The Development of a Predictive Autofocus Algorithm using a General Image Formation Model

Frederick Nicolls

Submitted to the Faculty of Engineering, University of Cape Town, in partial fulfilment of the requirements for the degree Master of Science in Engineering.
Cape Town,
November 1995

Declaration

I declare that this dissertation is my own work. It is being submitted for the degree of Master of Science in Engineering at the University of Cape Town. It has not been submitted before for any degree or examination at this or any other university.

F.C. Nicolls
(Signature of Candidate)

Acknowledgements

I would like to thank the following people and institutions for their contribution towards this thesis.

Gerhard de Jager and Trevor Sewell for their guidance and help.

Dane Gerneke for expertly driving the microscope.

The Foundation for Research and Development for their financial assistance.

Leica Cambridge Ltd for information and software provided.

Everybody in the UCT image processing laboratory for the stimulating and interesting working environment.

Abstract

This work outlines the development of a general imaging model for use in autofocus, astigmatism correction, and resolution analysis. The model is based on the modulation transfer function of the imaging system in the presence of aberrations, in particular defocus. The extension of the model to include astigmatism is also included.

The signals used are related to the ratios of the Fourier transforms of images captured under different operating conditions. Methods are developed for working with these signals in a consistent manner.

The model described is then applied to the problem of autofocus. A general autofocus algorithm is presented and results given which reflect the predictive properties of this model.

The imaging system used for the generation of results was a scanning electron microscope, although the conclusions should be valid across a far wider range of instruments. It is however the specific requirements of the SEM that make the generalisation presented here particularly useful.

Contents

CONTENTS

List of Figures

LIST OF FIGURES

List of Tables

Section LIST OF TABLES

Chapter 1

Introduction

This thesis discusses a proposed method of autofocus for a large class of imaging systems. Furthermore, it extends the autofocus principle to that of astigmatism correction for those systems which exhibit this aberration and have a means of correcting for it.

The imaging system for which results are presented is a type 1 scanning electron microscope (SEM), operating in the secondary electron detection mode. The instrument used was the Leica S440. This microscope operates on a fully computerised platform, with electronic controls and digital framestores. It is therefore easy to capture images under very specific conditions which are in a form suitable for subsequent processing.

The content of this dissertation is by no means restricted to electron microscopy. The methods developed are equally applicable to any imaging system which conforms to the conditions of isoplanatism and linearity. The field of light optics invariably falls into this category.

1.1 Overview of relevant autofocus methods

Automatic focusing of imaging systems has always been a desirable and often achievable goal. The sections which follow give a short overview of autofocus methods that have been considered and used in the past, particularly those that are relevant to the methods that are developed in this thesis. Some of the methods described come from the field of light optical imaging systems, and some from electron microscopy. There is by no means a conflict here; it is shown in a later chapter that, in terms of the mechanism of image formation, many of these systems are directly analogous. The discussion is broken into two sections: those methods which are particular to a specific type or class of instrument, and those which are more generally applicable.

1.1.1 System specific methods

There are a number of proposed autofocus methods that are applicable to only a particular instrument or class of instrument. They are usually based on specific characteristics which may not be present in every imaging system. Although there are a large number of such methods, three in particular will be elaborated upon, namely the use of analogue video signals, the use of inherent system characteristics, and the use of additional subsystems.

Use of video signal

For situations where the images are to be displayed in real-time on a video display, there is a video signal available. This allows for essentially analogue methods which operate on the video signal. For example, in the SEM there have been a number of autofocus implementations based on the derivative of this signal [31]. A typical technique is to maximise the peak value of the derivative.

Such gradient methods however suffer from a severe disadvantage in that they are sensitive to noise. Hence methods based on the power spectral density have been introduced. These usually aim to maximise the power in a given frequency band. This is done by band-pass filtering the video signal and then measuring the resulting power. By scanning the object to be imaged in several directions it is possible to extend such approaches to astigmatism correction as well. Computationally this procedure is intensive, and has not really been feasible until recently.

Use of system-specific characteristics

Another approach involves utilising specific characteristics of an imaging system to extract relevant defocus information. In the transmission electron microscope, for example, it is possible to obtain such information by tilting the electron beam and measuring the resulting displacement of the image [16]. Also, in the SEM a method has been proposed where the focal length is continuously changed while scanning an object. The resulting image is then analysed for particular features which identify the optimal focal length [17].

Use of additional subsystems

Perhaps the most widely-used autofocus technique involves finding the distance to the object to be imaged by means of a dedicated subsystem. These range-finder approaches are typically based on the principles of stereoscopy or radar. The majority of hand-held cameras employ such subsystems.

1.1.2 Generally applicable methods

A second class of autofocus methods involves the use of techniques that can be applied to imaging systems in general. This is not to say that an implementation on one system can be applied unmodified to another, merely that the principles used are usually universally applicable. There are two major categories. The first uses a predefined focal measure and searches some image-dependent parameter space for the condition of best focus. The second uses knowledge of the image formation process to infer the degree of defocus from some given images.

Use of focus measures

A common approach to autofocus is that of defining a focus measure that exhibits a maximum when the image is in best focus [7, 26, 10, 30, 2]. A search is then made through the range of possible focal lengths, and when the optimal position is found the focus is set to that. Since these methods usually require very little knowledge of the specific imaging system they are often termed passive.

In developing such an algorithm the emphasis shifts to optimising the focus measure. In particular, positive features of a good focus measure are that

- It exhibits a sharp and well-defined peak at the position of best focus.
- It decreases (preferably monotonically) as the defocus increases.
- It is resistant to noise in the images.

The preference that the measure decrease monotonically would allow for easy implementation of an intelligent search procedure to find the best focus. In practice this is usually impossible to achieve. The image often has high energy frequency content in sidelobes around a main central lobe, which has the effect of both shifting the peak of the focus measure and introducing local maxima [30].

Some examples of focus measures that have been considered in the literature are the

- Variance of pixel values
- Energy of image gradient
- Self-entropy of the phase of the Fourier transform

As can be seen from these examples, the focus measure can be as simple or complex as desired, and the choice becomes a very important design parameter.

A severe disadvantage with this scenario is that the search tends to be blind. The criterion function may give no indication as to how far from the position of best focus an image was taken, nor in which direction the focal length should be adjusted to reduce the defocus. This restricts the search to being essentially linear, which is inefficient. Although there are methods of reducing the search space, this factor does represent a fundamental limitation. This weakness is however often tolerated for the convenience of an extremely general algorithm.

Use of system characteristics

With high-speed processing being easily available, methods which analyse captured images and relate them back to characteristics of the specific image formation process are becoming attainable. Here, knowledge about the precise way in which an imaging system affects the captured images is used to obtain defocus information, and focus can then be corrected accordingly. Such methods are often called active, because they require detailed knowledge of the system and thus usually need calibration before they operate accurately.

There is substantial overlap between using these methods for autofocus, and work that is done on the subject of inferring depth from the defocus information in an image. It should be apparent that if defocus in an image can be used to measure the distance to an object, then it can just as easily be used to correct the focal length accordingly. There are a number of papers on the subject of recovery of depth information by means of depth-from-defocus techniques [23, 29, 6, 36].

The way in which these methods work is to process one or more images which were captured under different imaging conditions. The internal focal disparity between these images, or the difference between one of the images and some known reference, is then used to determine something about the current state of the imaging system in relation to these images. This information is then used to find the distance of the imaged object from the focal plane. For example, Pentland [23] proposed a method whereby two images are taken of a scene, one with a pinhole camera (and therefore in focus), and one with a defocused camera. The resulting images were then used to find the point-spread function of the second camera, and a model used to relate this PSF to the distance of the object in the images from the cameras. This illustrates a usual requirement for these methods: a means is required for translating the characteristic data into the required variable, usually the camera-to-object distance. There are two approaches to this translation: the use of models, and the use of lookup tables.

Models: For optical imaging systems, the characteristic feature which is most often used is the system point-spread function (PSF). The usual approach is to use the thin lens approximation and the aperture function to derive a relation between the width of the PSF and the

distance from focus [23, 6, 36]. It is also often assumed that for the polychromatic case in the presence of aberrations this PSF takes on a roughly Gaussian shape, and the relation is derived for this situation [23, 36, 18]. By the use of these models the width of the PSF can be related back to the distance of the object from the focal plane.

Lookup tables: For any imaging system more complex than a simple camera it may be impossible to develop a sufficiently accurate model. It then becomes necessary to isolate some feature of the characteristic data, and to tabulate the variation of this feature with changing levels of defocus. Later, when the algorithm is in operation, this table is used to perform the required translation. It must be noted that the dimensionality of this table depends on the number of changeable parameters that will have an effect on the resulting images. Data must therefore be stored for every possible imaging configuration. This lookup table approach has also been noted in the literature [6].

Advantages of applying such methods to autofocus is that the entire search process that was inherent in the passive methods is bypassed. The cost of this, however, is calibration of the instrument before the translation between the image characteristic data and the desired distance can be made.

1.2 Requirements for SEM

In developing an autofocus algorithm for the SEM, a number of specific details must be taken into account. Most importantly, the microscope has an extremely large number of operating parameters that can be varied when images are formed. Accelerating voltage, probe current, magnification, focal length and stigmator coil current are but a few of the settings which must be decided upon before capture can begin, all of which have a considerable effect on the resulting images. It is for this reason that the passive methods that were described earlier have been the dominant paradigm in electron microscope autofocus thus far. The model-based approach seems not to have been investigated, and the parameter space is simply so large that storage of lookup-tables is unfeasible.

Of secondary importance is the fact that formation of a complete image in the SEM takes a few seconds if it is to be fairly noise-free. This places a restriction on the number of images that need to be captured in total for a single autofocusing operation if the system is to remain interactive. It is fortunate that, because the SEM captures images by scanning, there is always the possibility of sampling only portions or even lines of any given specimen. If the focusing algorithm can utilise this feature to reduce the time needed for image capture, then all the better.

Finally, as with all algorithms, it is advantageous that it be universally applicable. This means that the assumptions made and the methods employed must be valid and effective across all makes and models of SEM, and if possible extending also beyond the field of microscopy. Included in this requirement is a preference for a minimum degree of calibration, which would necessarily reduce the accuracy of the algorithm through time.

1.3 Overview of proposed method

In light of the final requirement of the previous section, the system-specific methods that were discussed are not suited to this investigation. The loss of generality that they represent outweighs any advantages that they might offer. Also, the general search methods for autofocus that were outlined have been applied to the SEM in the past, and although they work effectively the linear nature of the search is debilitating in terms of speed. In the words of Pentland [23], "The search is unnecessary, for there is a smooth *gradient* of focus as a function of depth". Furthermore, the search for best focus is usually impeded by the presence of astigmatism, and special care must be taken to ensure that the algorithm converges to the desired solution [7].

Thus, for the purposes of this thesis, the model-based approach in conjunction with the use of system characteristics is explored. As has been mentioned, this avenue does not seem to have been adequately investigated, and it demonstrates many strengths.

The basic assumptions and characterisations are the same as those made in the depth from defocus work: the system will be characterised by its point-spread function, or equivalently through the Fourier transform by the modulation transfer function. This is a very natural starting point in lieu of the vast knowledge base constructed around these concepts in the literature.

