shown in Figure 5-23.

The GBM-IPRA is effectively a multivariable optimisation algorithm that seeks to find the most suitable set of mode coefficients that can simultaneously satisfy the intensity requirements at the grating and image planes. A solution to the phase retrieval problem is one that yields a set of mode coefficients a_{mn} that can simultaneously satisfy the target intensity requirements at both (grating and image) planes. However the phase at the image plane may be very different from the initial setting. It is interesting to examine the mode coefficients a_{mn}, the absolute values of which are shown in Figure 5-25. The first two plots in Figure 5-25 show $|a_{mn}|$ for reconstructions of the target grating and

image fields, E_G and E_F before beginning the IPRA, i.e. with uniform phase assigned to the fields at each plane. The third plot shows $|a_{mn}|$ after 1000 iterations of the IPRA. The grating target amplitude A_G contains a single on-axis Gaussian beam so only a few low-order modes contain power (Note that the incident Gaussian at the grating does not have the same width as the fundamental of the GBM mode-set, which acts as the basis for describing the fields). Since the far-field target amplitude, A_F is devoid of any power inside the circular array of eight beams, there is effectively no power in the lower-order modes. Notice that the symmetry of A_F is reflected in the distribution of power in the mode coefficients. The plots of $|a_{mn}|$ in Figure 5-25(*c*) is quite similar to that in Figure 5-25(*b*) but with higher power in some lower-order modes. Notice also while in Figure 5-25(*a*) and (*b*) power is confined to a small number of modes, in Figure 5-25(*c*) a small but finite amount of power exists in nearly all of the modes. Thus the solution that was found makes use of most of the available modes in some way.

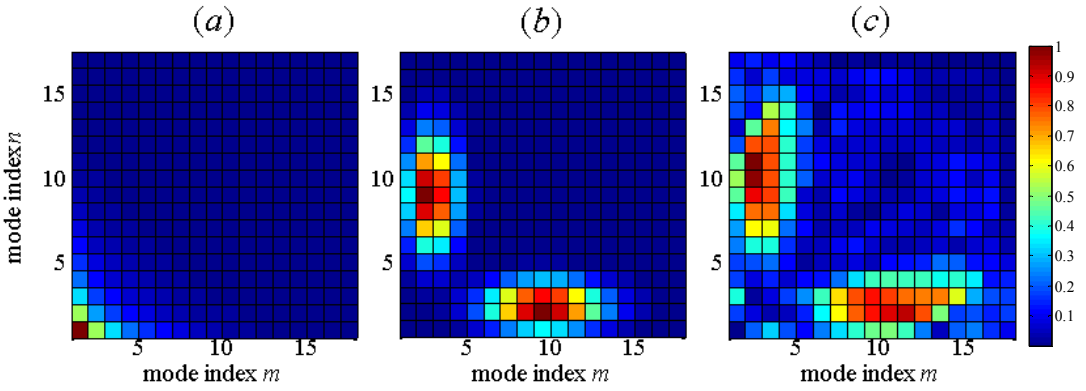


Figure 5-25. The absolute values of mode coefficients a_{mn} for the 18×18 GBM mode-set used to find a solution to the beam-splitting problem to produce a circular array of eight far field Gaussian beams. (*a*) $|a_{mn}|$ for reconstructed grating-plane target intensity $|A_G|^2$, (*b*) $|a_{mn}|$ for reconstructed far-field target intensity $|A_F|^2$ and (*c*) $|a_{mn}|$ for the solution found after 1000 iterations of GBM-IPRA.

We now examine more closely the quality of the far-field intensity for the solution found. Figure 5-26 and Figure 5-27 show close ups of the intensity and phase distributions in the vicinity of the two beams labelled A and B in Figure 5-23. Neither of these beams exhibit the circular symmetry ideally required. Also strong intensity fluctuations exist across the beams – normally referred to as speckles [5.34]. Ideally the phase should be uniform across each beam, but inspection of the phase across beams A and B shows that is not the case. Regions of phase with a constant gradient are indicated by contours with equal spacing and these occur at the centre of the beam. However closely spaced contour lines, for example in Figure 5-26(*b*), indicate regions of steep

phase, which are associated with regions of low intensity. These rapid phase fluctuations at the outer parts of the beam that give rise to the intensity fluctuations seen about the centre of the Gaussian beam. The intensity fluctuations observed can be divided into two types: 1) fluctuations originating from neighbouring sample points with a phase difference close to π between them and 2) fluctuations caused by spiral phase singularities. A fluctuation of the first type corresponds to a change in sign and has an intensity value close to zero, whereas a fluctuation of the second type is actually a zero in the wavefront. A phase singularity, or speckle, occurs at a zero location at the point where the plane of observation intersects an optical vortex in the propagating field. This type of defect introduced into the wavefront during iterative phase retrieval is usually associated with beam shaping problems [5.34].

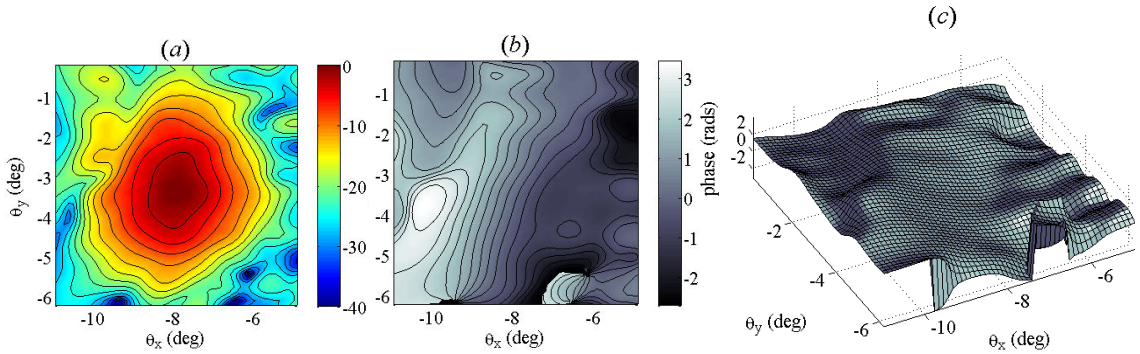


Figure 5-26. (*a*) Log-scale plot of intensity (dB) and (*b*)-(*c*) phase distributions in the vicinity of the output beam labelled A in Figure 5-23.

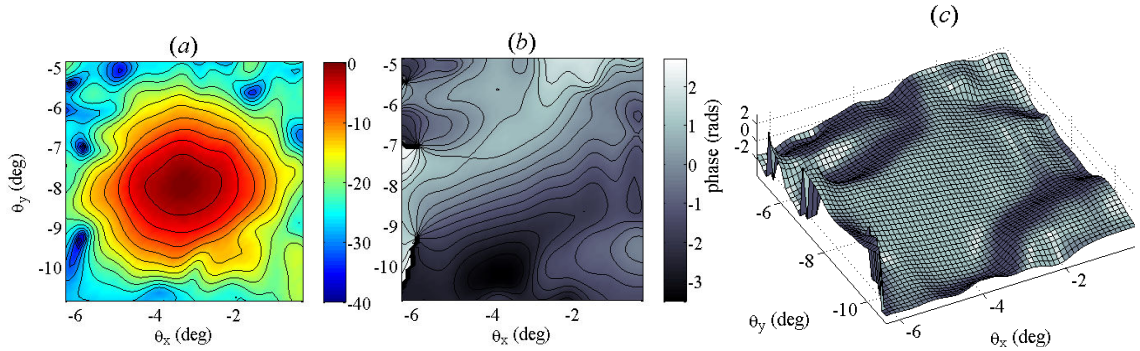


Figure 5-27. (*a*) Log-scale plot of intensity (dB) and (*b*)–(*c*) phase distributions in the vicinity of the output beam labelled B in Figure 5-23.

The phase grating solution was found with the GBM-IPRA so now verify the solution by propagating the grating field to the Fourier plane using FFT. Figure 5-28(*a*) shows the far-field intensity of the solution found by the GBM-IPRA. The intensity drops to zero near the edges of the frame, indicating that the phase grating solution found by the IPRA succeeds in containing all of the power transmitted from the grating within the region of interest. However a Fourier transform of the Gaussian-illuminated phase

grating shows that in fact power is present outside the signal window. The reason for the discrepancy between the Fourier transformed wavefront and the Gaussian beam mode propagated wavefront is that, the modes used for propagating to the far field are scaled such that the highest-order mode fits just inside the signal window. Thus any power that exists at propagation angles outside the signal window cannot be accounted for in the current GBM decomposition simply because there are no modes defined at those points. In other words the GBM basis set is not complete, which is the drawback of using as small a GBM basis set as possible for computational efficiency.

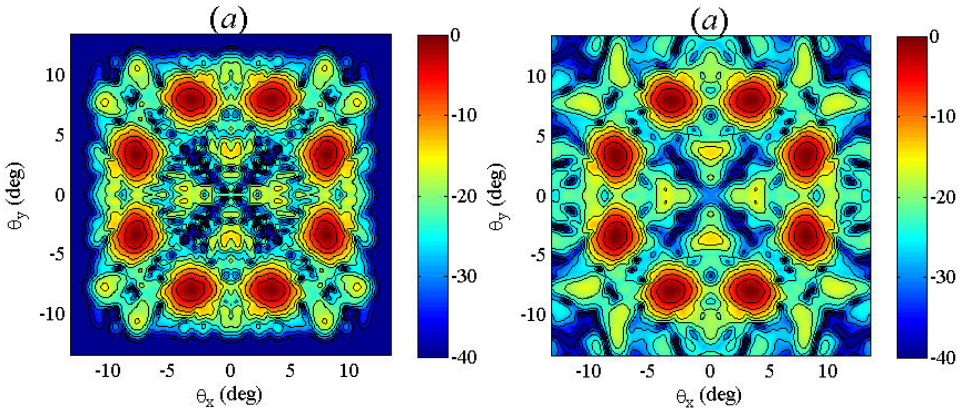


Figure 5-28. The far field intensity from the Gaussian-illuminated phase grating solution found using the GBM-IPRA with propagation performed using (*a*) Gaussian beam mode propagation (with the same mode set used during the GBM-IPRA) and (*b*) a fast Fourier transform.

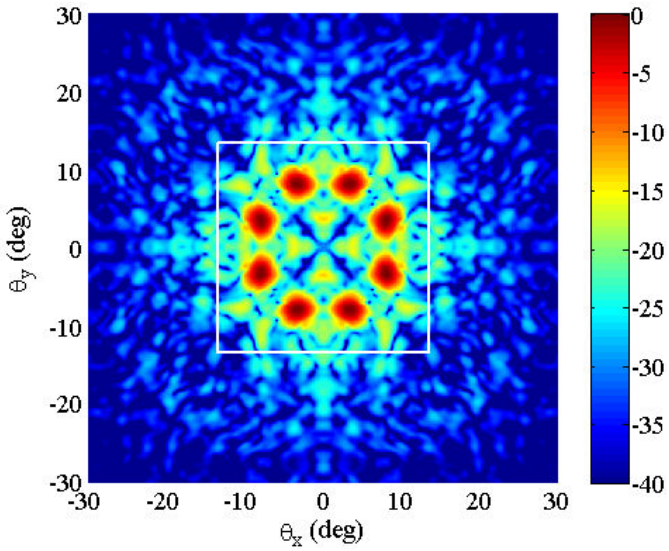


Figure 5-29. Intensity of the Fourier transformed grating wavefront. The white box indicates the region of interest in the far field (the image window) inside which all of the Gaussian beam modes used in the IPRA fit.

Figure 5-29 shows a wider view of the Fourier plane intensity and illustrates clearly that power extends beyond the edges of the signal window (bounded by the white box),

which is not accounted for with propagation using Gaussian beam modes. A better solution than that found by the GBM-IPRA might be possible if more modes, of wider extent were used, i.e. if the signal window was made larger. The reason why this was not done here was because of computational limitations and the goal was to develop an efficient computational approach and then analyse its limitations.

Comparing solutions obtained with FFT-IPRA and GBM-IPRA

The previous section described the use of a novel Gaussian Beam Mode Analysis technique for the solution of the phase retrieval problem in the case of phase gratings. It was shown that a very efficient algorithm could be developed based on a GBM basis set of limited size. In order to analyse the performance of this technique in terms of its accuracy and speed we compare the GBM approach with the more traditional fast Fourier transform approach. Therefore we next applied a FFT-based IPRA to find a solution to the same problem to see how the GBM-based algorithm compares. Another reason was to check whether the intensity fluctuations (associated with speckles and phase vortices) observed in the GBM-based approach would appear with a FFT-based method, or whether in fact they were due some limitation in the GBM-IPRA (e.g. insufficient number of modes).

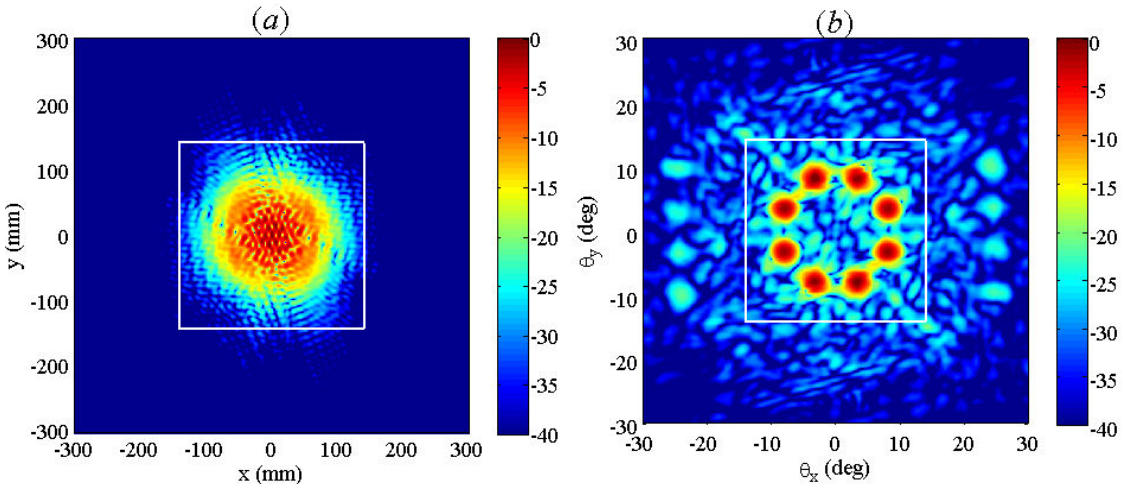


Figure 5-30. (*a*) Grating plane and (*b*) Fourier plane intensity distributions of the solution obtained with 1000 iterations of the FFT-based IPRA corresponding with those for the GBM approach described in the previous section. The white squares represent the signal windows (regions of interest) at each plane.

Figure 5-30 shows the grating and image plane intensities after 1000 iterations of the FFT-IPRA, which was executed using the same initial signal phase as was used with the GBM-IPRA. The signal window (white square) at the grating plane represents the intended dimensions of the grating, so if the solution is to be used to create a diffractive

phase element the phase outside this area are not included. Notice that only a very small fraction of power at the grating plane strays outside the window, thus the intensity correlation (between target and estimated intensity) evaluated within the window and over the entire plane yield similar values of 90.55% and 90.32%, respectively (intensity correlation refers to the value returned by integrating the target intensity multiplied by the solution intensity). At the Fourier plane a much higher proportion of power exists outside the signal window. Thus the intensity correlations evaluated within the window yields a value very different to when it is evaluated over the entire Fourier plane: c_F = 88.62% and $c_{F\text{-}WIN}$ = 93.65%, respectively.

The grating plane intensity and phase distributions returned by the FFT-IPRA are shown in Figure 5-31. At the grating plane the intensity correlation between the solution and target intensity is 90.71%, compared to 87.81% for the solution obtained with the GBM-based IPRA. The four-fold symmetry of the target signal intensity has not been preserved in the solution found with the FFT-IPRA. Instead the amplitude and phase exhibit 2-fold rotational symmetry about the origin.

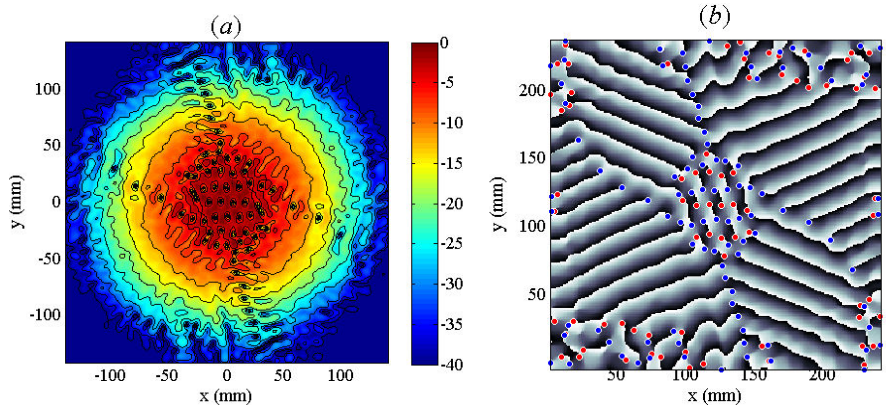


Figure 5-31. Grating plane (*a*) intensity and (*b*) phase distributions after 1000 iterations of the FFT-based IPRA. Blue and red markers in (*b*) represent the locations of positively and negatively charge optical vortices in the phase distribution.

The far field intensity (Figure 5-32) possesses the same two-fold rotational symmetry seen at the grating plane. The intensity correlation between the far field solution intensity and the target intensity within the signal window is 95.86% (compared to 93.06 for the solution obtained with the GBM-based algorithm). The far field intensity obtained by the FFT-IPRA contains speckle (points of zero-intensity) inside the image plane Gaussian beams, as occurred in the GBM solution (but at different points). Since the solutions obtained with both methods produce wavefronts with speckles, and hence

optical vortices, we can infer with some confidence that the GBM-based algorithm is not to blame for their presence.

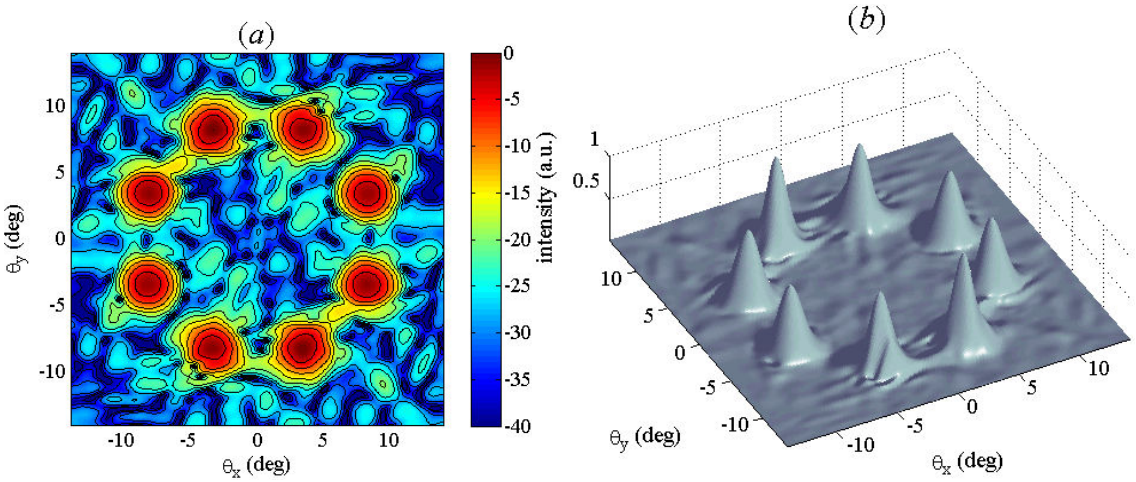


Figure 5-32. The Fourier plane intensity after 1000 iterations of the FFT-based IPRA. The false-coloured plot in (*a*) is displayed in log-scale to highlight the existence of zero-intensity points that occur towards the edges of the beams.

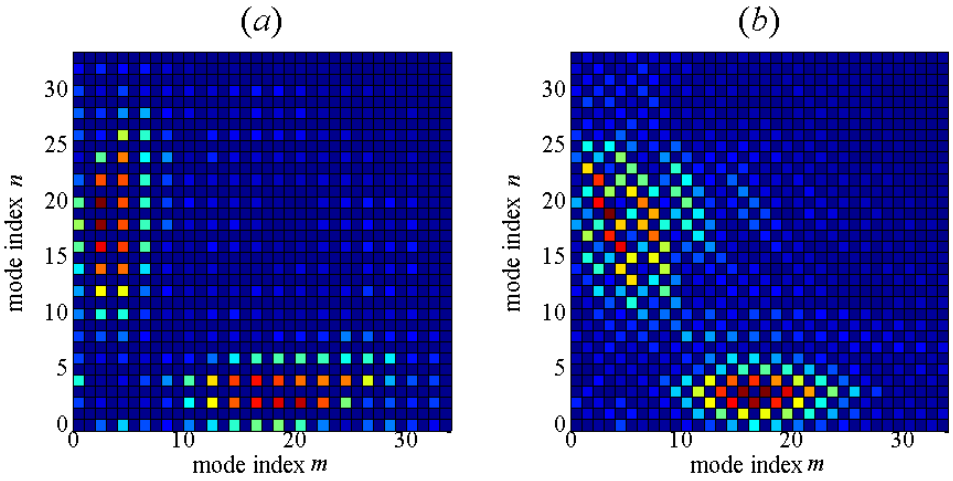


Figure 5-33. Mode coefficient amplitudes for (*a*) GBM-IPRA solution and (*b*) FFT-based IPRA. Amplitude coefficients in (*a*) are non-zero only at positions where indices *m* and *n* are even.

A Gaussian beam mode decomposition of the Fourier plane solution obtained from the FFT-IPRA was performed to compare both algorithms in terms of mode coefficients, the absolute values of which are shown in Figure 5-33. Since the FFT-based solution does not have the two-fold reflection symmetry of the GBM-based solution but rather central symmetry, an expanded mode-set that includes all (even- and odd-numbered) modes with indices in the range $m = n = 0$ to 34. Not surprisingly a plot of the mode coefficient $|a_{mn}|$ still has a checkerboard appearance since only half of all modes contain power. The absence of power in the remaining 612 modes is due to the central reflection symmetry observed in the grating and far field plane wavefronts. Thus modes for which

mode indices are either both even or both odd would in general be expected to contain some power. Apart from this difference the distribution of power between mode coefficients is similar for both solutions.

Tracking algorithm progress (speed to obtain a solution)

Next we now compare the progress of the GBM-based algorithm with that of the FFT-based algorithm. When performing iterative phase retrieval, algorithm progress is ascertained by examining the quality of the current solution (at a given iteration). Typically, in phase retrieval literature, the metric of choice for determining solution quality is the RMS-error between the target and estimated intensities. In our implementation intensity correlation (between the target and solution intensities) was used as the quality metric and it was evaluated at both the grating and far field planes.

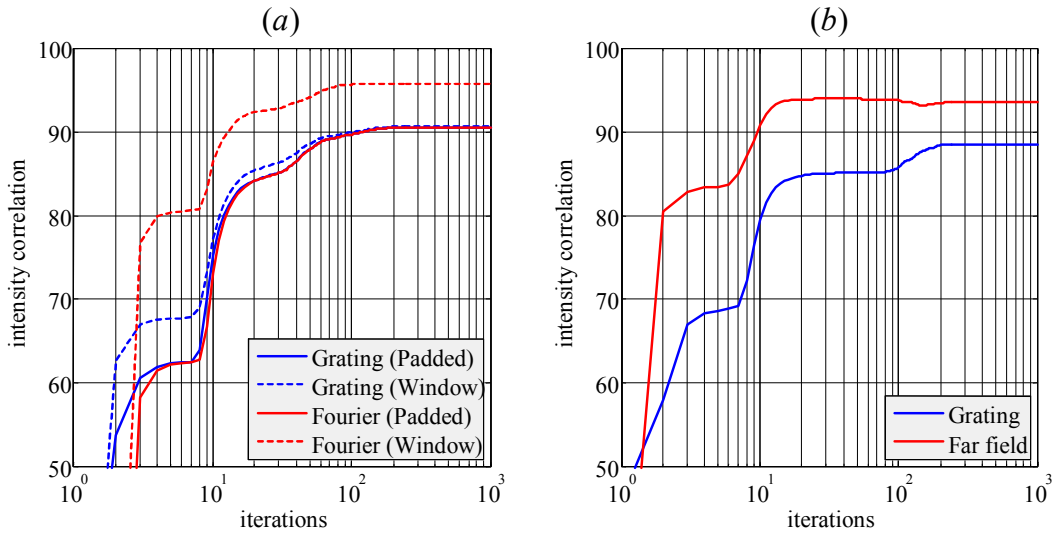


Figure 5-34. Algorithm progress in terms of intensity correlation between target and estimated intensity distributions at the grating and far field (Fourier) planes for trial runs consisting of 1000 iterations of (*a*) the FFT-based IPRA and (*b*) the GBM-based IPRA. Both algorithms were started with the same initial far field phase distribution.

Figure 5-34 shows the progress of the two algorithms in terms of intensity correlations (between the target and phase retrieval solution intensities) made at the two planes. The final solution found by GBM-IPRA is not quite as good (indicated by a lower final intensity correlation value, 93% compared to 96%) as that achieved by the FFT-IPRA, presumably because of the modest number of modes used in the algorithm. Referring to Figure 5-34(*a*), during iterations with the FFT-IPRA the intensity correlation was evaluated at two regions in the grating and Fourier planes. The solid lines correspond to

values obtained by comparing target and estimated intensities over the entire grating and Fourier planes, respectively. The dashed lines correspond to values obtained by comparing intensities within the signal windows only. As the algorithm proceeds the correlation level with the target pattern monotonically increases or remains constant (at which point the algorithm stagnates) as expected [5.19]. The difference in values for the correlations over just the signal windows occur because when the merit function is evaluated inside the signal window any power diffracted outside this field of view is not accounted for. It is useful to note that the correlation values evaluated at the grating signal window provide a good indicator of algorithm progress.

This last point is important for assessing the progress of the GBM-IPRA algorithm. Because of limited computational resources the mode set used in GBM-IPRA is optimised such that the modes chosen fit neatly inside the regions of interest at the grating and image planes. Thus our field of view at both planes is restricted to regions equivalent to the signal windows used in the FFT-IPRA. The intensity correlation values in Figure 5-34(*b*) obtained with the GBM-IPRA are in fact equivalent to the values within the signal windows in Figure 5-34(*a*) obtained with the FFT-IPRA. From the above discussion the values of intensity correlation evaluated for the signal window at the grating plane is a more accurate indicator of algorithm progress. In the case of the GBM-IPRA solution the far field intensity correlation contains a local minimum somewhere between the 100^{th} and the 200^{th} iteration, which may be due to our limited field of view.

Notice that the intensity correlation plots in Figure 5-34(*b*) reach three plateaus during the course of iterations. These plateaus correspond to local maxima in the solution space and although the algorithm can sometimes proceed to a better solution it usually takes a large number of iterations to get out of the local maximum and move towards an improved solution. In other words the algorithm is very effective at finding a local maximum solution, and does so quickly (usually within a few hundred iterations). However once the algorithm finds one of the possibly many local solutions it can get stuck, or stagnate, at this local solution and may not be able to reach a better solution through continued iteration. The source of and solution to stagnation in iterative phase retrieval algorithms is the subject of much research in the field of phase retrieval [5.35].

Phase Unwrapping

Phase is extracted from a complex-valued wavefront $E(x,y)$ using the four-quadrant arc tangent operator as follows

$$\phi(x, y) = \arctan\{Im\{E(x, y)\}, Re\{E(x, y)\}\} \tag{5.21}$$

so the phase values are *wrapped* within the interval $(-\pi,+\pi]$. Therefore although the true phase $\Phi(x)$ of an electromagnetic signal may span many multiples of π one can only access the wrapped phase $\Phi_w(x)$. The wrapping operation that produces the wrapped phase $\Phi_w(x)$ is defined as

$$\Phi_w(x) = \Phi(x) + 2\pi k(x) \tag{5.22}$$

where $k(x)$ is an integer-valued function that forces $-\pi < \Phi_w(x) \leq \pi$. The result of the wrapping process is the presence of discontinuities in the wrapped phase map called phase wraps. Phase unwrapping is defined as any procedure that obtains an estimate $\phi(x)$ of the true phase $\Phi(x)$ from the wrapped phase values $\Phi_w(x)$. In other words the wrapped values of $\Phi_w(x)$ must be unwrapped to obtain an estimate $\phi(x)$. In applications where the wrapped phase $\Phi_w(x)$ is extracted from a measured signal that is subject to noise the phase unwrapping process seeks to recover the true phase $\Phi(x)$ from the noisy principal value $\Phi_w(x)$ defined as

$$\Phi_w(x) = \Phi(x) + n(x) + 2\pi k(x) \tag{5.23}$$

where $n(x)$ represents the noise in measurements. However, since we are concerned with unwrapping synthesised phase here, which is noise-free, term $n(x)$ can be omitted.

Phase unwrapping is a process typically used in fields where measured phase corresponds to some physical quantity of interest such as terrain elevation (in interoferometric Synthetic Aperture Radar), potential map, temperature, stress, wavefront distortion in adaptive optics, etc. In such applications the principal noisy phase values obtained by measurements are wrapped into the range $(-\pi, \pi]$ and must be unwrapped to obtain a true phase map corresponding to that physical data.

In the context of phase grating design the wrapped phase solution obtained by phase retrieval corresponds to the surface height $h(x)$ of a transmission or reflection grating. Figure 5-35(*a*) shows the wrapped grating phase $\phi(x, y)$ found by the GBM-IPRA. Because $\phi(x, y)$ is restricted to values in the interval $[-\pi, +\pi)$ it contains a substantial number of phase wraps at points where $\phi(x, y)$ makes vertical jumps between $+\pi$ and $-\pi$. The x and y cuts in Figure 5-35(*b*-*c*) taken through the centre of the phase

map at $(x, y) = (0, 0)$ show approximately fourteen phase wraps in each direction, with a mean separation between discontinuities of approximately 20 mm. If the wrapped phase $\phi(x, y)$ were translated directly into a surface height $h(x, y)$ the limited precision afforded by the milling process would make accurate machining of regions in the vicinity of phase wraps extremely difficult and would result in a less than ideal finished grating surface. Furthermore it is our intention to produce a reflection grating for off-axis illumination. However if a reflection grating were made with a surface derived from the wrapped phase recessed regions near phase jumps would be shadowed by raised areas and so would not contribute to the phase modulation of the incident beam. To minimise errors due to limited machining accuracy and to avoid shadowing requires finding an equivalent smoother phase function with considerably fewer phase jumps. Thus our motivation for employing phase unwrapping techniques is not to obtain a 'true' phase, but rather to obtain a smoother equivalent phase function with significantly fewer phase wraps that is easier to manufacture and less sensitive to shadowing effects.

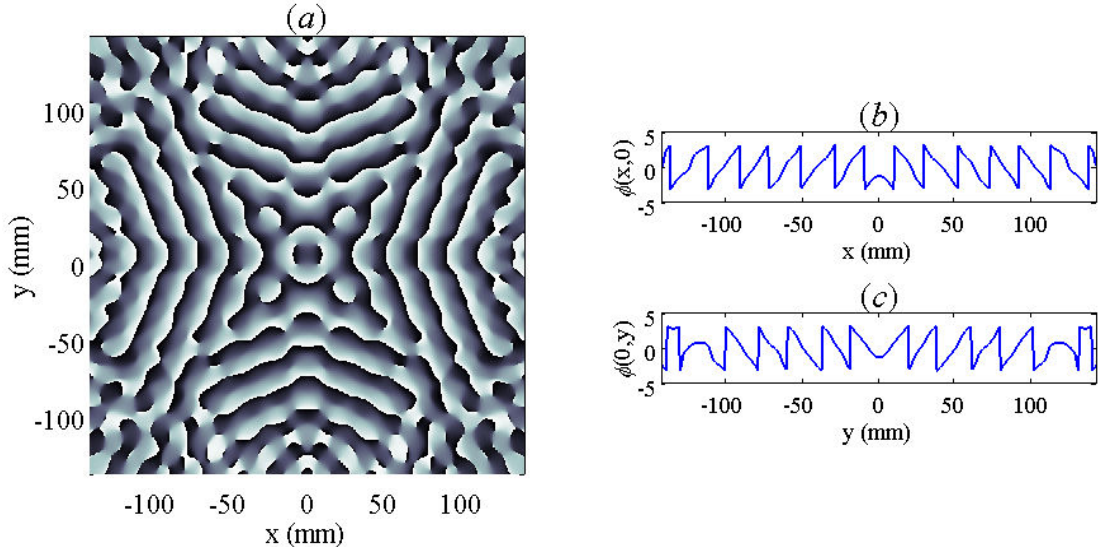


Figure 5-35. (*a*) The two-dimensional grating plane phase $\phi(x, y)$ found by the GBM-IPRA, (*b*) horizontal and (*c*) vertical cuts through the point $(x, y) = (0,0)$. There are approximately fourteen phase discontinuities in each cut.

The phase wrapping operation given by Eq. (5.22) implies that phase unwrapping involves detecting the positions of all 2π phase jumps in the wrapped phase $\Phi_w(x)$ and then adding appropriate $k(x)$ multiples of 2π at those points. We first consider the basic ideas of one-dimensional phase unwrapping to introduce the concepts involved.

First the points x_W at which phase wraps occur are identified by differentiating $\phi_w(x)$ with respect to x and locating impulses in the phase gradient $\nabla\phi_w(x)$ with values

close to $\pm 2\pi$. The unwrapped (true) phase is then recovered by adding $\pm 2\pi$ to $\nabla\phi_w(x)$ at points x_W before integrating the phase gradient. A one-dimensional wrapped phase $\phi_w(x)$, such as those shown in Figure 5-35(*b*-*c*) can be easily unwrapped using Itoh's method [5.36]. This technique is based on the notion that the unwrapped phase $\phi(x)$ can be obtained by integrating the phase gradient of the wrapped phase $\phi_w(x)$

$$\Phi(x) = \Phi(x_0) + \int \nabla\Phi_w(x) \cdot dx \tag{5.24}$$

where $\Phi(x_0) = \Phi_w(x_0)$ and $\nabla\Phi_w(x)$ is the gradient of the wrapped phase $\Phi_w(x)$. Since a one-dimensional line integral can only follow one path, unwrapping one-dimensional phase data is straightforward and well defined. Two-dimensional phase unwrapping is based on the same idea, but with the phase gradient integrated along closed paths.

The one-dimensional problem is extended to two-dimensions as follows. Assuming we know the phase and its gradient at an initial point r_0, then the phase at point r is obtained from the path integral

$$\Phi(r) = \int_C \nabla\Phi \cdot dr + \Phi(r_0) \tag{5.25}$$

where C is *any* path in a domain D connecting points r and r_0 and $\nabla\Phi$ is the gradient of phase $\Phi(r)$. Indeed all two-dimensional phase unwrapping involves integrating phase gradients. If integration is independent of the path C then any simple phase unwrapping technique can be applied to two-dimensional phase unwrapping. Thus Itoh's one-dimensional method can be applied in a column-by-column (vertical) or row-by-row (horizontal) fashion and both would give the same result.

Simple 2-D phase unwrapping can fail for a number of reasons because of problematic regions within the measured phase map. Such regions of phase have undesirable characteristics such as due to low signal-to-noise ratio, areas of low signal level (where the phase becomes random), and under sampled regions. In such turbulent regions the presence of residues means that integration becomes highly path-dependent. Fortunately since our phase data is synthesized and not extracted from a measured signal it does not suffer from problems encountered when attempting to unwrap measured phase data, such as noise, height-induced layover, shadows, low signal power, etc. In fact the only reason why simple phase unwrapping techniques might fail to unambiguously unwrap the phase data is because of the existence of phase singularities. Unfortunately though the phase solution returned by the GBM-IPRA contains many phase singularities

Consider the wrapped phase at the grating plane in Figure 5-36. Simple 1-D phase unwrapping is now applied to unwrap the phase within the two 60mm×60mm region inside the superimposed green and red squares.

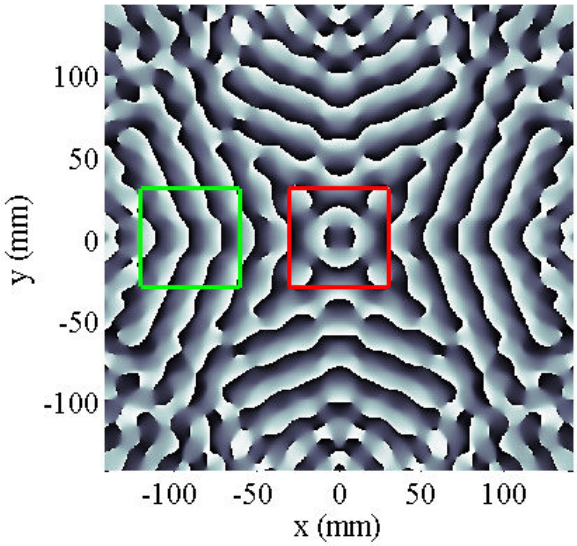


Figure 5-36. The grating plane phase data of the solution found by the GBM-IPRA. One-dimensional phase unwrapping (Itoh's method) will be applied to the square regions inside the red and green frames.

First the phase within the green square is unwrapped. The wrapped phase in Figure 5-37(*a*) is unwrapped one-dimensionally, first column-by-column in Figure 5-37(*b*) and then row-by-row Figure 5-37(*c*) with identical results: all phase wraps have been successfully removed. In other words integration is independent of the path taken so the phase has been unambiguously unwrapped.

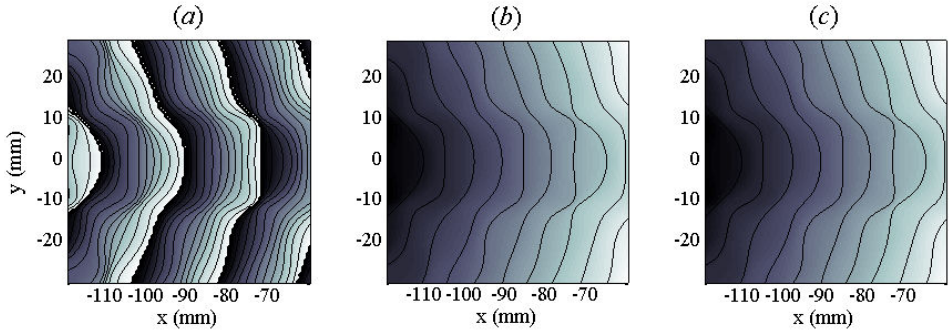


Figure 5-37. The wrapped phase in (*a*) possesses four continuous (unbroken) vertically aligned phase wraps that all end on edges of the frame. If we unwrap (one-dimensionally) each column the resulting unwrapped phase in (*b*) is identical to that in (*c*) which was produced by unwrapping each row one-dimensionally. Thus the phase has been unambiguously unwrapped.

Next we attempt to unwrap the phase within the red square of Figure 5-36. Figure 5-38(*a*) shows the wrapped phase and the result of unwrapping first vertically and then

horizontally are shown in Figure 5-38(*b*-*c*), yielding different results. In other words the different integration paths taken produce inconsistent results, i.e. integration is path-dependent. More importantly – from the point of view of phase grating manufacture – the phase wraps have not been removed but merely replaced by either horizontal or vertical phase wraps.

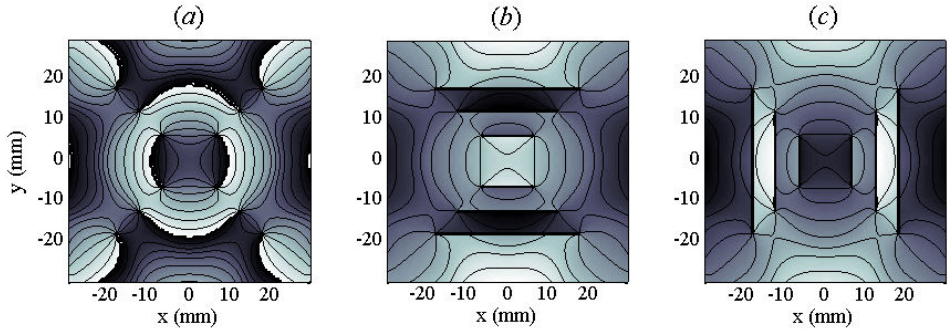


Figure 5-38. The wrapped phase in (*a*) contains 8 variously oriented discontinuities (phase wraps) that are all broken, i.e. they start (or end) at individual points. In this case the result of unwrapping (*b*) column-by-column is different when done (*c*) row-by-row. The original phase wraps have replaced by vertical, or horizontal, phase wraps, which begin or end on the same set of points each time.

In general, the value of a line integral in the form of Eq. (5.25) is dependent on the path *C*. Path-independence can be determined by evaluating any one of four equivalent conditions. Typically the only condition evaluated to detect path dependence in two-dimensional arrays is

$$\int_C \nabla\Phi(r) \cdot dr = 0 \qquad (5.26)$$

If the integrated phase gradient around a closed path equals zero the evaluation of Eq. (5.25) is independent of path taken and phase unwrapping is a trivial process. In general, however two-dimensional problems violate this condition so integration is path dependent. The task of phase unwrapping is then to find an appropriate integration path.

Residues: the source of path dependence

Referring to Figure 5-38 where one-dimensional phase-unwrapping failed to unambiguously unwrap the phase, in all three phase maps (*a*)–(*c*) the phase wraps begin (or end) at the same set of isolated points. It is from these points that error accumulates in the phase unwrapping process and cause the inconsistent results observed when phase unwrapping is evaluated using different integration paths. These isolated points are called discontinuity sources (because the phase discontinuities, or phase wraps, begin

and end at these points) or, more commonly residues. The result of integrating around a small closed loop that encircles a single residue is nonzero and thus condition (5.26) is violated. In other words path-dependency in phase unwrapping is due to the existence of phase residues in the wrapped phase. For 2-D phase maps with *simple* topography (i.e. containing only edge dislocations) Itoh's one-dimensional method can simply be extended to two dimensions, otherwise more sophisticated means must be employed.

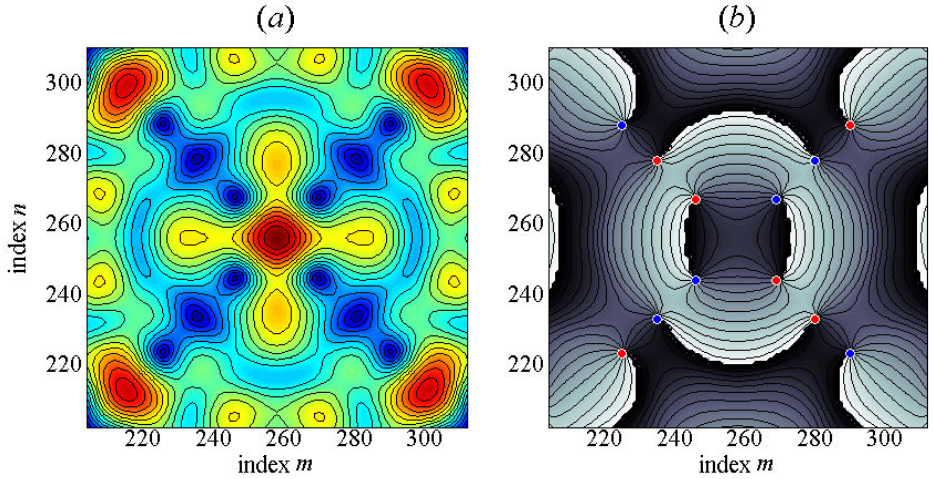


Figure 5-39. (*a*) Amplitude and (*b*) phase in the problematic region enclosed by the red box in Figure 5-36. In the phase map (*b*) maximum and minimum values (+π and −π) are shown as white and black contours, respectively.

Figure 5-39 shows the amplitude and phase in the problematic region of Figure 5-36. All discontinuities in the phase begin (or end) at single isolated points called residues, which are indicated with circular markers in Figure 5-39(*b*). The amplitude in the vicinity of each residue is zero, as indicated by dark blue (low-level) contour lines in Figure 5-39(*a*), i.e. residues occur at zeros in the complex field $E(x, y)$ because at these points the phase is undefined. Note also that the phase accumulated as one encircles counter-clockwise a single residue is always ±2π radians. Although in Figure 5-39 each residue is the source of just a single phase discontinuity, it is possible (though extremely unlikely) that residues can produce multiple phase wraps, in which case the accumulated phase is an integer multiple of ±2π radians. The integer multiple of 2π is used to assign a 'charge' to a particular residue. Since it is extremely unlikely that residues encountered will accumulate anything other than ±2π radians, typically residues are assigned charges of ±1 only. Thus we refer to a positive (negative) residue as one that accumulates +(−)2π phase as it is encircled anti-clockwise. The polarity of charged residues are indicated by red (positive) and blue (negative) circular markers in

Figure 5-39(*b*). The result of integrating counter-clockwise about a closed loop containing a vector field $\mathbf{F}(r)$ with a single positive (negative) phase residue is $\pm 2\pi$

$$\int \mathbf{F}(r) \cdot dr = \pm 2\pi \tag{5.27}$$

Thus integration is reduced to evaluating the charge of the phase residue enclosed by the path.

Extending this concept, the evaluation of a large closed-path integral containing several residues reduces to summing small closed-path integrals about individual residues and therefore

$$\int \nabla\Phi(r) \cdot dr = 2\pi \times (\text{sum of enclosed residue charges}) \tag{5.28}$$

which is referred to as the residue theorem for phase unwrapping. Charges can be balanced by connecting pairs of oppositely charged residues with branch cuts and connecting any remaining isolated residues to edges of the phase map. By balancing residue charges all closed paths will enclose either an equal number of positive and negative residues, or none at all, and the line integral will always evaluate to zero. Thus the phase inside the closed-path can be unambiguously unwrapped as long as integration does not encircle unbalanced residue charges – a condition that is guaranteed by ensuring that unwrapping does not cross any branch cuts. The branch cuts, or unwrapping barriers, must be defined explicitly and the many ways to perform charge balancing and thus of choosing path selection are the subject of all path-following phase unwrapping methods.

Detecting Residues in 2-D Phase Arrays

To utilize residues in path-following phase unwrapping techniques we must first locate all of the residues in a given 2-D phase data array. To precisely locate a phase residue requires integrating the phase gradient $\nabla\Phi$ about the smallest possible closed-path

$$q = \int \nabla\Phi(r) \cdot dr \tag{5.29}$$

The only possible values for a residue charge, q is zero or $\pm 2\pi$ thus indicating the absence, or presence of either a positive or a negative residue within the closed path. When operating on discrete data, the smallest closed path is a 2×2 array of pixels, so a residues location can be determined to within each 4-pixel element. For an array of wrapped phase values $\Phi(m, n)$, where m and n are array indices, we sum the phase gradients between pairs of neighbouring pixels in the 4-pixel array, i.e.

$$q = \sum_{i=1}^{4} \nabla\Phi_i \tag{5.30}$$

The discrete phase gradients $\nabla\Phi_i$ are obtained by wrapping (into the interval $[-\pi, \pi)$) the differences of the wrapped phase values Φ_i as

$$\Delta\Phi_1 = wrap\{\Phi(m, n+1) - \Phi(m, n)\}$$
$$\Delta\Phi_2 = wrap\{\Phi(m+1, n+1) - \Phi(m, n+1)\}$$
$$\Delta\Phi_3 = wrap\{\Phi(m+1, n) - \Phi(m+1, n+1)\}$$
$$\Delta\Phi_4 = wrap\{\Phi(m, n) - \Phi(m+1, n)\}$$

where phase differences $\Delta\Phi_i$ are calculated by proceeding counter-clockwise about the 2×2 pixel array. Now associating the phase differences with the true phase ϕ such that

$$\Delta\phi_1 = \phi(m, n+1) - \phi(m, n)$$
$$\Delta\phi_2 = \phi(m+1, n+1) - \phi(m, n+1)$$
$$\Delta\phi_3 = \phi(m+1, n) - \phi(m+1, n+1)$$
$$\Delta\phi_4 = \phi(m, n) - \phi(m+1, n)$$

the residue charge is

$$q = \sum_{i=1}^{4} \Delta\phi_i \tag{5.31}$$

Recognising that the phase differences $\Delta\phi_i$ can be associated with partial derivatives of $\phi(m, n)$ in x and y as follows

$$\phi_y(m, n) \quad \Rightarrow \phi(m, n+1) - \phi(m, n) = \Delta\phi_1$$
$$\phi_x(m, n+1) \Rightarrow \phi(m+1, n+1) - \phi(m, n+1) = \Delta\phi_2$$
$$\phi_y(m+1, n) \Rightarrow \phi(m+1, n+1) - \phi(m+1, n) = -\Delta\phi_3$$
$$\phi_x(m, n) \quad \Rightarrow \phi(m+1, n) - \phi(m, n) = -\Delta\phi_4$$

where subscripts x and y refer to partial differentiation in those directions. The residue charge is then

$$q = \sum_{i=1}^{4} \Delta\phi_i = [\phi_x(m, n+1) - \phi_x(m, n)] - [\phi_y(m+1, n) - \phi_y(m, n)] \tag{5.32}$$

Making the following associations

$$\frac{\partial^2\phi}{\partial x \partial y} \Rightarrow \phi_{xy}(m, n) = \phi_x(m, n+1) - \phi_x(m, n)$$

$$\frac{\partial^2\phi}{\partial y \partial x} \Rightarrow \phi_{yx}(m, n) = \phi_y(m+1, n) - \phi_y(m, n)$$

we see that

$$q = \sum_{i=1}^{4} \Delta\phi_i = \frac{\partial^2 \phi}{\partial x \partial y} - \frac{\partial^2 \phi}{\partial y \partial x} \tag{5.33}$$

In other words the four-pixel element contains a single residue only if the cross derivatives are not equal. A residue is defined by four pixels and the convention adopted is that the upper left pixel of the 2×2 array is marked as the residue. Thus if a residue is located at pixel (i, j) it is implied that all four pixels (i, j), $(i+1, j)$, $(i, j+1)$ and $(i+1, j+1)$ define the residue. Identifying all residues in a phase map involves searching all 2×2 pixel elements in the 2-D array of phase values for residues. Around each loop of four pixels the closed-path integral of phase gradients (sum of wrapped phase differences) is evaluated. When a residue is located its position is recorded in a "residue map", with a value of +1 or −1 to indicate its polarity. The positions of the residues associated with the wrapped phase of the phase grating were identified by creating a residue map. Figure 5-40 shows the amplitude and phase at the grating plane with the positions of residues indicated by red and blue circular markers to indicate positively and negatively charged residues, respectively.

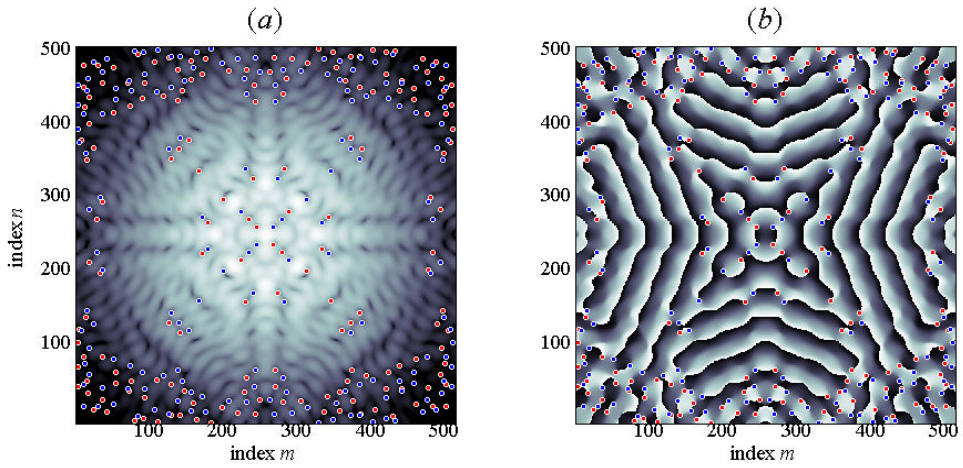


Figure 5-40. Grating plane (*a*) amplitude (in log-scale with colour axis scaled to a lower limit of –50dB) and (*b*) wrapped phase. The blue and red markers indicate the locations of positive and negative residues, as extracted from the residue map.

In most situations only the phase of the measured signal is available, however in the case where the user has access to both the amplitude and phase this extra information can be used to guide the simple one-dimensional phase unwrapping by creating a binary mask, with zeroes in regions of negligible field magnitude and hence irrelevant phase data and with ones in regions of relatively high intensity. For example

the intensity in the four corners of the grating plane (outside a disc of radius equal to twice the Gaussian beam radius W_G) is negligible. Even if the phase distribution in these low-power regions varies rapidly and contains undesirable features (screw, mixed edge-screw dislocations), since the field is very weak here these phase features have little influence on the incident beam. Thus the phase in these regions can be masked out and the remaining phase within the bulk of the Gaussian beam unwrapped. In Figure 5-40 there are 166 positive and 166 negative residues in total. Statistically the number of positive and negative residues should be equal. In practise this is usually not the case since some residues may lie outside the frame of observation. For this example however because of the symmetry of the phase, the number of positive and negative residues is in fact equal. Notice that approximately half of the residues (84 positive and 84 negative) occur outside the main bulk of the Gaussian beam (beyond a radius of $2W_G$), in low-intensity regions. The remainder (82 positive and 82 negative) residues that occur within regions of relatively high intensity mean that masking out phase in low intensity regions (outside a disc of radius $2W_G$) will not remove the path dependence but it does help to reduce computational costs by ignoring phase of little significance.

Local and global phase unwrapping

All phase unwrapping procedures can be classed as either local or global techniques [5.37]. Both involve integrating phase gradients in some way. Path-following algorithms solve the phase unwrapping problem by approximately, or exactly, minimising discontinuities in the unwrapped phase. The common mechanism by which these algorithms unwrap phase is by generating integration paths by means of localised pixel-by-pixel operations and thus are referred to as local phase unwrapping techniques. The other class of phase unwrapping methods are path-independent methods that employ mathematical techniques to find an unwrapped phase that fits the wrapped phase on a global, rather than pixel-by-pixel, scale. They assumes that true phase gradients $\nabla\phi(\mathbf{r})$ are corrupted by noise $n(\mathbf{r})$ and therefore measured phase gradients are given by

$$\nabla\Phi(\mathbf{r}) = \nabla\phi(\mathbf{r}) + n(\mathbf{r}) \tag{5.34}$$

Global algorithms then proceed by minimising the squared error ε^2 of the phase gradients in a least squares (LS) sense, i.e.

$$\varepsilon^2 = \int (\nabla\phi(\mathbf{r}) - \nabla\Phi(\mathbf{r}))^2 \, dA \tag{5.35}$$

where dA is an element of the region A.

Global phase unwrapping

We first examine a global phase unwrapping algorithm that is based on the physical interpretation of least-squares (LS) phase unwrapping [5.38]. A phase map that contains singular points (residues) produces rotational phase-gradient fields, which are responsible for the path integration of the phase gradient vector being strongly path-dependent. In this case the phase gradient vector, $\boldsymbol{S}$ is expressed as the sum of a scalar potential ϕ (phase) and a vector potential $\boldsymbol{A}$ (generated by the singular points), as

$$\boldsymbol{S} = \nabla\phi + \nabla\times\boldsymbol{A}$$

In other words the task of phase unwrapping is how to extract the scalar potential ϕ (the phase) from the mixed potentials of $\boldsymbol{S}$. The essence of least squares phase unwrapping is to extract the irrotational phase gradient vector components from the rotational components and determine the phase (scalar potential) that best describes the vortex-free phase gradient vector fields in a least square sense. The assumption made is that the true phase ϕ has no singular points and that the rotational phase gradient fields due to the singular points are noise-induced artefacts. Therefore by correcting for noise-corrupted phase components, the true phase can be unwrapped using any simple phase unwrapping technique (such as a 2-D version of Itoh's method).

Given a vortex-infested complex field $E(r)$ containing residues it can be decomposed into two components

$$E(r) = E_V(r) \cdot E_F(r)$$

a rotational term containing all of the singular points: the residues or vortices, and is therefore referred to as the vortex-only field, $E_V(r)$ and an irrotational term containing only the scalar potential, which is thus referred to as the vortex-free field, $E_F(r)$. The irrotational phase gradients are due to the irrotational field so the phase ϕ is that extracted from $E_F(r)$. Thus solving for the vortex-free field yields

$$E_F(r) = E(r) \cdot E_V^*(r) = E(r) \cdot E_A(r)$$

where $E_A(r) = E_V^*(r)$ is the complex conjugate of E_V and is referred to as the vortex-annihilating field. It is exactly the same as $E_V(r)$ but with the topological charges of each vortex, or residue, reversed. When the original vortex-infested field $E(r)$ is multiplied by $E_A(r)$ each of the original vortices is accompanied by a neighbouring vortex of opposite charge in its immediate vicinity. The rotational phase-gradient fields associated with each cancel and the bipolar vortex pair is annihilated.

Since the turbulent regions are due to residues, or vortices, with charges of ±1 only, if all of the residues within the phase are located we can then create a field containing phase distribution due to all of the residues, or vortices, therefore the vortex-only field. The vortex-only field $E_V(r)$ is created by adding a spiral phase term

$$\phi_S(x, y) = (m)\arctan\left(\frac{y - y_v}{x - x_v}\right)$$

to the phase $\phi_V(r)$ of the vortex-only field $E_V(r)$ for every residue found in $E(r)$. The topological charge, m (= ±1) associated with a new residue is set opposite to the charge of its counterpart in $E(r)$.

Figure 5-41 shows the result of applying vortex-annihilation to the wrapped grating phase. The result, shown in Figure 5-41(*b*), is a phase solution that can be unwrapped by standard 1-D technique, which yields the 2-D unwrapped phase shown in Figure 5-41(*c*) with virtually no discontinuities. Notice that in general the phase gradient in Figure 5-41(*b*) is under-estimated, as indicated by fewer phase wraps across the plane, compared to the original wrapped phase (Figure 5-40). This is a common feature of unweighted LS phase unwrapping algorithms [5.39].

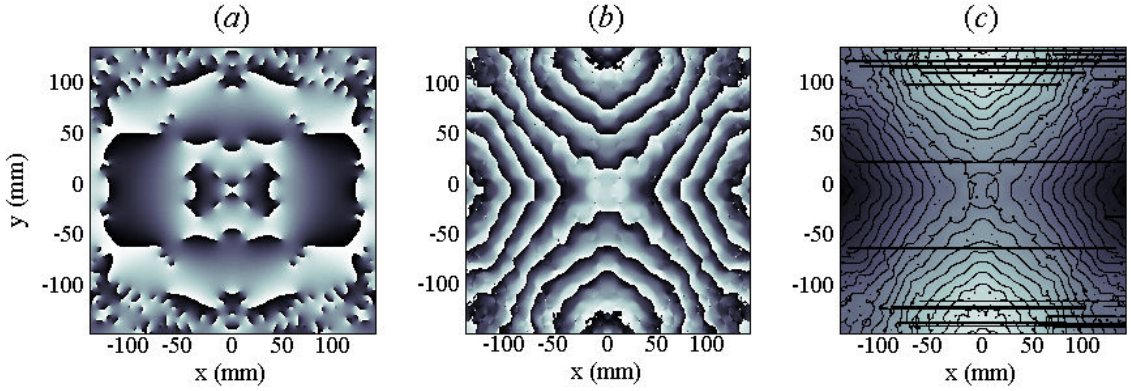


Figure 5-41. Phase unwrapping by direct elimination of vortex fields. (*a*) The rotational, phase Φ_V (*a*) of the vortex-only field E_V (in Figure 5-40) is subtracted from the wrapped field E (not shown) and results in the vortex-free field E_F, whose phase Φ_F in (*b*) can then be unwrapped using a 2-D version of Itoh's 1-D phase unwrapping method to produce the unwrapped phase map in (*c*) which is free from phase discontinuities except for a few horizontal streaks, due to slightly incorrect placement of a few residues in the vortex-only field E_V. These streaks can be removed by including an iterative optimisation step to refine residue locations.

The problem with the vortex-annihilation technique is that the model assumes that the only source of degradation in the measured phase gradients is due to noise. Aoki has shown [5.38] that while global methods can be applied successfully to unwrap noise-degraded phase maps, when used on phase maps in which residues are due to a mixture

of noise and object-intrinsic singular points the phase is unwrapped incorrectly. Thus while the Gaussian-illuminated wrapped phase produces the circular array of eight Gaussian beams in the far field shown in Figure 5-42(*a*), the unwrapped phase does not, as seen in Figure 5-42(*b*). Clearly then the residues in the wrapped phase are an integral part of the solution needed to generate the correct far field intensity. That is not to say that all solutions to this phase retrieval problem require residues and indeed many other solutions may exist that do not require the rotational phase gradients used by this solution.

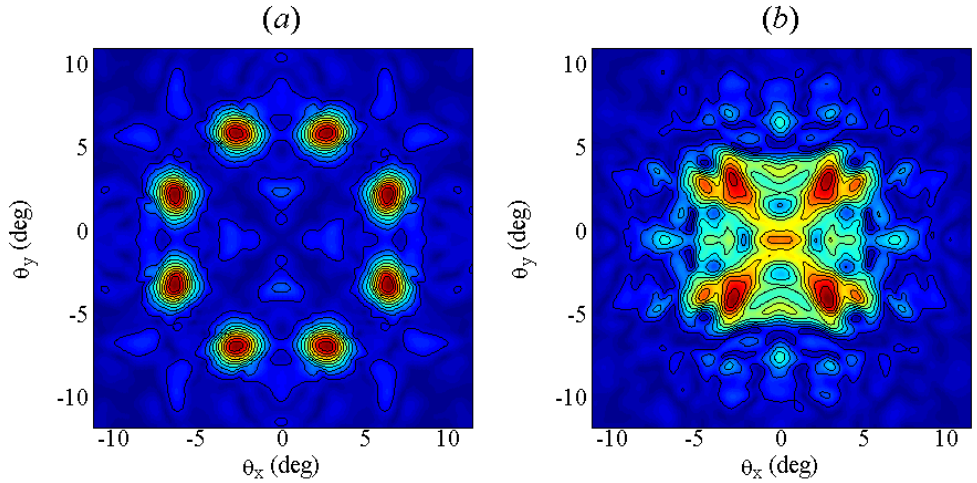


Figure 5-42. The circular array of eight diffraction orders at the Fourier plane is produced in (*a*) by the original vortex-infested field $E(r)$ but not in (*b*) by the vortex-free field $E_F(r)$ produced by direct elimination of the residues.

Path-following techniques

The rest of our discussion on phase unwrapping is restricted to various path-following (local) phase unwrapping techniques. If a phase image contains residues these must be located and balanced with a set of branch cuts so that any closed path always encloses an equal number of positive and negative residues (or no residues at all). Once branch cuts are in place the phase can be unwrapped along any path that does not cross a branch cut. Thus phase unwrapping is reduced to choosing a good set of branch cuts.

The choice of branch cuts is not so obvious and some criteria must be established to guide their placement. It is not enough to simply require that all residues be balanced. For example, if branch cuts cross over and in doing so completely isolate portions of the phase there is no way to relate the unwrapped phase in the isolated regions to the rest of the phase. A natural criterion for choosing branch cuts is that they be as short as possible to avoid crossover. While this criterion may, in some instances,

be statistically the best strategy it would be impractical to examine all possible choices in order to find the shortest branch cuts. There are $n!$ ways to pair n positive residues with n negative residues and even more (about $2n^2$) possibilities if arrangements that contain more than one pair of positive and negative residues are considered.

Four path-following techniques were investigated for unwrapping the phase for the phase grating to produce the sparse 8-beam circular array of far field Gaussian beams. The classic path-following approach by Goldstein *et al* [5.40] is a fast and effective method that involves identifying residues and balancing them with connecting branch cuts. A completely different approach that does not generate branch cuts or even identify residues relies on a quality map of the phase data to guide the integration path. Another method merges these two methods into a hybrid technique and uses a quality map to guide the placement of branch cuts. These fourth algorithms that were tested involve minimising the discontinuities in the unwrapped surface.

Goldstein's Branch Cut Algorithm

Goldstein's algorithm is effective at generating optimal (short) branch cuts. The idea is to connect with branch cuts nearby residues in pairs (or multiple pairs) of residues of opposite polarity, called dipoles. While similar algorithms are restricted to connecting only pairs of residues, Goldstein's algorithm generates more general branch cuts that can join multiple dipoles and thus is representative of "branch-cut" algorithms in general. The algorithm consists of three steps

1) Identify all residues in the phase map

2) Generate branch cuts

3) Path-integrate (unwrap phase) around the branch cuts

Step 1 is straightforward and involves creating a residue map as explained previously. Step 2 generates the branch cuts and is the substance of Goldstein's algorithm. Branch cuts are selected solely on residue positions and therefore the phase is not needed for this step. First a residue is located in the residue map. Then a search begins for residues in a 3×3 pixel square centred on the first residue. If one is found a branch cut is placed between the new charge and the charge at the centre of the 3×3 box. If the two residues are oppositely charged the net charge is zero and the pair is labelled as "balanced". If the residues have the same polarity the search of the 3×3 box continues for another residue. Whenever a new unconnected residue is found its ±1 charge is added to the sum

of the polarities of the other connected residues. If, when the search of the 3×3 box is complete, the net charge is not zero the 3×3 search box is moved to each of the connected residues in turn and the search repeated. If at the end of this search the cumulative charge is still non-zero, the search area is increased to a 5×5 pixel box. This process continues until either the net charge equals zero or the search box reaches the image border, in which case a branch cut is connected to the image border. Connecting residues to the border serves to balance or discharge them since any path integral cannot then encircle them.

Step 3 uses a flood-fill algorithm to perform the path integration. The algorithm begins by selecting a starting pixel and storing its phase in a "solution" array. Its four neighbouring pixels are unwrapped (and added to the solution array) and their indices added to an "adjoin" list (used to store pixel indices adjoining unwrapped pixels). The algorithm proceeds iteratively by selecting (and removing) a pixel from the adjoin list, and unwrapping and inserting its neighbours in the adjoin list, all the while avoiding branch-cut and unwrapped pixels. When the adjoin list is empty all pixels have been unwrapped. If the image contains a number of isolated regions a pixel in the next region is selected as a starting pixel and the process is repeated. After all non-branch cut pixels have been unwrapped the branch cut pixels are then unwrapped (since, strictly speaking, branch cuts lie between pixels), which is done last to avoid unwrapping across branch cuts.

In cases where isolated regions occur (due to corrupt phase) the unwrapped phase may have an incorrect multiple of 2π. It is impossible to unambiguously unwrap the phase in isolated regions and therefore the success of a particular phase unwrapping algorithm must be judged on a qualitative basis. However because branch cuts prevent path integration from encircling unbalanced residues one is assured that such unwrapping errors are confined to isolated patches of corrupt phase and that elsewhere the surface is correctly unwrapped.

One way to enhance Goldstein's algorithm is to connect with branch cuts closely spaced residues of opposite polarity, called "dipoles" and remove them as a pre-processing step before applying Goldstein's algorithm. Dipoles are connected using a nearest-neighbour procedure but only adjoining residues are considered. Although connected dipoles are not considered by Goldstein's algorithm their branch cuts remain to prevent the path integration step from encircling the unbalanced residues.

Figure 5-43 shows the branch cuts and unwrapped phase for the entire 2-D phase map. The result of unwrapping the entire 2-D phase map is that the original symmetry of the phase is lost. Worse still, there are now 169 isolated regions due to overlapping branch cuts, which results in a 2π error in the unwrapped phase.

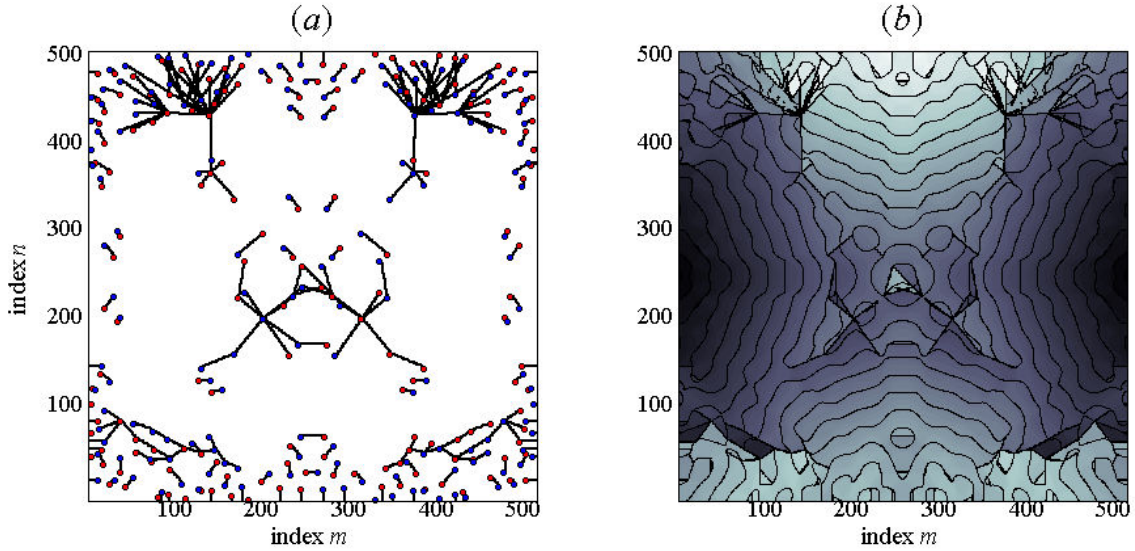


Figure 5-43. The branch-cuts in (*a*) generated by Goldstein's algorithm guide the integration path to produce the unwrapped phase in (*b*). Note that the dipole pre-processing step is not needed here, since no oppositely charged residue pairs were found to be in close enough proximity to each other. Because the branch cuts were generated for the entire residue map, the branch cut map and hence the unwrapped phase do not possess the four-fold symmetry that the wrapped phase had.

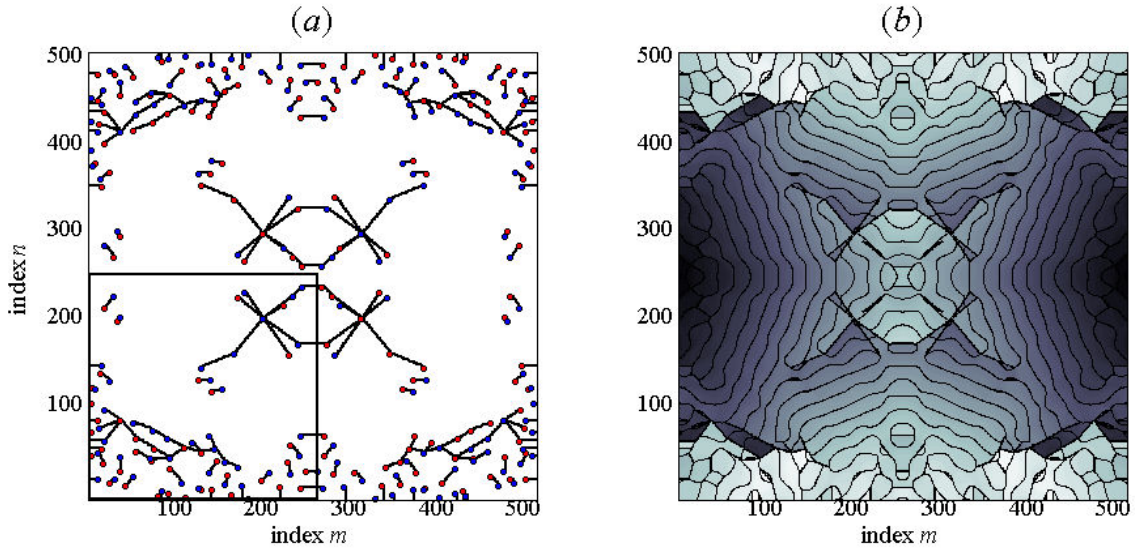


Figure 5-44. The result of unwrapping the third (lower-left) quadrant only with Goldstein's algorithm. Branch cuts are generated for residues in the third quadrant only and then the phase in this region unwrapped. The phase in the third quadrant is then reflected about the *x*- and *y*-axes to create the unwrapped phase of the remaining three quadrants. Now there are only two significant isolated regions near the centre of the image (above and below the x-axis).

Because the phase map exhibits two-fold reflection symmetry about the x- and y-axes it makes sense to unwrap the phase in a single quadrant only, e.g. the third (lower left) quadrant. This maintains the original symmetry, reduces computational overhead and, as seen in Figure 5-44, reduces significantly the number of isolated regions.

While Goldstein's algorithm is fast and generally satisfactory it can fail on some problems. The nearest-neighbour strategy used to place branch cuts minimises the lengths of branch cuts without regard to the quality of the image phase and therefore is not always the best approach. For example poorly placed branch cuts can isolate entire regions, resulting in an incorrect multiple of 2π in the unwrapped phase in those regions. Avoiding these problems requires exploitation of additional information from the phase data to guide the placement of branch cuts. The next two algorithms take this approach.

Quality-Guided Path Following

In the context of phase unwrapping a quality map is used to define the quality of each phase value in the 2-D phase image. Various quality maps can be derived from phase data including phase derivative variance, maximum phase gradients, and pseudocorrelation, as explained in [5.39].

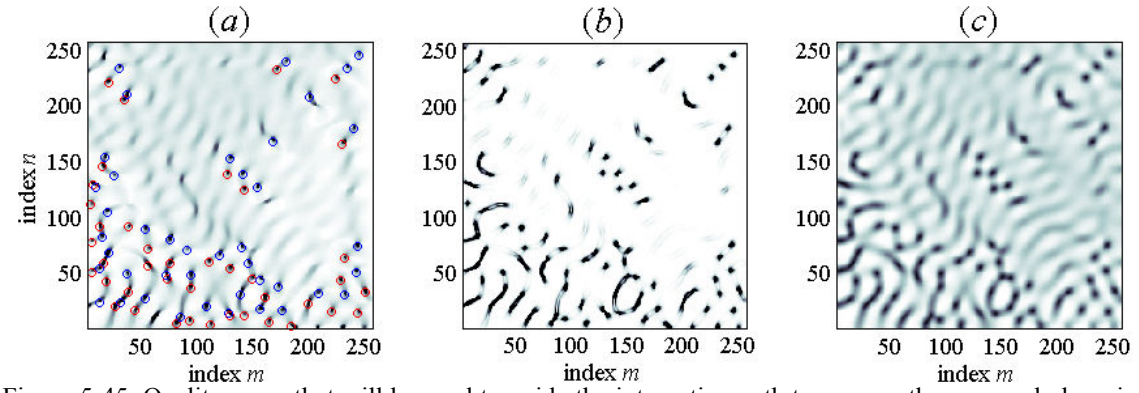


Figure 5-45. Quality maps that will be used to guide the integration path to unwrap the wrapped phase in the third quadrant. The quality maps are (*a*) minimum phase gradient, (*b*) minimum phase variance and (*c*) maximum pseudocorrelation. Low-quality pixels are shown as dark pixels and vice versa. The most significant thing to note is that residues, shown by the red and blue markers in (*a*), are located only in regions corresponding to low quality pixels.

Comparison of quality maps with the residue map for a particular phase image shows that regions of corrupted phase, where residues are located, tend to correspond to low-quality regions (Figure 5-45), which suggests that integration should follow high-quality pixels and avoid low-quality pixels. Quality-guided path following techniques do not

identify residues or generate branch cuts but depend solely on the assumption that a good quality map will successfully guide integration without encircling unbalanced residues.

The algorithm begins by selecting the highest-quality pixel. Its four neighbouring pixels are unwrapped and their indices stored in the "adjoin" list, which is maintained in order of quality values. The highest-quality pixel is removed from the list and its four neighbouring pixels unwrapped and added to the list (unless a neighbour has already been unwrapped). This process continues iteratively until all pixels have been unwrapped. The algorithm is a region-growing approach, whereby an unwrapped region grows beginning with the high-quality pixel and ending with lowest-quality pixel in that region. Thus the algorithm begins by confining unwrapping to regions of high-quality phase in between patches of corrupted phase, which are unwrapped last.

Effectively the algorithm is a modified version of the flood-fill procedure of Goldstein's algorithm (Step 3) the main difference being how the adjoin list is managed. Whereas the flood-fill procedure unwraps pixels in any order, in the quality-guided algorithm the adjoin list is sorted based on quality values. This adds significantly to execution time and necessitates a "list-trimming" procedure to keep the list size small. When the list size exceeds a predetermined number of entries, or bound (for an $m \times n$-pixel array a $m+n$ bound is used), the lowest-quality pixels are removed from the list and designated "postponed" for later consideration. The "minimum quality threshold" is set to the lowest-quality of the remaining pixels in the list and subsequently any pixels with quality values below this threshold are unwrapped and marked "postponed". When the adjoin list is empty the remaining pixels that adjoin the unwrapped pixels are below the minimum quality threshold. The threshold is therefore reduced and postponed pixels with quality values in excess of the new threshold are added to the list.