In all cases in the development of this work every effort has been made to keep the proposed methods as general as possible. Where a widely-made assumption would lead to a neat closed-form solution, this assumption is analysed and generalised to the point where it represents a minimal restriction. It is felt that this should be a primary driving factor in any research work.

An outline of the overall principle that is used will follow. This is meant only as an introduction, and in the chapters ahead the notions are formally and completely presented.

1.3.1 Starting points

Many classes of imaging systems can be very accurately characterised by means of the notion of a point-spread function. Effectively this is a function which represents the image which results when an infinitesimal spot is imaged by the system. For any real system this function will never be a spot, but will be spread out in some characteristic manner. The shape of the PSF is a good indicator of the configuration of the system at any time.

Equivalent to the PSF is the system modulation transfer function. This is just the modulus of the Fourier transform of the PSF. The MTF is also an indicator of the configuration of the system, but in terms of spatial frequency rather than spatial coordinates.

The starting point for the proposed method involves the use of two images, Fourier transforming them, and forming the ratio. It is shown in this thesis that if this is done properly, the result approximates the ratio of the MTFs corresponding to the two images. Since each of these MTFs is an indicator of the imaging configuration at the times of capture, it may also be expected to contain data pertaining to the level of defocus in each of the images.

1.3.2 Development

It is in what occurs after these ratios are formed that this work deviates from the usual depth-from-defocus approaches. In most cases the width of the resulting ratio signal is found, and related directly back to the distance of the images from best focus. Alternatively, in the more general case the ratio signal is compared with a table of stored signals, each of which corresponds to a different out-of-focus distance for a given set of imaging conditions.

The approach followed differs from these two scenarios. Instead, a small number of assumptions are made which represent what is felt to be the absolute minimum that ensures a solution. These assumptions stem from experimental data obtained from the SEM, all of which are presented in this text. It is likely that most imaging systems will conform to them. The key attributes of these assumptions are that they allow for an essentially parameter independent model of the imaging system to be developed. This model can then be used to make predictions about the distance of the images from best focus.

1.4 Outline of thesis

In order to analyse the effects of defocus on images, it is necessary to have a sequence of test images with progressively differing degrees of defocus. Such a test sequence has been captured using the S440 and a silver grid for a specimen. This sequence will be referred to

as **farfocus**, and the details appear in Appendix A.

A chapter by chapter overview of the thesis follows.

The second chapter describes the process of image formation in the scanning electron microscope, particularly in terms of the detection of secondary electrons. It justifies the linearity of the system, and goes on to explain how it can be characterised by a point-spread function. The outline of formation of the MTF ratio is then described.

The chapter which follows deals with noise. The noise that is present in the test images after they have been transformed is analysed and categorised. A theoretical discussion of ways in which this noise can be reduced and the difficulties associated with the procedures is then presented. Particular regard is given to the final quantity that is required, namely the ratio of the two corresponding MTFs. It is shown that under some conditions the usual noise reduction techniques by averaging can fail outright. Alternative methods for dealing with these cases are then considered and justified.

The findings of that chapter are then used to demonstrate effective ways of forming the desired ratio. An equation modelling the noise present in the frequency domain is developed, and this is used to finalise the division process. The method is then developed for collapsing the ratio into a single dimension, also while taking the noise into account. A very brief section is then dedicated to the effects of windowing on the Fourier transform.

Having developed the methods for forming the MTF ratios, it is shown how they can be used to estimate the corresponding MTF for any given out-of-focus level. This introduces a whole new topic, namely the characterisation of the imaging system for the changing defocus levels. It is shown that the electron beam can, to a good approximation, be regarded as a geometric cone with some very specific properties, and these properties are then used to develop the rudiments of a general image formation model. It should be noted that this model is created for, and is largely only applicable to, the problem of image formation in the presence of defocus. In the course of the discussion a Gaussian counterpart to the general model is introduced. This Gaussian alternative is developed throughout the dissertation, since it allows for closed-form solutions which are an aid to the understanding of the processes.

The next chapter utilises this model which has been developed and justified in the autofocus context. It describes the conditions under which the assumptions made are valid, and goes on to present the ways in which it can be used to design an autofocus algorithm. Some theoretical analysis is given to the uniqueness properties of the MTF ratio, and methods developed which allow useful manipulation of these ratios. A very general autofocus algorithm is then presented, and the reasons for the experimental failure of this procedure given. The generality is then slightly compromised, and an algorithm presented which circumvents the causes of failure in the previous case. The theoretical soundness of this method is analysed,

and the conditions of applicability demonstrated. Note that the Gaussian parallel continues into this section.

The final major chapter deals with the specifics of the general development put forward in the preceding chapter. Distance measures, which had up to now not been defined, are proposed and applied. Approximate sensitivity functions are derived for the use of these measures in the autofocus algorithm, and modifications are made which emphasise those factors on which the success of the algorithm depends. The method has at this stage been completely specified, and results using the **farfocus** image series are generated. Finally, it is described how the speed of the resulting algorithm can be improved.

All that remains is to tie up some of the details that would be required in an implementation. The effect of magnification in the resulting MTF is discussed, and it is explained how it can be used to improve the reliability of the method. The extension to astigmatism correction is outlined. The thesis is then concluded, and recommendations given about how this work might be extended.

Section 1.4: Outline of thesis

Chapter 2

Image Formation in the SEM

In the SEM, a beam of electrons is used to form an image of a specimen. The SEM is a point-source (type 1) scanning microscope, which means that at any time the illuminating beam is focussed to a small spot on the object. This results in a signal which can be detected. The spot is then scanned across the specimen and an image built up. Depending on the configuration and the detectors used, the signal can provide information on a number of physical characteristics of the specimen, such as topography and atomic composition.

For the purposes of this report, image formation in terms of the detection of secondary electrons (SEs) with an Everhart-Thornley (E-T) detector is discussed in more detail. For an account of image formation in terms of other signals, see any of the standard introductory texts [24, 11].

This chapter first describes the overall configuration of the SEM. The notion of an image field is defined, which allows the specimen to be regarded as a two-dimensional entity without any ambiguity. It is shown that if secondary electron imaging is considered, then a point spread function for the system can be determined. This point-spread function corresponds directly to the current density distribution of the beam at the position that the image was formed. The linearity of the detectors is then verified. The chapter closes with a brief outline of the basic principle that is used in the autofocus method developed in this thesis.

2.1 Detection of Secondary Electrons

A diagram of the layout of the SEM is shown in Figure 2.1.

An electron source emits electrons, usually by either thermionic emission or a strong electric field. An overall potential bias in the direction of the specimen accelerates these electrons,

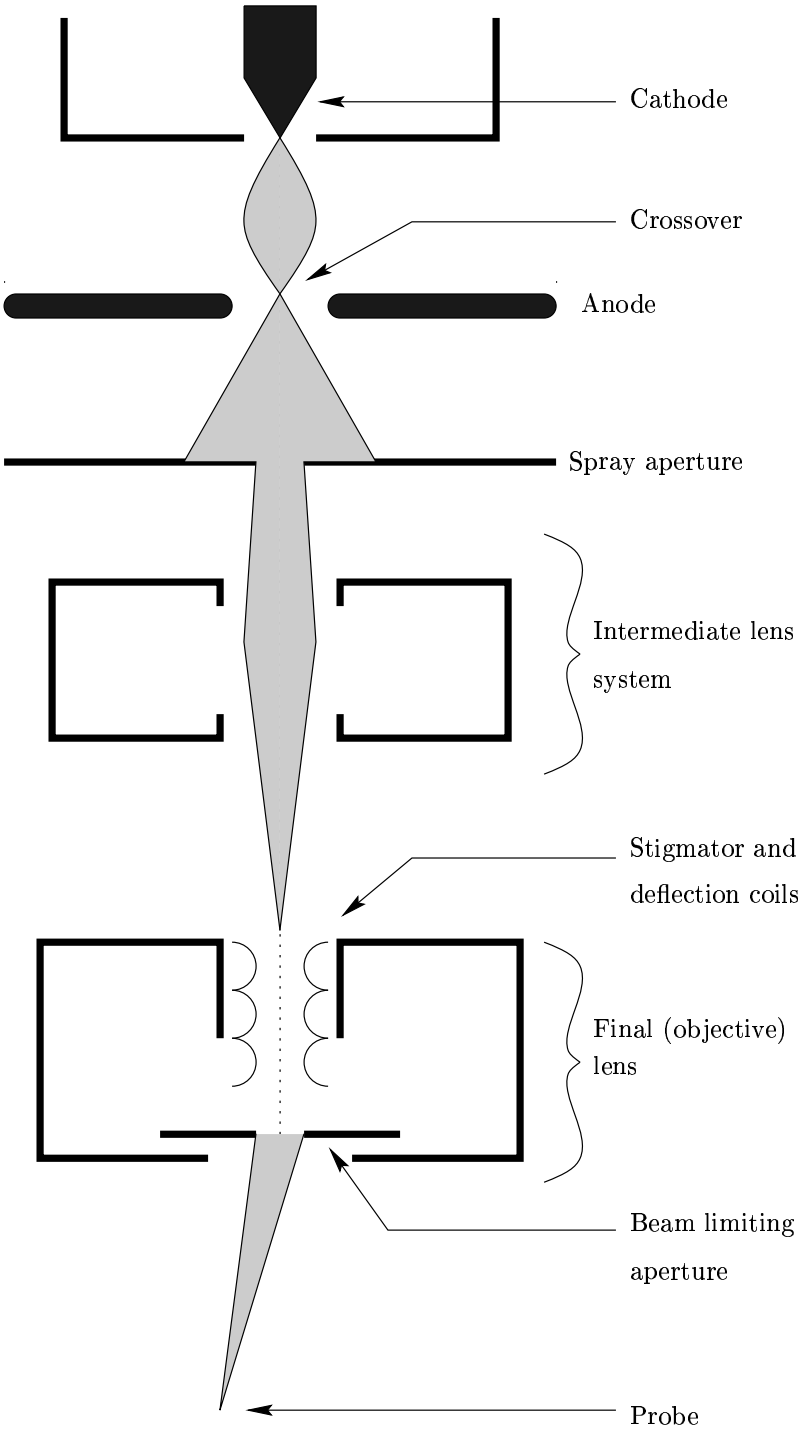


Figure 2.1: Layout of the SEM

which are made to pass through a two- or three-stage electron lens system. These lenses have the effect of demagnifying the minimum beam cross section at the gun (the crossover), so that a very narrow electron probe is formed at the specimen surface. A deflection coil system in front of the last lens can scan this probe across the specimen.

Where the electron beam impinges on the specimen, atomic interactions occur. The secondary electrons, which will be discussed here, are generated by inelastic excitation of atoms to such high energy levels that electrons can overcome the work function and escape. By convention, SEs are considered to be those electrons with an energy of less than 50eV.

The SEs are usually collected by means of an E-T detector. The detector consists of a positively biased grid which attracts the low energy secondaries, which are subsequently accelerated onto a scintillator biased at about +10keV. The light quanta generated are then recorded by a photomultiplier.

2.2 Linearity of the image formation process

For a single electron with energy E incident at a point on the surface of a specimen, there exists a probability $d^2\eta(E, \phi, \zeta, \chi, E_B)/dE_B d\Omega$ of a resulting electron being emitted within the energy interval $(E_B, E_B + dE_B)$ and inside the solid angle element $d\Omega$, with an exit angle ζ relative to surface normal and an azimuth χ (here ϕ is the tilt angle between surface normal and the incident electron) [24, p.128]. Thus every point where an electron enters the specimen will have an associated electron emission probability distribution.

Now, for a three-dimensional specimen being viewed from a particular direction (say along the z-axis) by electrons as shown in Figure 2.2, it is possible in a fairly natural way to define a corresponding two-dimensional image field as follows: Consider an incident electron with trajectory parallel to the z-axis and passing through the point $(x, y, 0)$. When this electron strikes the specimen, atomic interactions occur which contribute to the resulting signal. The two-dimensional image field will shortly be defined in terms of this response for any electron incident at position (x, y). Note however that such an approach eliminates the need to consider details of the effect of the z-value of the surface not being a single-valued function of the position (x, y).

Since all incoming electrons are assumed to come from the same direction, the angle between the surface normal and the incident electron is fixed for any entry point. The dependence on ϕ can then be absorbed into x, y, z. Also, if all incoming electrons are considered to have the same energy, then the dependence on E will fall away. Finally, as discussed above, if the specimen is regarded as a two-dimensional image field for incident electrons, then the effect of z is implicit in (x, y). Taking these factors into account and explicitly introducing positional

Section 2.2: Linearity of the image formation process

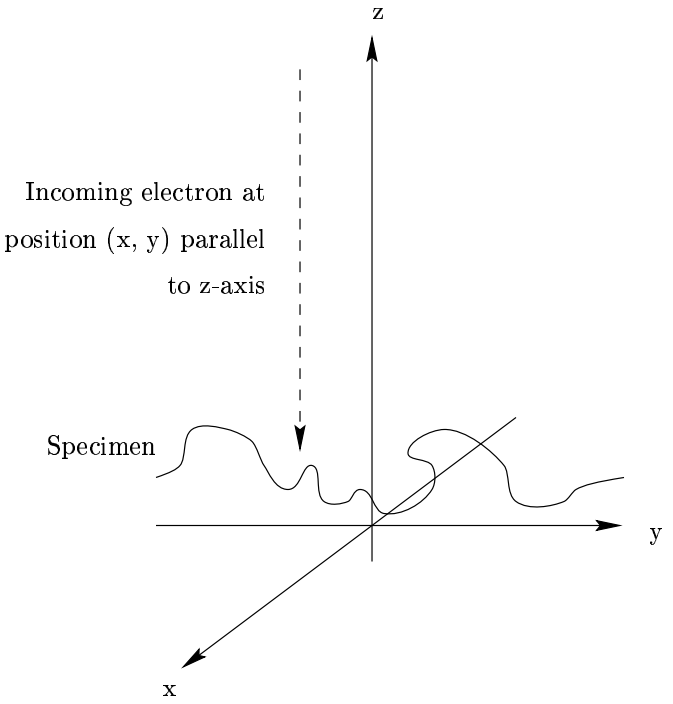


Figure 2.2: Assigning a 2-D image field to a specimen

dependence, the distribution can be rewritten

$$d^2\eta(x, y, \zeta, \chi, E_B)/dE_B d\Omega \tag{2.1}$$

The average number of electrons emitted (per incident electron) with energy in the secondary electron range (< 50eV) and with unspecified exit angle will then be

$$\delta(x, y) = \int_0^{50eV} \int_\Omega \frac{d^2\eta}{dE_B d\Omega} d\Omega dE_B \tag{2.2}$$

This quantity $\delta(x, y)$ thus gives the average number of secondary electrons emitted from the sample as a result of a single incident electron along a trajectory parallel to the z-axis and passing through $(x, y, 0)$. We define $\delta(x, y)$ to be the *image field* for such incident electrons of a particular energy E. Thus, for any particular region of the specimen, the number of emitted electrons is proportional to the incident electrons and the proportionality constant is δ. The image formation process is therefore seen to have a linear basis.

2.3 Detailed model of image formation

It was shown previously that the physics of image formation could be described by $n_{SE} = \delta(x, y)n_B$, where n_{SE} is the number of secondary electrons liberated due to n_B incident electrons at position (x, y) of the image field.

If dependence on time is introduced, then this becomes

$$i_{SE} = \delta(x, y)i_B \tag{2.3}$$

with i_{SE} the current resulting from an incident beam current i_B at (x, y).

If an infinitesimal area element $dx'dy'$ centred on (x', y') is considered, it can be said that the SE current emitted from this area is

$$di_{SE} = \delta(x', y')J_o(x', y')dx'dy' \tag{2.4}$$

where $J_o(x', y')$ is the incident current density at the point (x', y'). The total SE current emitted from the specimen is then

$$i_{SE} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x', y')J_o(x', y')dx'dy' \tag{2.5}$$

If the system is assumed isoplanatic, then the current density distribution of the beam can be written

$$J_o(x', y') = J(x' - x, y' - y) \tag{2.6}$$

where (x, y) is the centre of the incident beam current distribution. Note that the condition of isoplanatism requires that the aberration of the system be constant to a small fraction of a wavelength for all points in a region of the geometrical image that is large compared with the extent of the diffraction image of a point source formed by the system [14].

Now, for every point (x, y) of the centre of the beam, the total resulting secondary electron current is

$$i_{SE}(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x', y')J(x' - x, y' - y)dx'dy' \tag{2.7}$$

which is just the convolution product

$$i_{SE}(x, y) = \delta(x, y) \otimes J(-x, -y) \tag{2.8}$$

where $\otimes$ represents the convolution operator. Thus it can be seen that, under the conditions described here, it is possible to define a point-spread function $h'(x, y) = J(-x, -y)$ which has the property of the total secondary electron yield being a linear convolution of the image field

with this PSF.

2.4 Linearity of the detector

The discussion which follows is based on Reimer [24, p.178].

For every electron that is emitted from the specimen, there is a probability p_c of it being collected by the detector. Once collected, there is a probability p_{bs} of it being backscattered. On average, if the electron is not backscattered, it will cause n_{ehp} electron-hole pairs in the scintillator, of which a fraction q_{Sc} will be converted into light quanta. Therefore, on average the number of light quanta occurring in the detector per electron emitted from the specimen will be $p_c(1-p_{bs})n_{ehp}q_{Sc}$. If the gain of the system is given by G_D (say volts per light quanta per second), then the output signal from the microscope will be

$$f(x,y) = i_{SE}(x,y)p_{Sc}(1-p_{bs})n_{ehp}q_{Sc}G_D \tag{2.9}$$

with $f(x,y)$ the resulting signal for the beam centred on (x,y). Thus the signal can be seen to be a linear function of the number of SEs emitted from the specimen. Letting $G_{overall} = p_c(1-p_{bs})n_{ehp}q_{Sc}G_D$, we can then use Equation 2.8 to write

$$f(x,y) = G_{overall}\,\delta(x,y) \otimes h'(x,y) \tag{2.10}$$

so with $h(x,y) = G_{overall}\,h'(x,y)$,

$$f(x,y) = \delta(x,y) \otimes h(x,y) \tag{2.11}$$

It has thus been shown that, under the conditions described here, image formation in the SEM can be considered to be a linear convolution of two quantities:

- A specimen dependent component, namely a two-dimensional field of secondary electron yield coefficients
- A system dependent point-spread function which, in conjunction with the image field defined here, is effectively the scaled and reflected current density distribution of the electron beam.

This result is an extremely important one because it characterises the linearity of the image formation process in terms of observables.

2.5 Autofocus approach

The outline of the general approach to be taken in developing an autofocus system will now be discussed.

In the previous section it was shown that within certain approximations the SEM image formation process is linear, and that an image $f(x, y)$ can be modelled by a convolution of an image field $\delta(x, y)$ with a point-spread function $h(x, y)$ (equation 2.11).

In the Fourier transform domain this can be represented by a simple point-by-point multiplication

$$F(\omega_x, \omega_y) = \Delta(\omega_x, \omega_y)H(\omega_x, \omega_y) \tag{2.12}$$

where the upper case functions are the Fourier transforms of the respective lower case functions. ω_x and ω_y are the spatial frequency coordinates in the Fourier transform domain. If two images are now taken of exactly the same area of the specimen, but using different imaging conditions, then

$$\begin{aligned} F_1(\omega_x, \omega_y) &= \Delta(\omega_x, \omega_y)H_1(\omega_x, \omega_y) \\ F_2(\omega_x, \omega_y) &= \Delta(\omega_x, \omega_y)H_2(\omega_x, \omega_y) \end{aligned} \tag{2.13}$$

where the difference between H_1 and H_2 comes about because of this change in settings. For purposes of this thesis this change is considered to come about due to a change in focal length, although in general other factors such as changes in aperture can also be utilised [6]. The corresponding spatial domain PSFs h_1 and h_2 can be said to relate to the current density distribution of the beam at those positions where it intercepted the specimen. The ratio F_1/F_2 is now

$$\frac{F_1(\omega_x, \omega_y)}{F_2(\omega_x, \omega_y)} = \frac{H_1(\omega_x, \omega_y)}{H_2(\omega_x, \omega_y)} \tag{2.14}$$

which is an image independent quantity that varies according to the two PSFs used in the formation of the images. Note that the quantity $|H(\omega_x, \omega_y)|$ is just the MTF corresponding to the point spread function $h(x, y)$.

On an image capture level, F_1 and F_2 will therefore be identical except that the focal length of the microscope was changed for one of the images. The quantity $|F_1/F_2|$ can be used to extract information about the PSFs at the positions along the beam where the images were formed, and future chapters demonstrate that it is possible to use it to predict where the minimum beam crossover occurs with respect to the specimen position. The focus can then be corrected to have this crossover coincide with the specimen surface.

This thesis concentrates on the processes of formation and the subsequent extraction of the required information from the MTF ratio.

Section 2.5: Autofocus approach

Chapter 3

Noise

The effect of noise in the image formation process is now discussed. Models for the noise in the frequency domain are given, and methods proposed for reducing this noise in the formation of the ratio signal.

This chapter begins with a characterisation of the noise that appears in the SEM images. The noise in both the spatial domain and the frequency domain is considered. It is asserted that the spatial domain characterisation is inappropriate in this case. The noise in the frequency domain is then discussed in more detail. Cuts through the Fourier transform of the images for varying out-of-focus levels are presented. It is shown that the frequency domain signal can be considered to be contaminated by additive nonzero mean noise. A method is then developed for finding the histogram of this noise. The noise is independent of the defocus level of the instrument.

Methods of reducing the Fourier domain noise are then discussed. Some methods are described for using simple averaging to achieve this. It is shown that the magnitude in the frequency domain is the better representation to use in this case, because of the property of invariance to spatial shifts. The degrading effect that the electron beam has on the specimen is then introduced. More complex noise reduction techniques are then proposed which circumvent this problem.