Figure 5-46 shows the result of unwrapping the third quadrant of the grating plane phase using the quality-guided algorithm implemented with the following quality maps:

- Maximum phase gradient
- Minimum variance of phase derivatives
- Maximum Pseudocorrelation

The unwrapped phase in Figure 5-46(*c*) contains quite a few disconnected regions, such as the region indicated by the red arrow. The unwrapped phase in Figure 5-46(*b*) has the

least amount of discontinuity in the region of significant amplitude and is therefore considered the best solution. The reason is due to the uniformity of the quality map in the regions of phase that can be easily unwrapped. The only significant discontinuity present in Figure 5-46(*b*) is a vertical phase wrap at the bottom border (indicated by the red arrow), which is due to an almost continuous line of low-quality pixels in the quality map that separate two regions across this boundary.

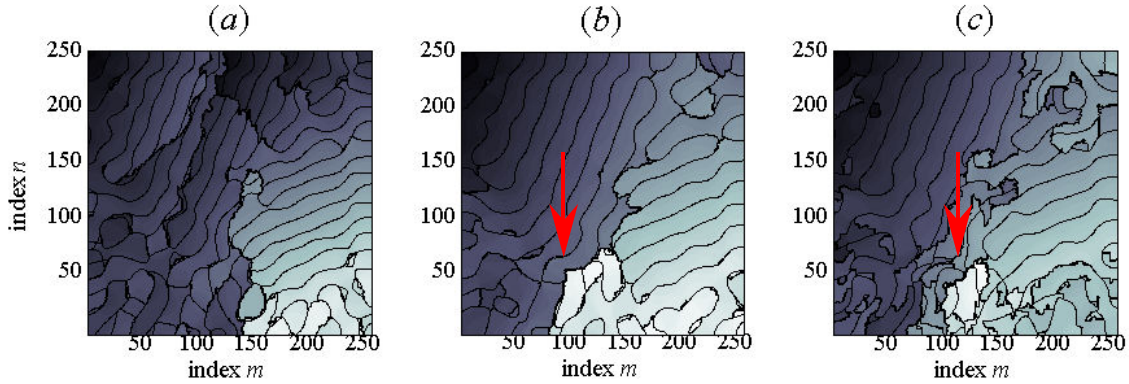


Figure 5-46. Results of applying quality-guided path-following phase unwrapping to the third quadrant of the wrapped phase, using the quality maps shown in Figure 5-45.

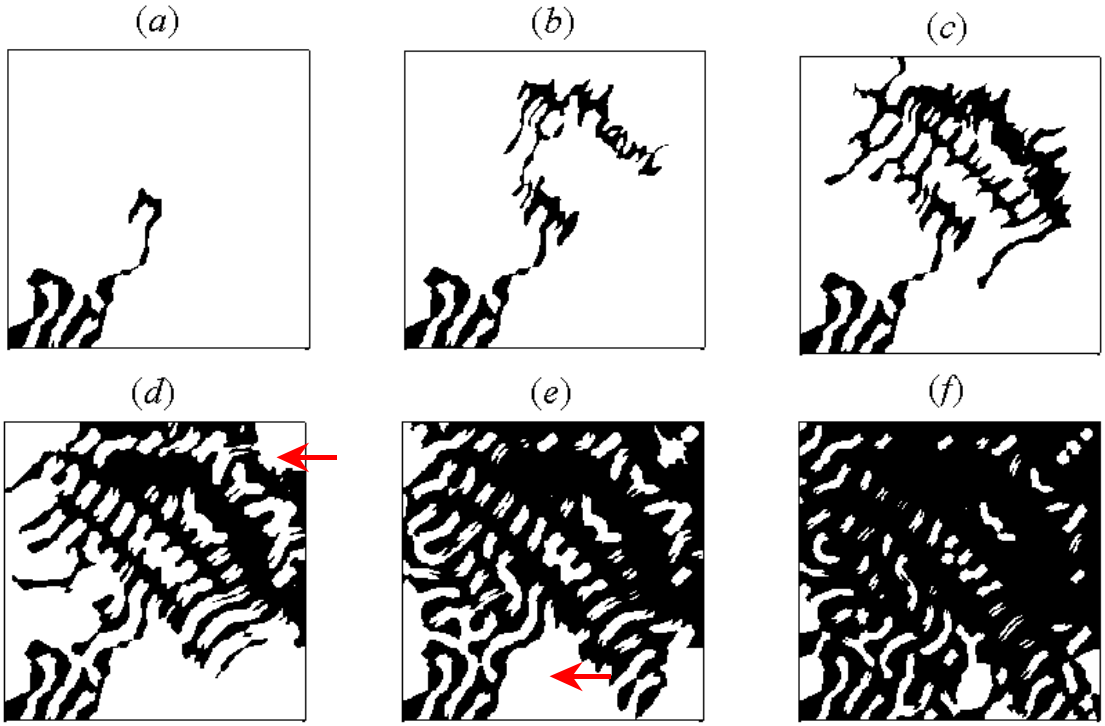


Figure 5-47. Each plot shows pixels that have been unwrapped at different stages during quality guided algorithm, with a quality map based on the minimum variance of the phase gradient in Figure 5-45(*b*). Black pixels correspond to pixels that have already been unwrapped and white pixels to those that have yet to be unwrapped. The six plots correspond to points in time where (*a*) 5%, (*b*) 10%, (*c*) 20%, (*d*) 40%, (*e*) 60% and (*f*) 80% of all pixels have been unwrapped. Unwrapping begins at the lower-left corner (*a*) and then follows high-quality pixels towards the centre of the image (*b*) where high-quality pixels are located. The unwrapped region grows steadily as the minimum quality threshold decreases.

Figure 5-47 shows how the quality-guided algorithm proceeds when path integration follows the quality map based on minimum variance of phase gradients. Notice in Figure 5-47(*e*) that a wall of low-quality pixels means that a large region of high-quality pixels in the lower part of the image (pointed to by the arrow) is unwrapped much later than the other high-quality pixels. This results in the phase discontinuity seen in the unwrapped phase of Figure 5-46(*b*). The same effect is also responsible for the discontinuity seen in the upper-right corner, as indicated in Figure 5-47(*d*) where another high-quality region (red arrow) unwrapped some time after the surrounding pixels have been unwrapped.

Because the guidance of path integration is dependent on the availability of a good quality map, without one the algorithm is useless. Excessive noise, for example, can corrupt a quality map and cause the integration path to wander back and forth across the image, in which case Goldstein's algorithm may provide superior results. When a reliably good quality map is available the algorithm often performs better than Goldstein's algorithm.

Mask Cut Algorithm

The quality-guided algorithm does not use information about residues to guide integration so there is no guarantee that unbalanced residues will not be encircled and incorrect multiple 2π errors introduced to the unwrapped phase. The mask cut algorithm combines the advantages of the quality-guided algorithm with those of Goldstein's algorithm to yield an algorithm that uses a quality map to guide branch cut placement.

The mask cut algorithm can be regarded as the reverse of the quality-guided algorithm. Instead of unwrapping regions of high-quality pixels first, the algorithm starts at a residue and "region grows" pixel masks through low-quality regions. The pixel masks perform the same role as branch cuts that connect residues and are therefore referred to as mask cuts. Mask growth terminates when its net charge is zero (when there are an equal number of positive and negative residues under the mask), or when it reaches an image border. Since masks are generated by a region-growing approach they tend to be thick and so must be thinned using a simple morphological operation (Figure 5-48). This mask-thinning process involves repeated steps through the image, each time removing pixels adjoining non-masked pixels (provided that their removal does not alter the masks connectivity) until no more pixels can be removed. After thinning, the phase around the mask cuts is unwrapped using the flood-fill procedure.

The mask cut algorithm consists of four steps

1) Identify residues

2) Generate mask cuts

3) Thin mask cuts

4) Path-integrate around mask cuts

Step 2 is the main substance of the mask cut algorithm and is essentially a combination of the quality-guided path follower and Goldstein's algorithm and as with the latter a "list trimming" procedure is essential. Since quality-guidance is used to place mask cuts, which are associated with regions of corrupt phase, the pixels retrieved from the adjoin list are now the lowest-, rather than highest-quality pixels.

When thinning mask cuts in Step 3, each pixel to be removed must be examined to ensure mask connectivity will not change upon its removal. If, during a walk around the 3×3 neighbourhood of the pixel under consideration, the number of transitions between mask and non-mask pixels is found to be greater than two then the centre pixel cannot be removed without disconnecting the mask. Despite thinning, mask cuts still tend to be thicker than the branch cuts of Goldstein's algorithm.

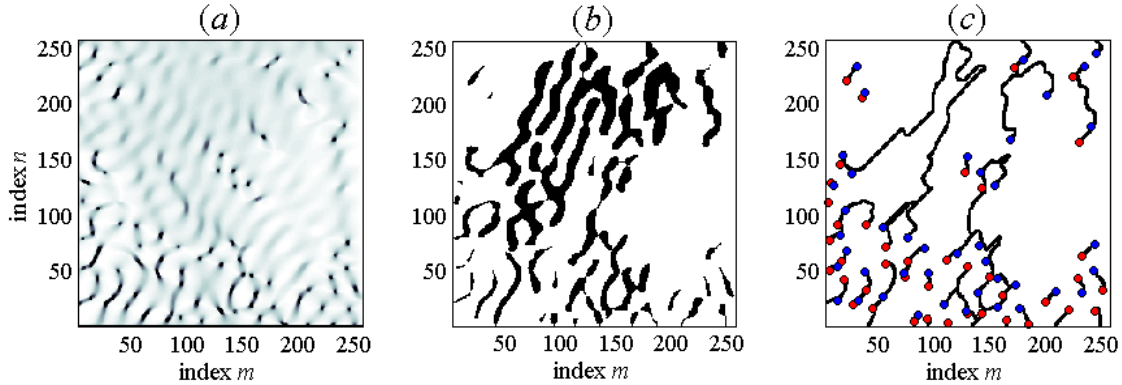


Figure 5-48. The quality map (*a*), derived from the minimum gradient, is used in collaboration with the residue positions to guide the placement of the mask cuts in (*b*) after which the mask cuts are thinned in (*c*). Notice that because each mask cut must be balanced (in terms of residual charge) the mask cuts tend to wander great distances in search of balancing charges. The result is that large regions of the phase image are isolated from each other thus producing phase discontinuities in the unwrapped phase.

Experimenting with different quality maps can provide varied results because two quality maps may place mask cuts in different regions, some better than others. Figure 5-49 shows the result of applying the mask-cut algorithm with three quality maps to unwrap the phase in the third quadrant of the wrapped grating phase. In all cases the results are inferior to the results obtained with the quality-guided algorithm. This is

because each isolated region of mask pixels is required to have a net residue charge of zero. The main source of this problem appears to be the group of 4 negative and 2 positive residues near the centre of the phase map, which result in the placement of mask cuts through the surrounding laminar phase, thus leading to unnecessary phase discontinuities.

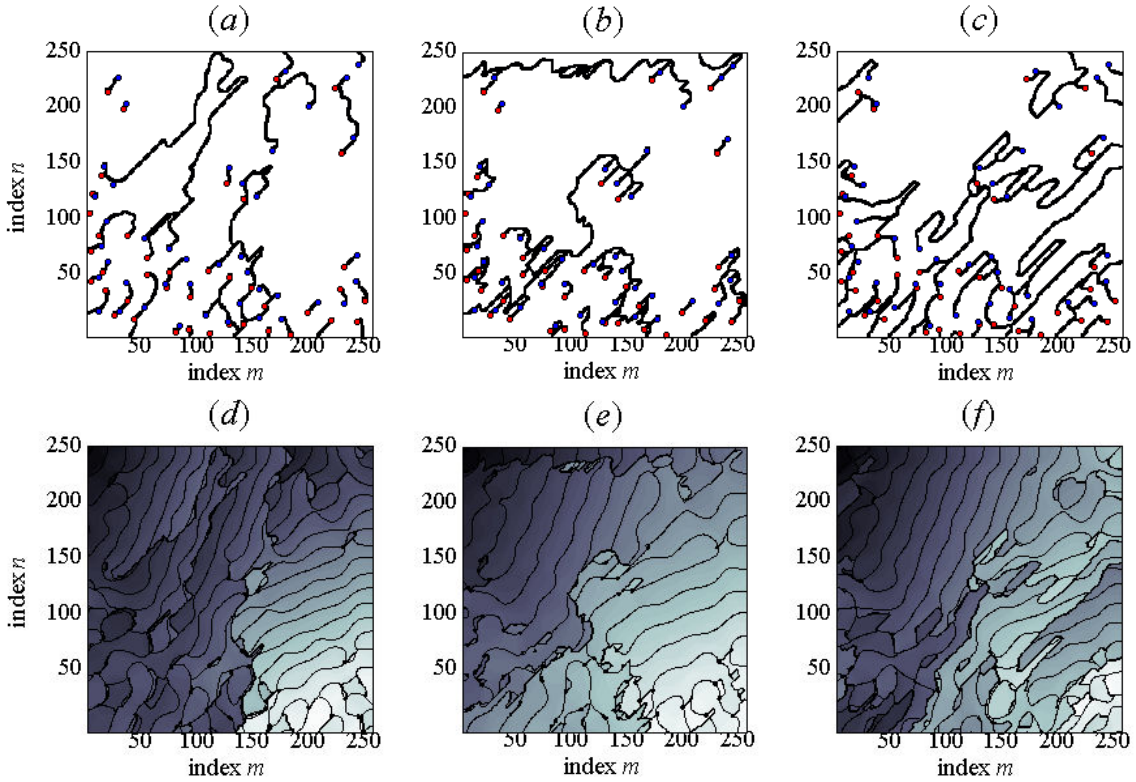


Figure 5-49. Results of phase unwrapping the third quadrant with the mask cut algorithm in which the mask cuts (a – c), generated using the quality maps shown in Figure 5-45, produce the unwrapped phases (d – f).

Note that the dipole pre-processing step used to improve the performance of Goldstein's algorithm is of no benefit to the mask cut algorithm since mask cut placement is influenced by the quality map rather than residues.

Flynn's Minimum Discontinuity Approach

The three algorithms examined so far employ diverse means to solve the phase unwrapping problem by approximately minimising discontinuities in the unwrapped phase: Goldstein's algorithm uses branch cuts; the quality-guided algorithm uses a quality map; the mask-cut algorithm grows mask cuts guided by a quality map. The final path-following algorithm to be examined provides a solution that minimises exactly, rather than approximately, discontinuities in the unwrapped phase. The

algorithm uses a tree-growing approach that traces paths of discontinuities in the phase, detects paths that form closed loops and add multiples of 2π to the phase within the loops to minimise the discontinuities. This process is repeated until no more loops are detected and it is guaranteed to converge on a solution with minimum discontinuity.

Figure 5-50 shows the result of applying Flynn's minimum discontinuity algorithm to the third quadrant of the wrapped phase. Although the unwrapped phase obtained using Flynn's minimum discontinuity algorithm has minimised the discontinuities in the unwrapped phase, to do so the amount of phase added through iterative additions of 2π to closed loops has resulted in a maximum phase difference across the grating of a massive ~468π radians, or equivalently ~234 wavelengths, which for the design wavelength of λ_0 = 3mm results in a transmission grating depth of approximately 1.3m! Clearly such a grating would be far from compact and so would be impractical in terms of grating manufacture. Furthermore the gratings phase modulation should be imposed on the incident beam at an infinitely thin plane, but diffraction through such a thick element may cause unexpected effects due to diffraction within the volume of the grating and not produce the expected far field output.

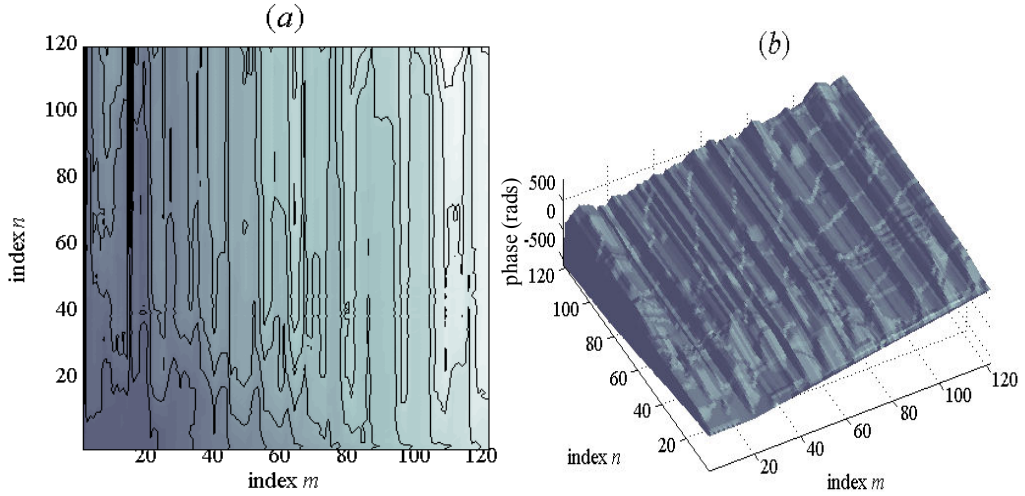


Figure 5-50. The third quadrant of the grating plane phase after phase unwrapping using Flynn's minimum discontinuity algorithm.

In conclusion to this discussion on phase unwrapping the quality-guided algorithm, based on a quality map derived from the variance of the phase gradient, produced the most desirable results: an unwrapped phase with the fewest discontinuities within the bulk of the Gaussian beam. The unwrapped phase of the entire grating plane is constructed by reflecting the phase of the third quadrant about the *x*- and *y*-axes and is shown in Figure 5-51. Note that if the wrapped phase had been anti-symmetric, as it

would had phase retrieval been performed using only odd-numbered Gaussian beam modes, a π phase shift is required between adjoining quadrants.

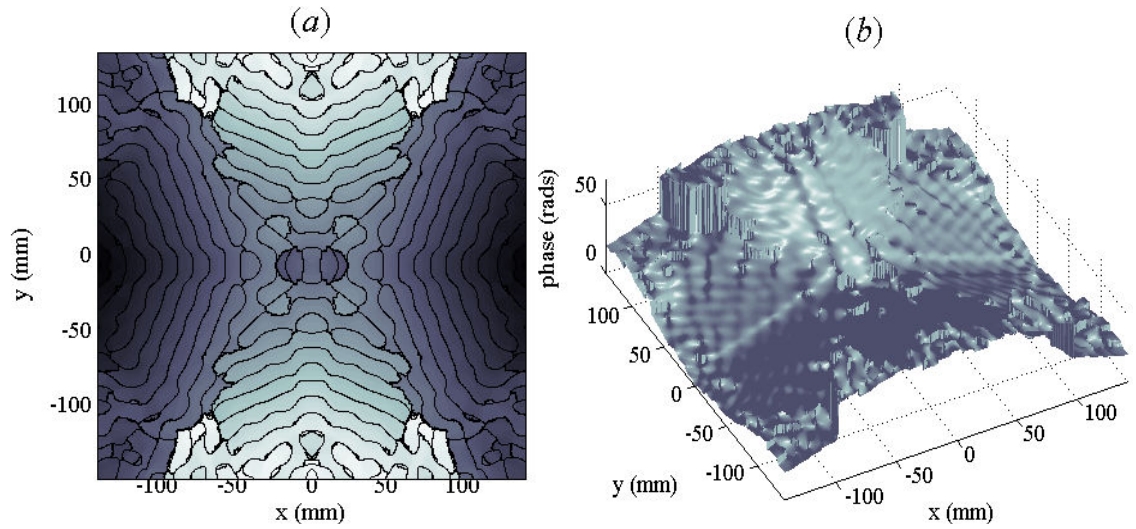


Figure 5-51. (*a*) False-coloured plot with contours superimposed and (*b*) surface plot of the unwrapped phase in its entirety, constructed by reflecting the unwrapped phase of the third quadrant shown in Figure 5-46(*b*) about the *x*- and *y*-axes. The maximum phase value is ~22π radians, which corresponds to a maximum grating height of approximately 11 wavelengths (for an on-axis reflection grating).

In the unwrapped phase shown in Figure 5-51 most remaining phase discontinuities are confined to the four corners. Since the incident Gaussian beam intensity in these regions is very low the phase in these regions contributes little to the far field image produced by the grating and so can be discarded without significantly affecting grating performance. To this end the phase outside a circular disc of radius $2W_G$, is masked out to produce the circular grating shown in Figure 5-52. The remaining phase within the disc contains only a small number of discontinuities and so is easier to manufacture.

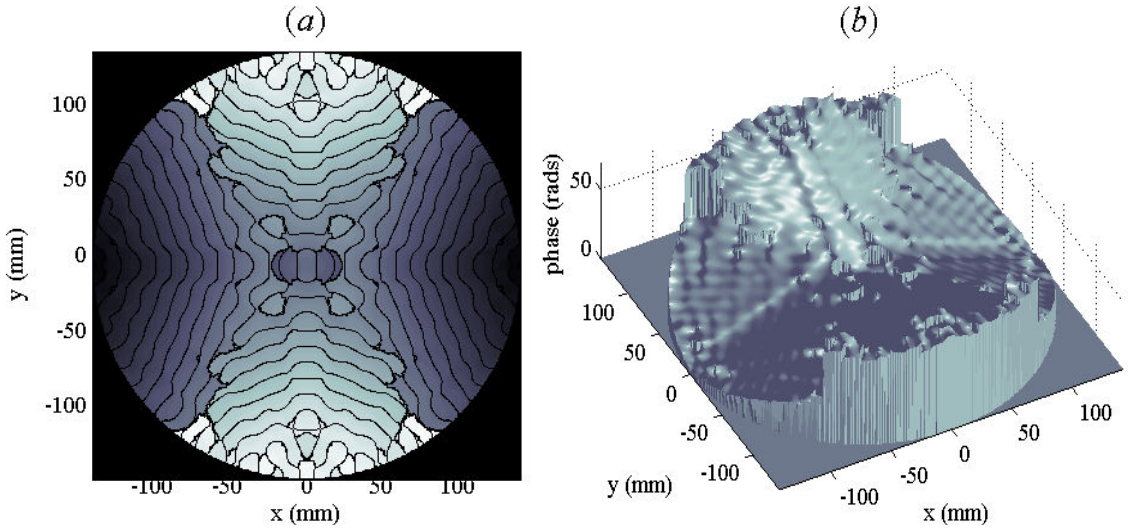


Figure 5-52. The 2-D unwrapped grating plane phase with the phase values outside a disc of radius $2W_G$ masked out.

Grating Fabrication and Experimental Measurements

Two phase gratings were produced using the unwrapped phase solution obtained using the quality-guided path-following algorithm described in the previous section. A reflection grating was manufactured from milling graded aluminium and a transmission grating from HDPE (which was assumed to have a refractive index of $n = 1.54$).

The grating surface height function $h(x, y)$ is defined such that the minimum height corresponds to the maximum cutting depth, and the maximum height corresponds to zero cutting depth. The surface height of a transmission grating is

$$h_{trans}(x, y) = \frac{\phi(x, y)}{k_0(n - 1)}$$

while the surface height of a reflective grating surface is

$$h_{ref}(x, y) = -\frac{\phi(x, y)}{k_0 \cos(\theta_{inc})}$$

where the negative sign was included because, while in transmission large phase lags are induced by forcing the wavefront to travel through a greater thickness of dielectric (to slow locally the propagating wavefront), in reflection the same phase lag is achieved by forcing the wavefront to travel a greater distance through free space. Of course if the minus sign is omitted, then the direction of phase modulation is reversed and the field transmitted from the grating is now $E^*(x, y)$, the complex conjugate of $E(x, y)$, the Fourier transform of which is $E^*(-u,-v)$, i.e. the far field wavefront is mirrored about the x and y axes, and the far field phase is the negative of what it otherwise would be.

Before manufacturing the two gratings we need to determine if a grating based on the unwrapped phase will actually produce the same results as one based on the wrapped phase. When designing a phase grating it is treated as an infinitely thin device. In other words it is assumed that the phase modulation is imparted on the incident wavefront at an infinitely thin plane. If the grating surface height $h(x, y)$ is derived from a phase function $\phi(x, y)$ that is limited to values in the interval $[-\pi, +\pi)$ this assumption is to a good approximation valid. In the case of the two Dammann gratings and the blazed beam-splitter described earlier this assumption was justified since the maximum peak-to-trough depth of the grating surface is on the scale of the design wavelength λ_0 (half a wavelength for the Dammann gratings; one wavelength for the blazed beam-splitter). However if the transmission and reflective gratings now being considered are derived from an unwrapped phase function that spans 21.94π their surfaces have

maximum heights (depths) of 60.93mm and 23.27mm, respectively (i.e. many wavelengths). Thus the assumption of an infinitely thin phase modulating device is no longer valid for these two gratings.

To see what effect, if any, a non-zero grating depth has on the far field intensity a number of simulations were conducted using the MODAL software. The far field intensity from a thin grating (derived from the wrapped phase) was calculated and compared with that produced by a thick grating (derived from the unwrapped phase). In these simulations the gratings were treated as reflection gratings since (at the time of writing) MODAL treats transmission elements as having zero thickness. In other words even if the phase input into MODAL spans multiples of 2π it is effectively wrapped back into the interval $[-\pi, +\pi)$, so two transmission gratings, one derived from a wrapped phase and the other from the equivalent unwrapped phase, are treated identically in MODAL. This does not occur when modelling reflection grating surfaces in MODAL. Before presenting results of the MODAL simulations for the thin and thick implementations of the reflection phase grating, we explain how reflection phase gratings are represented in MODAL.

Representing reflective grating surfaces in MODAL

In MODAL the surface of a reflective element is treated as the surface of an equivalent transmission element with the height divided by two, so a 2-D phase function $\phi(x, y)$ is translated into a reflective surface with a height function of

$$h_{ref}(x, y) = \frac{h_{trans}(x, y)}{2} = \frac{\phi(x, y)}{2k_0} \tag{5.36}$$

In other words normal incidence is assumed. In practice though reflective gratings operate in oblique incidence, i.e. with a non-zero angle of incidence θ_{inc}. The reflective surface needed to induce a phase modulation $\phi(x, y)$ on an incident wavefront has a height of

$$h_{ref}(x, y; \theta_{inc}) = -\frac{\phi(x, y)}{2k_0\cos(\theta_{inc})} \tag{5.37}$$

where the negative sign is included as explained previously. Therefore when representing a reflecting surface in MODAL, the phase data that is input into MODAL must be redefined to include the oblique angle of incidence in the phase data itself. Thus the phase that is input into MODAL to create a reflection grating is not $\phi(x, y)$ but rather

$$\phi_{ref}(x, y; \theta_{inc}) = -\frac{\phi(x, y)}{\cos(\theta_{inc})} \tag{5.38}$$

When read into MODAL this translates into a reflective surface with the correct height for illumination at the required angle of incidence, θ_{inc}. For the reflective grating considered here $\theta_{inc} = 45°$, so the phase data that must be input to MODAL is $\sqrt{2}\phi(x, y)$.

Comparison between output from thin and thick reflection gratings

First the thin and thick reflection gratings were modelled assuming normal incidence ($\theta_{inc} = 0$). Grating illumination was provided by an ideal collimated Gaussian beam and the reflected wavefront propagated to the far field. The far field amplitudes produced by the two gratings (see Figure 5-53) are almost identical in form and structure. Therefore the fact that the phase modulation of the thick grating is imparted gradually across the emerging wavefront, and not a single infinitely thin plane, has little effect for normal incidence. Although MODAL does not at present permit the modelling of transmission devices of differing thickness we can infer from these results that there would be very little difference between the far field intensities produced by a thin transmission phase grating and a thick version of the same grating. Thus it was decided that the phase grating(s) could be made from a design derived from the unwrapped phase function.

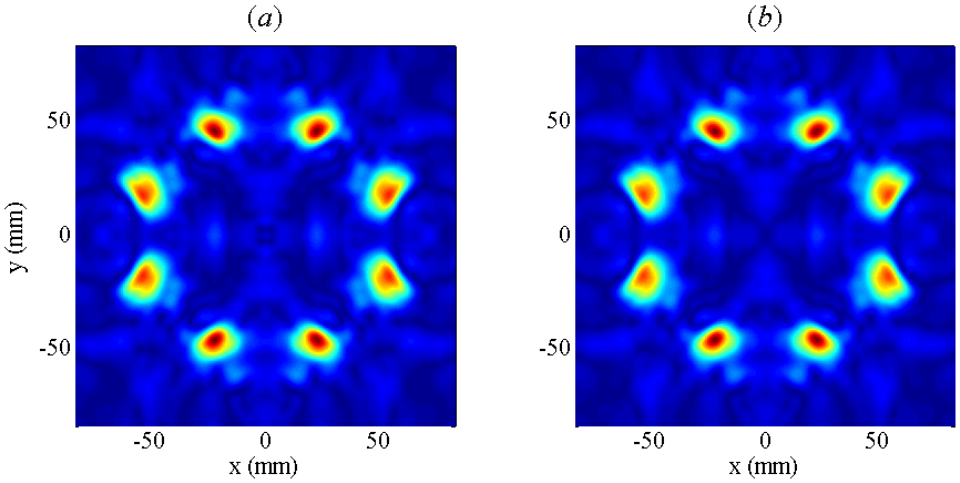


Figure 5-53. MODAL simulated far field amplitudes from (*a*) the wrapped and (*b*) the unwrapped reflective phase gratings with the incident and reflected beam paths along the grating normal.

The transmission and reflection phase gratings were machined on a CNC milling machine in the NUIM Experimental Physics mechanical workshop and are pictured in Figure 5-54. The two gratings were designed for illumination with a collimated

Gaussian beam of radius W_G = 71mm, and thus were given diameters of $4W_G \approx 284$mm. For the reflection grating the angle of incidence with respect to the grating normal, was set to $\theta_{inc} = 45°$ thus giving a 90° angle of throw (between incident and reflected beam paths at the grating).

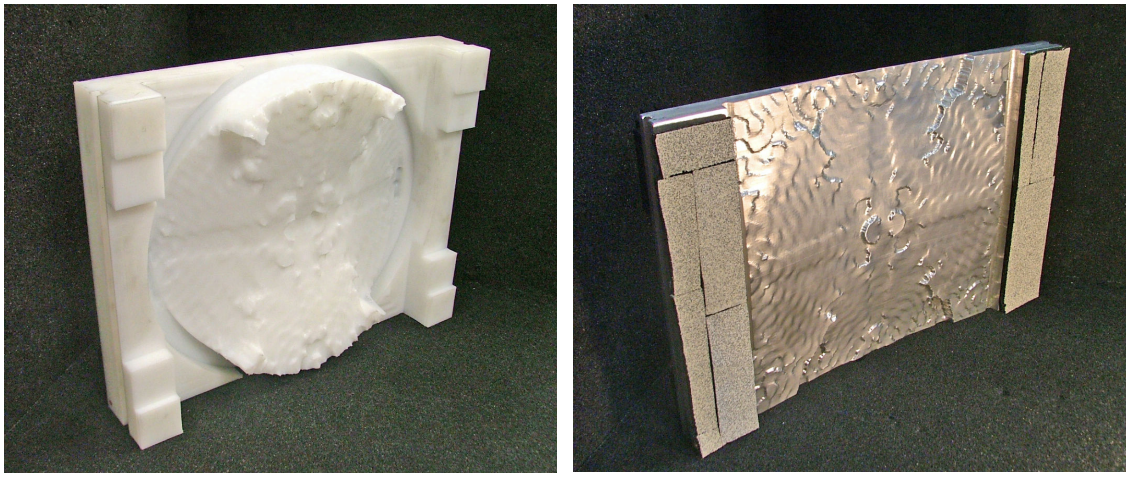


Figure 5-54. The finished transmission (left) and reflection (right) phase gratings that were each designed to generate a circular array of eight far field Gaussian beams. The transmission grating is derived from the unwrapped phase with the phase in low intensity regions (outside a disc of radius $2W_G$ = 71mm) masked out, while the reflective grating includes the phase transformation in low-intensity regions.

Note that projection effects were not taken into account in the design of the reflection grating (i.e. it was designed for normal incidence). However, through analysis of the measured and MODAL predicted beam patterns produced by the reflection grating, even for an oblique angle of incidence, we were able to gain insight into how the phase modulation actually produced the circular array of eight beams.

Measurements of the transmission Fourier phase grating

Three different Fourier optics test arrangements were simulated in MODAL to determine which would produce the highest quality output plane intensity from the transmission grating. The first 4-*f* arrangement was one of the arrangements used to measure beam patterns from the 3×3 and 5×5 Dammann gratings: two parabolic mirrors M_1 and M_2, each of focal length 350mm and 90° angle of throw. The resulting output plane amplitude shown in Figure 5-55(*a*) is, as usual, highly distorted. In the second arrangement M_2 was replaced with an ellipsoidal mirror (with a 500 mm focal length and 45° angle of throw), which results in the output plane amplitude being much less

distorted – see Figure 5-55(*b*). The third arrangement consists of two ellipsoidal mirrors. The longer focal length f_1 of mirror M_1 means that the Gaussian beam incident on the grating is larger than in the previous two arrangements, which may account for the uneven intensity distribution observed - see Figure 5-55(*c*). When illuminated with a smaller Gaussian beam, the phase transformation on the low intensity part of the beam at the four corners of the grating is essentially redundant, but with the larger incident Gaussian beam it now contributes to the wavefront transmitted from the grating. The highest quality image is obtained using the second arrangement (M_1 parabolic with 90° angle of throw and f_1 = 350 mm; M_2 ellipsoidal with 45° angle of throw and f_2 = 500 mm) so an experimental measurement was performed using only this arrangement.

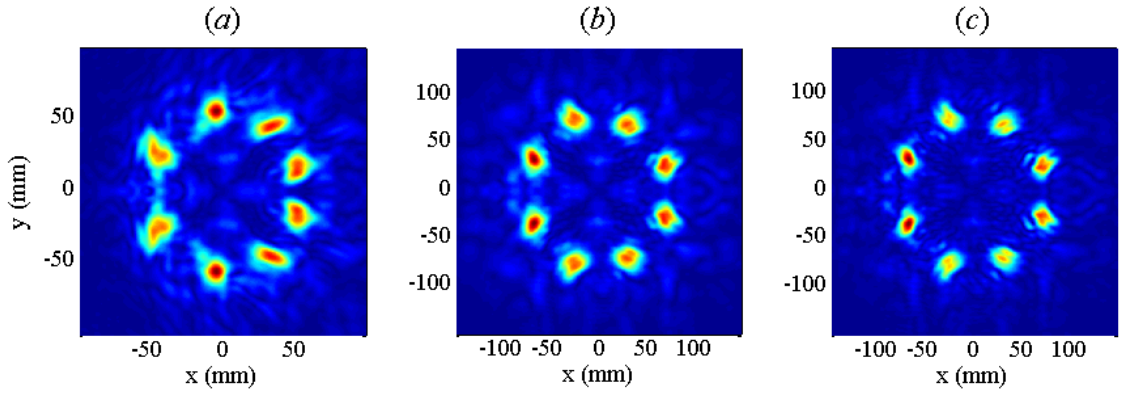


Figure 5-55. Output plane amplitude distributions from the transmission Fourier grating calculated with numerical simulations developed in MODAL for three different Fourier optics test arrangements: (*a*) two parabolic mirrors – hence the high level of distortion, (*b*) a parabolic mirror M_1 and an ellipsoidal mirror M_2 and (*c*) two ellipsoidal mirrors.

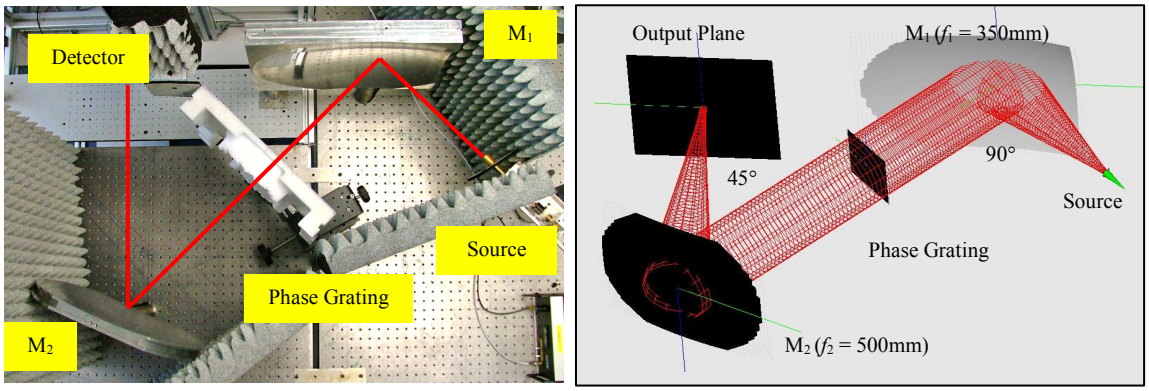


Figure 5-56. (**Left**) Photograph and (**Right**) screen shot from MODAL of the 4-*f* arrangement used to measure the circular 8-beam transmission phase grating. Parabolic mirror M_1 (f_1 = 350mm, 90°) collimates the source beam and ellipsoidal mirror M_2 (f_2 = 500mm, 45°) focuses the far field pattern from the grating onto the output (detector) plane.

Figure 5-56 shows a photograph as well as a screen shot generated by the MODAL software of the Fourier optics test arrangement with a combination of an ellipsoidal and a parabolic mirror. The experimentally measured output plane intensity is shown in Figure 5-57 and compares extremely well with the MODAL simulated intensity pattern shown in Figure 5-58. In both the simulated and measured intensities some distortion is introduced by mirror M_2, which accounts for the uneven vertical beam spacing between the beams on the left and the beams on the right of the array centre. Notice also that the upper left beams are the most intense and that maximum beam intensity drops off as we move towards the lower right corner of the image. This is most probably due to a slight misalignment of the grating between the two mirrors, which causes the illuminating Gaussian beam to be slightly off-centre, with respect to the grating. The measurement was made with Ecosorb surrounding the circular phase-modulating region of the phase grating. Another measurement made with no Ecosorb in place showed an almost identical measured intensity distribution. If a longer focal length mirror M_1 had been used to illuminate the grating Ecosorb would have been needed because surrounding the phase-modulating circular region of the transmission grating the HDPE block is machined flat and illumination of this part of device would have resulted in power being diffracted into an on-axis spot.