The section which follows that considers the effect of forming the ratio of MTFs on the noise in the images. The noise model that was described is used in this discussion. It is justified that after averaging a Gaussian noise distribution can be assumed. An in-depth discussion is presented on the noise that occurs in the ratio, and a closed-form solution for the probability density of this noise derived. Methods are then developed for estimating the true value of any point in the ratio. The use of the mean as an estimator is first discussed, and it is shown that it is inappropriate. The use of the median and mode as estimators is then considered,

and it is demonstrated that the former allows for a more accurate value to be obtained.

These findings are then used to develop an appropriate noise reduction method. Estimation of the parameters of the noise model is presented, and justification is given for the noise reduction technique that was briefly introduced earlier in the chapter. It is confirmed that this technique is a valid operation, and that it does indeed result in a reduction in the noise. The chapter concludes with a discussion of how the radial redundancy in the MTF can be used to effect yet more reduction.

3.1 Manifestation of noise in images

For our purposes we are interested in the way in which noise affects the images, rather than in the details of the processes which cause this noise. The appearance of this noise in the spatial and frequency domain is therefore presented here.

3.1.1 Noise in the spatial domain

From the spatial domain it is possible to extract information regarding the probability distribution of the noise, as well as whether it is additive or multiplicative. The simplest way to garner this information is to look at the image of a uniform, featureless object. The histogram of pixel intensities in this image will represent the distribution of the noise to within a constant offset. Furthermore, if the noise distribution (with respect to, say, the mean) is unchanged for a similar object but with a different average intensity, then it can be deduced that the noise is additive. Using such methods an elaborate classification of the noise can be performed, if it is required. For our purposes however, such a classification is unnecessary. Since the methods to be used operate on the Fourier transform of the images, the appearance of the noise in the frequency domain is of more concern.

3.1.2 Noise in the frequency domain

In order to examine the effects of noise in the frequency domain it is necessary to apply the Fourier transform to the image. This discussion is best continued by means of direct use of examples of images acquired from the SEM.

Images from the through-focus series **farfocus** will be used in this analysis. As has been mentioned, all the details of the test sequences appear in appendix A.

For each image the centre 256×256 pixels were extracted. The Fourier transform was then applied, and the modulus taken. No special use was made of windowing to reduce artifacts in

the transform domain, so in effect a rectangular window was in use (the use of windowing will be discussed in section 4.1). Radial cuts were then taken of the data from the centre on a line out to the edge, resulting in a one-dimensional signal which can easily be analysed. Finally, to reduce the high frequency fluctuations in the signals they were smoothed using a 9-point moving average filter. Two such cuts are shown in Figure 3.1 for images **far4** and **far7**. The

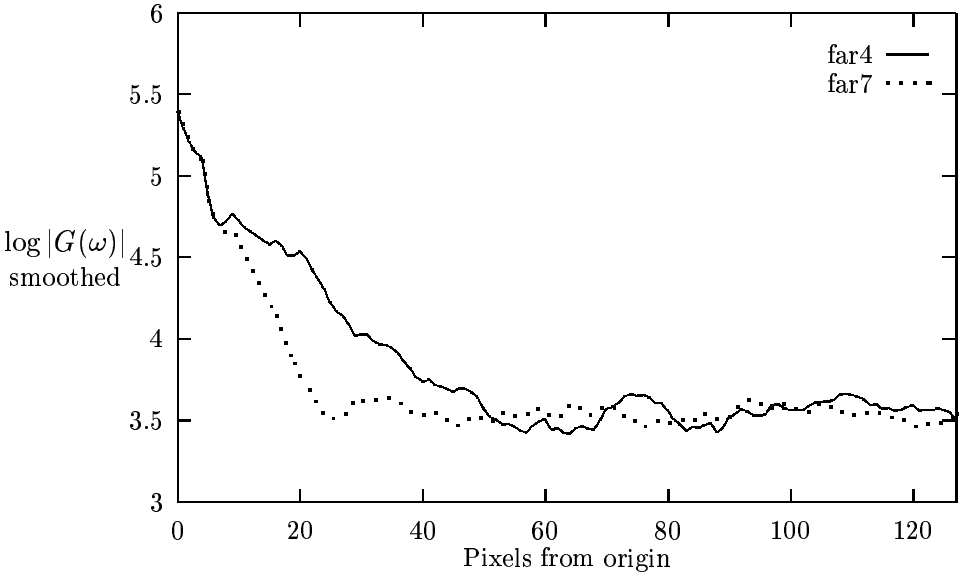


Figure 3.1: Logarithmic plots of radial cuts through magnitude of Fourier transforms of images **far4** and **far7**.

transformed data is roughly radially symmetric, so these profiles are representative of cuts taken at all angles from the centre of the image.

It can be seen that there is a main lobe of high intensity in the centre, which falls off to a constant value on the periphery. This lobe is wider for **far4** than for **far7**, reflecting that the former was taken closer to focus than the latter, and that less high frequency content has been lost in the conversion from object to image space. Most noteworthy however is the fact that beyond the range of the main lobe the transform falls to a constant, the value of which does not change with the degree of defocus. This suggests that the entire transformed image is riding on a white noise background, the magnitude of which remains constant for all focal lengths.

A further significant observation is that there is some fluctuation of the signal in this flat region, and by implication this extends into the main lobe which contains the significant information. In order to determine the character of these fluctuations, a histogram was plotted of the pixel values far from the centre where the effect of the main lobe is negligible.

Section 3.1: Manifestation of noise in images

The centre 256 × 256 pixels of image **far9** were extracted, and the transform and modulus taken. The four 64 × 64 blocks in the corners of this image were then used to extract the histogram information, and the result is given in Figure 3.2.

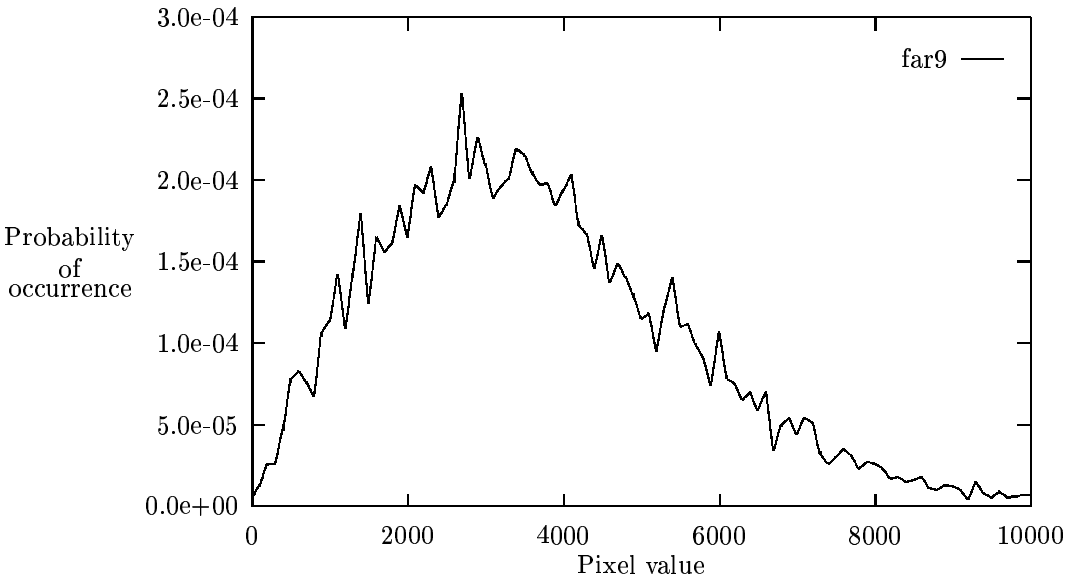


Figure 3.2: Histogram of pixel values on the periphery of magnitude of Fourier transform (image **far9**)

The distribution appears positively skewed. This is expected because of the operation of taking the magnitude, which precludes the possibility of negative pixel values.

Table 3.1 shows statistics based on each image in the **farfocus** image sequence. The same procedure was followed as for the previous case of calculating the histogram, except that the four blocks extracted from the corners had dimensions of only 8 × 8 pixels. The images further down in the table correspond to greater degrees of defocus. It can be seen that once the distance from focus is far enough so that the main lobe does not affect the outer pixels, the mean and standard deviation of these pixels do not vary significantly with defocus. One would not, however, expect this to be the case if some more fundamental instrument parameter (such as probe current or operating voltage) were changed.

It is this fluctuation of the signal in the frequency domain which is unwanted and needs to be effectively handled. Henceforth, when noise in the images is mentioned, it is this frequency domain fluctuation rather than the spatial domain contamination which is being referred to.

Image Name	Mean	Standard deviation	Image Name	Mean	Standard deviation
far1	6690.81	3732.01	**far10**	6016.54	3137.98
far2	4438.34	2470.30	**far11**	3639.31	1906.41
far3	3721.79	1979.48	**far12**	3727.93	1909.55
far4	3712.16	1945.06	**far13**	3672.52	1963.96
far5	3729.85	1955.25	**far14**	3724.17	1944.57
far6	3653.18	1854.78	**far15**	3684.85	1919.39
far7	3666.91	1957.87	**far16**	3872.38	2044.13
far8	3647.25	1915.06	**far17**	3781.28	2043.17
far9	3744.43	1869.51	**far18**	3747.84	1970.13

Table 3.1: Statistics derived from periphery pixels of magnitude transforms of images **far1** to **far9** (magnification 500×) and **far10** to **far18** (magnification 1000×)

3.2 Frequency domain noise reduction by averaging

The autofocus algorithm operates on the ratio of two MTFs which exist in the frequency domain. Any noise in the ratio will naturally have a degrading effect on the procedure, and it is therefore beneficial to minimise the magnitude of this contamination. There appear to be two stages at which the noise can be reduced: the first is in the Fourier transform of the images to be used informing the ratio, and the second is to reduce the noise in the ratio estimate itself. Each of these stages will be elaborated on in turn.

3.2.1 Reduction of noise before formation of ratio

The first stage at which the noise can be reduced is in the Fourier transforms of the images themselves, before the ratio is formed. This should result in a reduction in the noise in the ratio because the component signals are less contaminated. A simple and generally effective way to do this is to employ averaging over a number of transformed images. There are two possibilities that will be discussed here: the case where the area of the specimen remains constant through the averaging, and that where it is allowed to change.

In order to do this a set of statistically independent samples of the transform of the area of the specimen are required. This introduces some possible scenarios depending on the way in which the image capture stage is implemented.

- A number of images can be captured through time of the same area of the specimen. For this case, the image content remains identical, but the noise in each image is different. Assuming the noise to be Gaussian, forming the average (on a pixel by pixel basis) of

the magnitude of the Fourier transform of these images will result in a $\sqrt{n}$ improvement in the signal to noise ratio in the transform domain, where n is the number of images averaged.

- Since ultimately the signal that will be used is one-dimensional, a further option would be to simply scan the specimen along a line a number of times. The noise contributions will then be time-separated and therefore independent, and averaging n signals will again reduce the standard deviation of the noise by a factor of $\sqrt{n}$.

In both cases a reduction in the magnitude of the fluctuations in the frequency domain will occur. What comes out will be a noise-reduced estimate of the Fourier transform of the area of the specimen that was transformed.

It is apparent that the averaging could in fact be performed in the spatial domain before the transform is taken. Although this would be simpler, it is undesirable to require it for the following reason: there usually occurs some lateral image drifting in the SEM. This means that images captured progressively through time may in fact not be of exactly the same area of the specimen, and will therefore not be registered exactly. Averaging in such cases will cause a blurring of the signal detail. It can, however, be shown that the magnitude frequency domain is invariant to translations in the spatial domain. A shift in space simply causes a phase change in the Fourier transform, which is eliminated by the process of taking the modulus. This is demonstrated by the form of the Fourier transform pair

$$f(x-a, y-b) \Longleftrightarrow F(\omega_x, \omega_y) e^{-i\omega_x a} e^{-i\omega_y b} \tag{3.1}$$

with $F(\omega_x, \omega_y)$ the transform of $f(x, y)$. The exponential factors introduced by the image shift are removed when the modulus operation is applied. Because of this translation invariance, only the removal of the effects of noise in the frequency (rather than the spatial domain) will be considered.

There is, however, another problem particular to the SEM which encourages a deviation even from the methods suggested here: the electron beam can have a degrading effect on the specimen. Imaging the same area for a long period to reduce noise can therefore cause permanent damage to the specimen. Furthermore, charge build-up in the specimen will change its imaging characteristics; if an area of the specimen is not permitted sufficient time between electron exposures to allow charge to dissipate, the interaction between the beam electrons and the object will be affected. This will cause the appearance of the specimen to change (non-destructively) through time. The combined effect of these factors is that in general it is not favourable (or even possible) to use the same part of the specimen repeatedly for purposes of noise reduction.

3.2.2 Reduction of noise during or after formation of ratio

Since the system is characterised by a PSF which is invariant with respect to the images formed, there is no reason to use the same area of the specimen repeatedly. Any image formed on the SEM will exhibit the same MTF, as long as the imaging parameters are not changed. This presents two additional scenarios which will be analysed.

The first is to capture many image pairs, each taken at the same focal lengths, and to form the MTF ratio for each case. Since each image pair corresponds to the same area of the specimen, they are all estimates of this ratio. The piece of specimen imaged need not be the same from one pair to the next. The natural thing to do would then be to form the average of these estimates, effecting noise reduction in this manner. It will however be shown that this operation is invalid, and under a few general assumptions will result in no noise reduction at all.

The second scenario, which appears to offer an optimal solution, lies somewhere between this method and the method of averaging the numerator and denominator transforms that was presented in the previous section. Here, image pairs are captured as described before, each having possibly different content. The average of all the numerator transforms is then formed, and divided by the average of the denominator transforms. Although each of the numerator and denominator averages contain a combination of information relating to different specimen areas, it turns out that the ratio is a noise reduced estimate of the required quantity. This is discussed in detail in section 3.4.1.

Before the justification for statements made in this section can be given, a detailed discussion of the effects of noise in the division process must be provided.

3.3 Effect of division on noise

Having introduced the subject of noise that occurs in the Fourier transforms of images in the SEM, the effect on the noise of dividing one of these transforms by another has to be considered. This is important in the formation of the ratio of MTFs discussed in section 2.5.

In the previous sections it has been shown that, in the transform domain, the images formed have some very definite characteristics. There is a central lobe of low frequency data, which carries the information that was preserved in the conversion from object to image (and then to Fourier) space. Outside this lobe the signal falls to a roughly constant value which indicates the presence of a uniform white noise background in the images. Also, contaminating the entire signal is a random fluctuation, the distribution of which is given in Figure 3.2.

To avoid having to deal with two separate cases, only the more general case of the probability

distribution of the fluctuations having a Gaussian profile is considered. Even if this is not exactly correct, the distribution is close enough to Gaussian that some general conclusions may be drawn. If averaging is performed the assumption becomes even more accurate. This follows from the property of sums of random variables: if two variables are added the resulting probability distribution is given by the convolution of the probability distributions for each [22]. In conjunction with the central limit theorem this guarantees that if enough random variables are added together, then the resulting distribution will tend towards a Gaussian [5]. The probability distribution of the mean will then also be a Gaussian.

3.3.1 Division of two Gaussian random variables

Consider the random variable $Z = X/Y$, where X and Y are random variables with associated probability densities $p_X(x)$ and $p_Y(y)$. The probability distribution of Z is then

$$p_Z(z) = \int_{-\infty}^{\infty} |y| p_{XY}(yz, y) dy \tag{3.2}$$

where $p_{XY}(x, y)$ is the joint probability density of X and Y, and is equal to

$$p_{XY}(x, y) = p_X(x) p_Y(y) \tag{3.3}$$

for the case where they are statistically independent [22, p.196].

For X, Y normally distributed with means x_0 and y_0 and standard deviations σ_x and σ_y, the precise probability distributions are given by

$$\begin{aligned} p_X(x) &= \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}(\frac{x-x_0}{\sigma_x})^2} \\ p_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}(\frac{y-y_0}{\sigma_y})^2} \end{aligned} \tag{3.4}$$

Substituting equation 3.4 into 3.2 and letting $\sigma_x = \sigma_y = \sigma$, the distribution of Z becomes

$$p_Z(z) = \int_{-\infty}^{\infty} |y| \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}(\frac{yz}{\sigma} - \frac{x_0}{\sigma})^2} e^{-\frac{1}{2}(\frac{y}{\sigma} - \frac{y_0}{\sigma})^2} dy \tag{3.5}$$

Normalising the variables by setting $p = y/\sigma$, $x_0' = x_0/\sigma$, and $y_0' = y_0/\sigma$, and regrouping terms, this expression for $p_Z(z)$ can be written

$$p_Z(z) = K \int_{-\infty}^{\infty} |p| e^{-(\frac{1}{2}z^2 + \frac{1}{2})p^2 - 2(-\frac{1}{2}zx_0' - \frac{1}{2}y_0')p} dp \tag{3.6}$$

where

$$K = \frac{1}{2\pi} e^{-\frac{1}{2}x_0'^2} e^{-\frac{1}{2}y_0'^2} \tag{3.7}$$

From tables of integrals, a solution to this equation can be found (see Appendix D) to be

$$p_Z(z) = 2K \left\{ \frac{\nu(z)}{\mu(z)} \sqrt{\frac{\pi}{\mu(z)}} e^{\frac{\nu(z)^2}{\mu(z)}} \Phi\left(\frac{\nu(z)}{\sqrt{\mu(z)}} \right) + \frac{1}{2\mu(z)} \right\} \tag{3.8}$$

where $\mu(z)$ and $\nu(z)$ are

$$\begin{aligned} \mu(z) &= \frac{1}{2}z^2 + \frac{1}{2} \\ \nu(z) &= -\frac{1}{2}zx_0' - \frac{1}{2}y_0' \end{aligned} \tag{3.9}$$

and $\Phi(x)$ is the usual error function defined as

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{3.10}$$

In the following two sections this expression for $p_Z(z)$ is used to analyse the use of the mean of this distribution to reduce noise, and then the use of the median and mode.

3.3.2 The mean as an estimator

The expression for $p_Z(z)$ given in equation 3.8 will be analysed in two stages, firstly for the special case of zero-mean random variables, and then for the general case where the means can be nonzero.

Ratio of zero-mean normal random variables

If $x_0 = y_0 = 0$, then $\nu(z) = 0$ in equation 3.9 and the probability density for Z reduces to

$$p_Z(z) = \frac{1}{\pi} \frac{1}{z^2 + 1} \tag{3.11}$$

This is a case of the Cauchy probability distribution. It is discussed in some of the texts on probability because of its fairly singular characteristics [27]. A plot of this distribution is shown in Figure 3.3.

A characteristic of the Cauchy distribution is that its mean does not exist (although it does have a principal value of zero). This non-intuitive fact can be simply observed if an attempt is made to find the mean by usual computational methods. This is demonstrated in Appendix C.1. Of course the median and mode of the distribution exist and are both equal to zero. It can also be shown that the standard deviation of a Cauchy-distributed variable is infinite.

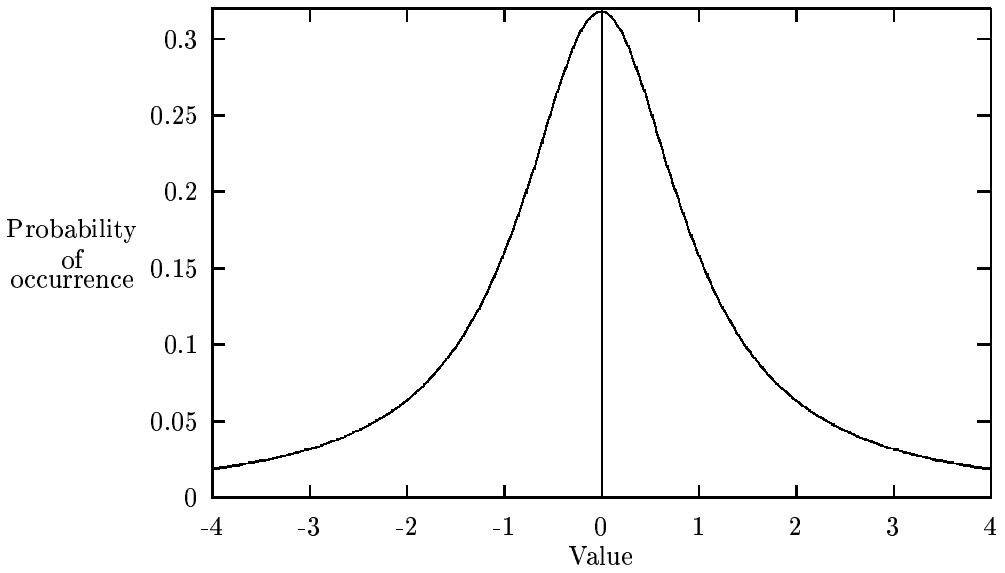


Figure 3.3: Plot of Cauchy probability distribution

Related, but of more interest to us, is the fact that the variable formed by taking the average of n variables with Cauchy distributions is itself Cauchy-distributed. In order to show this a fundamental theorem in statistics needs to be observed which states that, if the random variables X and Y are independent, then the density of their sum $Z = X + Y$ equals the convolution of their respective densities. Since convolution corresponds to multiplication in the Fourier transform domain, the transform pair

$$\frac{1}{z^2 + \alpha^2} \Longleftrightarrow \frac{\pi}{\alpha} e^{-\alpha|\omega|} \tag{3.12}$$

is used to show that the Fourier transform of the Cauchy distribution in equation 3.11 is

$$\mathcal{F}\{p_Z(z)\} = e^{-|\omega|} \tag{3.13}$$

Thus if $p_{nZ}(z)$ is the distribution of the sum of n variables with probability distribution $p_Z(z)$, then

$$\mathcal{F}\{p_{nZ}(z)\} = e^{-n|\omega|} \tag{3.14}$$

Performing the inverse transform the form of the resulting distribution is found to be

$$\begin{aligned} p_{nZ}(z) &= K\frac{1}{z^2 + n^2} \\ &= K\frac{1/n^2}{(z/n)^2 + 1} \end{aligned} \tag{3.15}$$

where K is chosen to normalise the distribution. Since this is the distribution for the sum of the variables, the form for the average, $p_{\overline{n}Z}(z)$, is obtained by compressing the distribution in the z-direction by a factor of n. Thus the distribution of the average after normalisation becomes

$$p_{\overline{n}Z}(z) = \frac{1}{\pi}\frac{1}{z^2+1} \tag{3.16}$$

which is identical to $p_Z(z)$.