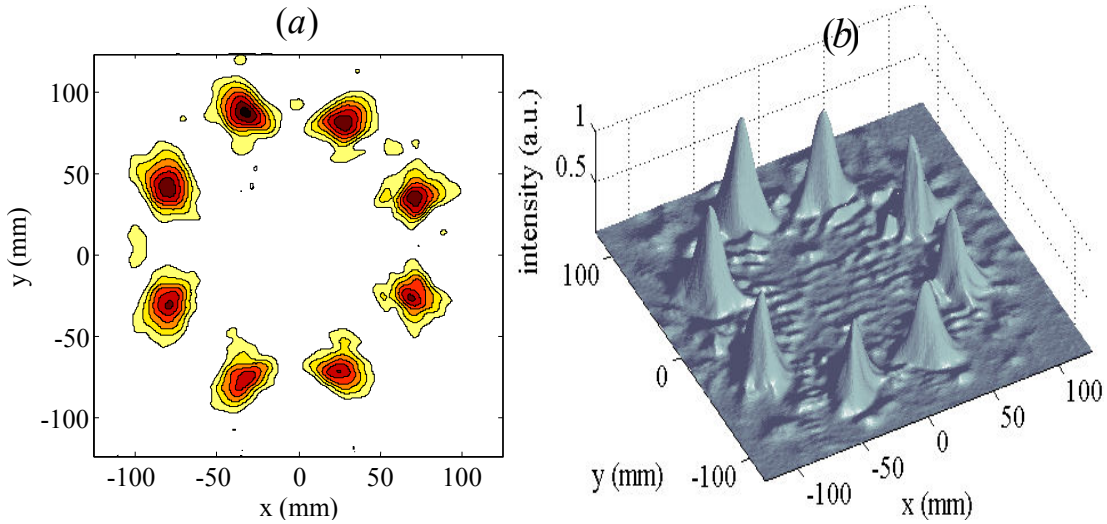


Figure 5-57. Experimentally measured output plane (*a*) amplitude and (*b*) intensity from the transmission phase grating with the Fourier optics test arrangement shown in Figure 5-56.

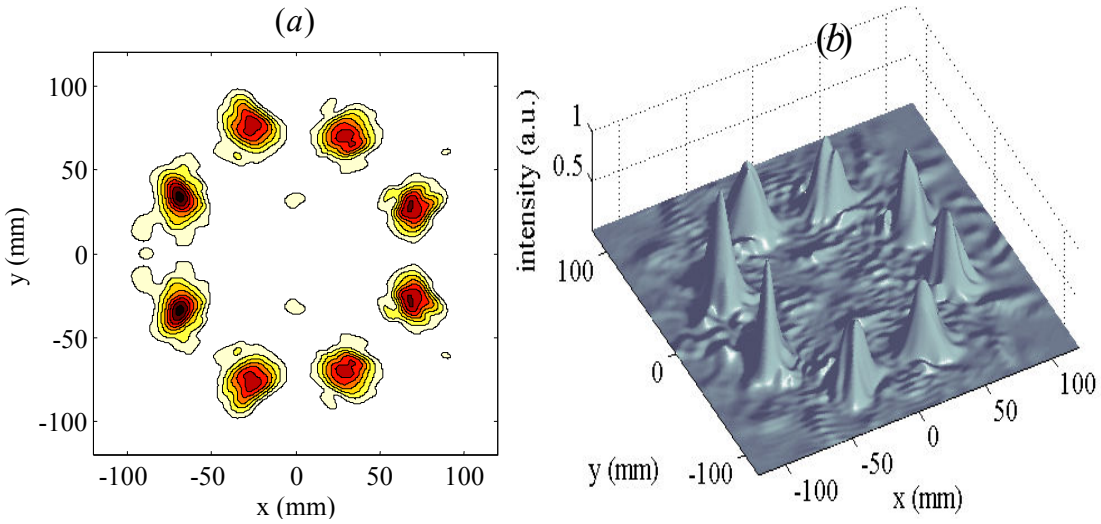


Figure 5-58. MODAL simulated output plane amplitude (*a*) and intensity (*b*) from the transmission phase grating with the Fourier optics test arrangement shown in Figure 5-56.

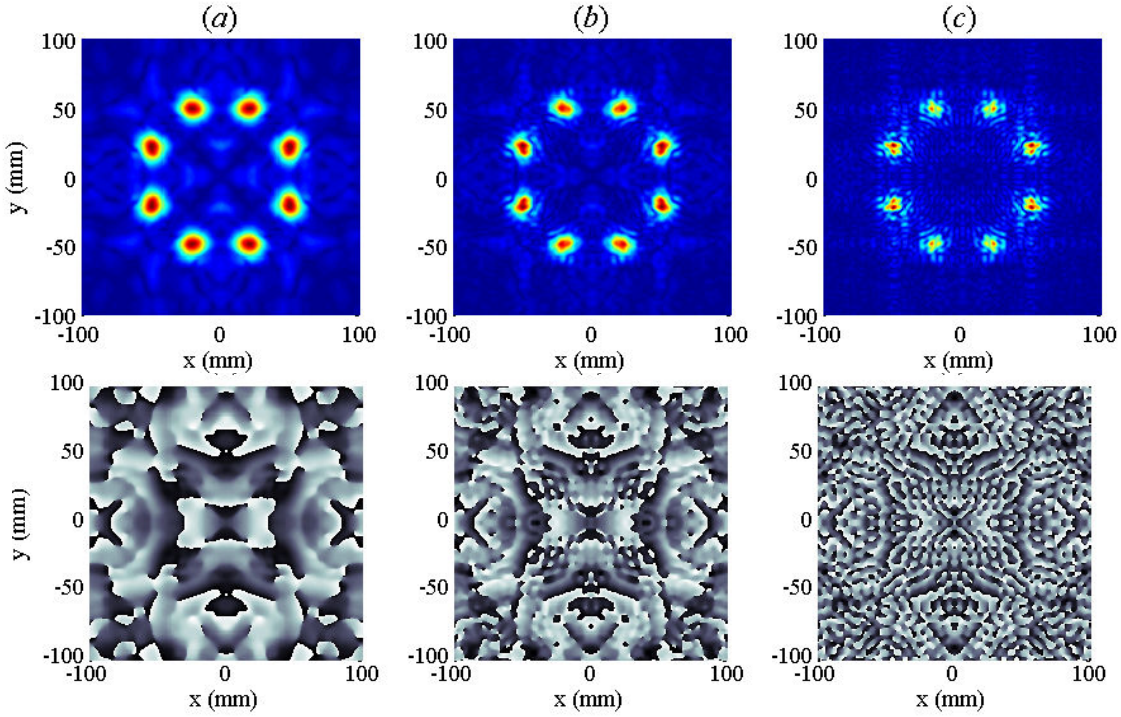


Figure 5-59. Fourier plane intensity (top) and phase (bottom) from the phase grating when illuminated with a Gaussian beam of radius (*a*) W_G = 71mm (*b*) W_G = 110mm and (*c*) $W_G = \infty$ (uniform illumination).

We return briefly to the simulations produced using MODAL shown in Figure 5-55. Notice that for the third test arrangement, the calculated output plane amplitude shown in Figure 5-55(*c*) contains a large amount of faint low-level intensity features. Furthermore the structure of the eight beams appears to be slightly fragmented. This is seen more clearly in Figure 5-59 where the intensity of the Fourier transformed phase

grating field is shown for illumination with a collimated Gaussian beam with increasing values of radius W_G.

Clearly from the results shown in Figure 5-59 the grating is very sensitive to changes in illuminating beam width. Figure 5-59(*a*) shows the Fourier plane intensity and phase for illumination of the grating with a Gaussian beam of radius W_G = 71mm, which is the radius of the target Gaussian beam amplitude used in the IPRA design of the grating and is similar to the size of the beam produced by one of the 350mm focal length parabolic mirrors in the experimental arrangement. For illumination with a larger incident Gaussian of radius W_G = 110mm, as shown in Figure 5-59(*a*), the output beams become narrower, as expected. However as well as the output beams becoming smaller, intensity fluctuations, or speckles, are also introduced, which are accompanied by phase fluctuations including spiral phase singularities. Finally, when uniformly illuminated, as in Figure 5-59(*c*), even more phase fluctuations are introduced and the intensity is dominated by speckles, even within the eight beams. The sensitivity of beam-shaping kinoforms (Fourier phase gratings) to variations in the width of the incident Gaussian beam was described in [5.41], where it was noted that the highest degree of correlation between the target far field intensity and that obtained from a beam-shaping device occurs only when illumination is provided by the distribution closest to that used during synthesis of the kinoforms phase. This same behaviour is observed with our beam-splitting grating, which only generates the target output (eight quasi-Gaussian shaped beams) when illuminated by an incident Gaussian beam of the same radius for which it was designed. Note that there is not a simple Fourier transform relationship between the width of the target signal beams here and the gratings illumination beam. In fact the target far field Gaussian beams are much wider than would be produced by the incident Gaussian beam at the output plane in the absence of the grating. If the target far field intensity had instead consisted of an array of Gaussian beams of the correct size (in terms of Fourier relationships) for the incident Gaussian, the phase grating should then not be as sensitive to the width of the input Gaussian beam.

Measurements of the reflection Fourier phase grating

The reflection phase grating was measured using four different Fourier optics test arrangements. The next few pages show the four arrangements (photographs and MODAL screen shots) as well as the experimentally measured output. The simulated intensity mages from MODAL are not presented since in all cases the measured and simulated intensities compared extremely well.

Test Arrangement 1

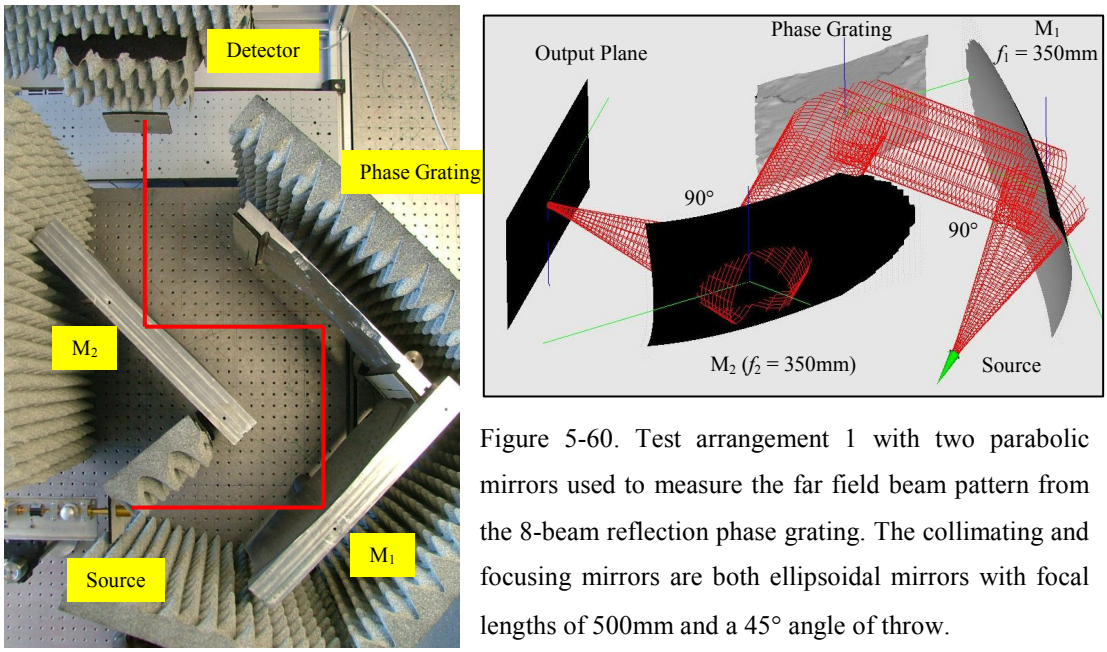


Figure 5-60. Test arrangement 1 with two parabolic mirrors used to measure the far field beam pattern from the 8-beam reflection phase grating. The collimating and focusing mirrors are both ellipsoidal mirrors with focal lengths of 500mm and a 45° angle of throw.

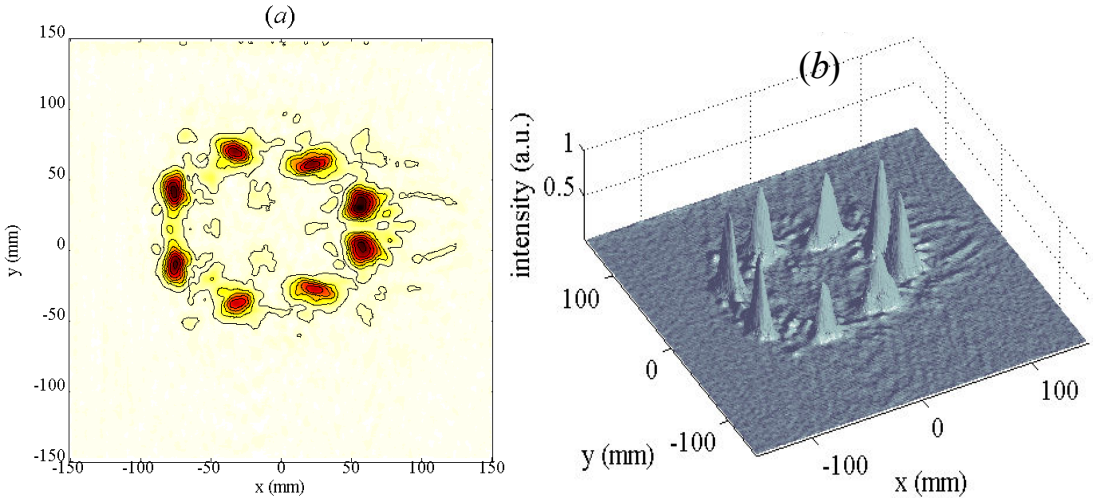


Figure 5-61. Experimentally measured output plane amplitude (*a*) and intensity (*b*) from the reflection Fourier phase grating using the test arrangement shown above. Power is evenly distributed between all eight Gaussian beams. The array of diffraction orders is not quite circular, but rather is elongated in the *x* direction (height of ~50mm and width of ~70mm). The two rightmost beams are the most intense.

Test Arrangement 2

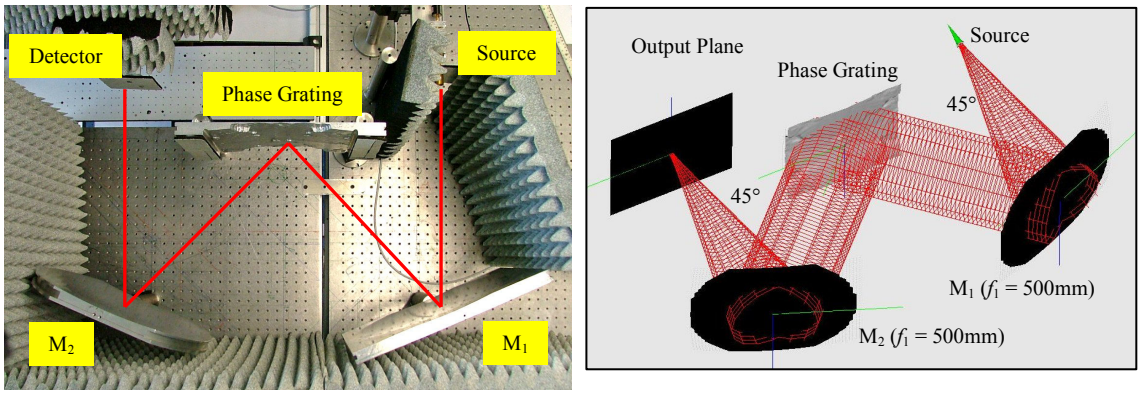


Figure 5-62. A 4-*f* Fourier optics test arrangement with two ellipsoidal mirrors used to measure the far field beam pattern from the 8-beam reflection phase grating. The collimating and focusing mirrors are both ellipsoidal mirrors with focal lengths of 500 mm and 45° angle of throw.

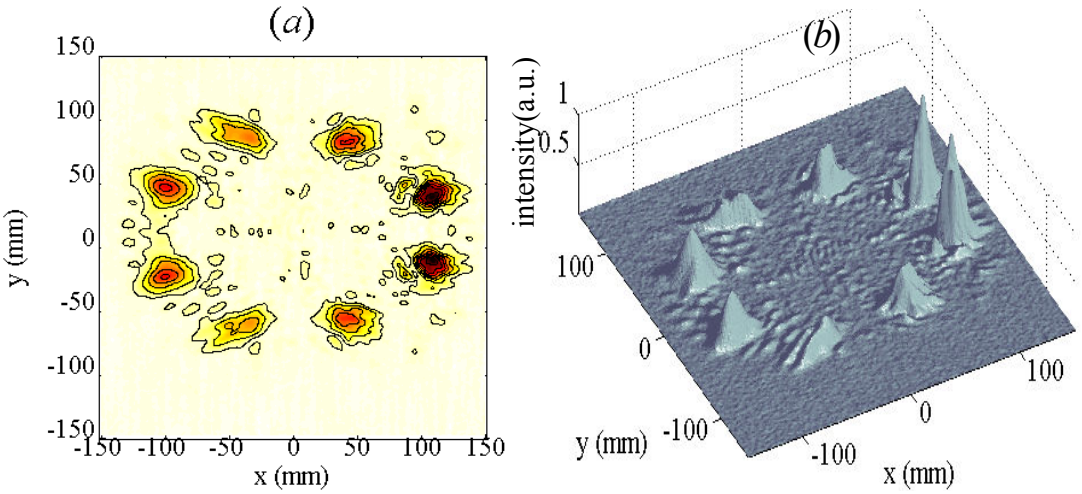


Figure 5-63. Experimentally measured output plane amplitude (*a*) and intensity (*b*). The width and height of the circular array are now ~70mm and 100mm due to the longer focal length of M_2.

Test Arrangement 3

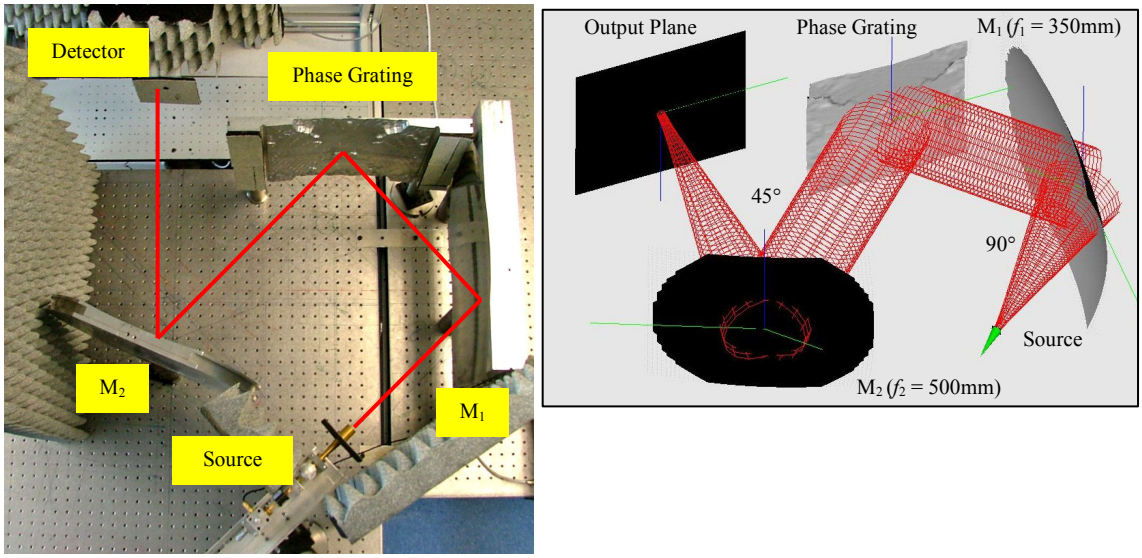


Figure 5-64. Fourier optics test arrangement combining parabolic mirror M_1 with ellipsoidal mirror M_2.

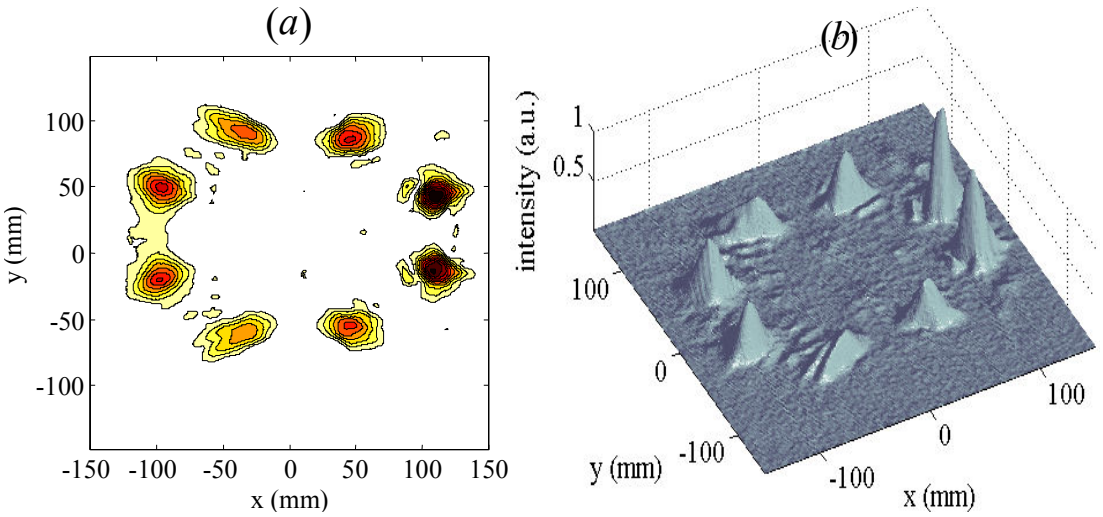


Figure 5-65. Experimentally measured output plane amplitude (*a*) and intensity (*b*). The width and height of the circular array are now ~70mm and 100mm due to the longer focal length of M_2.

Test Arrangement 4

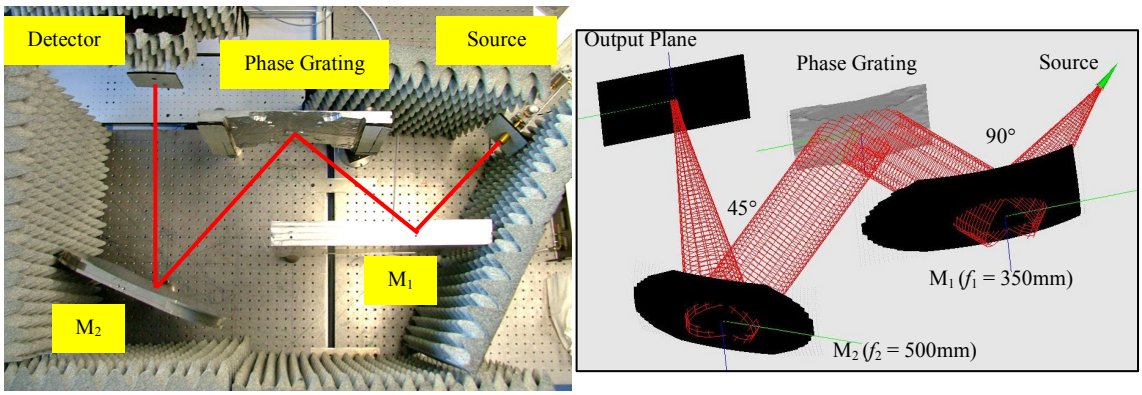


Figure 5-66. Same test arrangement as shown in Figure 5-64 except that M_1 is repositioned such that the source beam incident on M_1 comes from the opposite direction to that above. The far edge of M_1 may be obstructing some of the radiation reflected from the grating.

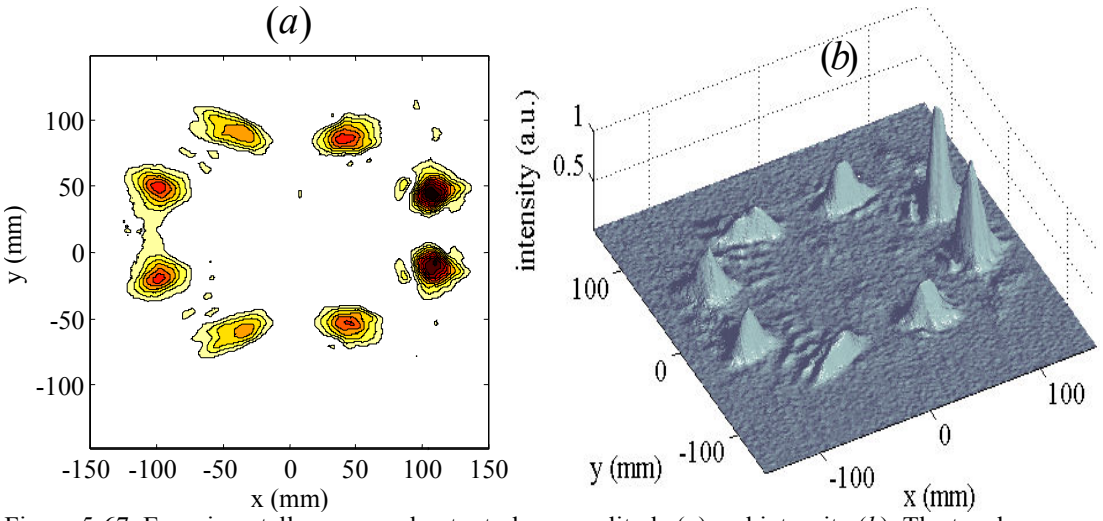


Figure 5-67. Experimentally measured output plane amplitude (*a*) and intensity (*b*). The two beams on the left side of the plot are less intense than in the image obtained with the alternative arrangement with the same pair of mirrors, which may be due to truncation by mirror M_1 in this set-up.

For all four Fourier optics test arrangements the sparse array of eight beams is observed, however image quality is substantially lower than in measurements obtained from the transmission grating. In each measurement of the reflection grating, the power distribution between the beams in the spot array is very uneven: two very intense beams on the right (at large positive *x*-values) dominate each measured spot array. The highest quality image (see Figure 5-61) was obtained using the first test arrangement (with the two 350mm focal length parabolic mirrors), but even in this image the two beams on the right (largest positive *x*-values) are more intense than any of the other eight beams.

MODAL was used previously to simulate the far field intensity from a thin reflection phase grating (derived from the wrapped grating phase) and from a thick reflection phase grating (derived from the unwrapped grating phase) and no difference was found. However in those simulations the reflection gratings were operated in normal incidence. However the manufactured reflection grating was actually measured with illumination at oblique incidence so the simulations were repeated but with the two (thin and thick) gratings now illuminated at oblique incidence (with an angle of incidence $\theta_{inc} = 45°$). Also the previous simulations illuminated the grating with an ideal Gaussian beam and propagated the reflected wavefront to an output plane in the far field. In these new simulations the system defined in MODAL includes the collimating and the collecting mirrors M_1 and M_2 and is fed by a corrugated cylindrical feed horn (as in the experimental measurements).

Since measurements on the reflection grating using the first test arrangement (with two 350mm focal length parabolic mirrors) provided the best results this was the system used for comparing the output from thin and thick variations of the reflection grating in MODAL. The resulting output plane intensity patterns from MODAL simulations are shown in Figure 5-68.

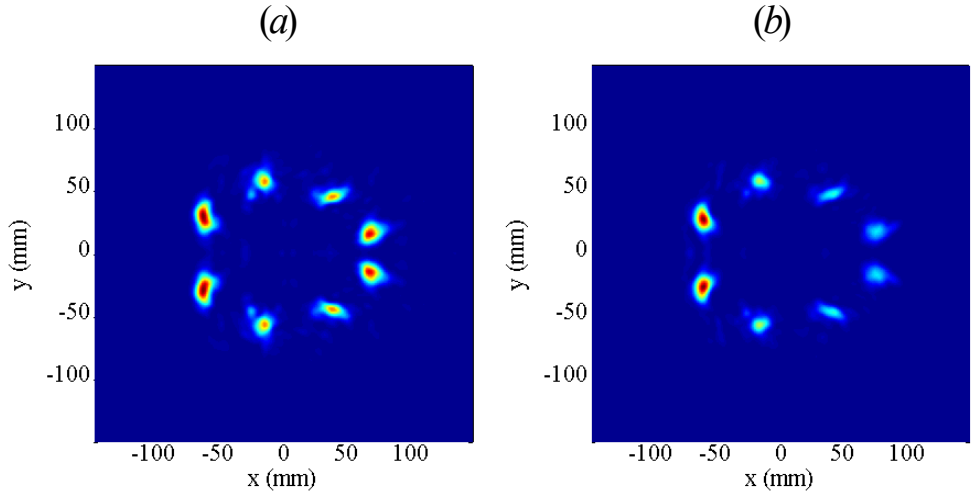


Figure 5-68. MODAL simulated intensity patterns at the output plane of the 4-*f* test arrangement with two parabolic mirrors (*f* = 350mm, 90° angle of throw) from a reflection grating designed for oblique incidence ($\theta_{inc} = 45°$) and derived from (*a*) wrapped phase and (*b*) unwrapped phase. Both beam arrays suffer from distortion from the parabolic mirror used to focus the diffracted wavefront to the output plane.

As usual both output plane images are subject to distortion due to the high angle of throw of the collecting/focusing mirror M_2. The most striking feature however is that the spot array is more elliptical than circular. This problem will be addressed shortly.

Analysis and Improvements in Design Approach

As regards the difference between the two images in Figure 5-68, it is clear that power is more evenly distributed between beams in Figure 5-68(*a*) from the thin grating, than it is in Figure 5-68(*b*) from the thick grating (the two spots on the left being more intense than any others). Furthermore besides distortion from the collecting mirror, the individual spot profiles are far less circularly symmetric than was observed in simulations of gratings designed for normal incidence. These defects can all be explained by considering the intensity distributions illuminating the two gratings.

In normal incidence the grating is illuminated by a circularly symmetric Gaussian beam, which is not the case for a grating operating in oblique incidence. Figure 5-69 shows the intensity calculated on the surface of the thin and thick reflection phase gratings designed for oblique incidence. Clearly the intensity with which the two gratings are illuminated is far from circular. The intensity illuminating the thin grating is more elliptical than circular, in other words a Gaussian beam whose waist radius in the *x*-direction is greater than that in the *y*-direciton: $W_x > W_y$. The intensity illuminating the thick grating appears to be a slightly off-centre elliptical Gaussian beam.

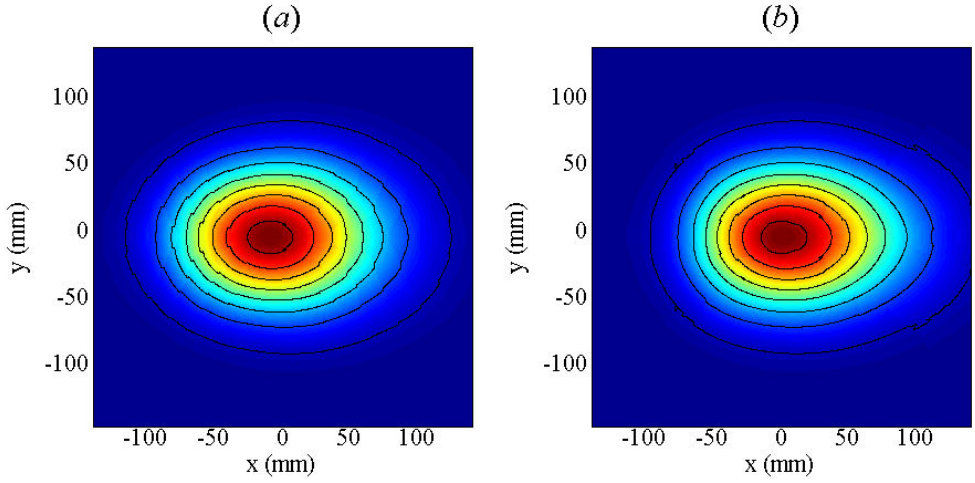


Figure 5-69. Intensity distributions at the surface of the reflection phase gratings designed for oblique incidence (θ_{inc} = 45°) derived from (*a*) wrapped phase and (*b*) unwrapped phase modulations. The intensity no longer has the profile of a circularly symmetric Gaussian beam, but is more elliptical.

Illumination of the phase grating with four different incident Gaussian beams: circular, offset circular, elliptical and offset elliptical, was modelled using Fourier transforms (FFT). Figures Figure 5-70 to Figure 5-73 show the grating intensity and the resulting Fourier plane intensity for the four cases.

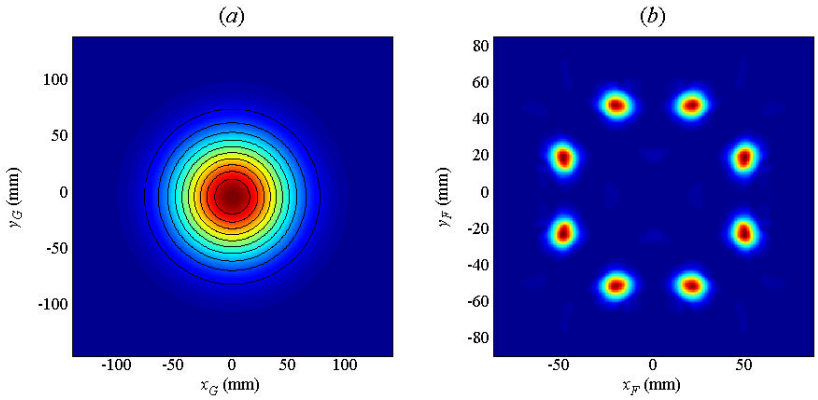


Figure 5-70. Illumination of the grating with a centred circularly symmetric Gaussian beam (*a*).

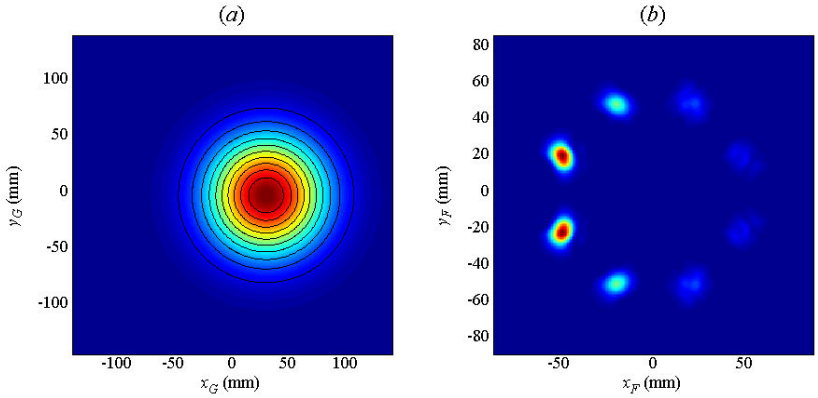


Figure 5-71. Illumination of the grating with an off-axis circularly symmetric Gaussian beam (*a*).

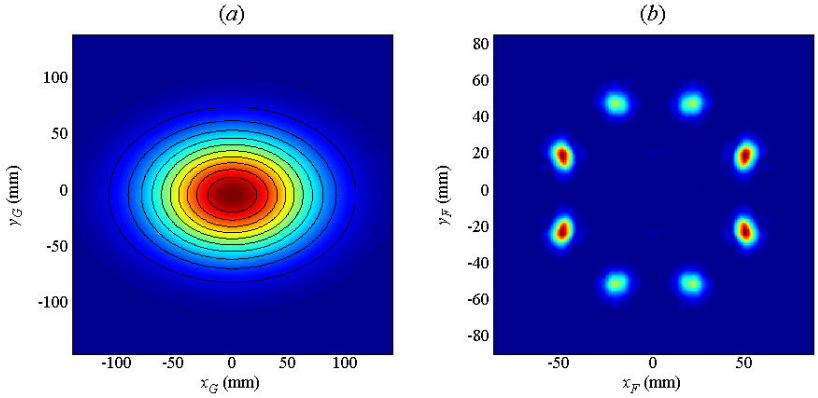


Figure 5-72. Illumination of the grating with a centred elliptical Gaussian beam (*a*). The Fourier plane spots are now slightly elongated in the y-direciton.

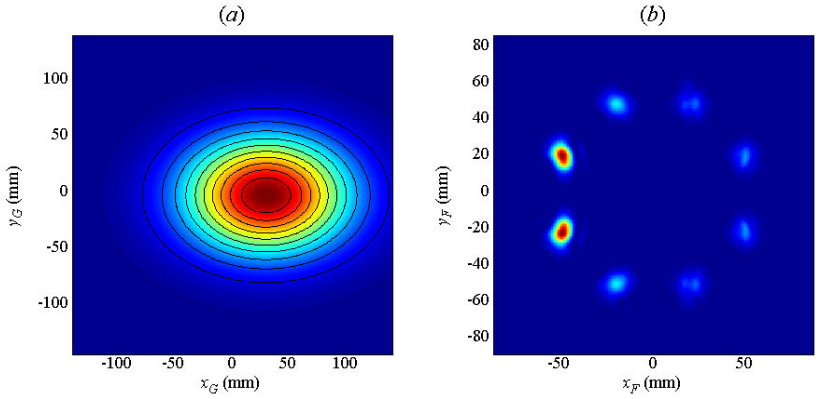


Figure 5-73. Illumination of the grating with an off-axis elliptical Gaussian beam (*a*) combines the defects of illumination shown in the previous two plots.

Illumination with a circularly symmetric Gaussian beam ($W_x = W_y$) produces the expected far field intensity: a circular array of equally intense circular Gaussian beams. For illumination with a horizontally offset circular Gaussian beam the two output spots on the left are significantly more intense than the rest of the output plane beams. This suggests that because one half of the grating is illuminated more so than the other half, the phase modulation form the brightly illuminated half contributes more to the far field distribution than the weakly illuminated side. For illumination with a centred elliptical Gaussian beam the far field Gaussian beams are elongated, as expected, but more importantly the four beams left and right of the array centre are much more intense than the four beams closest to the y-axis. This is the same behaviour observed in the MODAL predicted output plane intensity of Figure 5-68. Finally the grating is modelled for illumination with a horizontally offset elliptical Gaussian beam with the result that the two beams to the left are most intense and the upper and lower beams are even less intense, in agreement with both MODAL simulations (see Figure 5-68) and experimental measurements (see Figure 5-63, Figure 5-65 and Figure 5-67).

These simulations show that the grating is extremely sensitive to the shape as well as the position of the illuminating beam, much more so than would be expected from, for example, a periodic DPE such as a Dammann grating. The reason for the sensitivity is that the grating phase modulation is not periodic, so it is essential that all regions of the grating phase be illuminated with the same intensity profile as that used during the iterative phase retrieval algorithm. An iterative algorithm is described in [5.41] for synthesising grating phase so that the resulting grating is less sensitive to beam position.

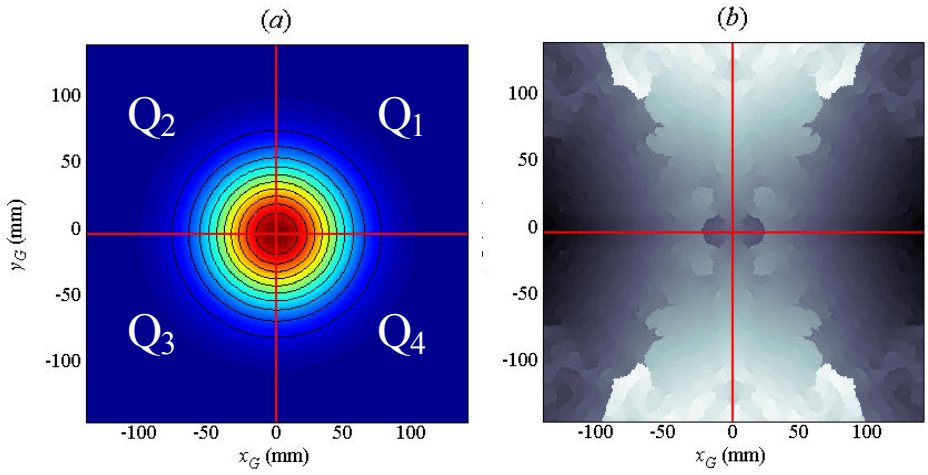


Figure 5-74. The grating phase in (*b*) when illuminated with a centred circularly symmetric Gaussian beam (*a*) can be divided into four quadrants Q_1 to Q, that are mirrored images of each other.

Next the unwrapped phase profile was divided into four quadrants (see Figure 5-74) and each propagated (using FFT) independently of the others to the far field. The result shown in Figure 5-75 is that each quadrant produces two of the eight far field Gaussian beams.

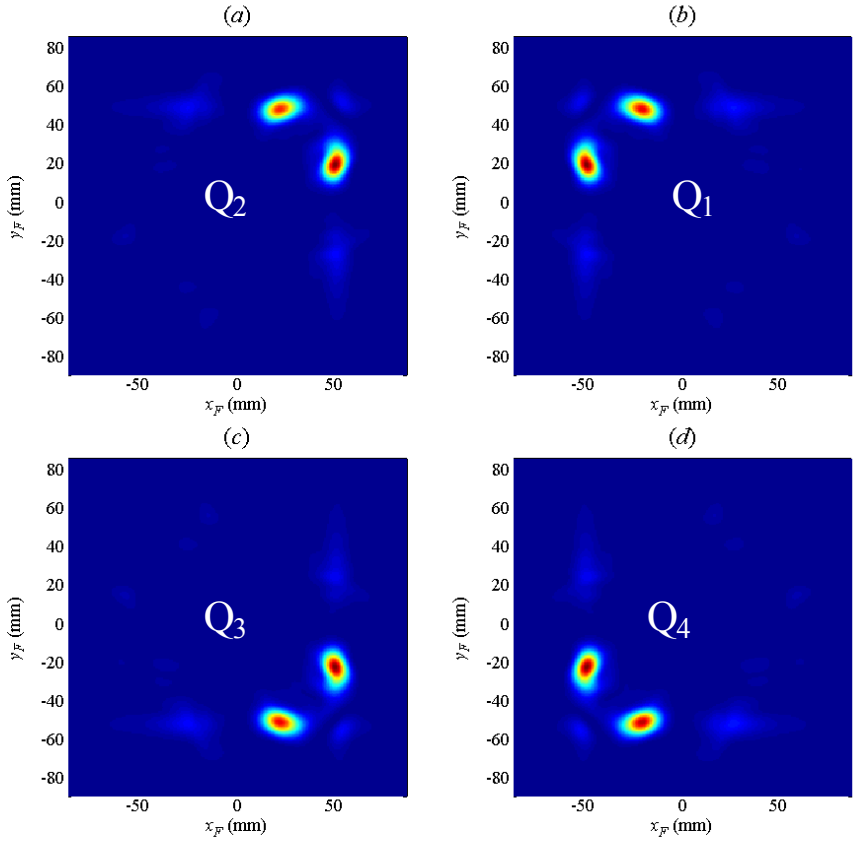


Figure 5-75. The intensity of the Fourier transform of each of the four quadrants at the grating plane shown in Figure 5-74. Each quadrant produces two of the eight far field quasi-Gaussian beams.

If we also draw two diagonals through the centre of the grating at (x, y) = (0, 0), such that each quadrant consists of two triangular facets (see Figure 5-76) then the grating is composed of eight such triangular facets, each of which is responsible for creating a single far field beam, thus explaining the sensitivity of the phase grating as a whole to the position and symmetry of the illuminating beam.

Since facet 1 in Figure 5-76(*a*) is much smoother than facet 2 it was tested whether a phase modulation in which the phase in facets 2, 3, 6 and 7 are replaced with the phase of facets 1, 4, 5 and 8 could still produce the desired far field intensity but with a phase grating that is smoother and therefore easier to manufacture. The resulting phase grating shown in Figure 5-76(*b*) has 4-fold symmetry and does not have the phase discontinuities of facet 2 (and its reflected counterparts) and so would be simpler to

machine. However it was found that the far field intensity from this newly assembled grating did not produce the target far field intensity as well as the original phase grating.

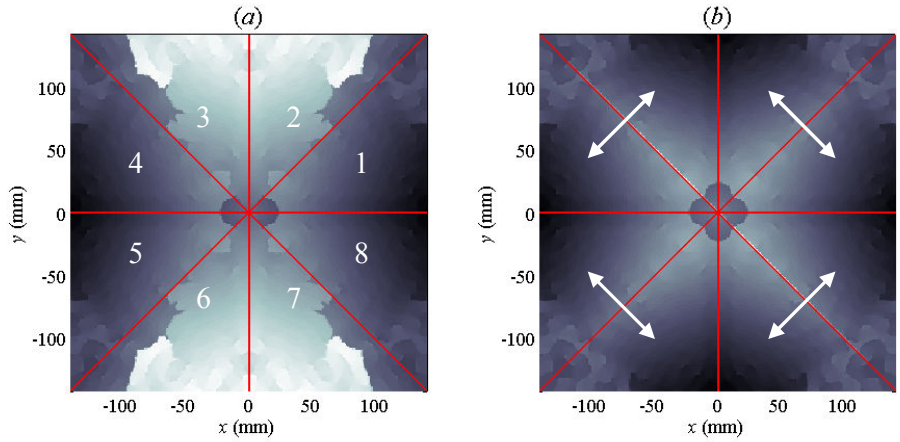


Figure 5-76. The unwrapped phase (*a*) can be divided into eight triangular facets (numbered 1 to 8). The phase modulation in each facet is responsible for generating a single output plane Gaussian beam. The phase in (*b*) was constructed by replacing the phase in facets 2, 3, 6 and 7 by reflecting the phase in neighbouring facets about the diagonals. By itself, the phase in (*b*) is a poor candidate solution for producing the target far field intensity, so it was used as the starting point for the FFT-IPRA.

The FFT-IPRA was started with this new 4-fold symmetric phase as its starting point. The algorithm stagnated to a solution (see Figure 5-77) after 15 iterations. The corresponding far field intensity is a slight improvement on the original solution. More importantly though the grating phase is an extremely smoothly-varying function, the eight facets of which are, apart from some ripples in phase along the lines dividing each of the eight facets, simply slanted planes. Such a phase would be a much simpler device to manufacture on a milling machine and would present few problems as a reflective element since the vast proportion of the surface is without sharp phase discontinuities.

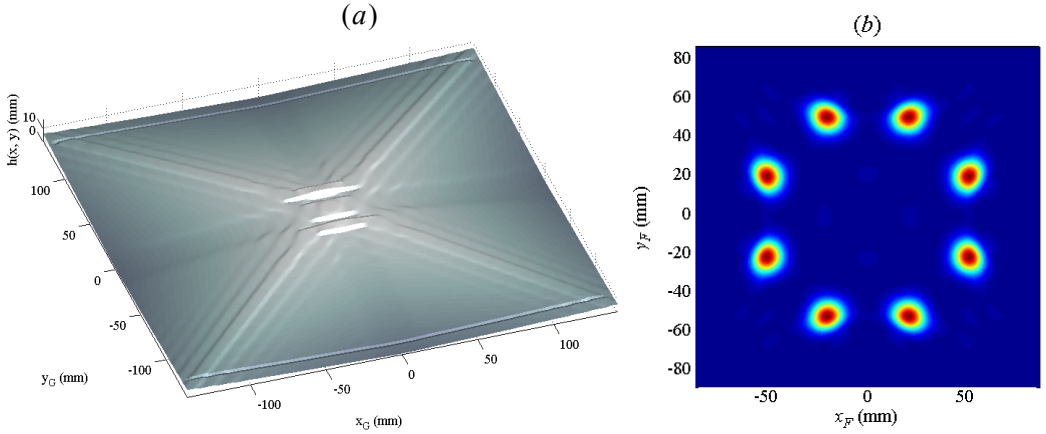


Figure 5-77. (*a*) The phase solution found after 15 iterations of the FFT-IPRA, which was started with the initial grating plane phase set to that shown in Figure 5-76(*b*). The plot in (*b*) shows the far field intensity corresponding to the Gaussian-illuminated phase in (*a*).

Accounting for projection effects

Referring to Figure 5-68 the horizontal elongation of the simulated far field intensity from the (wrapped and unwrapped) reflection phase gratings is due to projection effects because the grating is operated at oblique incidence. To compensate the grating must be projected onto a plane at an angle θ_{inc} to the grating plane, which translates to extending the grating width by $1/\cos(\theta_{inc})$. Thus for $\theta_{inc} = 45°$ the width of the reflective grating should be changed from 284mm to $\sqrt{2}(284\text{mm}) = 402\text{mm}$. The stretched reflection grating derived from the unwrapped phase is then as shown in Figure 5-78.

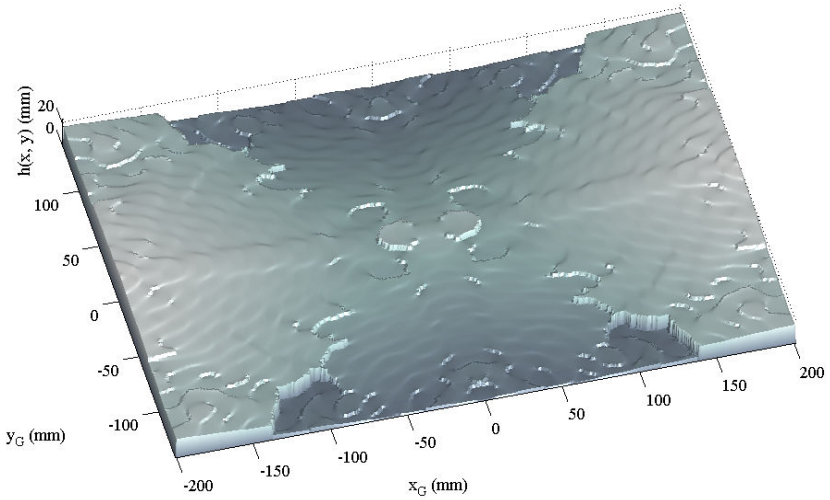


Figure 5-78. The reflection phase grating to produce a circular array of far field Gaussian beams, derived from the unwrapped phase and stretched in the *x*-direction to compensate for projection effects for oblique incidence (at an angle of 45°).

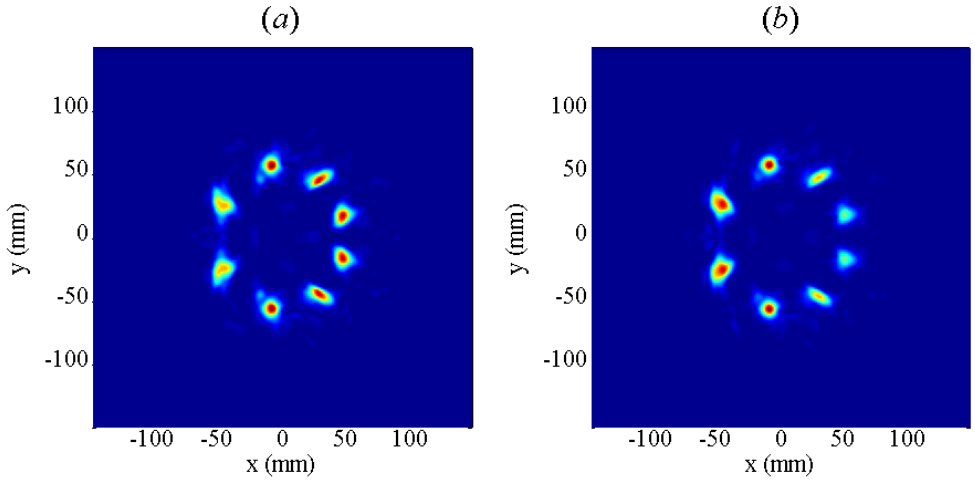


Figure 5-79. MODAL simulated intensity at the output plane of the 4-*f* test arrangement with two parabolic mirrors (*f* = 350mm, 90° angle of throw). The intensity patterns are produced by two reflective gratings (designed for oblique incidence $\theta_{inc} = 45°$) and appropriately stretched to account for projection effects. Images (*a*) and (*b*) correspond to output intensity patterns from phase gratings derived from wrapped and unwrapped phase, respectively.

Figure 5-79 shows the output plane intensity from MODAL simulations of the wrapped and unwrapped reflection phase gratings after the grating width has been set to 402 mm to account for projection effects. The beam array is now circular as required.

The output beams in Figure 5-79(*a*) from the wrapped grating are all of equal intensity, except for the two leftmost beams, which can be attributed to distortion by the collecting mirror. However in the pattern in Figure 5-79(*a*) produced by the unwrapped grating the two leftmost beams are more intense (than from the wrapped grating) and the two rightmost beams less intense. The reason for the lower quality image obtained from the unwrapped grating is due to the incident intensity on its surface. Figure 5-80 shows the intensity at the surface of the two gratings. The intensity profile at the unwrapped grating is slightly offset to the right of the grating centre. Thus the faceted regions on the left of the unwrapped grating receive less power than those on the right, which explains the difference in intensity between the rightmost and leftmost beams in Figure 5-79(*b*).

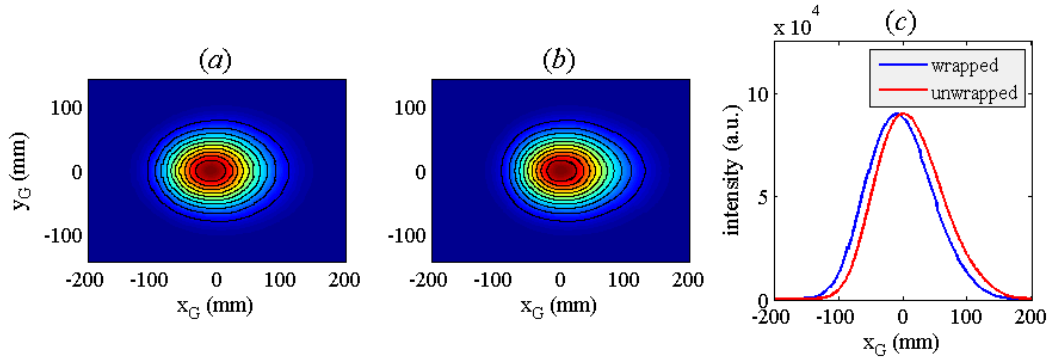


Figure 5-80. Intensity at the plane of (*a*) the wrapped reflective grating and (*b*) the unwrapped reflective grating as calculated using a simulation developed in MODAL.

Limitations in MODAL when modelling reflection phase gratings

Figure 5-79 suggests that a reflection grating derived from the wrapped phase front would have proved a better design than one derived from the unwrapped phase. However it must be pointed out that MODAL only includes a single reflection from a surface and also does not account for recessed regions on the surface that are shadowed by nearby raised areas that project forward. For smoothly varying reflecting surfaces such as mirrors clearly this does not present a problem. However for a surface with steep-sided features this may not be the case and for one designed for oblique incidence inaccurate beam pattern predictions may result. The reflection grating was designed to operate with a 90° angle of throw and since the wrapped phase contains many

discontinuities the predicted intensity from a grating derived from the wrapped phase may not be as accurate as that shown in Figure 5-79(*a*).

The unwrapped phase contains substantially fewer sharp phase discontinuities so the fact that MODAL does not include shadowing effects and multiple reflections is not such an issue when modelling reflection from a grating derived from the unwrapped phase. Thus although Figure 5-79(*a*) indicates that better performance is achieved with the wrapped reflection grating, it is not certain how accurate this prediction is. Thus although the image in Figure 5-79(*b*) is lower in quality than the image in Figure 5-79(*a*), because MODAL does not model the interaction of the propagating wavefront as accurately in the second case, in practise the measured beam pattern from a wrapped reflection grating may not resemble that predicted by MODAL.

Simulating Truncation Effects using Gaussian Beam Mode Analysis

Now we investigate what effects, if any, truncation at collecting mirror M_2 had on the measured output plane intensity patterns from the Fourier phase grating to produce a ring of 8 Gaussian beams in a number of the test arrangements used. As before truncation analysis was performed in terms of Gaussian beam mode analysis.

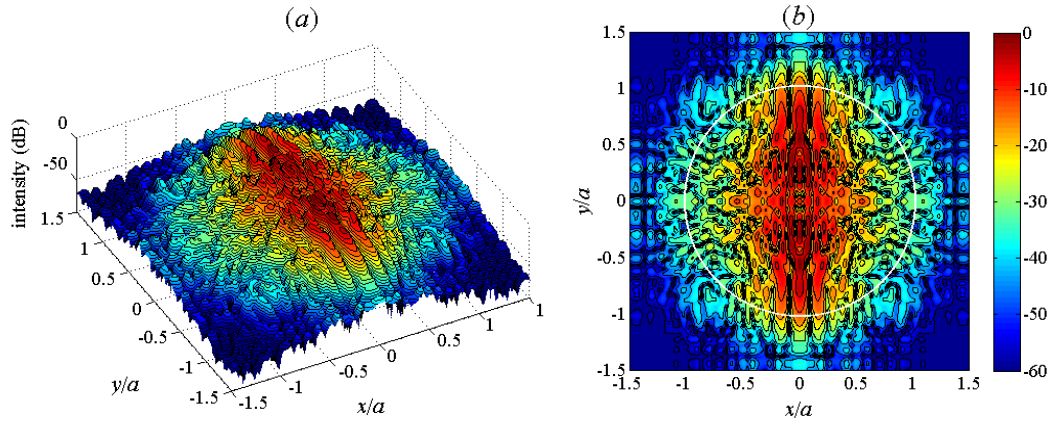


Figure 5-81. Log-scaled plots of the beam intensity from the 8-beam Fourier grating in the plane of mirror M_2 with radius a = 142.37mm. Mirrors M_1 and M_2 both have focal lengths of 350mm. The white circle in (*b*) represents the perimeter of the truncating circular aperture at M_2.

First we consider the 4-*f* Fourier optics test arrangement (used to test the reflection phase grating) in which parabolic mirrors with 350mm focal lengths were used for mirrors M_1 and M_2 (see Figure 5-60). The aperture of M_2 is treated as a circular aperture with radius a = 142.37mm. Figure 5-81 shows the Gaussian beam mode approximation of the beam intensity in the plane of M_2 before and after truncation by a circular

aperture representing the collecting surface of M_2. Notice that the beam pattern is elongated in the y-direction relative to the x-direction. Therefore more power is lost in the y-direction than in the x-direction after truncation.

Figure 5-82 shows the simulated output plane intensity distribution with and without truncation at mirror M_2. Figure 5-83 shows x-cuts from Figure 5-82 taken through the centre of the first (top) and second rows of Gaussian beams. Clearly, truncation at M_2 causes reduced intensity in the upper and lower pairs of Gaussian beams, compared to relatively little change in intensity in the two central rows of beams (in agreement with experimental results shown in Figure 5-61).

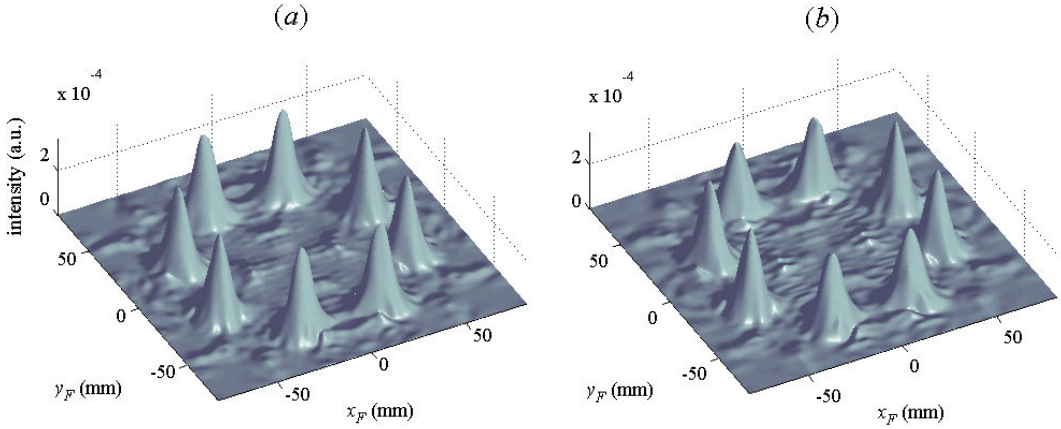


Figure 5-82. Output plane intensity from the 8-beam Fourier grating in a 4-f set-up where mirrors M_1 and M_2 both have focal lengths of 350mm (a) without and (b) with truncation included at mirror M_2.

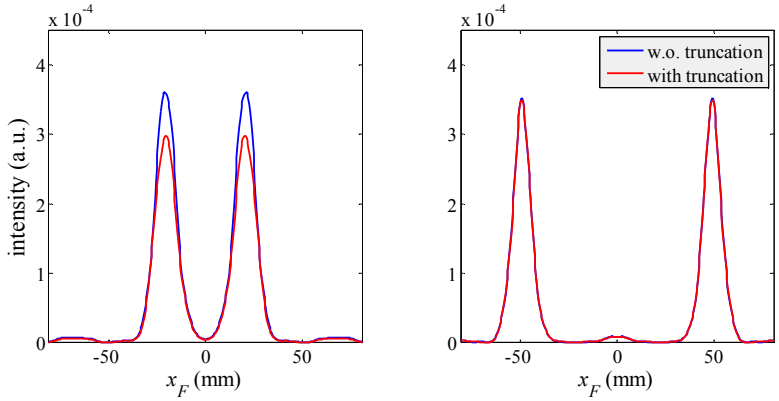


Figure 5-83. Horizontal cuts through the centre of (left) the first and (right) the second rows of Gaussian beams at y_F = 50mm and y_F = 21mm in the output plane of the 350mm focal length mirror.

Next we replace mirror M_2 with a 500mm focal length ellipsoidal mirrors (to represent the test arrangement used to measure the transmission phase grating – see Figure 5-56). The aperture of this mirror is treated as a truncated circular aperture with radius a = 202mm but with a height of only 335mm ≈ 1.65a. Thus, more power at the top and bottom of the mirror plane ($|y| > 0.825a$) is truncated, as shown in Figure 5-84.

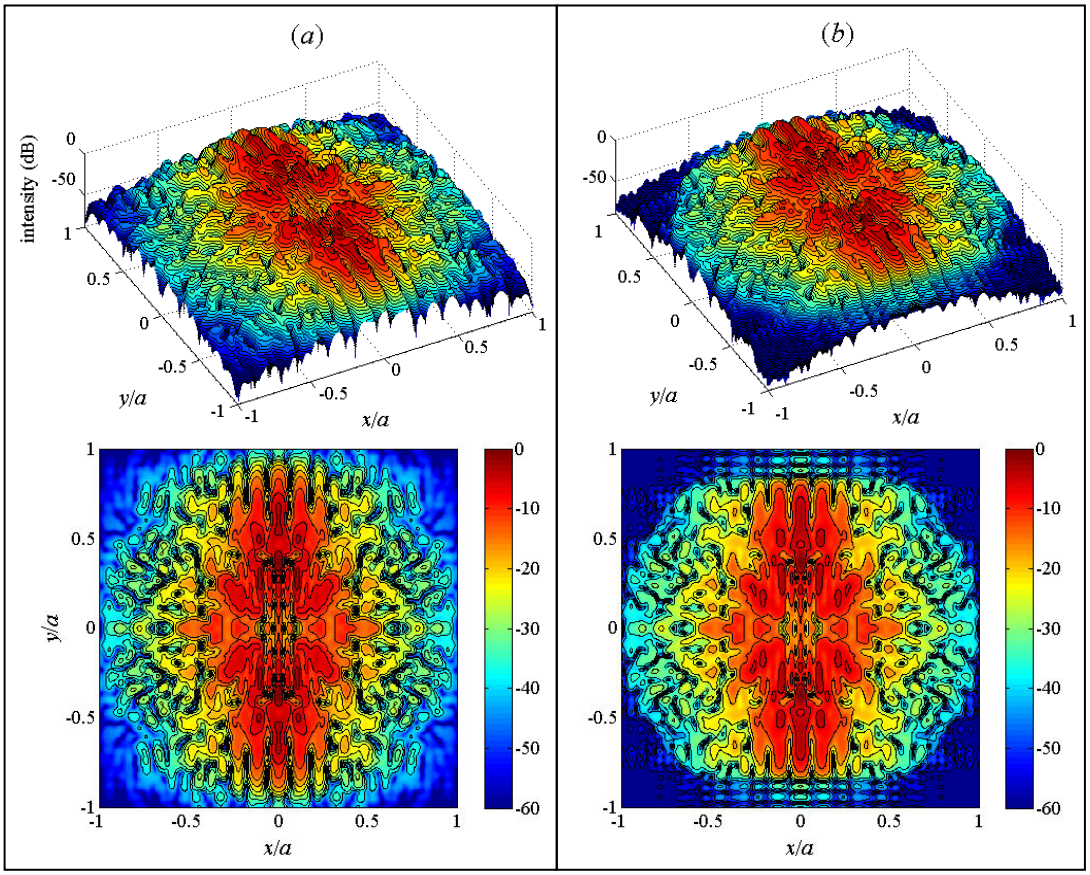


Figure 5-84. Log-scaled plot of beam intensity in the plane of mirror M_2 (*a*) before and (*b*) after truncation with a truncated circular aperture of radius a and a height of $1.65a$.

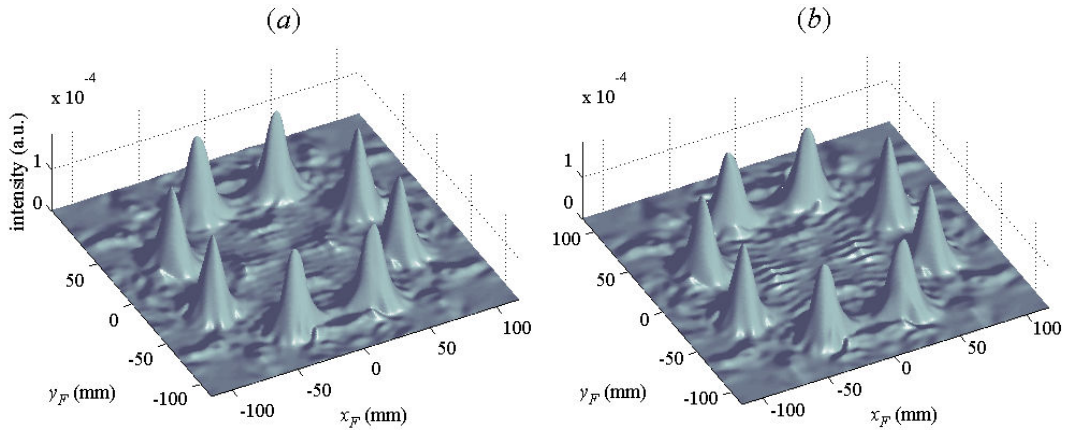


Figure 5-85. Output plane intensity (top) and amplitude (bottom) from the 8-beam Fourier grating in a Fourier optics test arrangement with mirrors M_1 and M_2 of focal lengths f_1 = 350mm and f_2 = 500m (*a*) without and (*b*) with truncation included at mirror M_2.

Figure 5-85 shows the calculated output plane intensity after truncation with the non-circular aperture of mirror M_2. As well as the upper and lower pairs of Gaussian beams being less intense than those on the left and right, truncation also introduces some low level intensity ringing in the central part of the image: horizontally aligned striped

features inside the circle of beams in Figure 5-85(*b*), which are presumably due to edge diffraction at the truncating aperture of M_2.

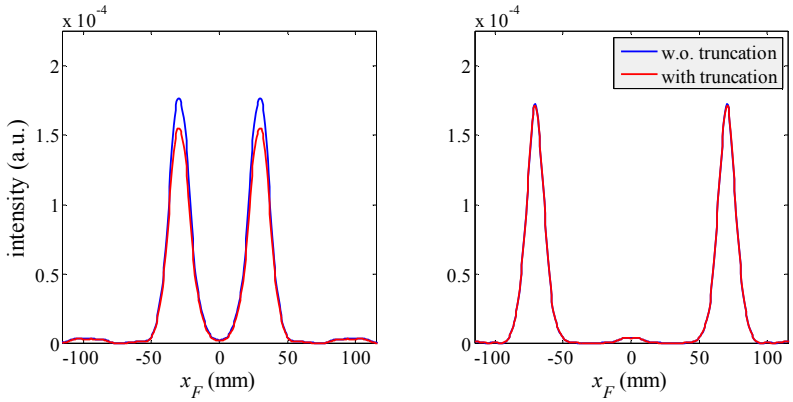


Figure 5-86. Horizontal cuts through the centre of the first and second row pairs of Gaussian beams at (*a*) y_F =70mm and (*b*) y_F = 30mm in the intensity plots of Figure 5-85. Truncation results in lower intensity in the top and bottom rows of beams, thus introducing substantial non-uniformity into the array.

Finally the effect of focusing onto the output plane with a hypothetical 500mm focal length mirror M_2 whose truncating aperture can be represented as a circular aperture (of radius a = 202mm) was simulated. Figure 5-87 shows x-cuts through the output plane intensity at the centre of the first and second rows of Gaussian beams. In this case truncation at M_2 does not reduce the beam uniformity because the top and bottom rows have the same intensity that they have without truncation at M_2.

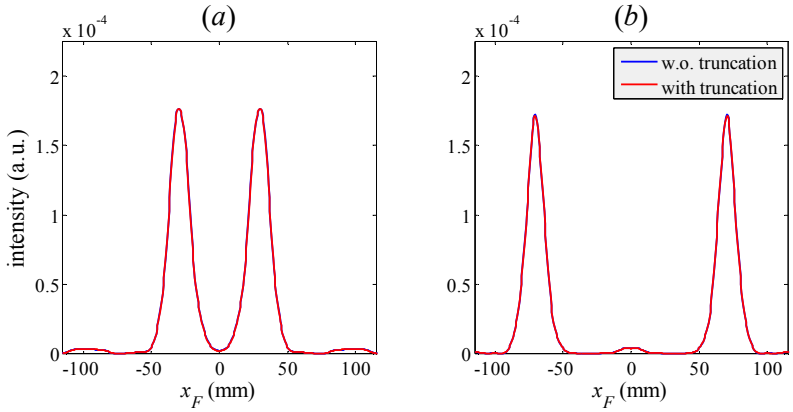


Figure 5-87. Horizontal cuts through the centre of the first and second row pairs of Gaussian beams at (*a*) y_F =71mm and (*b*) y_F = 30mm in the output plane intensity patterns. The fully circular aperture at M_2 means that now almost the entire wavefront transmitted from the grating is collected by M_2 and focused onto the output plane.

5.5 Chapter Conclusions

We have demonstrated the successful application of Gaussian beam mode analysis to the general phase retrieval problem: finding a suitable phase distribution to produce a general far field intensity pattern that is not necessarily periodic in nature. A two-dimensional phase grating to produce a sparse two-dimensional non-separable array of eight equally intense Gaussian beams was designed using a GBM-based iterative phase retrieval algorithm. The GBM-IPRA found a solution with a phase modulation that produced the required circular array of diffraction orders in the correct positions. Furthermore the distribution of power between the eight diffraction orders was relatively uniform. The problem was also tackled using the FFT-based IPRA and the solution found was similar to that found by the GBM-based algorithm. However both algorithms yielded less than ideal solutions. In particular the far field diffraction orders were not circular in profile due to a turbulent phase distribution across the face of each beam. Furthermore, points of zero intensity occurred within each output beam, which reduced the solution quality. The particular phase modulation was deemed to be an inappropriate candidate for direct translation into a surface relief profile so several phase unwrapping techniques were experimented with to find an equivalent smoother phase modulation that would be easier to manufacture. Two test gratings (in transmission and reflection) were machined in the department workshop. The far field diffraction patterns from the two gratings were experimentally measured with a number of Fourier optics test arrangements using different combinations of mirrors with different focal lengths and angles of throw. The measurements compared well with the results of numerical simulations developed in MODAL. The simulations of the grating with truncation included at the second mirror are in good agreement with the intensity measurements obtained, which (particularly in the case of the reflection grating) consistently exhibited lower intensity in the upper and lower pairs of Gaussian beams, relative to the other four beams. We were able to explain the sensitivity to illuminating beam position and size which the grating exhibited by analysing the diffraction pattern from small regions within the grating phase transformation.

Again MODAL was used for numerically simulating experimental arrangements with thin phase gratings. Any differences between simulated and measured images indicated possible misalignment errors in the experimental arrangement. MODAL was also most useful for modelling reflection phase gratings.

Chapter 6.

Conclusions

6.1 Gaussian Beam Mode Analysis

A major focus of the work described in this thesis was to apply Gaussian Beam Mode Analysis (GBMA) to the description of a variety of problems. If one were not constrained by limited computational resources a modal analysis using an infinite number of Gaussian beam modes would result in perfect reconstruction of a given wavefront. Of course in practise GBMA is performed using a finite number of modes and best use must be made of these modes. To illustrate this point examples of GBMA of one- and two-dimensional top-hat functions were presented in §2.8. It was shown that a set of Gaussian beam modes can be scaled (by appropriate choice of the value of W_0) in a variety of ways, only one of which will yield optimum correlation between the input field and GBM-reconstructed field.

In terms of verification of spatially filtered transmission imaging experiments, a GBMA model was developed that was able to produce intensity patterns that correlated extremely well with the experimentally acquired intensity images. The use of truncation analysis to account for the limited aperture sizes of the mirrors that were used in the system provided more insight into how image formation occurred within the system. Gaussian beam mode analysis was also applied to the study of phase gratings. For the analysis of Dammann gratings, which produce wavefronts with sharp phase modulations, the relationship between spatial frequency content of a complex-valued wavefront and the spatial periods of Gaussian beam modes was examined and used to determine mode-set size (number of modes) and scaling for a given problem. When used to analyse such wavefronts, the choice of highest-order mode index needed to describe the wavefront to a specified accuracy (that permits reconstruction of the highest spatial frequency components of the field) is then simply a matter of matching the spatial frequency of the highest-order mode to that of the input field. Although these rules were developed with phase gratings in mind, they can be used to accurately describe any arbitrary two-dimensional wavefront.

One of the attractions of using Gaussian beam mode analysis, over other scalar wave diffraction techniques is that it is a computationally efficient analysis tool. This is true in more traditional applications of GBMA, where reasonably accurate wavefront reconstruction can be achieved using only a small number of modes. Propagation of the reconstructed field through optical components, stops, apertures, etc. is then

straightforward and computationally efficient. However, it is clear from some of the examples described in this thesis (particularly in Chapters 3 and 4) that this advantage, over say Fresnel integrals, diminishes when applied to wavefronts with increasingly complicated beam profiles. The accurate modal description of a beam with a wide range of spatial frequencies requires the use of low- as well as high-order Gaussian beam modes. Furthermore, in order to be able to accurately describe the higher-order beam modes involved, the sample-spacing at the plane where reconstruction is performed must be sufficiently small. To ensure adequate sampling of Gaussian-Hermite beam modes the sample size must be equal to one-fifth the quasi-sinusoidal period of the highest-order mode in use.

The requirement to include many higher-order modes and also to densely sample these modes increases computational load. The pre-processing described in §2.8, that is used to determine which subsets of modes (classified according to symmetry properties) to use for decomposition of a given one- or two-dimensional field, reduces computational overhead by simply omitting redundant operations, i.e. by eliminating from consideration the evaluation of mode coefficients of modes that cannot contribute any power to the field. The other method that was used to make GBMA more efficient is to reduce the time taken to evaluate mode coefficients. As outlined in §2.8 this is achieved by avoiding numerical integration of the overlap integral and instead employing singular-valued decomposition (SVD) to calculate mode coefficients. Indeed the use of SVD was essential in facilitating the development of the GBM-based iterative phase retrieval algorithm (IPRA) described in Chapter 5. By using SVD to calculate the pseudo-inverse of a rectangular matrix of Gaussian-Hermite beam modes, the mode coefficients of the two-dimensional grating- and output-plane fields could be calculated in less than a second on a desktop computer. This allowed many more iterations to be evaluated in a fraction of the time needed otherwise. However, as was seen in §3.3, when large numbers of Gaussian beam modes and a high sampling density was required, the matrices involved became extremely large and SVD could no longer be used (because of the maximum matrix size permitted by MATLAB). In such cases, where one must resort to the much slower process of calculating the overlap integral to evaluate mode coefficients, alternatives to Gaussian beam mode analysis (e.g. Fast Fourier transforms) would seem the better option, as far as computational efficiency is concerned.

6.2 Imaging Experiments

One of the main aims of the project with which the author of this thesis was involved was to evaluate the usefulness of THz imaging, in particular for its application to the field of medical imaging. This goal was pursued through extensive experimentation with a number of different imaging systems. The majority of imaging work concentrated on transmission imaging using a system consisting of off-axis reflectors. Although many images were acquired, several difficulties were encountered. Firstly, we only had at our disposal the facilities to record beam intensity profiles. This limitation severely limited the use of commonly-used image recovery (deconvolution) techniques. In the intervening time since the experiments described in this thesis were performed, the THz Optics group at NUI Maynooth has acquired a vector network analyzer (VNA) which allows for measurement of both amplitude and phase at multiple frequencies. Future imaging experiments will no doubt benefit from the ability to measure a full complex-valued wavefront. As well as allowing for more accurate image recovery, it will also aid in the verification of numerical techniques by for example allowing for more detailed comparisons between experimentally obtained and numerically simulated data produced by simulations developed using MODAL.