The previous discussion demonstrates the possibility of an apparently reasonable distribution having some non-intuitive properties. It is a widely known principle that for normally distributed noise added to a signal, the signal-to-noise ratio can be improved by a factor of $\sqrt{n}$ by means of averaging over n samples. Here, however, is a case of a distribution which looks similar to a Gaussian distribution, but where averaging will have no effect whatsoever.

Ratio of nonzero-mean normal random variables

In the preceding section the case of forming the ratio of zero-mean random variables was discussed. It was observed that the probability distribution of the resulting variable was such that it had no mean or standard deviation, and that averaging could in this case not improve the signal-to-noise ratio. In this section it will be demonstrated that a similar situation occurs for the more general case of nonzero-mean Gaussian variables.

Appendix C.2 provides the proof that even if the restriction of zero-mean is dropped, the expected value of the ratio of two normally-distributed random variables is still undefined. The indication is thus that there would continue to be some complications in attempting noise reduction by means of sample averaging. It is not expected that the corresponding results of the previous section will carry over entirely unmodified, and that averaging will have no effect whatsoever on the probability distribution. However, the fact that the mean does not exist suggests that no theoretical justification exists for assuming that the operation will produce the desired result.

It is concluded that in this situation it is incorrect to attempt to reduce noise by means of averaging over a number of samples. The sample mean is not an estimator of the distribution mean, because this latter quantity does not exist. This is not just a theoretical complication with no practical relevance; simulation shows that the process of averaging does not result in a reduction in the spread of values that the variable takes on.

There are other measures of location that can be used to describe the distribution of a random variable. In the section which follows, the median and mode are discussed as estimators of the ratio of the noise-free values.

3.3.3 The median and mode as estimators

The median of a distribution of a random variable X is any number $\text{Med}(X)$ having the property that

$$P_X\{X < \text{Med}(X)\} \leq \frac{1}{2} \quad \text{and} \quad P_X\{X > \text{Med}(X)\} \leq \frac{1}{2} \tag{3.17}$$

The mode of a random variable is its most probable value (the value for which the probability distribution is a maximum). The use of these two location measures will now be discussed in evaluating the actual value of the MTF ratio in the limit as noise becomes insignificant.

Plots of the probability distribution for $Z = X/Y$ are shown in Figure 3.4 for two cases of $x_0/y_0 = \frac{1}{2}$ and $x_0/y_0 = 2$. In each plot the distributions are given for three values of noise. It is noted that in all cases the distributions are positively skewed and unimodal. Also, as the magnitude of the noise becomes lower the location of the peak of the distribution becomes a better estimate of the value x_0/y_0. Furthermore, it appears to be the case that this observation is more accurate for the case of $x_0/y_0 < 1$.

In order to verify this statement and to assess the limits within which the mode is accurate as an estimate of x_0/y_0, a plot of the mode versus ratio is shown for changing noise. This appears in Figure 3.5. It can be seen that the mode is consistently less than the actual ratio, but as the noise decreases it becomes a better estimate.

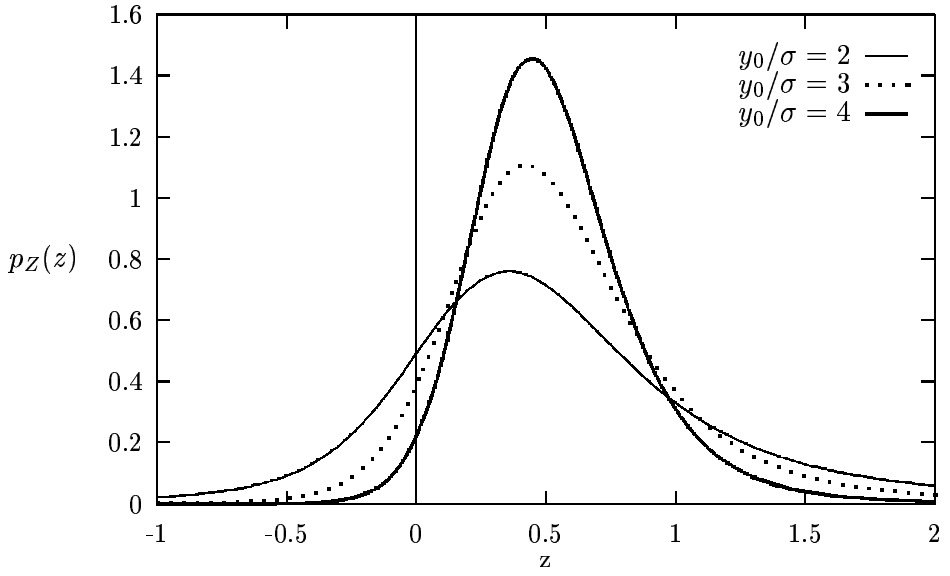
Since the distributions are positively skewed, it is expected that the median of the distribution will be greater than the mode. Thus it is expected that the median will be a better estimate of the ratio than the mode, which consistently under-predicted. A plot corresponding to the previous one but for the case of the median is given in Figure 3.6.

It is evident that for low values of noise and for $x_0/y_0 < 1$, the median is an effective estimate of the required ratio. It is therefore concluded that when the ratio $Z = X/Y$ of two Gaussian random variables with means x_0 and y_0 is formed, the median of the resulting distribution is a good estimate of the ratio x_0/y_0 under conditions of low noise and x_0/y_0 small.

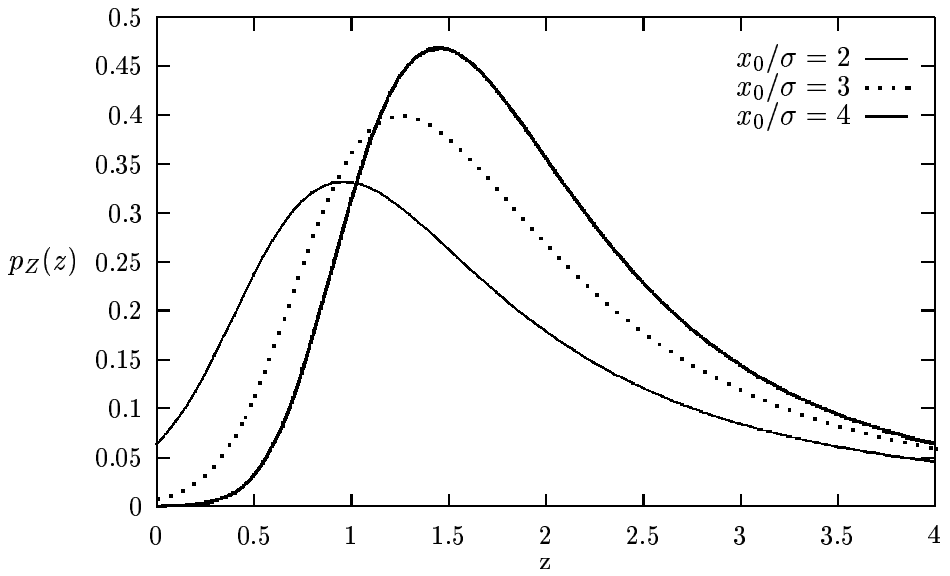
3.4 Methods of reducing noise

With the theoretical foundation now established, the discussion of noise reduction in the frequency domain can be continued. The method of reducing noise in the formation of the ratio is now clarified, and the subsequent collapsing of the signal into a single dimension is discussed.

In section 3.1.2 the noise that occurs in the frequency domain was presented. It was shown that once the images are Fourier transformed, what remains is a constant-level background



(a) $p_Z(z)$ for $x_0/y_0 = \frac{1}{2}$



(b) $p_Z(z)$ for $x_0/y_0 = 2$

Figure 3.4: Probability distribution of Z for two values of x_0/y_0 and changing noise

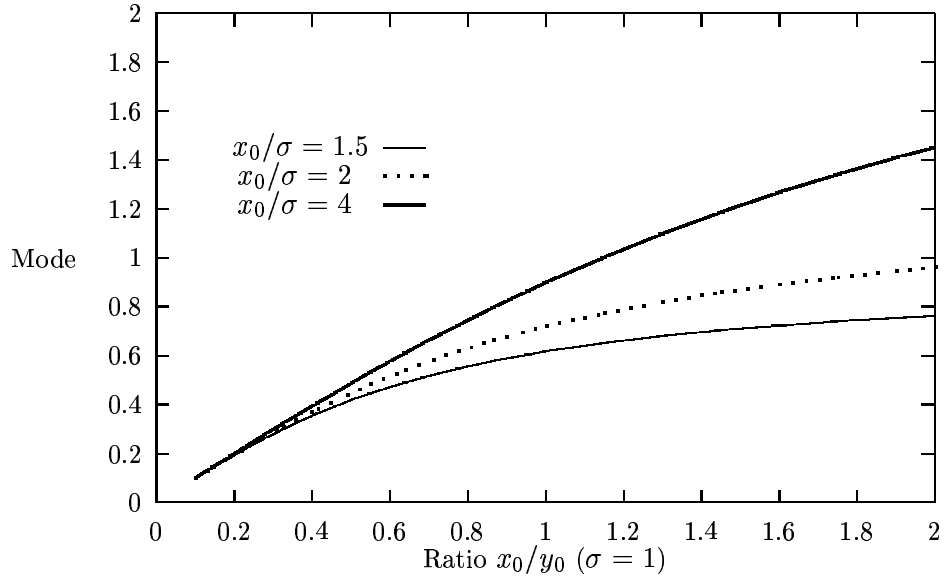


Figure 3.5: Mode vs. x_0/y_0 ratio for changing noise

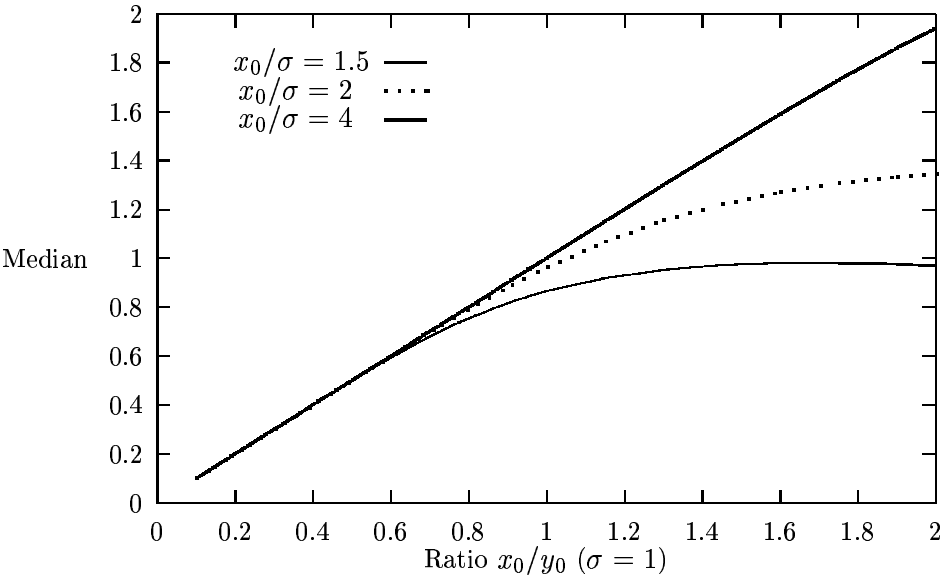


Figure 3.6: Median vs. x_0/y_0 ratio for changing noise

superimposed on the actual signal. Furthermore, this entire resulting signal is contaminated by random noise. The representation in the frequency domain can be modelled by the equation

$$|F_{image}(\omega_x, \omega_y)| = |F(\omega_x, \omega_y)| + n(\omega_x, \omega_y) + C \tag{3.18}$$

with $|F_{image}(\omega_x, \omega_y)|$ the modulus of the transform of the image $f(x, y)$, $|F(\omega_x, \omega_y)|$ the desired modulus of the transform without contamination, $n(\omega_x, \omega_y)$ a zero-mean random noise variable and C the overall offset. In order to successfully utilise the signal, $|F(\omega_x, \omega_y)|$ needs to be approximated from $|F_{image}(\omega_x, \omega_y)|$ by minimising the effect of the two undesirable factors C and $n(\omega_x, \omega_y)$.