Another problem with the transmission imaging experiments was the large amount of distortion introduced by the optics in the imaging system. This problem could be alleviated by reducing the angles of throw of the mirrors, or using an in-line system of lenses. However, standing waves could be an issue with the latter option. Alternatively a system, equivalent to the one that was used for near-field reflection imaging experiments, could be devised that requires no optics and hence would not incur the associated beam distortions. The preferred option would be to illuminate the object with a tightly-focused beam so as to concentrate beam power onto a small area. This would also remove the problem that was inherent in the particular transmission arrangement described in Chapter 3, which could only accommodate relatively small objects because of the small beam size produced by the optic that collimated the source beam. Such a system would have uniform contrast and illumination across the object plane and thus allow objects of any size to be imaged. The main conclusion to be drawn from the study of transmission imaging is that ones ability to obtain useful information in transmission is severely limited by the presence of water within an object, due to waters high absorption and reflectivity at these wavelengths. In contrast, the presence of

water in a reflection imaging system allows for contrast between objects because a sample with high water content appears highly reflective, whereas objects with low water content allow incident radiation to be transmitted through and so not detected. After various iterations of the experimental set-up, the results obtained from the near-field reflection experiments proved promising. The final experimental set-up, in which the source and detectors were fed using sections of bare waveguides, yielded the best results. Dynamic range could have been improved by simply including a thin reflecting sheet between the source and detector waveguides so as to reduce direct coupling between the two. Although standing wave effects were significant a routine to reduce their impact (by summation of images recorded at various distances) was developed and resulted in images that revealed structural information on the objects under test.

6.3 Phase Gratings

A significant part of the work described in this thesis was the investigation of phase gratings. Chapters 4 and 5 provided in-depth background information on the development of multiplexing, or beam-splitting, phase gratings in particular. The operation of Dammann gratings was described and the output from these devices explained in terms of Fourier analysis. A review of techniques used for the design of phase gratings by means of multivariable optimisation was presented in §4.3. All of these methods rely on some means of being able to determine the performance of a particular phase grating solution. Many different merit functions for determining grating performance are found in the literature on phase grating design. Some of the most commonly used are described in §4.2. As with Gaussian beam mode analysis, accounting for symmetry properties in phase grating design can reduce computational complexity. In §4.4 methods to account for both reflection and translational symmetry are described. Furthermore, an elegant interpretation of the operation of phase gratings with translational symmetry is presented in terms of Fourier theory.

In relation to phase gratings, Gaussian beam mode analysis was used in two capacities. Previously Dammann gratings were described in terms of a small number of Gaussian beam modes. In this thesis this approach was developed by using larger sets of Gaussian beam modes so as to be able to describe more closely the smallest feature sizes and hence the high spatial frequency content of these discrete-level diffractive

phase elements. The result is a model that more accurately predicts the far-field output, including higher off-axis features that were not accounted for in the previous work, but which are important for qualifying grating performance since grating diffraction efficiency is defined by the relative intensity of the required diffraction orders to the remaining orders.

In the examination of Fourier phase gratings, both Fourier techniques and Gaussian beam mode analysis were used for phase grating design through an iterative phase retrieval algorithm (IPRA). For a phase grating that is described in terms of Gaussian beam modes, the IPRA effectively acts as an efficient multivariable optimisation technique. The benefit of using this approach (over other optimisation techniques described in Chapter 4) is that the algorithm is easy to implement and a reasonably good solution can usually be found in a relatively small number of iterations (compared to stochastic methods). The GBM-based IPRA was used to find a solution that produces a sparse, circular array of eight uniformly intense Gaussian beams. The solution found by the algorithm was very good and managed to produce the required array, as well as completely eliminating the on-axis diffraction order. There is however room for improvement in the solution.

Analysis of the solution showed that the grating was extremely dependent on the size and position of the illuminating Gaussian beam. Thus when illuminated with a non-ideal beam (as was the case in some of the experimental arrangements) the observed intensity image was quite different from the ideal output. It was discovered that the choice of far-field beam size used in the algorithm was not appropriate for the size of the illuminating beam. The result of this mismatch in input and output beam sizes was that the problem was more akin to a beam-shaping problem, rather than a beam-splitting problem. An improved solution, that would produce a phase grating whose output is less dependent on input beam size and shape, but which would still produce the required output beam sizes could be achieved in one of two ways. One option would be to search for a beam-splitting problem that generates an array of far-field Gaussian beams appropriate to the size of the illuminating beam. Once a solution is found a quadratic ‘defocusing’ phase front could be added at the grating to create an array of Gaussian beams with a larger radius.

The other method would be to treat the problem as a beam-shaping problem. As was seen when performing phase unwrapping, the gratings phase has many vortices associated with it. This is characteristic of local solutions to beam-shaping phase

retrieval problems when additional iterations no longer yield improvements to the best-found solution. The problem of course, as with all deterministic algorithms, is that once a local solution has been found it is difficult for the algorithm to proceed to a better solution.

One proposed reason for stagnation in beam-shaping iterative phase retrieval is the presence of phase vortices. When the iterative algorithm is proceeding towards a solution, the number of phase vortices within the solution phase naturally decreases with increasing iterations. When the number of vortices no longer falls a local solution has been reached. The algorithm can only proceed to an improved solution if the vortex population can be made to decrease even further. This is achieved by globally altering the phase through *vortex-annihilation* (as was described in the section on Phase Unwrapping in §5.4.2). After vortex-annihilation, the algorithm proceeds to a better solution, until the next local solution is encountered.

Apart from limited success of IPRA the main deteriorating factor in image quality was the distortions introduced by the off-axis optics used in the test arrangements. Two possible solutions to this problem involve modifying the existing solution. One solution is to do avoid the distortions by removing the off-axis reflectors altogether. This can be achieved by combining the focusing function of a lens/mirror with the beam-splitting/-shaping function of a DPE. The result would be a compact, self-imaging DPE that requires no alignment. The other option is to redesign the phase grating so as to compensate for the distortions introduced by the focusing element. In this case one must know the amplitude and phase of the distorted wavefront at the output focal plane of the focusing element. Such a task is well suited to MODAL, thus making MODAL invaluable not only in the analysis of pre-defined optical systems but also as a tool for the design of quasi-optical components.

Appendices

Appendix A.

This Appendix provides details of how scalar wave propagation was implemented in numerical simulations using both Fresnel integrals and Fourier transforms. Also included here is the derivation of the complex beam parameter $q(z)$, which is required to apply the ABCD matrix method to propagation with Gaussian beam mode analysis.

A.1 Fresnel Integrals

In order to correctly calculate diffraction effects one must solve the wave equation subject to the boundary conditions imposed by the obstacle(s) causing diffraction

$$\frac{\partial^2 E}{\partial x^2}+\frac{\partial^2 E}{\partial y^2}+\frac{\partial^2 E}{\partial z^2}=\frac{x^2\,\partial^2 E}{\omega^2\,\partial t^2}$$

However this procedure is extremely complicated and only a few solutions exist for simple cases, such as for an infinite straight edge as calculated by Sommerfield. A simpler method involves making some approximations, which when applied are found to agree closely with experimental measurements.

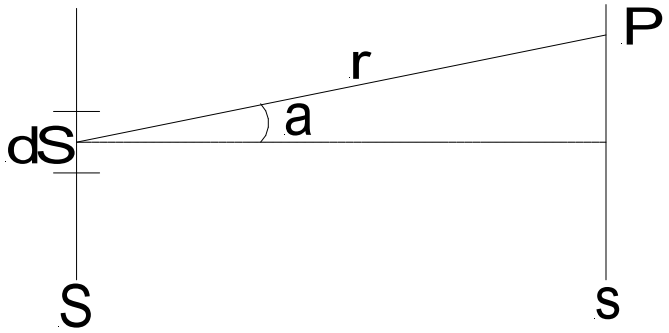


Figure A-1. Geometric construction showing contribution from a single interval dS from a field at plane Σ to the field strength at the point P at the plane σ.

Fresnel integrals are used to calculate scalar wave diffraction of a wavefront propagating from an input plane to an output plane (in either the near- or far-field of the input plane). Consider the construction shown in Figure A-1 where an aperture of length ΔS, situated in a plane Σ, is the source of a wavefront of wavelength λ with constant amplitude E_0 and with a phase that varies as $\exp[-i\,kr]$, where the magnitude of the wavevector $\boldsymbol{k}$ is given by $k = 2\pi/\lambda$. The contribution to the wave front at a point **P** on the screen σ has the form

$$dE_{\mathbf{P}} = \text{const} \times \frac{E_0\, e^{i(kr-\omega t)}}{r} dS \tag{A.1}$$

This is essentially a spherical wavelet as in the Huygen's treatment. The constant term depends on the source and the effects of polarisation of the field and can be derived rigorously by applying Green's Theorem in a full electromagnetic vector treatment. Since we are interested in relative, rather than absolute field intensity distributions this constant is omitted from further discussion.

If the aperture ΔS is divided into n equally spaced points separated from **P** by distances $[r_1, r_2, \ldots, r_n]$, the contribution to the field strength at **P** due to these points is given by the summation

$$E_{\mathbf{P}} = \left[E_0 \frac{e^{i(kr)}}{r_1} dS_1 + E_0 \frac{e^{i(kr)}}{r_2} dS_2 + \ldots + E_0 \frac{e^{i(kr)}}{r_n} dS_n \right] e^{-i\omega t} \tag{A.2}$$

$$E_{\mathbf{P}} = e^{-i\omega t} \sum_{\mathrm{i}\,=\,1}^{n} E_0 \frac{e^{i(kr)}}{r_{\mathrm{i}}}\, dS_{\mathrm{i}} \tag{A.3}$$

If the number of points n in the aperture ΔS is increased so that the spacing between points becomes vanishingly small the sum in equation (A.2) can be replaced by an integral such that the field strength at **P** is then given by

$$E_{\mathbf{P}} = \int E_0 \frac{e^{i(kr\,-\,\omega t)}}{r} dS \tag{A.4}$$

$$E_{\mathbf{P}} = \int E_0 \frac{\cos(kr - \omega t)}{r}\, dS + i \int E_0 \frac{\sin(kr - \omega t)}{r}\, dS \tag{A.5}$$

The power of a field crossing a surface is an extremely rapidly varying function of time ($\sim 10^{10}$ Hz for a wavelength of 3mm) making measurements of its instantaneous value impractical, so instead irradiance is measured. Irradiance is the average energy per unit area per unit time and is loosely referred as the "amount" of light illuminating a surface. The power per unit area that crosses a surface is represented by the magnitude of the Poynting vector,

$$\overline{\mathbf{S}} = c^2 \varepsilon_0\, \overline{\mathbf{E}} \times \overline{\mathbf{B}} = c^2 \varepsilon_0 \left[\mathbf{E}_0 \times \mathbf{B}_0 \cos^2(kr - \omega t) \right] \tag{A.6}$$

where the **E** and **B** fields are given by

$$\overline{\mathbf{E}} = \mathbf{E}_0 \cos(kr - \omega t), \quad \overline{\mathbf{B}} = \mathbf{B}_0 \cos(kr - \omega t) \tag{A.7}$$

The irradiance is the time-averaged value of the Poynting vector expressed as

$$\mathrm{I} = \left\langle \mathbf{S} \right\rangle_T = c^2 \varepsilon_0 \left| \mathbf{E}_0 \times \mathbf{B}_0 \right| \left\langle \cos^2(kr - \omega t) \right\rangle_T \tag{A.8}$$

where $\langle f \rangle_T$ denotes the time-averaged value of a function f over a time interval T. For $T >> \tau$, $\langle \cos^2(kr - \omega t) \rangle_T = ½$, and since the E-field is more effective at doing work than the B-field the irradiance can be written as

$$\mathrm{I} = \frac{c\varepsilon_0}{2}|\bar{\mathbf{E}}|^2 \tag{A.9}$$

If we are interested only in relative intensity, the constant term is omitted and the intensity at point P is proportional to the field modulus at that point

$$\mathrm{I_P} = \mathbf{E_P}\mathbf{E_P}^* \tag{A.10}$$

with $\mathbf{E_P}^*$ being the complex conjugate of $\mathbf{E}$.

$$\mathrm{I_P} = \left[\int E_0 \frac{\cos(kr)}{r}\, dS\right]^2 + \left[\int E_0 \frac{\sin(kr)}{r}\, dS\right]^2 = C^2 + S^2 \tag{A.11}$$

where the integrals C and S are the Fresnel's integrals, each of which can be computed separately. Since the field intensity is proportional to the squared magnitude of the E-field, the amplitude at point P is of course given by

$$\mathrm{A_P} = \sqrt{C^2 + S^2} \tag{A.12}$$

Meanwhile the value of the phase is given by the argument of the E-field as

$$\phi_\mathrm{P} = \mathrm{Arg}\{\mathbf{E}\} = \tan^{-1}(S/C) \tag{A.13}$$

In practice when numerically evaluating Fresnel integrals the fields at both input and output planes are represented by regularly sampled arrays so the integrals in Eq. (A.11) are replaced by summations (the rectangular rule) of the form

$$C = \sum_{-L/2}^{+L/2} E_0 \frac{\cos(kr)}{r} \tag{A.14}$$

$$S = \sum_{-L/2}^{+L/2} E_0 \frac{\sin(kr)}{r} \tag{A.15}$$

where, for a one-dimensional input plane of length L, summations are calculated from $x = -L/2$ to $x = +L/2$. The process of calculating the contributions by all points on the input plane to the field strength at point P is repeated for all points in the observation plane. A more accurate formulation based on the trapezoidal rule is also possible.

Consider the case where the opaque screen Σ containing the single small aperture of length dS is illuminated by plane waves from a very distant point source S. If the plane of observation σ, is a screen parallel with, and very close to, Σ an easily recognisable image of the aperture along with some fringing will be projected onto the

screen σ. This phenomenon is known as Fresnel or near-field diffraction. As the screen is moved further from the aperture the fringes will change continuously until at a great distance from the aperture the observed pattern will no longer resemble the aperture. If the distance is further increased the only change in the pattern will be its size but not its shape, a phenomenon called Fraunhofer or far-field diffraction. As a rule of thumb, Fraunhofer diffraction occurs at an aperture (or obstacle) of width a when

$$R > a^2/\lambda$$

where R is the smaller of the two distances S to Σ and Σ to σ. A practical method of ensuring far-field diffraction occurs is by introducing a lens between source and aperture and between aperture and screen, thereby effectively placing both source and screen at infinity (as shown in Figure A-2).

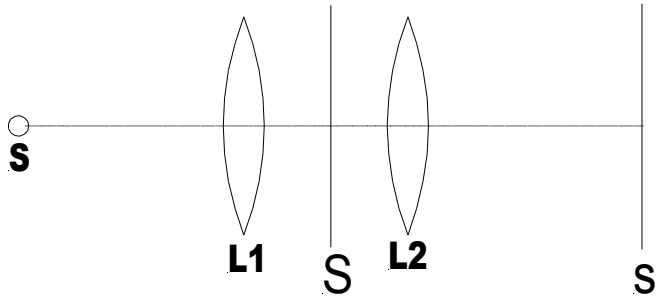


Figure A-2. The inclusion of two lenses, one before and one after the aperture ensures that the observation plane S is in the Fraunhofer (far-field) diffraction region of the aperture.

The Huygens-Fresnel Principle does not account for variations in amplitude with changing off-axis angles θ over the surface of secondary wavefronts. In Fraunhofer diffraction the distance from aperture to plane of observation is so large that changes in θ are negligible. However where Fresnel (near-field) diffraction is concerned the approximations made under the Huygens-Fresnel Principle is insufficient to adequately describe the observed diffraction effects, thus an obliquity, or inclination factor $K(\theta)$ must be introduced, which is defined such that

$$K(0) = 1, K(\pi/2) = 0$$

The obliquity factor used in simulations was

$$K(\theta) = \frac{1 + \cos\theta}{2} = \frac{1 + (z/r)}{2} \qquad \text{(A.16)}$$

A practical problem with Fresnel integrals is that in two dimensions computing equations (A.14) and (A.15) can become restrictive since one must integrate over four dimensions: the two transverse dimensions (x and y) at the input and output planes. In

terms of programming this would normally require four nested for loops. However computational overhead can be significantly reduced by taking advantage of MATLAB's ability to perform fast matrix calculations. To calculate the wavefront at point (x_i, y_j) on the output plane a two-dimensional matrix whose entries are the propagation distances r from all points on the input plane to (xi, y_j) is first created. Matrix multiplication is then used to calculate the value of $E(i, j)$ due to contributions from all points on the input grid. This process is repeated for all points on the output grid, thus requiring only two nested for loops.

A.2 Fourier transforms for computing scalar wave diffraction

The Fourier transform of an optical wavefront is equivalent to calculating the far-field, or Fraunhofer diffraction pattern due to that wavefront. The simple lens constitutes a Fourier-transform computer, which is capable of transforming a complex two-dimensional pattern into a two-dimensional transform at the speed of light. The diffraction pattern of a spatial object formed by a lens can be shown to be a two-dimensional Fourier transform, or spectrum, of the input [A.1]. Spatial filtering of the input can be achieved using masks or filters at the Fourier plane so as to manipulate the final image produced by a second lens – a technique that was exploited in transmission imaging experiments presented in Chapter 3. A lens is a Fourier transforming device since it produces the far-field, or Fourier Transform (FT) of the input wavefront at its output plane. In this section practical aspects relating to the use of Fourier transforms for computing scalar wave diffraction of paraxial beams with particular emphasis on the use of Fast Fourier transform algorithms are described. The fundamentals of Fourier theory as an analysis tool for optical simulation are not covered but can be found in many good resources including [A.2] and [A.3].

A discrete Fourier transform (DFT) algorithm is used to compute the Fourier transform of discretely sampled one- or two-dimensional signals. A fast Fourier transform (FFT) algorithm is an efficient means of computing the DFT, many variations of which exist today. Simulations involving FFT computations described in this thesis were performed using the MATLAB functions `fft.m` and `fft2.m` (and their inverses `ifft.m` and `ifft2.m`), which are based on the algorithm of Cooley and Tukey [A.5]. The actual way that these functions must be used to compute the Fourier transform of an input signal, or field is non-trivial, however.

A DFT operates solely on a discretely sampled array of (possibly complex) numbers (representing the signal values) without any information about the field of reference in which the input field is defined. Therefore the Fourier spectrum that results from applying a DFT to an input field is also without any frame of reference and is just a series of numbers. Hence the user must interpret how the sampling interval expressed in the chosen units for the coordinate frame in the object plane relate to the sampling interval and appropriate units for the coordinate frame in the image plane. The Fourier transform of an optical distribution that is defined in terms of spatial coordinates (x, y) will produce a spectrum of the input beam, with values defined in terms of spatial frequencies (u, v). As will be seen, the physical spacing between samples in the input plane (i.e. Δx and Δy) determines the range of spatial frequencies in which the Fourier spectrum is defined. Furthermore the maximum spatial frequency is inversely proportional to sample spacing. An important consideration when using an N-point FFT (like the ones used in MATLAB) is that the function takes as input an N-point array and returns an output array of the same size. Thus the DFT of the input field by itself may produce a crudely sampled Fourier spectrum. However, when dealing with scalar diffraction one is only interested in the central part of the spectrum (that is confined to a narrow angular spread) within which the paraxial approximation is valid. This means that in order to be able to extract that part of the spectrum that is of interest with reasonable resolution (e.g. with the same number of data points as the input field), the Fourier spectrum must be over-sampled. This is achieved by appending extra zeros to the input field. So-called zero-padding can be done in one of two ways: trailing zeros can be added to the end of the input field array, or the input array can be inserted into a zero-valued array of the appropriate size. The latter seems a more intuitive approach, however the FFT functions in Matlab are designed for zero-padding with trailing zeros.

Use of the FFT functions in MATLAB is now illustrated by calculating the Fourier transform of a specific field: that from a binary-level phase grating when illuminated with a collimated Gaussian beam (the amplitude and phase distributions of which are shown in Figure A-3). Given the two-dimensional field E at the grating plane we wish to calculate its far-field diffraction pattern, which is given by the Fourier Transform of E.

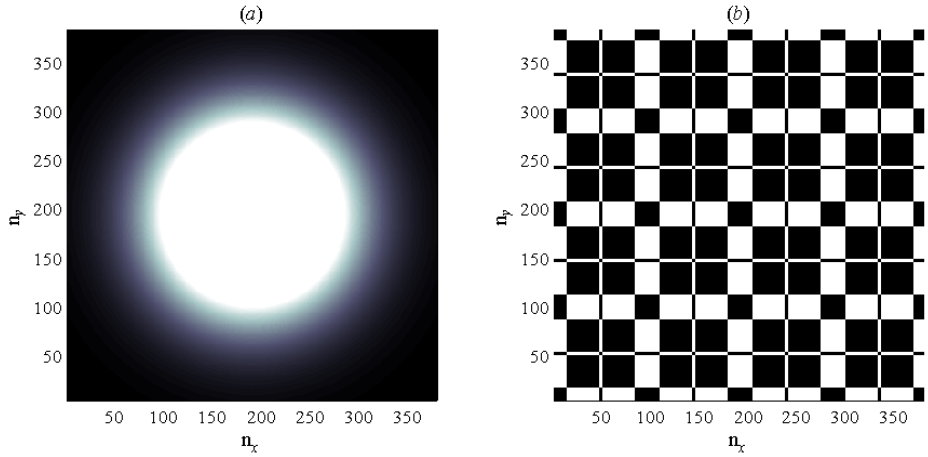


Figure A-3. (*a*) Amplitude and (*b*) phase distributions associated with a Dammann grating (a binary phase grating) when illuminated with a collimated Gaussian beam. The grating field E_G is represented as a 2-D array of size $[n_x, n_y] = [380, 380]$. The colour axis in (*a*) has been scaled to the range $[0,e^{-1}]$ to make the beam width ($2\times W_G$) more obvious.

Referring to the phase grating field as E_G the far-field, or Fourier plane field, E_F is equal to the Fourier transform of E_G which is implemented in MATLAB with the syntax

```
>> Ef = fft2(Eg);
```

where `Eg` is a 2-D array of numbers representing the grating field distribution. Figure A-4(*a*) shows the amplitude distribution resulting from the above operation. Most power is concentrated in the four corners with very little contained at the centre. This plot serves to illustrate the way in which a DFT outputs Fourier-transformed data: (in one-dimension) the zero-frequency component is located at one end, the positive frequency spectrum occurs next, followed by the negative frequency components in positions which do not correspond to a proper ordering in Fourier space.

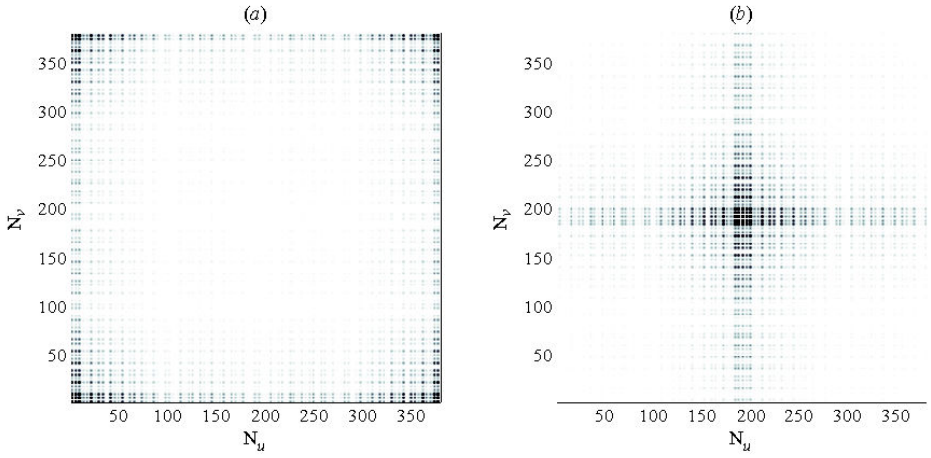


Figure A-4. Plots of (negative) amplitudes of arrays produced by taking the 2-D FFT of E_G (*a*) before and (*b*) after quadrants have been swapped into their correct positions using the function `fftshift.m`.

In one-dimension the array output from the DFT is arranged as shown in the upper part of Figure A-5. The output array consists of N elements, with array indices $i = [0,\ldots,N-1]$. Each element corresponds to the Fourier transformed signal value at a unique spatial frequency f_i. Because of the periodicity implied in the Fourier transform the maximum and minimum frequencies $+f_c$ and $-f_c$ are equal so only one is included in the output of the DFT. Here f_c is the Nyquist critical frequency and all output frequencies f_i lie in the range $[-f_c,\ldots,+f_c]$. The frequencies can be rearranged into correct ascending order by swapping the lower-half and the upper-half of the output array so that the zero-frequency component f_0 is now centred (lower part of Figure A-5). Similarly a re-ordering of the elements in a two-dimensional array is achieved by swapping the 1st quadrant with the 2nd and the 3rd with the 4th. Quadrant swapping is implemented in MATLAB with the function `fftshift.m`, which rearranges the quadrants of the Fourier spectrum such that zero-frequency component of the Fourier spectrum as illustrated in Figure A-4(*b*) for the amplitude of the Fourier transform of the Dammann grating field. Power is now concentrated at the centre of the spectrum and falls off with increasing off-axis distance, as expected.

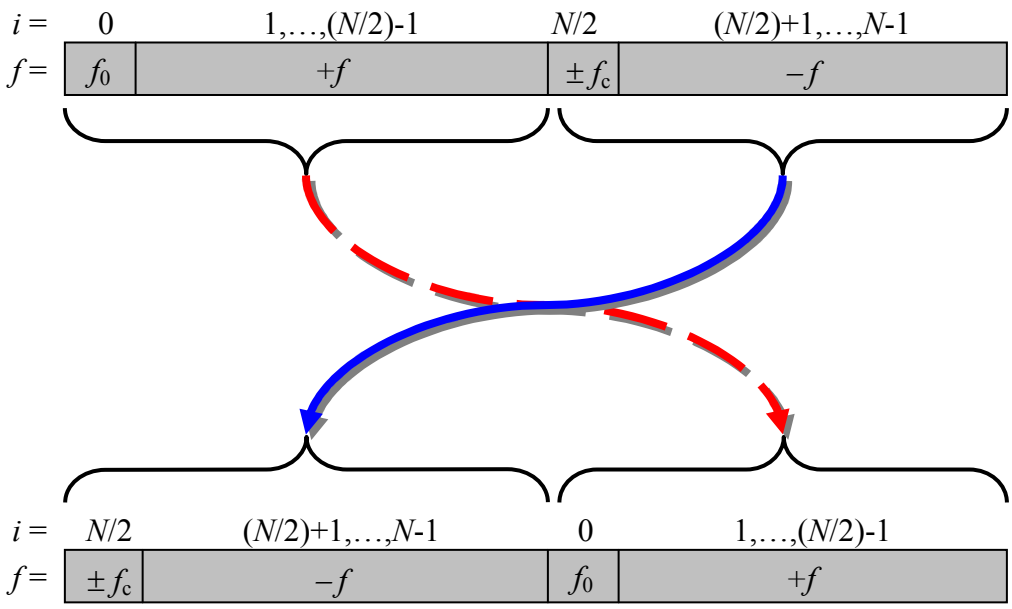


Figure A-5. Rearranging the ordering of elements in a 1-D array (upper) output by a DFT, showing the relationship between array indices i and the spatial frequencies f_i – the coordinates in which the Fourier spectrum is defined. Frequency re-ordering (or quadrant swapping in 2-D) produces the correct ascending order (lower).

Depending on how zero-padding of the input is performed an additional quadrant swap may be required. According to Wilson [A.4], as well as the quadrant swap that one must perform on the Fourier transformed field, an additional quadrant swap must be performed on the input field before calculating the DFT, using the following syntax

```
>> Ef = fftshift(fft2(fftshift(Eg)));
```

This extra quadrant swap is due to the way that MATLAB implements the fast Fourier transform: it assumes that zero-padding of the input is performed by adding trailing zeros to the input array. A more natural method is to pad the input array symmetrically (with the same number of zeros appended before and after the input array) so that the input signal is located at the centre of the padded array. If this is the case an extra quadrant shift is required to make the padded array suitable for use with the FFT function `fft2.m`.

Figure A-6(*a*) shows the phase distribution ϕ_F that was extracted from the Fourier transformed grating field E_F when quadrant swapping was performed only after the DFT was calculated. Figure A-6(*b*) shows ϕ_F after quadrant swapping was performed on both the input and output planes with the above syntax. Although the pre-DFT quadrant swap has no effect on the amplitude of the spectrum it is clearly needed to give the correct phase distribution: in this case large regions of uniform phase – as opposed to the rapidly varying phase distribution shown in Figure A-6(*a*).

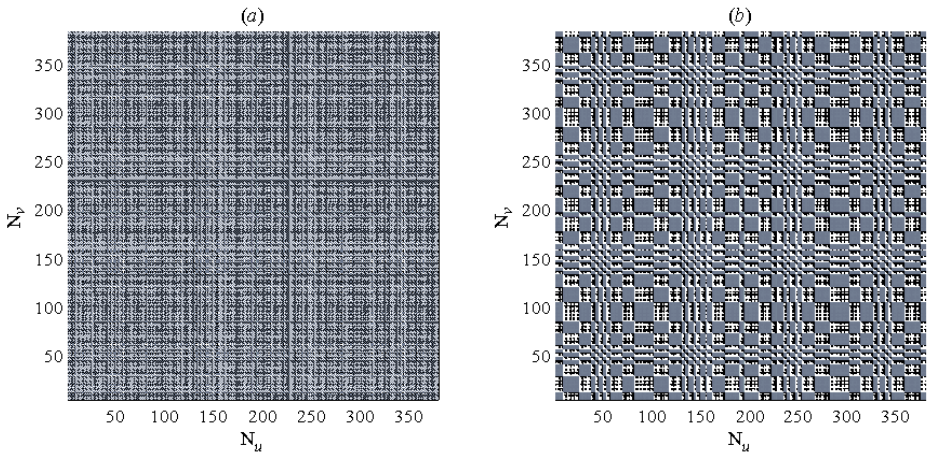


Figure A-6. Phase distributions extracted from the array produced (*a*) without and (*b*) with quadrant swapping at the object (grating) plane before the DFT is calculated.

So far the amplitude and phase distributions of the Fourier transformed grating field have been plotted against sample numbers. Obviously it would be more useful to be able to display the position of features in these images in terms of more meaningful units. In the object, or input, domain the grating profile is defined in spatial coordinates

(x_i, y_i), in units of mm, while in the Fourier domain we can describe the feature positions in terms of spatial frequency (u, v), in units of mm^{-1}. In one dimension, the input sample spacing Δx_i is given by

$$\Delta x_i = \frac{L}{n-1}$$

where L is the aperture width (in which the input field is defined) and n is the number of samples. Any discretely sampled field is bandwidth limited and should be sampled with a sufficiently high sample rate to avoid aliasing, which occurs when frequency components outside the frequency range are incorrectly translated into that range. From the sampling theorem, aliasing can be avoided by sampling at a rate equal to twice or more of the maximum frequency component in the signal, referred to as the Nyquist critical frequency, f_c. Thus if the sample spacing Δx fulfils the following criterion

$$\Delta x_i \leq \frac{1}{2f_c} \tag{A.17}$$

then aliasing will not occur and the Fourier transform of the input field has zero amplitude outside the frequency rage $[-f_c, +f_c]$.

The Fourier transform of an input array with N elements produces an output array with the same number of elements. Thus the N-point Fourier spectrum consists of discrete samples that are separated in the spatial frequency domain by

$$\Delta f = \frac{\max\{f\} - \min\{f\}}{N\text{-}1}$$

where $\min\{f\} = -f_c$ and $\max\{f\} = +f_c - \Delta f$ thus

$$\Delta f = \frac{2f_c}{N}$$

Now $f_c = \frac{1}{2}(N\Delta f)$, which upon substitution into equation (A.17) yields the following expression equivalent to the sampling criterion but in terms of spacing in the spatial and frequency domains

$$N \leq \frac{1}{\Delta f \Delta x_i} \tag{A.18}$$

After the zero-frequency component f_0 has been centre (see Figure A-5), the spatial frequencies spanned by the Fourier spectrum are then

$$f = q\Delta f, \qquad q = [-(N/2), \ldots, (N/2)-1]$$

and in two dimensions spatial frequencies are denoted either by (f_x, f_y) or by (u, v).

Now we relate the spatial coordinates (x_i, y_i) in the input, or object, plane to the spatial coordinates (x_o ,y_o) in the image, or output, plane. In one-dimension the discrete Fourier transform $E_o(q)$ of an input field $E_i(p)$ is expressed as

$$E_o(q) = \sum_{p=0}^{N-1} E_o(p)\, e^{-i\frac{2\pi}{N}(pq)} \quad p = [0,\ldots N\text{-}1],\ \ q = [0,\ldots,N\text{-}1] \tag{A.19}$$

evaluated at spatial frequencies $f(q)$. The one-dimensional Fresnel diffraction integral is

$$E_o(x_o) = \frac{e^{ikz}}{i\lambda z}\, e^{\frac{ik}{2z}x_o^2} \int_{-\infty}^{+\infty} \left[E_i(x_i)\, e^{\frac{-ik}{z}(x_o x_i)} \right] e^{\frac{ik}{2z}x_i^2}\, dx_i$$

which, when valid, is appropriate for computing the distribution E_o in the Fresnel or near-field region of the input plane field E_i. If the output plane is situated at a large distance z from the input plane such that

$$z >> \max\left\{\frac{k x_i^2}{2}\right\}$$

then the quadratic term on the right-hand side is approximately equal to unity, in which case the output plane distribution $E_o(x_o)$ is simply given by the Fourier transform of the input function $Ei(x_i)$, i.e.

$$E_o(x_o) = \frac{e^{ikz}}{i\lambda z}\, e^{\frac{ik}{2z}x_o^2} \int_{-\infty}^{+\infty} E_i(x_i)\, e^{\frac{-ik}{z}(x_o x_i)}\, dx_i$$

Then equating the exponential term of this, the Fraunhofer limited diffraction integral with that of the discrete Fourier transform in equation (A.19) as follows

$$\exp\left[-i\frac{2\pi x_o x_i}{\lambda z}\right] = \exp\left[-i\frac{2\pi(pq)}{N}\right]$$

where the wavenumber $k = 2\pi/\lambda$. The spatial frequencies, $f(q)$ can then be written in terms of the spatial coordinates in the output plane as

$$f = \frac{x_o}{\lambda z}$$

Solving for the number of points, N in the DFT yields

$$N = \frac{\lambda z (pq)}{x_i\, x_o} \tag{A.20}$$

Where the spatial coordinate arrays at the input and output planes are related to their respective sample spacing Δx_i and Δx_o as follows

$$x_i = p\Delta x_i, \qquad x_o = q\Delta x_o$$

which, upon substitution into equation (A.20) yields the following expression relating the number of points used in the DFT to the desired sample spacing at the input and output planes

$$N = \frac{\lambda z}{\Delta x_i \Delta x_o} \tag{A.21}$$

Clearly this expression implies that output plane sample spacing Δx_o is inversely proportional to both input plane sample spacing, Δx_i and N. Thus although the value of Δx_i may be fixed for a given input field, higher output plane resolution (a smaller value of Δx_o) can be achieved by simply increasing N, the size of the array that is fed as input to the DFT. Given a one-dimensional input array E_i with n samples, a padded N-point input array is created by simply appending (N-n) zeros to array E_i. The output plane sample spacing can also be related to spatial frequency spacing by equating (A.18) with (A.21) to yield

$$\Delta x_o = \lambda z \Delta f \tag{A.22}$$

When dealing with diffraction patterns produced by phase gratings, the diffraction order positions are specified (by the grating equation) in terms of angles. Thus it useful to be able to specify the Fourier spectrum in terms of angular coordinates, (θ_x, θ_y) as well as spatial frequencies (u, v). Furthermore since a lens acts as a Fourier transformer angular coordinates can be converted to the spatial coordinates (x_o, y_o) of the output plane of a lens as follows

$$x_o = f \tan\theta \approx f\,\theta$$

where here f is the focal lengths of a lens. The spacing $\Delta\theta$ between samples in the Fourier spectrum in terms of angular coordinates is derived as follows. At a finite propagation distance z from the input plane, $\tan\theta = x_o/z$ thus

$$\Delta x_o/z = \Delta(\tan\theta) = (\sec^2\theta)\,\Delta\theta$$

where $\sec^2\theta = [1 + \tan^2\theta] = [1 + (x_o/z)^2]$ and therefore

$$\Delta\theta = \frac{\Delta x_o/z}{[1 + (x_o/z)^2]} \approx \frac{\Delta x_o}{z} \tag{A.23}$$

The maximum and minimum angles subtended by the Fourier spectrum are then given by -(N/2)$\Delta\theta$ and (N/2-1)$\Delta\theta$. However the paraxial approximation is only valid within a narrow angular spread: the central part of the spectrum (within approximately ±15°), referred to as the paraxial region. If the paraxial region is to be sampled with n points, then an appropriate value of $\Delta\theta$ must be used. Solving for Δx_o in equation (A.23) and

upon substitution into equation (A.21), the required array size N with which the DFT must be computed is given by

$$N = \frac{\lambda}{\Delta\theta \Delta x_i} \tag{A.24}$$

which ensures an angular spacing of $\Delta\theta$ between samples within the paraxial region of the Fourier spectrum. Figure A-7 shows the paraxial region of the Fourier spectrum produced by the phase grating in Figure A-3 in which zero-padding of the grating field was used in order to yield a 2-D output array within the angles

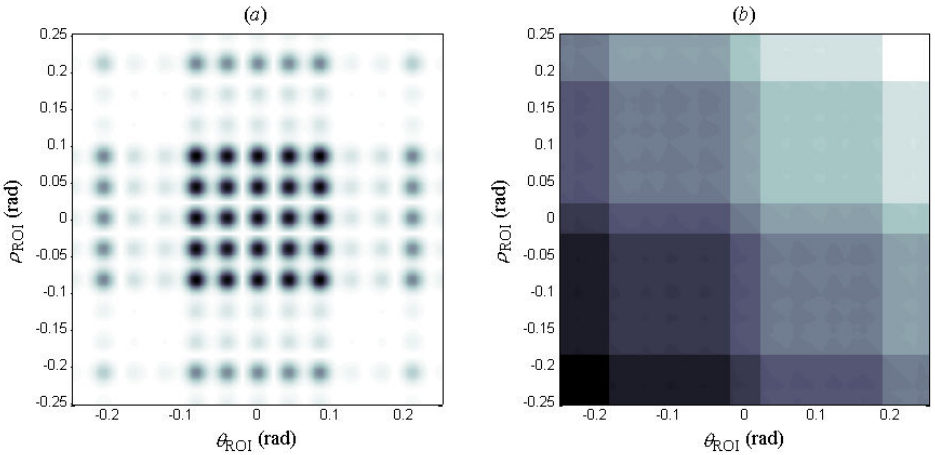


Figure A-7. (*a*) Amplitude and (*b*) phase distributions of the Fourier transform of the phase grating shown in Figure A-3. Zero-padding was used to reproduce the paraxial region (at angles $|\theta_x|$, $|\theta_y| \leq 0.25$ rad) with the same number of sample points that were used to describe the unpadded grating field.

Fresnel (near-field) propagation with Fourier Transforms

The field transmitted through a rectangular aperture (for example that of a phase grating) defined at a plane in Cartesian coordinates (x_i, y_i) is described by a complex electric field $E_i(x_i, y_i)$. The resulting field $E_o(x_o, y_o)$ at a finite propagation distance z from the input plane is calculated by integrating over the input field as follows

$$E_o(x_o, y_o) = \frac{e^{ikz}}{i\lambda z} \int\int E_i(x_i, y_i) e^{\frac{ik}{2z}\left[(x_i - x_o)^2 + (y_i - y_o)^2\right]} dx_i dy_i$$

The expression in square brackets inside the integral (resulting from the binomial expansion of r_{oi}) can be expanded as

$$\left[(x_o - x_i)^2 + (y_o - y_i)^2\right] = (x_o^2 + y_o^2) - 2(x_o x_i + y_o y_i) + (x_i^2 + y_i^2)$$

Factoring the term $\exp\left[\frac{jk}{2z}(x_o^2 + y_o^2)\right]$ outside the integral yields

$$E_o(x_o, y_o) = \frac{e^{ikz}}{i\lambda z} e^{\frac{ik}{2z}(x_o^2 + y_o^2)} \int\int E_i(x_i, y_i) e^{\frac{-ik}{z}(x_o x_i + y_o y_i)} e^{\frac{ik}{2z}(x_i^2 + y_i^2)} dx_i dy_i \tag{A.25}$$

which (aside from multiplicative factors) is the Fourier transform of the product of the complex field just to the right of the aperture with a quadratic exponential phase term

$$\phi_i(x_\mathrm{i}, y_\mathrm{i}) = e^{\frac{ik}{2z}(x_\mathrm{i}^2 + y_\mathrm{i}^2)}$$

This result given by equation (A.25) is referred to as the *Fresnel diffraction integral*. When the approximation is valid, the observer (at the output plane) is said to be in the region of Fresnel diffraction, or equivalently in the near field of the aperture. In one dimension equation (A.25) takes the form

$$E_\mathrm{o}(x_\mathrm{o}) = \frac{e^{ikz}}{i\lambda z} e^{\frac{ik}{2z}x_\mathrm{o}^2} \int \left[E_\mathrm{i}(x_\mathrm{i})\, e^{\frac{-ik}{z}(x_\mathrm{o}x_\mathrm{i})} \right] e^{\frac{ik}{2z}x_\mathrm{i}^2} dx_\mathrm{i}$$

In MATLAB the one-dimensional Fast Fourier Transform pair, as implemented with the function `fft.m`, is expressed as

$$X(q) = \sum_{p=0}^{N-1} x(p)\, e^{-i\frac{2\pi}{N}(p-1)(q-1)} \tag{A.26}$$

where x and X are two one-dimensional vectors of length N, elements of which are indexed with integers p and q, respectively, where $p = (0, 1, 2, \ldots, N\text{-}1)$ and $q = (0, 1, 2, \ldots, N\text{-}1)$. However since vector indexing in MATLAB begins at element 1 instead of zero equation (A.26) becomes

$$X(q) = \sum_{p=1}^{N} x(p) e^{-i\frac{2\pi}{N}(pq)} \tag{A.27}$$

The Fresnel diffraction integral that was expressed previously for continuous functions can now be rewritten to handle discrete data sets (to make use of the one- and two-dimensional FFT functions `fft.m` and `fft2.m` in MATLAB). The integral becomes a summation in which spatial coordinates x_i and x_o are indexed by integers p and q. The one-dimensional Fresnel diffraction integral can thus be expressed in terms of the discrete Fourier transform pair as

$$E_\mathrm{o}(q) = \frac{e^{ikz}}{i\lambda z} e^{\frac{ik}{2z}x_\mathrm{o}(q)^2} \sum_{p=1}^{N} \left[E_\mathrm{i}(x_\mathrm{i}(p))\, e^{\frac{-ik}{z}(x_\mathrm{o}(q)x_\mathrm{i}(p))} \right] e^{\frac{ik}{2z}x_\mathrm{i}(p)^2} \tag{A.28}$$

This discrete Fresnel transform can be evaluated by taking the Fourier transform of field $E_\mathrm{i}(x_\mathrm{i}, y_\mathrm{i})$ after multiplication with the exponential phase term

$$\phi_\mathrm{i}(p) = e^{\frac{i\pi}{\lambda z}x_\mathrm{i}(p)^2}$$

The phase term outside the summation is a constant for each point $x_\mathrm{o}(q)$, which we will refer to as

$$\phi_o(q) = \frac{e^{ikz}}{i\lambda z} e^{\frac{ik}{2z} x_o(q)^2}$$

and is included for completeness when calculating E_o, which over the output plane is given by

$$E_o(x_o, y_o) = \phi_o(x_o, y_o) \cdot \Im\{\phi_i(x_i, y_i) E_i(x_i, y_i)\}$$

where $\Im\{\ \}$ represents the Fourier transform of the bracketed quantity.

Over very small propagation distances the Fresnel Transform yields inaccurate results and requires the use of a double Fresnel Transform. This involves calculating the Fresnel transform of the input field at a large propagation distance s from the input plane. The resultant field is then transformed a distance $(s–z)$ in the opposite direction back towards the input plane. In simulations that involved the use of Fresnel transforms the desired propagation distance was first compared to the size of the input plane in order to determine whether or not it was necessary to use a double Fresnel transform.

A.3 The Complex Beam Parameter, *q*(*z*)

The notion of a complex Gaussian source is required to be able to apply the ABCD matrix method (described in §2.5) to the analysis of quasioptical systems. From this concept one can also derive expressions presented in §2.3.2 for the rate of evolution of the radius $W(z)$ and radius of curvature $R(z)$ of a propagating Gaussian beam.

In cylindrical coordinates the paraxial wave equation is

$$\frac{\partial u^2}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r}\frac{\partial^2 u}{\partial \varphi^2} - 2ik\frac{\partial u}{\partial z} = 0 \qquad \text{(A.29)}$$

where r represents perpendicular distance from the axis of propagation (the z-axis), φ represents angular coordinates and $u = u(r, \varphi, z)$. Assuming axial symmetry, the third term in equation (A.29) equals zero so the axially symmetric paraxial wave equation is

$$\frac{\partial u^2}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} - 2ik\frac{\partial u}{\partial z} = 0 \qquad \text{(A.30)}$$

the simplest solution of which has the form

$$u(r, z) = A(z)\ \exp\left[\frac{-ikr^2}{2q(z)}\right] \qquad \text{(A.31)}$$

where $A(z)$ and $q(z)$ are complex functions of z which can be solved by substituting equation (A.31) into (A.30) to give

$$-2ik\left(\frac{A}{q}+\frac{\partial A}{\partial z}\right)+\frac{k^2r^2A}{q^2}\left(\frac{\partial q}{\partial z}-1\right)=0 \tag{A.32}$$

Since this equation must hold true for all values of r and z and given that the first term depends on z only while the second term depends on both r and z, both terms must individually equal zero, thus yielding the following simultaneous conditions which must be satisfied

$$\frac{\partial A}{\partial z}=-\frac{A}{q} \qquad \text{and} \qquad \frac{\partial q}{\partial z}=1 \tag{A.33}$$

The second of these has the solution

$$q(z)=q(z_0)+(z-z_0) \tag{A.34}$$

Defining the reference position in z to be $z_0 = 0$ gives

$$q(z)=q(0)+z \tag{A.35}$$

where function $q(z)$ is called the complex beam parameter or Gaussian beam parameter. In equation (A.31) q appears as $1/q$ so we can write

$$\frac{1}{q(z)}=\mathrm{Re}\left\{\frac{1}{q(z)}\right\}-i\mathrm{Im}\left\{\frac{1}{q(z)}\right\}=\frac{1}{q_r(z)}-i\frac{1}{q_i(z)} \tag{A.36}$$

where subscripts r and i denote the real and imaginary parts of the quantity $1/q(z)$. The exponent in $u(r,z)$ can then also be separated into real and imaginary parts as follows

$$\exp\left[\frac{-ikr^2}{2q(z)}\right]=\exp\left[-\left(\frac{kr^2}{2q_i(z)}\right)-i\left(\frac{kr^2}{2q_r(z)}\right)\right] \tag{A.37}$$

The imaginary term has the form of a phase variation produced by a spherical wave front with radius of curvature R (where we can assume the parabolic approximation). In the limit as $r << R$ the phase delay is approximately

$$\phi(r,z)\cong\frac{\pi r^2}{\lambda R(z)}=\frac{kr^2}{2R(z)} \tag{A.38}$$

which, when equated with the imaginary term of exponential (A.37), yields the following association between the real part of $1/q(z)$ and the radius of curvature of the beam

$$q_r(z)=\frac{1}{R(z)} \tag{A.39}$$

The real part of the exponent in equation (A.37) has a Gaussian variation with off-axis distance r, i.e.

$$\exp\left[-\left(\frac{r}{W(z)}\right)^2\right] = \exp\left[-\frac{kr^2}{2q_i(z)}\right] \tag{A.40}$$

where we define $W(z)$ to be the off-axis distance or radius where beam magnitude falls to $1/e$ of its on-axis value. The imaginary part of $1/q$ is thus related to beam radius as follows

$$q_i(z) = \frac{\lambda}{\pi W^2(z)} \tag{A.41}$$

The radius of curvature $R(z)$ and spot size $W(z)$ of a free-space Gaussian beam at a plane z can thus be derived from the complex radius $q(z)$ which is now given, by combining equations (A.39) and (A.41), as

$$\frac{1}{q(z)} = \frac{1}{R(z)} - i\frac{\lambda}{\pi W^2(z)} \tag{A.42}$$

At the plane $z = 0$, $u(r,0) = A(0)\exp[-ikr^2/2q(0)]$ and if the beam radius W_0, at $z = 0$ is chosen such that

$$W_0 = \sqrt{\frac{2q(0)}{ik}} \tag{A.43}$$

the relative field distribution at this plane is

$$u(r,0) = A(0)\exp\left[\frac{-r^2}{W_0^2}\right] \tag{A.44}$$

Now solving for $q(0)$ in equation (A.43) and using equation (A.35) yields another important expression for $q(z)$,

$$q(z) = \frac{i\pi W_0^2}{\lambda} + z \tag{A.45}$$

Taken together equations (A.42) and (A.45) allow one to determine the beam radius and radius of curvature at any z plane as follows

$$W(z) = W_0\left[1 + \left(\frac{\lambda z}{\pi W_0^2}\right)^2\right]^{0.5} \tag{A.46}$$

$$R(z) = z + \frac{1}{z}\left(\frac{\pi W_0^2}{\lambda}\right)^2 \tag{A.47}$$

The minimum beam radius is the beam waist radius W_0, which occurs at $z = 0$, where the radius of curvature is infinite.

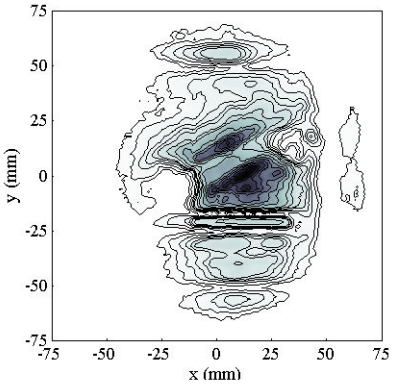
Appendix B.

Selected Near-Field Transmission Imaging Results

This appendix contains a selection of images obtained using the near-field transmission imaging arrangement described in Chapter 3 that were not included in the main text to conserve space. For each object measured a photograph, grey-scale plot of intensity (displayed in a linear scale) with contours overlaid and a close-up of the photograph of the object with intensity contours overlaid are shown. All measurements shown were made with a step size of 0.1mm over an area of 150mm × 150mm. Table B-1 lists the objects imaged in each of the figures shown on the following four pages.

Figure	Object
B-1	Small Brass Key in Envelope
B-2	Small Penknife in Envelope
B-3	Leaf
B-4	Ivy Leaf
B-5 *	Bacon
B-6	Bacon
B-7	Bacon
B-8	Bacon
B-9	Bacon
B-10	Bacon (Strip of Bacon Fat)
B-11	Chicken Skin
B-12	Lamb's Liver
B-13	Smoked Ham
B-14 *	Pork
B-15 *	Pork
B-16 *	Pork
B-17	Pork ('L' shaped piece)
B-18	Pork (triangular shaped piece)

Table B-1. Figure numbers and corresponding imaged objects they refer to. Figures whose number has an asterisk contain two images: the first obtained for a fresh sample and the second after the sample was allowed to dry to reduce water content.



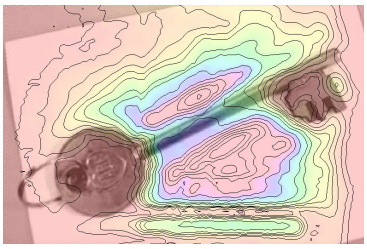
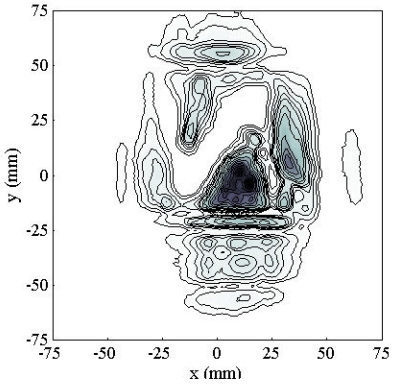


Figure B-1



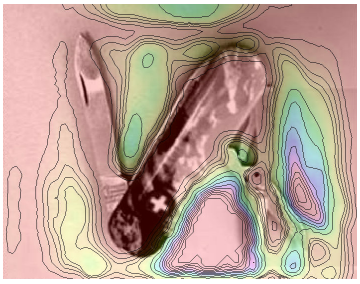
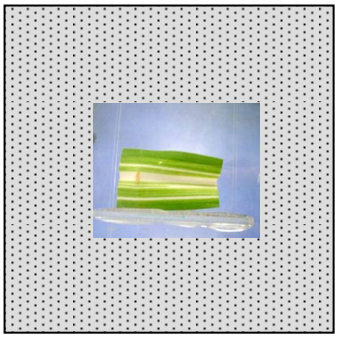
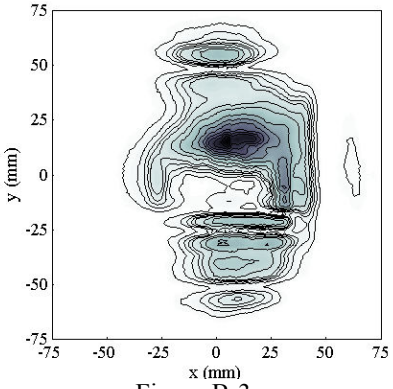


Figure B-2





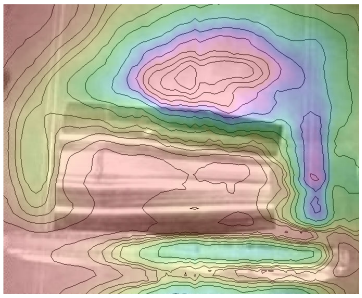
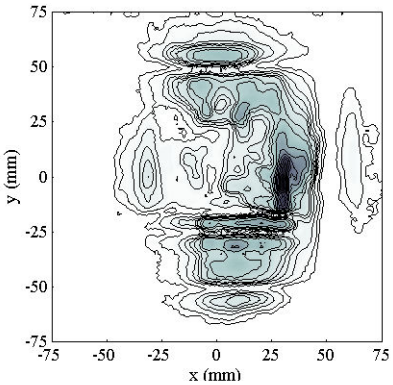


Figure B-3



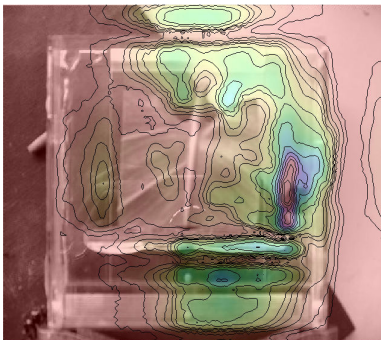


Figure B-4

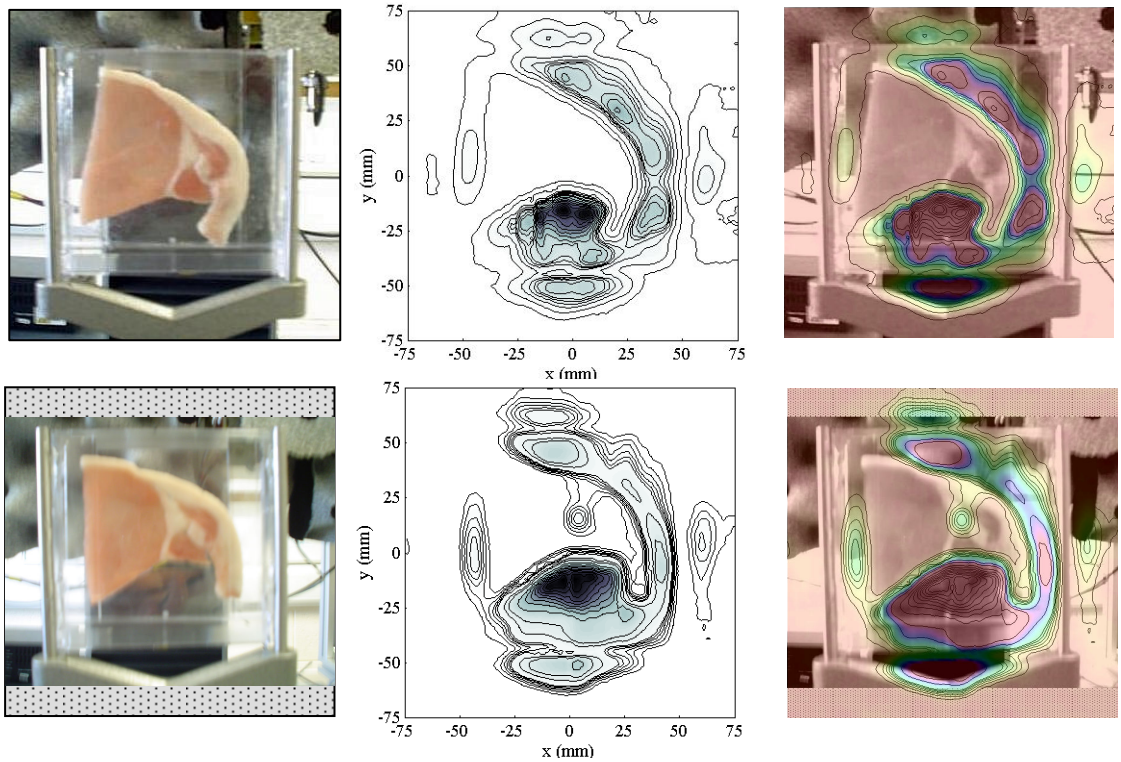


Figure B-5

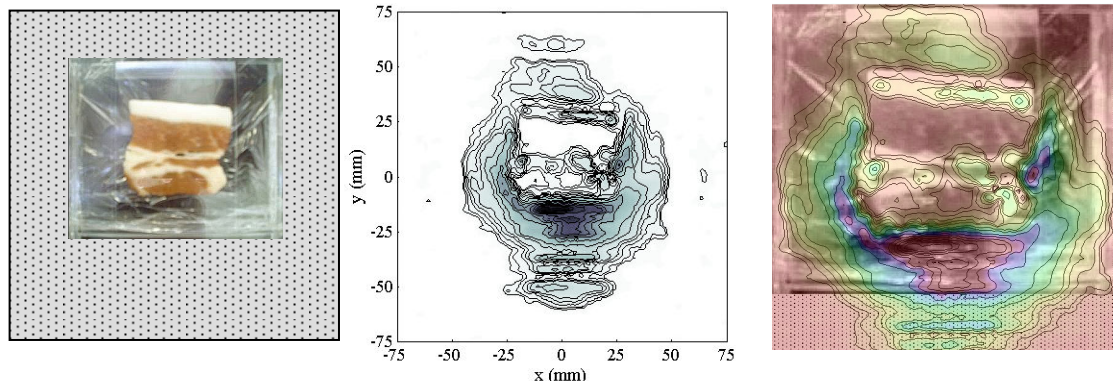


Figure B-6

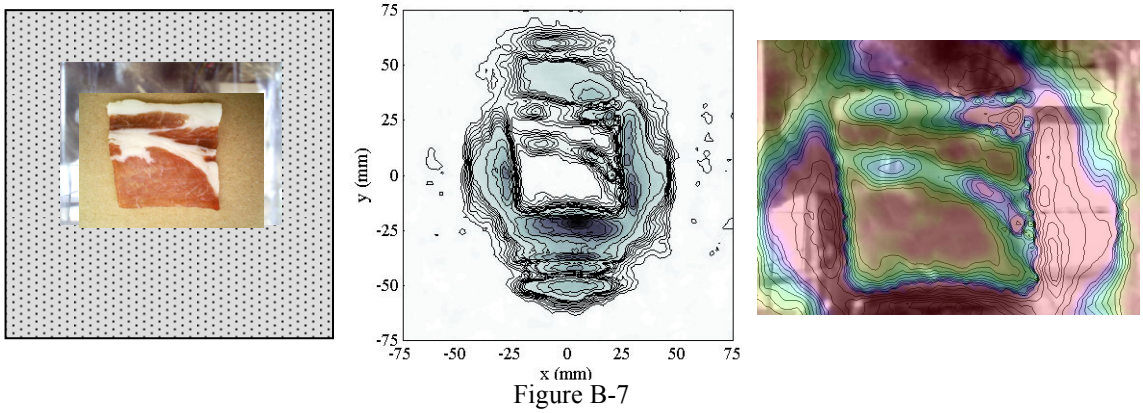


Figure B-7

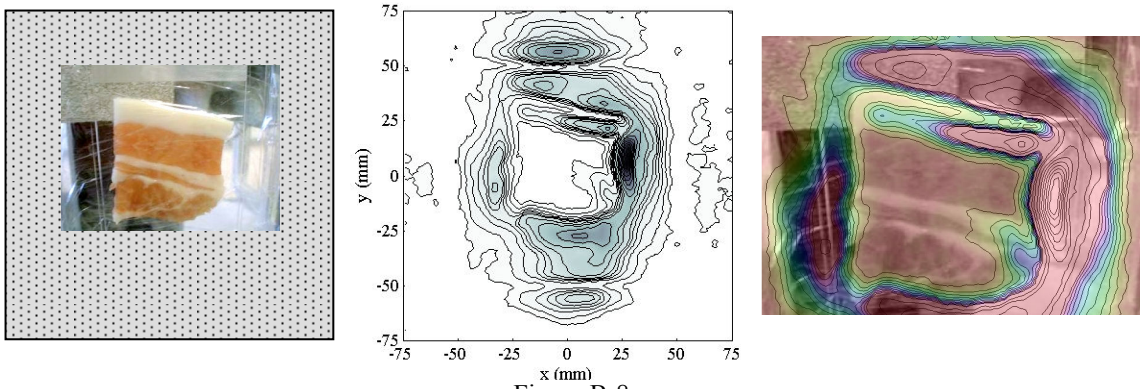


Figure B-8

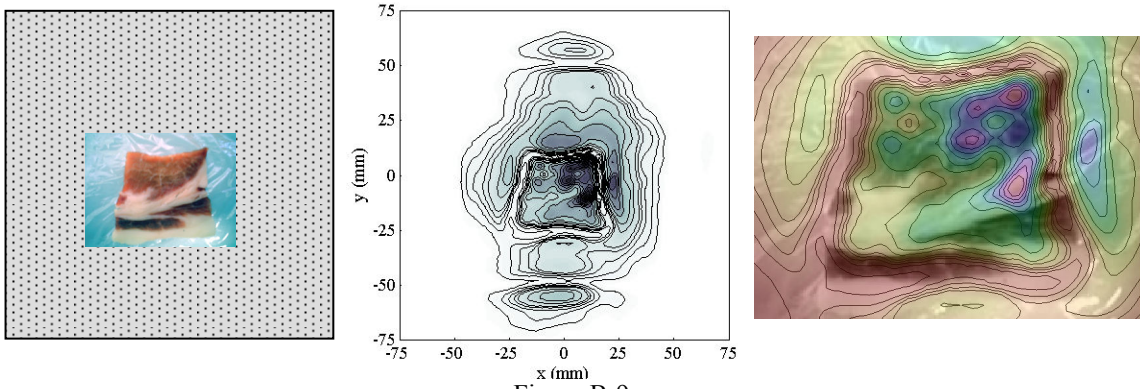


Figure B-9

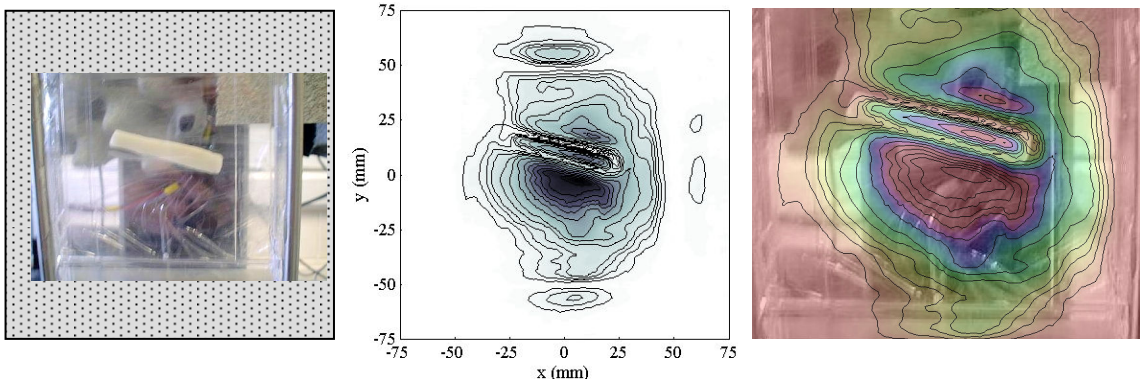


Figure B-10

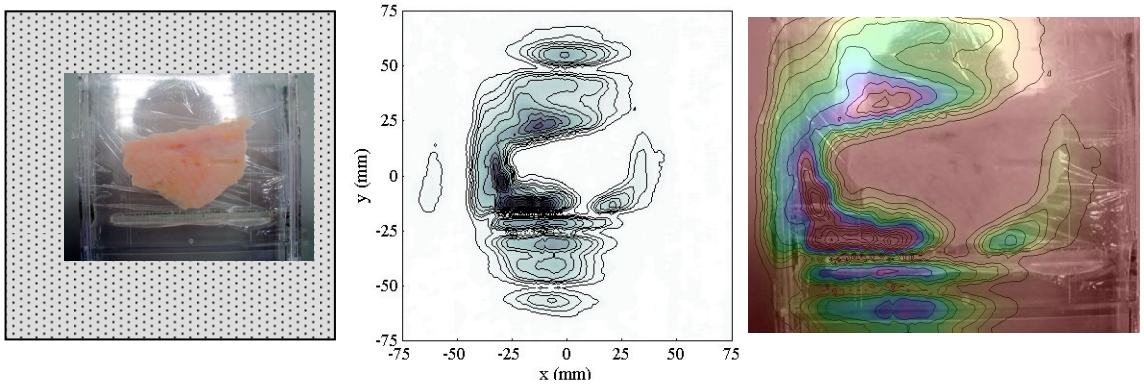


Figure B-11

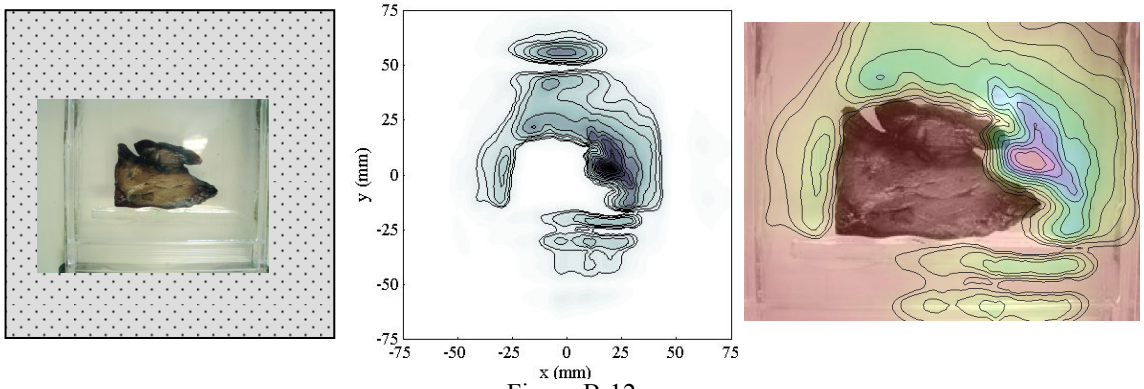


Figure B-12

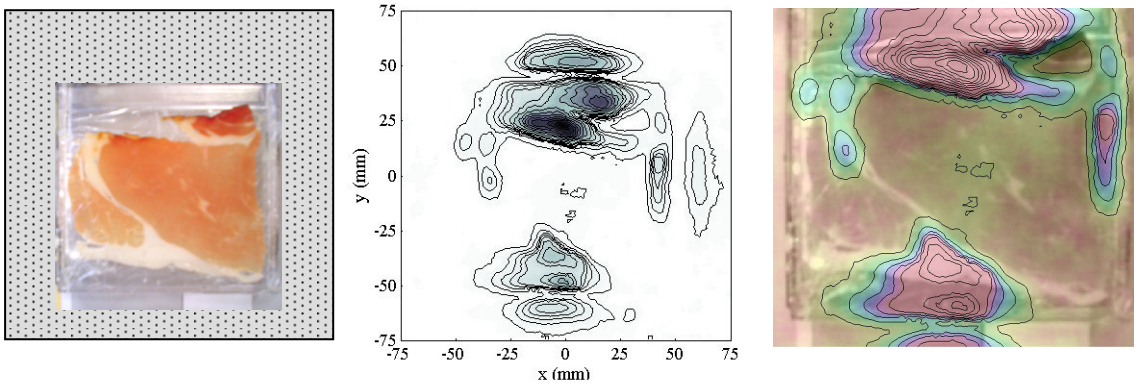


Figure B-13

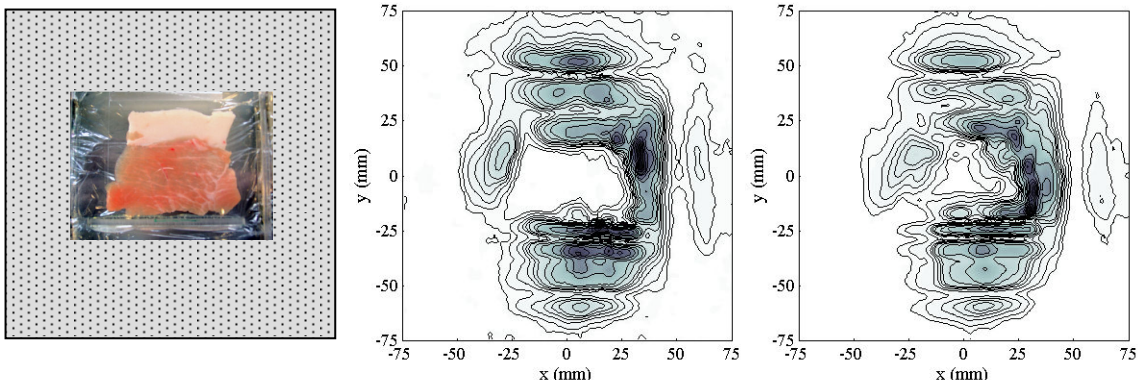


Figure B-14

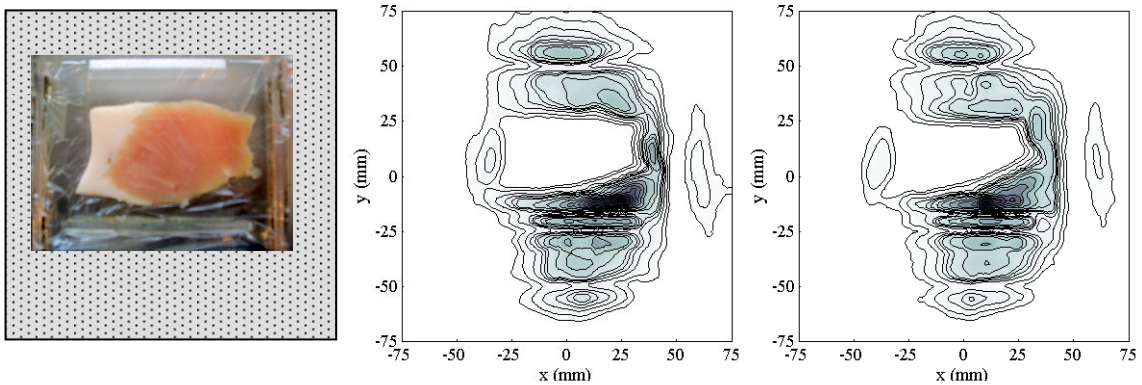
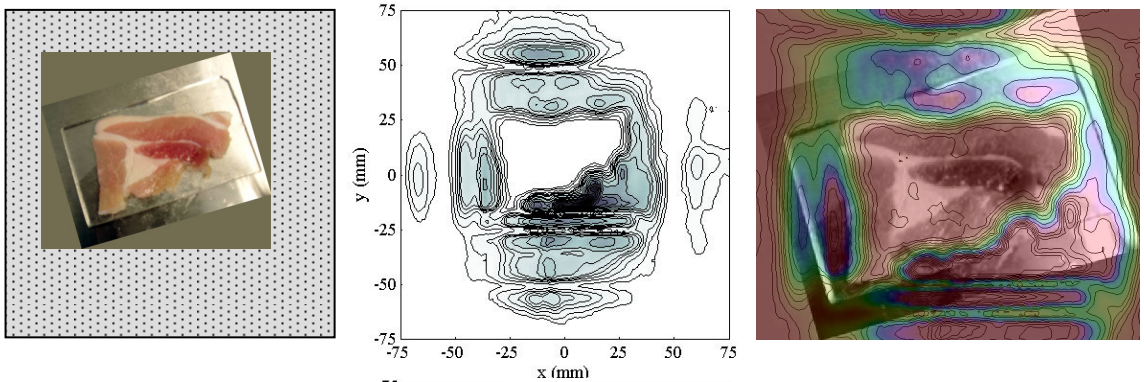


Figure B-15



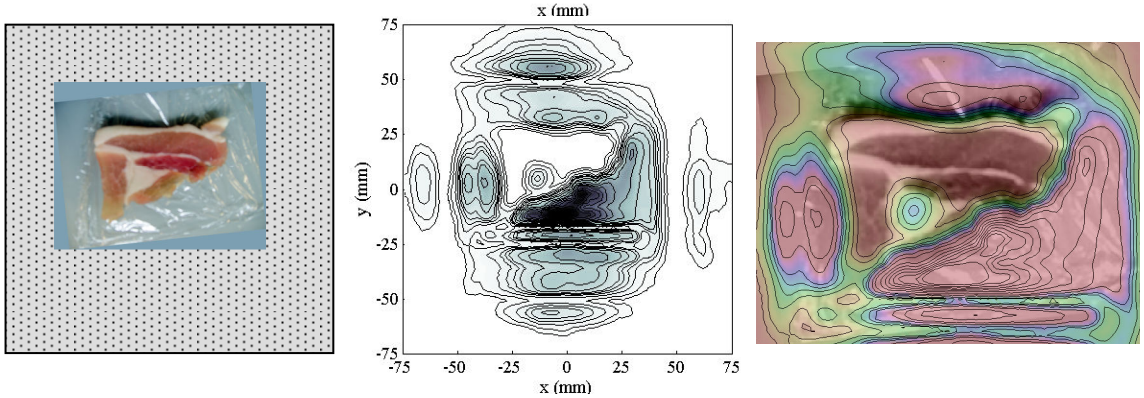


Figure B-16

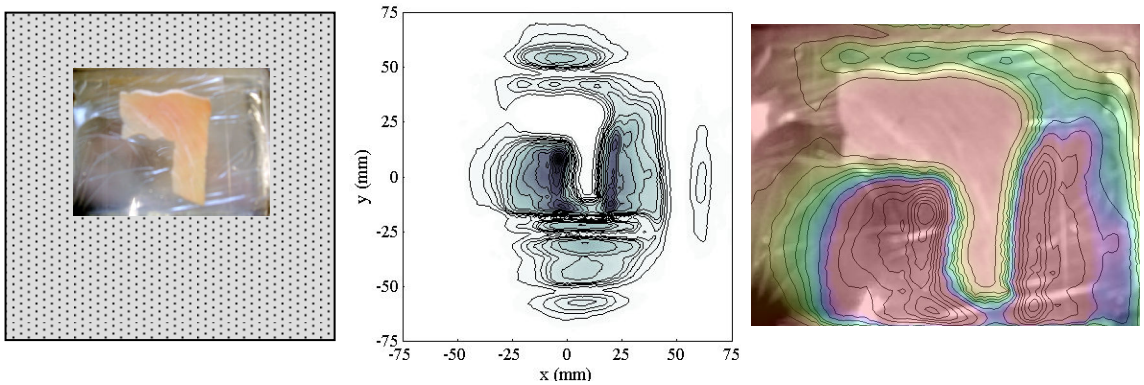


Figure B-17

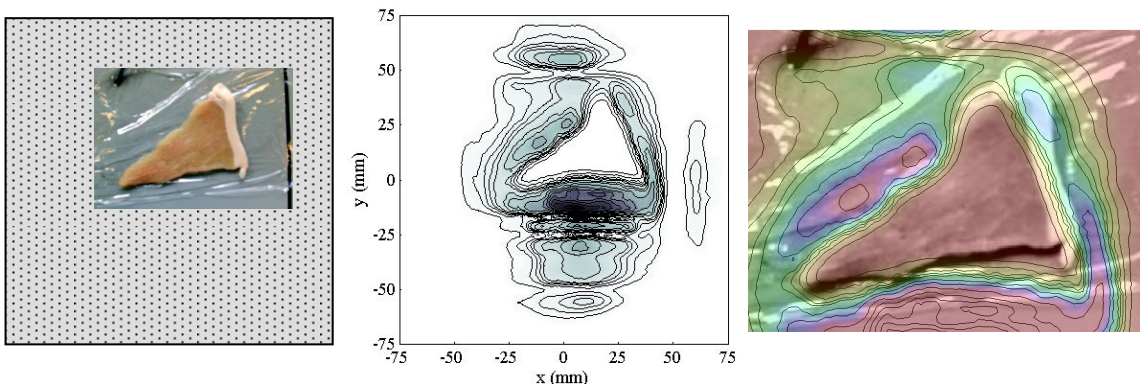


Figure B-18

Bibliography

Chapter 1

[1.1] J.M. Chamberlain, & R.E. Miles, eds., NATO ASI Series E: Applied Sciences, Vol. 334, Kluwer Academic Publisher, London (1997)