The first factor can be eliminated fairly simply: as long as the value C is known, it can just be subtracted out from the signal. In practice this value can be found by considering the transform of an image which is known to have a large degree of defocus. In this case the signal $|F(\omega_x, \omega_y)|$ is approximately zero everywhere, except at very low frequencies. Hence for large ω_x, ω_y,

$$|F_{image}(\omega_x, \omega_y)| = n(\omega_x, \omega_y) + C \tag{3.19}$$

Since $n(\omega_x, \omega_y)$ is zero-mean, the average over a large number of such pixels will be an approximation to the value C.

A simple approach to reducing the effect of noise $n(\omega_x, \omega_y)$ would be to run a smoothing kernel over the entire image $|F_{image}(\omega_x, \omega_y)|$. The problem with this is that it might change the signal in such a manner that it is no longer a good approximation to $|F(\omega_x, \omega_y)|$. The better solution was introduced in section 3.2.2 and is here discussed in detail.

3.4.1 Reducing noise in formation of ratio

Assuming the imaging system to be isoplanatic and linear, the image formation process can essentially be viewed as a convolution of a point-spread function by an image field. This point-spread function is independent of the position in the image, and depends only on the configuration of the imaging system. This makes it possible to use a number of image pairs with differing content but which have been formed with the same PSF.

There are a number of ways in which these pairs can be obtained. They could be completely independent samples, simply captured at the same focal lengths. Alternatively, they could be obtained from tessellating a numerator and a denominator image into subimages, and regarding each subimage as a separate sample. Whatever the means, it will be assumed for this discussion that there are p image pairs available. The transforms of these images will be designated $N_{k(image)}(\omega_x, \omega_y)$ and $D_{k(image)}(\omega_x, \omega_y)$ for the numerator and denominator images

respectively (where $k = 1, 2, \ldots, p$). From equation 3.18,

$$|N_{k(image)}(\omega_x, \omega_y)| = |N_k(\omega_x, \omega_y)| + n_\sigma(\omega_x, \omega_y) + C \tag{3.20}$$

Defining $N'_k(\omega_x, \omega_y)$ to be $|N_{k(image)}(\omega_x, \omega_y)| - C$ it can be written that

$$N'_k(\omega_x, \omega_y) = |N_k(\omega_x, \omega_y)| + n_\sigma(\omega_x, \omega_y) \tag{3.21}$$

This quantity $N'_k(\omega_x, \omega_y)$ can be calculated directly from the data once the constant C is known. The desired signal is then given by

$$|N_k(\omega_x, \omega_y)| = N'_k(\omega_x, \omega_y) - n_\sigma(\omega_x, \omega_y) \tag{3.22}$$

The same procedure can be used on the denominator subimages to find values for the expression

$$|D_k(\omega_x, \omega_y)| = D'_k(\omega_x, \omega_y) - n_\sigma(\omega_x, \omega_y) \tag{3.23}$$

The signals $|N_k(\omega_x, \omega_y)|$ and $|D_k(\omega_x, \omega_y)|$ are now the quantities which are assumed to conform to the linear convolution models

$$\begin{aligned} |N_k(\omega_x, \omega_y)| &= |F_k(\omega_x, \omega_y) H_n(\omega_x, \omega_y)| \\ |D_k(\omega_x, \omega_y)| &= |F_k(\omega_x, \omega_y) H_d(\omega_x, \omega_y)| \end{aligned} \tag{3.24}$$

with $H_n(\omega_x, \omega_y)$ and $H_d(\omega_x, \omega_y)$ the Fourier transforms of the PSFs that were used in the formation of the numerator and denominator images.

To find a noise-reduced estimate of the MTF ratio $|H_n(\omega_x, \omega_y)/H_d(\omega_x, \omega_y)|$ a simple method that might be considered would be to form p independent estimates of the ratio

$$\left(\frac{|H_n(\omega_x, \omega_y)|}{|H_d(\omega_x, \omega_y)|} \right)_k = \frac{|N_k(\omega_x, \omega_y)|}{|D_k(\omega_x, \omega_y)|} \tag{3.25}$$

and then to calculate the average. This would result in the expression

$$\frac{|H_n(\omega_x, \omega_y)|}{|H_d(\omega_x, \omega_y)|} = \frac{1}{p} \sum_{k=1}^{p} \frac{|N_k(\omega_x, \omega_y)|}{|D_k(\omega_x, \omega_y)|} \tag{3.26}$$

However, considering the discussion in section 3.3.2 it is apparent that the noise in each estimate of $|H_n(\omega_x, \omega_y)/H_d(\omega_x, \omega_y)|$ has the approximate probability density given in equation 3.8, and the mean of these estimates is undefined. A valid way of estimating the ratio will now be considered.

Taking the modulus and forming the average over all subimages, equations 3.24 become

$$\begin{aligned}\sum_{k=1}^{p}|N_k(\omega_x,\omega_y)| &= \frac{1}{p}|H_n(\omega_x,\omega_y)|\sum_{k=1}^{p}|F_k(\omega_x,\omega_y)| \\ \sum_{k=1}^{p}|D_k(\omega_x,\omega_y)| &= \frac{1}{p}|H_d(\omega_x,\omega_y)|\sum_{k=1}^{p}|F_k(\omega_x,\omega_y)|\end{aligned} \tag{3.27}$$

The MTFs $|H_n(\omega_x,\omega_y)|$ and $|H_d(\omega_x,\omega_y)|$ have here been taken out of the summation because they are assumed independent of the subimage. The MTF ratio can then be written

$$\frac{|H_n(\omega_x,\omega_y)|}{|D_k(\omega_x,\omega_y)|} = \frac{\sum_{k=1}^{p}|N_k(\omega_x,\omega_y)|}{\sum_{k=1}^{p}|D_k(\omega_x,\omega_y)|} \tag{3.28}$$

This demonstrates how the MTF ratio can be estimated by forming the average of all the numerator and denominator modulus transforms before performing the division, even in the case where the image content for the subimages is not the same.

The reason why $\sum_{k=1}^{p}|N_k(\omega_x,\omega_y)|/\sum_{k=1}^{p}|D_k(\omega_x,\omega_y)|$ is a better estimate of the MTF ratio than $|N_k(\omega_x,\omega_y)|/|D_k(\omega_x,\omega_y)|$ is the same as for the case of averaging across similar images to reduce noise. The noise contributions are statistically independent in each case, so the average of the images has less noise than each independent image. This holds, despite the fact that the signal content is not in each case equivalent. When the ratio is formed the most probable value for any pixel is the mode of the corresponding distribution. Since after averaging there is less noise in the quantities forming the ratio, this mode becomes a better estimate of the actual pixel value. This fact is clearly demonstrated in Figure 3.5.

The use of this technique in the extraction of MTFs has been noted by Lim [19, p.559] without justification. This discussion however shows incontrovertibly the reasons why it is an improvement on the ratio estimated without the averaging.

3.4.2 Reducing noise after formation of ratio

The point-spread function of a monochromatic imaging system is circularly symmetric in the absence of coma, astigmatism and distortion. This can be deduced from the discussion in Appendix E: For this case the modified aperture function is

$$\psi_{mod}(\varepsilon,\eta) = Me^{ikK_s(\varepsilon^2+\eta^2)^2}A(\varepsilon,\eta) \tag{3.29}$$

which can be seen to be circularly symmetric in the (ε,η)-plane. Since circular symmetry is preserved by the Fourier transform, and the PSF of the system is dependent on the scaled

Fourier transform of ψ_{mod}, the PSF will be symmetric about the origin. This will also be the case for the MTF, which is also related to the PSF only through the Fourier transform. Hence the ratio of two such MTFs will have circular symmetry.

In lieu of this symmetry, there is redundancy in the 2-dimensional MTF ratio data. In the absence of noise, a 1-dimensional signal formed by considering the ratio along a line from the centre outwards will fully specify the MTF ratio. For the case of noise, the additional information present can be used to improve the estimate of this signal. Another advantage with using the 1-D representation is that it is easier and faster to manipulate.

If astigmatism is introduced, this symmetry ceases to exist. However, since the point-spread function of an astigmatic beam is a physical observable (and is often assumed to be a roughly elliptical Gaussian [7]), discontinuities would not be expected in the distribution. Additionally, since the PSF is band-limited in space, the MTF can also have no discontinuities. When the ratio of MTFs is considered, the same can be said for all points in the ratio where the denominator MTF is nonzero. This stems from a theorem in real analysis which states that if two functions f and g are continuous, then the ratio f/g is continuous for all points where $g \neq 0$ [9, p.252].

If an outward radial line from the centre of the MTF ratio is considered, the profile of the data along this line would thus not be expected to change dramatically as the angle of the line is slowly varied. Thus the MTF ratio can be assumed approximately circularly symmetric as long as only the data within a small angle of the required line is used. Figure 3.7 demonstrates a possible area from which data can be used in making an estimate of the MTF ratio along the solid line.

Given the radial redundancy it should therefore be possible, in the presence of both spherical aberration and astigmatism, to use closely-spaced neighbours to a given pixel to reduce noise in the estimate of its value. These neighbours should be at the same distance from the centre of the MTF ratio, but at slightly different angles from the line along which the ratio is required. In this way a noise-reduced profile of the MTF ratio can be formed given the angle at which this profile is needed.

The complications that arose in the previous section in forming averages of ratio signals also occur in this case. The independent samples to be used in the averaging process have all been subject to the division process which invalidates the procedure of taking the mean. For the reasons discussed in section 3.3.3, the median of the samples should therefore be used to form the noise-reduced estimate.