[1.2] R.M. Woodward, V.P. Wallace, *et al*, J. Invest. Derm., Vol. 120, pp. 72-78 (2003)

[1.3] V.P. Wallace, A.J. Fitzgerald, *et al*, B. J. Derm., Vol. 120, pp. 424-432 (2004)

[1.4] A.J. Fitzgerald, V.P. Wallace, *et al*, Radiology, Vol. 239, pp. 533-540 (2006)

[1.5] D. Grischkowsky, S. Keiding, M. Van Exter & C. Fattinger, J. Opt. Soc. Am. B, Vol. **7**, pp. 2006 (1990)

[1.6] M.C. Nuss, P.M. Mankiewich, M.L. O'Malley, E.H. Westerwick and P.B. Littlewood, Phys. Rev. Lett., Vol. 66, pp. 3305 (1991)

[1.7] J.E. Pedersen and S. Kieding, IEEE J. Quantum Electron, Vol. 28, pp. 2518 (1992)

[1.8] H. Harde and D. Grischkowsky, J. Opt. Soc. Am. B, Vol. 8, pp. 1642 (1991)

[1.9] P.H. Siegel, '*Terahertz Technology in Biology and Medicene*', 2004 IEEE MTT-S Intl. Miscrowave Symp. Digest, Fort Worth, TX, pp. 1575-1578, (June, 2004)

[1.10] P.H. Siegel and R.J. Dengler, '*Terahertz Heterodyne Imaging for Biological Applications*', NASA/NCI Fundamental Technologies for Biomolecular Sensors, Chicago, IL (July, 2003)

[1.11] D.L. Woolard *et al*, eds. *Terahertz Sensing Technology*, Volume 1, World Scietific Publishing Co. Pte., Ltd., Singapore (2003)

[1.12] D. Mittleman, ed. *Sensing with Terahertz Radiation*, Springer Series in Optical Sciences, Springer-Verlag, Berlin (2003)

[1.13] A.J. Fitzgerald, B.E. Cole, P.F. Taday, '*Nondestructive Analysis of Tablet Coating Thickness Using Terahertz Pulsed Imaging*', J. of Pharmaceutical Sciences, Vol. 94, #1, pp. 177-183(January, 2005)

[1.14] J.A. Murphy, C. O'Sullivan, N. Trappe, W. Lanigan, R. Colgan, & S. Withington, '*Modal analysis of the quasi-optical performance of phase gratings*', Int. Journal IR & Millimeter Waves, Vol. 20, pp. 1569-1486, (1999)

[1.15] S. Withington, J.A. Murphy & K.G. Isaak, '*On the representation of mirrors in beam waveguides as inclined phase-transforming surfaces*', Infrared Physics and Technology, Vol. 36, pp. 722-734, (1995)

[1.16] J.A. Murphy & S. Withington, '*Perturbation analysis of Gaussian beam mode scattering at off-axis ellipsoidal mirrors*', Infrared Physics and Technology, Vol. 37, pp. 205-219, (1996)

[1.17] J. Lavelle, '*The Design and Optimisation of Quasioptical Telescopes*', PhD Thesis, Dept. of Experimental Physics, NUI Maynooth (2008)

[1.18] S. Monk, J. Arlt, D.A. Robertson, J. Courtial, M.J. Padgett, '*The generation of Bessel beams at millimetre-wave frequencies by use of an axicon*', Optics Communicaions, Vol. 170, pp. 213-215 (1999)

[1.19] R.J. Mahon, W. Lanigan, J.A. Murphy, N. Trappe, S.Withington, W. Jellema, '*Novel techniques for millimetre wave imaging systems operating at 100 GHz*', Proc. of SPIE, Florida, USA (2005)
[1.20] G.F. Delgado and J.F. Johansson, '*Quasioptical LO injection in an imaging receiver: An electro-optic approach*', in Proc. NRAO, Tuscon, Az. A.S.P. Conf. Series, Vol. 75, Multi-Feed Systems for Radio Telescopes Workshop, pp. 198-206 (May 1994)
[1.21] T. Klein *et al*, '*LO Beam Array Generation at 480 GHz by use of Phase Gratings*', Eigth International Symposium on Space Terahertz Technology, Harvard University, pp. 482-488 (March, 1997)

Chapter 2

[2.1] H. Kogelnik, '*Coupling and conversion coefficients for optical modes*', Proc. of Symp. on Quasi-Optics, New York, NY. Vol. 14 of Microwave Research Institute Symposia Series (June 8-10, 1964)
[2.2] P.F. Goldsmith, '*Quasi-optical techniques at millimetre and submillimeter wavelengths*', in Infrared and Millimeter Waves, K.J. Button, Ed. New York: Academic, Vol. 6, Ch. 5 (1982)
[2.3] R.J. Wylde, '*Millimetre-wave Gaussian beam-mode optics and corrugated feed horns*', Proc. IEEE H, Vol. 131, pt. H, No. 4, pp. 258-262 (Aug. 1984)
[2.4] J.A. Murphy, '*Aperture Efficiencies of large axisymmetric reflector antennas fed by conical horns*', IEEE Trans. Antennas Propagation, Vol. 36, pp.570-575 (Apr. 1988)
[2.5] J.C.G. Lesurf, 'Millimeter-Wave Optics, Devices and Systems', New York: Adam Hilger (1990)
[2.6] V. Yurchenko, J.A. Murphy, J.M. Lamarre, J. Brossard, '*Gaussian Fitting Parameters of the ESA PLANCK HFI Beams*', Int. J. of Infrared and Millimeter Waves, Vol. 25, pp. 601-616 (2004)
[2.7] C. O'Sullivan, J.A. Murphy, G. Cahill, R. May and S. Withington, 'Novel Applications of Gaussian beam mode analysis', in Proc. SPIE Vol. 6120, Terahertz and Gigahertz Electronics and Photonics V, R.J. Hwu & K.J. Linden eds., 61200J (2006)
[2.8] P.F. Goldsmith, '*Quasioptical Systems: Gaussian Beam Quasioptical Propagation and Applications*', Chp.3 Wiley, (1998)
[2.9] Jenkins & White, '*Fundamentals of Optics*', (McGraw Hill, 2nd ed. 1953)
[2.10] A.E. Siegman, '*Lasers*', University Science Books (1986)
[2.11] M.L. Gradziel, D. White, J.A. Murphy, S. Withington, '*Improving the Efficiency of Quasi-Optical Analysis and Design of Terahertz Systems*', 15th Int. Symposium on Space Terahertz Technology, Northampton, MA, USA, (April 27-29, 2004)
[2.12] J.A. Murphy and A. Egan, '*Examples of Fresnel diffraction using Gaussian modes*' Eur. J. Physics, 14:12-127, (1993)
[2.13] J.A. Murphy, A. Egan & S. Withington: '*Truncation efficiency in beam waveguides using Gaussian beam mode analysis*', IEEE Trans. Antennas & Propagation, **41**, pp. 1408-1413 (1993)
[2.14] D.H. Martin & J.W. Bowen, '*Long-Wave Optics*', (IEEE Transactions on Microwave Theory and Techniques, Vol. 41, No. 10, 1993), pp. 1676-1690

[2.15] R. Penrose, '*A generalized inverse for matrices*', Proc. of the Cambridge Philosophical Soc., Vol. 51, pp. 406-413 (1955)
[2.16] MATLAB Help Documentation
[2.17] J.D. Gaskill, '*Linear Systems, Fourier Transforms, and Optics*' (John Wiley & Sons, Inc., 1987), pp. 12.
[2.18] http://www.mathworld.com
[2.19] Gonzalez, Woods, Eddins, '*Digital Image Processing Using MATLAB*' (Pearson Prentice Hall, 2004), pp. 490-492
[2.20] Pedrotti & Pedrotti, '*Introduction to Optics*' (Prentice Hall, 1993), pp. 529-533
[2.21] C. O'Sullivan, J.A. Murphy, G. Cahill, M.L. Gradziel, N. Trappe, D. White, V. Yurchenko, W. Jellema, 'Developments in Quasi-Optical Design for THz', Proc. SPIE-04, Glasgow, UK, 21-35 June 2004, #52498-39, pp. 320-331 (2004)
[2.22] C. O'Sullivan, E. Atad-Ettedgui, W. Duncan, D. Henry, W. Jellema, J.A. Murphy, N. Trappe, H. Van de Stadt, S. Withington, & G. Yassin, '*Far-IR Optics Design and Verification*', Int. Journal IR & Millimeter Waves, Vol. 23, #7, pp. 1029-1045 (2002)
[2.23] G. Yassin, S. Withington, C. O'Sullivan, J.A. Murphy, T. Peacocke, W. Jellema, & P. Wesselius, '*Electromagnetic modelling of Submillimetre-wave Systems*', Proc. of 13th Int. Symposium on Space Terahertz Technology, Harvard, USA, pp. 525 (2002)
[2.24] MODAL homepage: http://physicsresearch.nuim.ie/modal/modal.html
[2.25] J.A. Murphy, R. Colgan, C. O'Sullivan, B. Maffei & P. Ade, '*Radiation patterns of multi-moded corrugated horns for far-IR space applications*', Infrared Physics and Technology, Vol. 42, pp 515-528 (2001)
[2.26] E. Gleeson, J.A. Murphy, S.E. Church, R. Colgan, C. O'Sullivan, '*Electromagnetic modelling of Few-Moded Winston Cones in the Far Infrared*', Proc. of the Experimental Cosmology at millimetre wavelengths: 2K1BC Workshop, Breuil-Cervinia, Italy, July, 2001, Vol. 616, pp. 295-297 (2002)
[2.27] V.B. Yurchecnko, J.A. Murphy, & J.M. Lamarre, '*Fast Physical Optics Simulations of the Multi-Beam Dual-Reflector Submillimetre-Wave Telescope on the ESA PLANCK Surveyor*', Int. J. Infrared and Millimeter Waves, Vol. 22, pp. 173-184, (2001)

Chapter 3

[3.1] J.P. Loughran, '*Terahertz Imaging for Medical Applications*', M.Sc. Thesis, Dept. of Exp. Physics, NUI Maynooth (2005)
[3.2] E. Cartwright, '*The Quasioptical Analysis of Sub-Millimetre Wave Instrumentation for Astronomy and Medical Imaging*', M.Sc. Thesis, Dept. of Exp. Physics, NUI Maynooth (2005)
[3.3] M. Izzetoglu, S.C. Brunce, *et al*, '*Functional Brain Imaging using Near Infrared Technology*', Engineering in Medicene and Biology Magazine, IEEE, Vol. 26, No. 4, pp. 38-46 (July-Aug 2007)

[3.4] S. Coyle, C.E. Markham, T.E. Ward and W.P. Lanigan, '*A mechanical mounting system for functional near-infrared spectroscopy brain imaging studies*', SPIE OPT Ireland Conference, Dublin, Ireland, pp. 618-627 (April, 2005)
[3.5] K. Humphreys, J.P. Loughran, J.A. Murphy, C. O'Sullivan, M. Gradziel, W. Lanigan, T. Ward, '*Medical Terahertz Imaging: Review of current technology and future potential*', 26th Annual Int. Conf. IEEE Engineering in Medicene and Biology, San Fancisco, pp. 1302-1305 (1-5 Sept. 2004)
[3.6] N. Trappe, S. Kehoe, E. Butler, J.A. Murphy, T. Finn, S. Withington, W. Jellema, '*Analysis of standing waves in submillimeter-wave optics*', Proc. of SPIE Vol. 6472, Terahertz and Gigahertz Electronics and Photonics VI (2007)
[3.7] W. Lanigan, '*Automated Fourier Optics Test facility for the evaluation of Phase Gratings at 100 GHz*', MSc Thesis, Dept. of Exp. Physics, NUI Maynooth (1998)
[3.8] R. Hennessy, '*Detector System for Single- and Multi-Mode Beam Measurements at Sub-Millimetre Wavelengths*', M.Sc. Thesis, Dept. of Exp. Physics, NUI Maynooth (2002)
[3.9] P.F. Goldsmith, '*Quasioptical Systems: Gaussian Beam Quasioptical Propagation and Applications*', Wiley, (1998)
[3.10] C. Fennessy, '*The Design of a Microwave Optics Test Facility*', M.Sc. Thesis, Dept. of Exp. Physics, St. Patricks College Maynooth (1996)
[3.11] J.A. Murphy, '*Distortion of a simple Gaussian beam on reflection from off-axis ellipsoidal mirrors*', Int. J. of Infrared and Millimeter Waves, Vol. 8, #9, pp. 1165-1187 (Sept. 1987)
[3.12] H. Anton, Calculus with Analytical Geometry, 4th ed., John Wiley & Sons Inc. (1992)
[3.13] S. Withington, J.A. Murphy, K.G. Isaak, '*Representation of mirrors in beam waveguides as inclined phase-transforming surfaces*', Infrared Physics & Technology, Vol. 36, #3, pp 723-734, (April 1995)
[3.14] J.A. Murphy, S. Withington, '*Perturbation analysis of Gaussian-beam-mode scattering at off-axis ellipsoidal mirrors*', Infrared Physics & Technology, Col. 37, #2, pp 205-219, (March 1996)
[3.15] T.J. Finn, N. Trappe, J.A. Murphy and S. Withington, '*The Gaussian beam mode analysis of off-axis aberrations in long wavelength optical systems*', Infrared Physics & Technology, Vol. 51, #4, pp 351-359, (March 2008)
[3.16] A. Michelson, *Studies in Optics*. U. of Chicago Press (1927)
[3.17] C. O'Sullivan, JA Murphy, G. Cahill, R. May & S. Withington, '*Novel applications of Gaussian beam mode analysis*', Proc. of SPIE Vol. 6120, Terahertz and Gigahertz Electronics and Photonics V, R.J. Hwu & K.J. Linden eds., pp. 61200J.1-61200J.12 (2006)
[3.18] N. Trappe, '*Quasi-optical analysis of the HIFI instrument for the Herschel Space Telescope*', Ph.D. Thesis, Dept. of Exp. Physics, NUI Maynooth (2002)
[3.19] J.A. Murphy, S. Withington, N. Trappe, R. Colgan, '*Gaussian Beam Mode Analysis of Standing Waves in Quasi-Optical Systems*', Proc. of the Millenium Conf. on Antennas and Propogation AP2000, RSA SP-444, Davos, Switzerland, (April, 2000)

[3.20] J.A. Murphy, N. Trappe, S. Withington, '*Gaussian beam mode analysis of partial reflections in simple quasi-optical systems fed by horn antennas*', Infrared Physics & Technology, Vol. 44, #4, pp. 289-297 (August, 2003)
[3.21] N. Trappe, J.A. Murphy, S. Withington, W. Jellema, '*Gaussian Beam Mode Analysis of Standing Waves Between Two Coupled Corrugated Horns*', IEEE Transactions on Antennas and Propogation, Vol. 53, pp. 1755-1761 (2005)
[3.22] L. Young, '*Terahertz Imaging System with Medical Applications'*, M.Sc. Thesis, Dept. of Exp. Physics, NUI Maynooth (2007)
[3.23]A. Dobroiu, M. Yamashita, *et al*, '*Terahertz imaging system based on a backward-wave oscillator*', Applied Optics, Vol. 43, #30, pp. 5637-5646, (2004)
[3.24] J.F. Kennedy, & Joaquim M.S. Cabral, '*Recovery Processes for Biological Materials*', John Wiley & Sons Ltd, (1993)
[3.25] N. Trappe, R. Mahon, W. Lanigan, J.A. Murphy, S. Withington, '*The quasi-optical analysis of Bessel beams in the far infrared*', Infrared Physics & Technology, Vol. 46, pp. 233-247 (2005)
[3.26] W. Lanigan, R. Mahon, J.A. Murphy, N. Trappe, '*Quasi-optical analysis of binary optical components at 100 GHz*', Joint 29th Int. Conf. On Infrared and Millimeter Waves and 12th Int. Conf. On Terahertz Electronics, Karsruhe, Germany, pp. 589-590 (Sept. 27 – Oct. 1, 2004)
[3.27] G. Chattopadhyay, K.B. Cooper, *et al*, '*A 600 GHz Imaging Radar for Contraband Detection*', Proc. of 19th Int. Symposium on Space Terahertz Technology, Groningen, The Netherlands (April 28-30, 2008)
[3.28] R.E. Fischer, B. Tadic-Galeb, *Optical System Design*, SPIE Press, McGraw-Hill, Chp. 12: *Basics of Thermal Infrared Imaging*, pp. 233, (2000)
[3.29] R. Appleby & H.B. Wallace, '*Standoff Detection of Weapons and contraband in the 100 GHz to 1 THz region*', IEEE Transactions on Antennas and Propagation, Vol. 55, #11, pp. 2944-2956, (2007)
[3.30] R. Appleby, R.N. Anderton, *et al*, '*Mechanically scanned real time passive millimetre wave imaging at 94 GHz*', Proc. SPIE Vol. 5077, Passive Millimeter-Wave Imaging Technology VI and Radar Sensor Technology VII, pp. 1-6 (2003)
[3.31] C. Groppi, C. Walker, *et al*, '*Large Format Heterodyne Arrays for Terahertz Astronomy*', Proc. of 19th Int. Symposium on Space Terahertz Technology, April 28-30, 2008, Groningen, The Netherlands, pp. 57 (2008)
[3.32] J.W. Lamb, '*Miscellaneous Data on Materials for Millimetre and Submillimetre Optics*', Int. J. of Infrared and Millimetre Waves, Vol. 17, #12, pp. 1997-2034 (1996)

Chapter 4

[4.1] J.R. Leger, D. Chen and K. Dai, '*High modal discrimination in a Nd:YAG laser resonator with internal phase gratings*', Optics Letters, Vol. 19, No. 23, p1976-1978 (1994)
[4.2] J.R. Leger, D. Chen and G. Mowry, '*Design and performance of diffractive optics for custom laser resonators*', Applied Optics, Vol. 34, No. 14, p2498-2509 (1995)

[4.3] R. Gusten *et al*., *'A 16-element 480 GHz Heterodyne Array for the Heinrich-Hertz-Telescope (HHT)*', Multi-feed Systems for Radio Telescopes, ASP Conf. Series, Vol. 75, p222-229 (1995)
[4.4] P.F. Goldsmith, '*Quasioptical Systems: Gaussian Beam Quasioptical Propagation and Applications*', IEEE Press, New York (1998)
[4.5] T. Klein, G.A. Ediss, R. Gusten, C. Kasemann, '*Phase Gratings as LO-Distributors in Submm Heterodyne Arrays*', Proc. on the 11th International Symposium on Space Terahertz Technology, p313-325, (May, 2000)
[4.6] K.F. Schuster *et al*, '*The IRAM 230 GHz Multibeam SIS Receiver*', Imaging at Radio through Submillimetre Wavelengths, APS Conf. Series, Vol. 217, p25-32 (2000)
[4.7] W.B. Veldkamp, J.R. Leger and G.J. Swanson, '*Coherent summation of laser beams using binary phase gratings*', Optics Letters, Vol. 11, No. 5, p303-305 (May 1986)
[4.8] D.C. O'Shea *et al*, '*Diffractive Optics: Design, Fabrication and Test*', SPIE tutorial texts, Vol. TT62 (2003)
[4.9] J. Jahns, M.M. Downs, *et al*, '*Dammann gratings for laser beam shaping*', Optical Engineering, Vol. 28, No. 12, p1267-1275 (1989)
[4.10] U.U. Graf and S. Heyminck, '*A Novel type of Phase Grating for THz Beam Multiplexing*', Proc. of the 11th Int. Symp. on THz Space Technology, Ann Arbor (2000)
[4.11] H. Dammann and K. Gortler, '*High-Efficiency in-line Multiple Imaging by means of Multiple Phase Holograms*', Optics Communications Vol. 3, No. 5 (1971)
[4.12] H. Dammann and E. Klotz, '*Coherent Optical Generation and Inspection of Two-Dimensional Periodic Structures*', Optica Acta, Vol. 24, No. 4, p505-515 (1977)
[4.13] U. Killat, C. Clausen, and G. Rabe, '*Binary Phase Gratings for Couplers Used in Fiber-Optics Communications*', Fiber and Integrated Optics, Vol. 3, No. 2-3, p221-235 (1980)
[4.14] U. Killat, G. Rabe, W. Rave, '*Binary Phase Gratings for Star Couplers with High Splitting Ratios*', Fiber and Integrated Optics, Vol. 4, No. 2, p159-167 (1982)
[4.15] T.H. Barnes *et al*, '*Reconfigurable free-space optical interconnections with a phase-only liquid-crystal spatial light modulator*', Applied Optics, Vol. 31, No. 26, p.5527-5535 (1992)
[4.16] J. Lesurf, 'Millimetre-Wave Optics, Devices and Systems', Institute of Physics Press (1990)
[4.17] N. Yoshikawa and T. Yatagai, '*Phase optimization of a kinoform by simulated annealing*', Applied Optics, Vol. 33, No. 5, p863-868 (1994)
[4.18] N. Trappe, '*Quasi-Optical Analysis of the HIFI Instrument for the Herschel Space Observatory*', Ph.D. Thesis, NUI Maynooth (2002)
[4.19] U. Krackhardt *et al*, '*Upper bound on the diffraction efficiency of phase-only fanout elements*', Applied Optics, Vol. 31, No. 1, p27-37 (1992)
[4.20] S. Jacobsson, *et al*, '*Partly illuminated kinoforms: a computer study*', Applied Optics, Vol. 26, No. 14, p2773-2781 (1987)

[4.21] M.R. Feldman and C.C. Guest, '*Iterative Encoding of High-Efficiency Holograms for Generation of Spot Arrays*', Optics Letters, Vol. 14, No. 10, p479-481 (1989)
[4.22] E. Hecht, Optics, 3rd Ed., Addison Wesley, Ch.10, p.438 (1998)
[4.23] R. Gusten *et al*, '*CHAMP – The Carbon Heterodyne Array of the MPIfR*', SPIE Conference on Advanced Technology MMW, Radio and Terahertz Telescopes, Proc. SPIE, p167-177 (March, 1998)
[4.24] J.A. Murphy, C. O'Sullivan, N. Trappe, W. Lanigan, R. Colgan, & S. Withington, '*Modal analysis of the quasi-optical performance of phase gratings*', Int. Journal IR & Millimeter Waves, Vol. 20, pp. 1469-1486 (1999)
[4.25] C. O'Sullivan, J.A. Murphy, N. Trappe, W. Lanigan, R. Colgan & S. Withington, '*The Gaussian Beam Mode Analysis of Phase Gratings*', Proc. of 10th Int. Symposium on Space Terahertz Technology, Charlotsville (March, 1999)
[4.26] W. Lanigan, N. Trappe, JA Murphy, R. Colgan, C. O'Sullivan & S. Withington, '*Quasi-optical multiplexing using reflection phase gratings*', Proc. of the 11th Int. Symposium on Space Terahertz Technology, Ann Arbor, MI, pp 616-625, (May, 2000)
[4.27] R. May, J.A. Murphy, W. Lanigan, '*Phase Gratings for the sub-millimetre waveband*', Joint 29th Int. Conf. On Infrared and Millimeter Waves and 12th Int. Conf. On Terahertz Electronics, Karsruhe, Germany, (Sept. 27 – Oct. 1, 2004)
[4.28] C.E. Groppi *et al*, '*Desert STAR: a 7 pixel 345 GHz Heterodyne Array Receiver for the Heinrich Hertz Telescope*', Proc. SPIE, ISSU 4855, p330-337 (2003)
[4.29] R.J. Koshel, '*Enhancement of the downhill simplex method of optimisation*', Proc. SPIE Vol. 4832, Int. Optical Design Conf, p270-282 (2002)
[4.30] K.V. Price, R.M. Storn and J.A. Lampinen, '*Differential Evolution: A Practical Apporach to Global Optimisation*', Springer (2005)
[4.31] R. Hooke and T.A. Jeeves, '*Direct Search solution of numerical and statistical problems*', Journal of the Association for Computing Machinery, Vol. 3, p297-314 (1961)
[4.32] J.A. Nelder and R. Mead, '*A Simplex Method for Function Minimization*', Computer Journal, Vol. 7, p308-313 (1965)
[4.33] Y. Huang and W.F. McColl, '*An Improved Simplex Method for Function Minimization*', IEEE Int. Conf. on Systems, Man and Cybernetics 1996, Beijing, China, Vol. 3, p1702-1705 (1996)
[4.34] J.C. Lagarias, JA Reeds *et al*, '*Convergence properties of the Nelder-Mead Simplex Method in Low Dimensions*', SIAM J. Optim., Vol. 9, No. 1, p112-147 (1998)
[4.35] M.C. Haenue, '*Submillimetre-Wave Local Oscillator Multiplexing using Phase Gratings*', M.Sc. Thesis, NUI Maynooth (1995)
[4.36] T. Dresel, '*Optimization of symmetric Dammann gratings with a multidimensional error feedback algrotihm*', Optics Communications, Vol. 129, p19-26 (1996)
[4.37] J.M. Johnson and Y. Rahmat-samii, '*Genetic Algorithm Optimization and its Application to Antenna Design*', in Proc. IEEE Antennas Propagat. Soc. Int. Symp., Seattle, WA, pp. 326–329 (June, 1994)

[4.38] S. Kirkpatrick, CD Gelatt, MP Vecchi, '*Optimization by Simulated Annealing*', Science, Vol. 220, NO. 4598, p671-220 (1983)
[4.39] A. Corona, M. Marchesi *et al*, '*Minimizing Multimodal Functions of Continuous Variables with the "Simulated Annealing" Algorithm*', ACM Transactions on Mathematical Software, Vol. 13, No. 3, p262-280 (1987)
[4.40] J. Turunen, *et al*, '*Optimisation and fabrication of grating beamsplitters*', J. Phys. D: Appl. Phys, Vol. 21, p102-105 (1988)
[4.41] J. Turunen, A. Vasara and J. Westerholm, '*Kinoform phase relief synthesis: a stochastic method*', Opt. Eng. Vol. 28, p1162-1167 (1989)
[4.42] N. Yoshikawa and T. Yatagai, '*Phase optimization of a kinoform by simulated annealing*', Applied Optics, Vol. 33, No. 5, p863-868 (1994)
[4.43] N.C. Evans and D.L. Shealy, '*Design and Optimization of an Irradiance Profile-Shaping System with a Genetic Algorithm Method*', Applied Optics, Vol. 37, No. 22, p5216-5221 (1998)
[4.44] X. Chen, K Yamamoto, '*An Experiment in Genetic Optimization in Lens Design*', J. Mod. Opt. Vol. 44, No. 9, p1693-1703 (1997)
[4.45] I. Ono, S. Kobayashi, and K. Yoshida, '*Global and multi-objective optimization for lens design by real-coded genetic algorithms*', Proc. SPIE 3482, p110-121 (1998)
[4.46] S. Boxwell, S.G. Fox and J.F. Roman, '*Design and Optimization of Optical Components using Genetic Algorithms*', Optical Engineering, Vol. 43, No. 7, p1643-1646 (2004)
[4.47] S. Doyle, D. Corcoran and J. Connell, '*Automated Mirror Design Using an Evolution Strategy*', Optical Engineering, Vol. 38, No. 2, p323-333 (1999)
[4.48] R.L. Morrison, '*Symmetries that Simplify the Design of Spot Array Phase Gratings*', J. Opt. Am. A, Vol. 9, No.3 (1992)
[4.49] J.D. Gaskill, '*Linear Systems, Fourier Transforms, and Optics*', John Wiley and Sons, Inc., (Series in Pure and Applied Optics), (1978)
[4.50] J. Jahns *et al*, '*Dammann Gratings for Laser Beam Shaping*', Optical Engineering, Vol. 28, #12, (1989)
[4.51] T. Klein *et al*, '*LO Beam Array Generation at 480 GHz by use of Phase Gratings*', Eigth International Symposium on Space Terahertz Technology, Harvard University, pp. 482-488 (March, 1997)
[4.52] A.J. Lee *et al*, '*Point-by-point inscription of narrow-band gratings in polymer ridge waveguides*', Appl. Phys. A., Vol. 90, pp 273-276 (Sept. 2007)
[4.53] Yun-Sik Jin, Geun-Ju Kim and Seok-Gy Jeon, '*Terahertz Dielectric Properties of Polymers*', Journal of the Korean Physical Society, Vol. 49, No. 2, pp. 513-517 (August 2006)
[4.54] William R. Folks, Sidhartha K. Pandey, Glenn Boreman, '*Refractive Index at THz Frequencies of Various Plastics*', in *Optical Terahertz Science and Technology*, OSA Technical Digest Series (CD) (Optical Society of America, 2007), paper MD10

Chapter 5

[5.1] R. Gusten *et al*, '*A 16-element 480 GHz Heterodyne Array for the Heinrich-Hertz-Telescope*', Multi-feed Systems for Radio Telescopes, ASP Conf. Series, Vol. 75, p222-229 (1995)
[5.2] R. Gusten *et al*, '*CHAMP – The Carbon Heterodyne Array of the MPIfR*', SPIE Conf. On Advanced Technology MMV Radio and Terahertz Technology, SPIE Vol. 3357, p167-177 (March, 1998)
[5.3] T. Klein *et al*, '*LO Beam Array Generation at 480 GHz by use of Phase Gratings*', 8th Int. Symp. On Space Terahertz Technology, p482-488, (March, 1997)
[5.4] U.U. Graf and S. Heyminck, '*Fourier Gratings as Submillimeter Beam Splitters*', IEEE Transactions on Antennas and Propagation, Vol. 49, #4, p.542-546 (April, 2001)
[5.5] T. Klein *et al*, '*Phase Gratings as LO-Distributors in submm Heterodyne Arrays*', 11th Int. Symposium on Space Terahertz Technology, p313-325, (May, 2000)
[5.6] H. Dammann and K. Gortler, '*High-Efficiency in-line Multiple Imaging by means of Multiple Phase Holograms*', Optics Communications Vol. 3, No. 5, pp.312-315 (1971)
[5.7] U. Killat, G. Rabe, W. Rave, '*Binary Phase Gratings for Star Couplers with High Splitting Ratios*', Fiber and Integrated Optics, Vol. 4, No. 2, p159-167 (1982)
[5.8] R.L. Morrison, '*Symmetries that Simplify the Design of Spot Array Phase Gratings*', J. Opt. Am. A, Vol. 9, No.3 (1992)
[5.9] H. Dammann, '*Blazed synthetic phase-only holograms*', Optik Vol. 31 p95-104 (1970)
[5.10] H. Dammann, '*Spectral characteristics of stepped-phase gratings*', Optik Vol. 53, p409-417 (1978)
[5.11] L.B. Lesem, P.M. Hirsch and J.A. Jordan, '*The Kinoform: A new wavefront reconstruction device*' IBM J. Res. Dev, Vol. 13, p150 (1969)
[5.12] F. Wyrowski, '*Diffractive optical elements: iterative calculations of quantised, blazed phase structures*', J. Opt. Soc. Am. A., Vol. 7, No. 6 p961-969 (1990)
[5.13] S. J. Walker and J. Jahns, '*Array generation with multilevel phase gratings*', J. Opt. Soc. Am. A, Vol. 7, No. 8, p1509-1513 (1990)
[5.14] D.C. O'Shea, *et al*, Diffractive Optics: Design, Fabrication and Test, SPIE tutorial texts, Vol. TT62 (2003)
[5.15] Rabadi, W.A. and Myler H.R., '*Iterative Image Reconstruction*: *A Wavelet Approach*', IEEE Signal Processing Letters, Vol. 5, #1, 1-3, (1998)
[5.16] M.A. Seldowitz, J.P. Allebach and D.W. Sweeney, '*Synthesis of digital holograms by direct binary search*', Applied Optics, Vol. 26, No. 14, p2788-2798 (1989)
[5.17] J.N. Mait, '*Understanding Diffractive Optic Design in the Scalar Domain*', J. Opt. Soc. Am. A, Vol. 12, No. 10, p2145-2158 (1995)
[5.18] R.W. Gerchberg and W.O. Saxton, '*Phase Determination from Diffraction and Image Plane Pictures in the Electron Microscope*', Optik, Vol. 34, No. 3, p275-284 (1971)

[5.19] R.W. Gerchberg and W.O. Saxton, '*A Practical Algorithm for the Determination of Phase from Image and Diffraction Plane Pictures*', Optik, Vol. 35, No.2, p237-246 (1972)
[5.20] J.R. Fienup, '*Recontruction of an object from the modulus of its Fourier tranform*', Optics Letters, Vol. 3, No. 1, p27-29 (1978)
[5.21] J. Turunen, F. Wyrowski (eds.), '*Diffractive Optics for Industrial and Commercial Applications*', Akademie Verlag (1997)
[5.22] J.R. Fienup, '*Phase retrieval algorithms: a comparison*', Applied Optics, Vol. 21, No. 15, p2758-2769 (1982)
[5.23] F. Wyrowski and O. Bryngdahl, '*Digital Holography as part of Diffractive Optics*', Rep. Prog. Phys. P.1481-1571 (1991)
[5.24] J.H. Seldin and J.R. Fienup, '*Numerical Investigation into the Uniqueness of Phase Retrieval*', J. Opt. Soc. Am. A, Vol. 7, No.3, p412-427 (1990)
[5.25] H. Dammann and E. Klotz, '*Coherent Optical Generation and Inspection of Two-Dimensional Periodic Structures*', Optica Acta, Vol. 24, No. 4, p505-515 (1977)
[5.26] N. Trappe, '*Quasi-Optical Analysis of the HIFI Instrument for the Herschel Space Observatory*', Ph.D. Thesis, NUI Maynooth (2002)
[5.27] J. Jahns et al, '*Dammann Gratings for Laser Beam Shaping*', Optical Engineering, Vol. 28, No. 12, (1989)
[5.28] R. Gusten *et al*, '*CHAMP – The Carbon Heterodyne Array of the MPIfR*', SPIE Conf. On Advanced Technology MMV Radio and Terahertz Technology, SPIE Vol. 3357, p167-177 (March, 1998)
[5.29] A.J. Lee *et al*, '*Point-by-point inscription of narrow-band gratings in polymer ridge waveguides*', Appl. Phys. A., Vol. 90, pp 273-276 (Sept. 2007)
[5.30] Yun-Sik Jin, Geun-Ju Kim and Seok-Gy Jeon, '*Terahertz Dielectric Properties of Polymers*', Journal of the Korean Physical Society, Vol. 49, No. 2, August 2006, pp. 513-517
[5.31] William R. Folks, Sidhartha K. Pandey, Glenn Boreman, '*Refractive Index at THz Frequencies of Various Plastics*', 2007 Optical Society of America
[5.32] J.D. Gaskill, '*Linear Systems, Fourier Transforms, and Optics*', John Wiley and Sons, Inc., (Series in Pure and Applied Optics), (1978)
[5.33] D.C. O'Shea, T.J. Suleski, A.D. Kathman & D.W. Prather, 'Diffractive Optics: Design, Fabrication and Test', SPIE Tutorial Texts in Optical Engineering, Vol. TT62 (2004)
[5.34] H. Aagedel, M. Schmid, *et al*, '*Theory of speckles in diffractive optics and its application to beam shaping*', J. Modern Optics, Vol. 3, #7, pp. 1409-1422 (1996)
[5.35] J.R. Fienup & C.C. Wackerman, '*Phase retrieval stagnation problems and solutions*', J. Opt. Soc. Am. A, Vol. 3, #11, pp. 1897-1907, (1986)
[5.36] K. Itoh, '*Analysis of the Phase Unwrapping Problem*', Applied Optics, Vol. 21, No. 14, pp. 2470 (1982)
[5.37] A. Collaro *et al*., '*Phase Unwrapping by Means of Genetic Algorithms*', J. Opt. Soc. Am. A, Vol. 15, No. 2 (1998) pp. 407-418

[5.38] T. Aoki *et al*., '*Two-Dimensional Phase Unwrapping by Direct Elimination of Rotational Vector Fields from Phase Gradients Obtained by Heterodyne Techniques*', Optical Review, Vol. 5, No. 6 (1998) pp. 374-379
[5.39] D.C. Ghiglia & M.D. Pritt, 'Two-Dimensional Phase Unwrapping: Theory, Algorithms and Software', John Wiley & Sons Inc. (1998)
[5.40] R.M. Goldstein, H.A. Zebker & C.L. Werner, '*Satellite radar interferometry: two-dimensional phase unwrapping*', Radio Science, Vol. 23, #4, pp. 713-720 (1988)
[5.41] S. Jacobsson, S. Hard & A. Bolle, '*Partly illuminated kinforms: a computer study*', Applied Optics, Vol. 26, #14, pp. 2773-2781 (1987)
[5.42] C. O'Sullivan, S. Withington & J.A. Murphy, '*Coherent and incoherent phase retrieval using Gaussian Beam Modes*', 8th Int. Conf. on Terahertz Electronics, pp 153-156, Darmstadt, Germany, (Sept. 2000)

Appendix A

[A.1] F. Pedrotti and L. Pedrotti, Introduction to Optics, 2nd Ed., Prentice-Hall (1993)
[A.2] J.W. Goodman, 'Introduction to Fourier Optics' 3rd Ed., Roberts & Company Publishers (2005)
[A.3] J.D. Gaskill, 'Linear systems, Fourier transforms, and optics', Wiley (1978)
[A.4] R.G. Wilson, Fourier Series and Optical Transform Techniques in Contemporary Optics: An Introduction, John Wiley & Sons (1995)
[A.5] J.W. Cooley and J.W. Tukey, 'An Algorithm for the Machine Computation of the Complex Fourier Series', Mathematics of Computation, Vol. 19, pp. 297-301 (April, 1965)