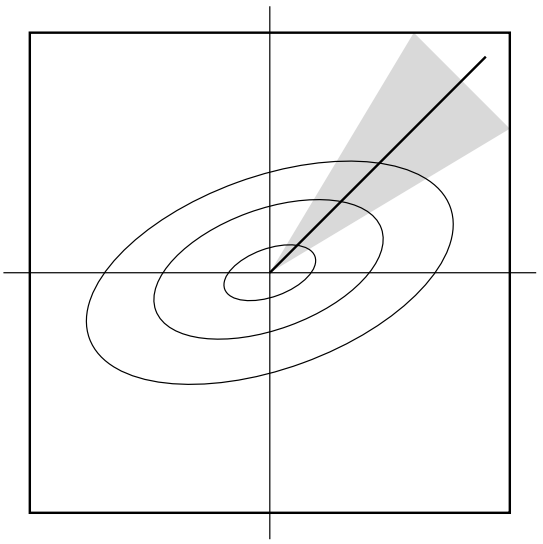


Figure 3.7: Area from which data is used in forming an estimate of the MTF along a line

Chapter 4

Determinism in Image Formation

The image formation process is now analysed to identify underlying consistency which can be made use of in an autofocus context. Before continuing with this, however, the effect of the Fourier transform having to be discrete for calculation purposes is discussed.

The chapter opens with a short discussion of the effects of windowing in the discrete Fourier transform. This is for completeness only; windowing is not considered in any detail in this thesis. The fact that results obtained later do not seem affected by windowing is seen as justification enough for this omission.

A complete procedure for forming the MTF ratios using all the noise reduction techniques is then presented. Procedures to follow in exception situations (such as division by zero) are also defined. It is shown how this procedure can be used to find the effective MTF of the system for any defocus level. Examples of two such MTFs are shown.

The section following that analyses the MTFs obtained for trends in the MTF width with respect to defocus level. Two cases are considered, one using the assumption of a Gaussian MTF and the other using a template. It is shown that for both of these characterisations the width of the MTF varies approximately linearly with changing focus level. Furthermore, it is demonstrated that the MTF profile along the length of the beam is essentially constant except for a scaling in the overall width. This is referred to as the self-similarity property.

These findings are then used in the construction of an extremely general beam model. This model uses only the assumptions of linearity of width with respect to defocus, and self-similarity. The model is additionally extended under the optional restriction of a Gaussian beam profile.

4.1 Artifacts from the discrete Fourier transform

So far the discussion has evolved around infinite extent signals which are analytic everywhere. In order to calculate Fourier transforms, however, these signals have to be truncated and sampled. The effects of these operations on the process of forming the ratio will briefly be discussed.

By the nature of the discrete Fourier transform, every finite signal is assumed to be periodic, replicating itself in prior and subsequent intervals. If the N-point transform is considered, this replication occurs at N-point intervals. Figure 4.1(a) shows the effective infinite extent signal that gets transformed when only a finite portion of a signal is considered. It can be seen that, with regard to the Fourier transform, the simple action of extracting a portion of the signal in effect multiplies the original by a rectangular window.

The major problem with this scenario is that the resulting signal is no longer continuous. When this signal is then transformed, frequency components are introduced which might obscure the desired transform coefficients.

An effective way to minimise this effect is to make use of special windowing when extracting the finite signal from its infinite extent counterpart. Figure 4.1(b) shows a typical window that is used for this purpose, and the resulting signal as transformed by the DFT. It can be seen that the discontinuities inherent in the previous case have been eliminated here.

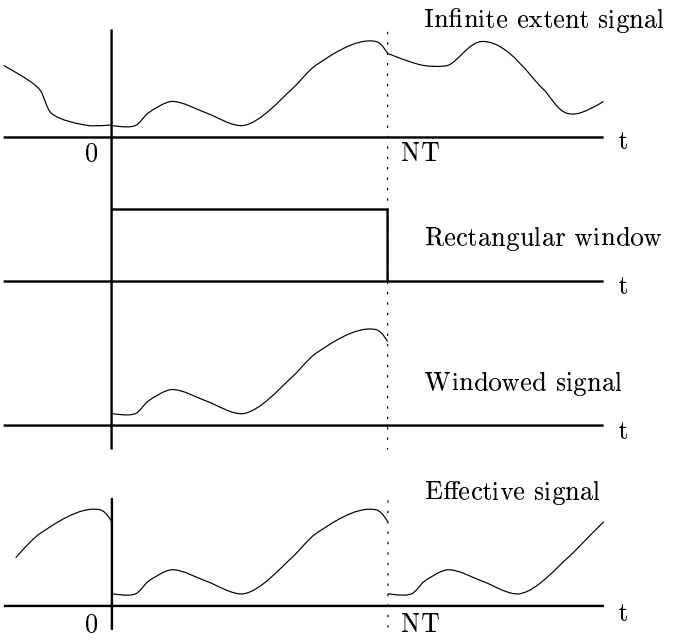
The action of windowing will always have the effect of altering the signal which is being transformed, since it is impossible to take the DFT of an infinite extent signal. Different windows will have varying effects on the transform, and a window must therefore be chosen which retains the required information for any given purpose. A complete analysis of the effect of sampled windows for DFTs has been made [13].

In the frequency domain this multiplication by a window can be effected by convolving the infinite transform with the transform of the window function. This has the effect that if noise is ignored, the Fourier transform of the windowed signal will be

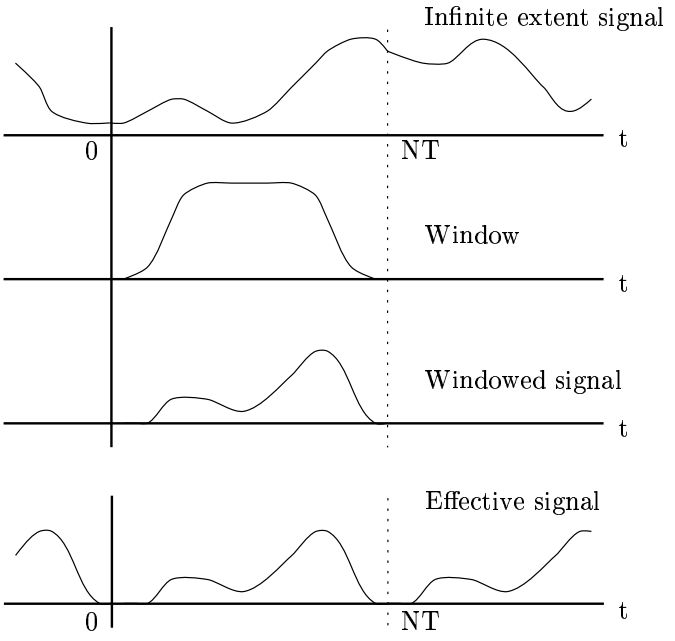
$$[F(\omega_x, \omega_y)H(\omega_x, \omega_y)] \otimes W(\omega_x, \omega_y) \tag{4.1}$$

where F is the image field, H the transfer function, and W the window function in the frequency domain. In the case of a rectangular window, this function is of the form $\sin(x)/x$. It is evident that if the ratio of two such windowed and transformed signals is taken, then the dependence on F is not entirely eliminated, some of it remaining due to the windowing.

No discussions about specific windows and their effects will be given in this thesis. In all cases the implied rectangular window will be used. The justification for this decision will be



(a) Rectangular window



(b) Customised window

Figure 4.1: Effect of finite extent processing of infinite extent signals

confirmed in subsequent chapters, where it will be seen that the results do not appear to be compromised by the window used.

4.2 Procedure for forming MTF ratio

In previous sections valid ways of dealing with the noise in the images and their Fourier transforms were presented. A brief discussion of windowing and the effects thereof was also given. This section will discuss the overall strategy for obtaining a signal which is representative of the ratio of two MTFs derived from two different out-of-focus conditions.

Each image is tiled into a number of square subimages. Since the test images obtained from the microscope are 1024×768 pixels, 12 subimages of dimension 256×256 were chosen. If the subimage is chosen to be larger, the resolution in the resulting Fourier transform will be higher, but fewer subimages will be available. The number 256 was chosen mainly because it is a power of 2 and because it is fairly small. Both of these factors facilitate a transform that will be quick to calculate. Each of these subimages is transformed and the modulus taken, and a pixel-by-pixel average of the results created.

As has been mentioned, the effect of windowing was ignored. It was however felt that the major effect of the rectangular window could possibly already have been taken into account in the constant offset factor C in equation 3.18.

This factor C is then found by averaging a number of pixels in the periphery of an image which is known to be considerably out of focus. Once found, this value is subtracted from the subimage average that was formed. The signal that results represents either $\sum_{k=1}^{p} |N_k(\omega_x, \omega_y)|$ or $\sum_{k=1}^{p} |D_k(\omega_x, \omega_y)|$ in equation 3.28. The numerator and denominator subimages are then chosen, and that equation can be used to form the noise-reduced 2-D representation of the MTF ratio.

Following the formation of the ratio, the angle must be specified at which the 1-dimensional MTF ratio profile is required. The median of all the points at each distance from the origin which lie within some angle of the required line is then found. Naturally some quantisation of distances has to be performed here, since all pixels do not lie at integral distances from the origin.

It should be noted that in the division process, division by zero is flagged as positive or negative infinity, depending on whether the numerator is positive or negative. These singularities are usually eliminated in the formation of the 1-D signal by the process of taking the median. Division of zero by zero is here defined to be zero, since this will result in the zeros of the ratio being the same as those of the numerator, which may be useful in later manipulations.

The resulting signal is then considered to be the noise-reduced profile of the MTF ratio along a line at that specific angle.

Calculation of the MTF

By dividing each transformed out-of-focus image by the transform of the in-focus image, the dependence on the specimen is effectively removed and what remains is the ratio of the two MTFs used to form the original images. However, insofar as the denominator MTF is a uniform constant, this quantity is just the MTF of the microscope for the out-of-focus level at which the numerator image was taken. This corresponds to the assumption that the image taken at best focus is a good approximation of the image field itself, or equivalently, that the image was formed using a PSF that very closely approximated a point. It should be noted that in fact the dependence on the specimen is not entirely removed by this process; the windowing discussed earlier prevents this from being exactly possible. For now this fact is ignored. The favourable results obtained later in this chapter suggest that this omission is not too detrimental.

This procedure was followed for the eight images **far2** to **far9**, dividing their transforms by the transform of **far1**. The division process used is the one described above, where a constant value is subtracted from each transformed subimage average before ordinary division is done. This results in 8 representations of the MTF of the system for the 8 corresponding out-of-focus levels. Two such MTF representations are shown in Figure 4.2 and Figure 4.3.

Generating 1-D representations

To reduce noise further, the radial averaging technique that was discussed is employed. Here, pixels in an entire sector surrounding a line are considered, and for all pixels the same distance from the origin the median is formed. In doing so the points further from the origin are subjected to a greater degree of averaging (and hence noise reduction) than those nearer the origin. However, since the signal is so much higher nearer the origin, the signal to noise ratio here is fairly high in any case and less averaging is needed. For the case of an astigmatic image where radial symmetry does not exist, there is a limit to the angular size of the sector that can be used in this process before this averaging begins to significantly corrupt the signal.

For each of the MTFs generated in the previous section, this radial averaging process was performed. Four of the resulting MTF profiles are shown in Figure 4.4 for the first four out-of-focus levels.

Also shown on these plots is the best-fit Gaussian to each of these transfer functions. They are best-fit in the least-squares sense, which means they minimise the sum of the differences

Section 4.2: Procedure for forming MTF ratio

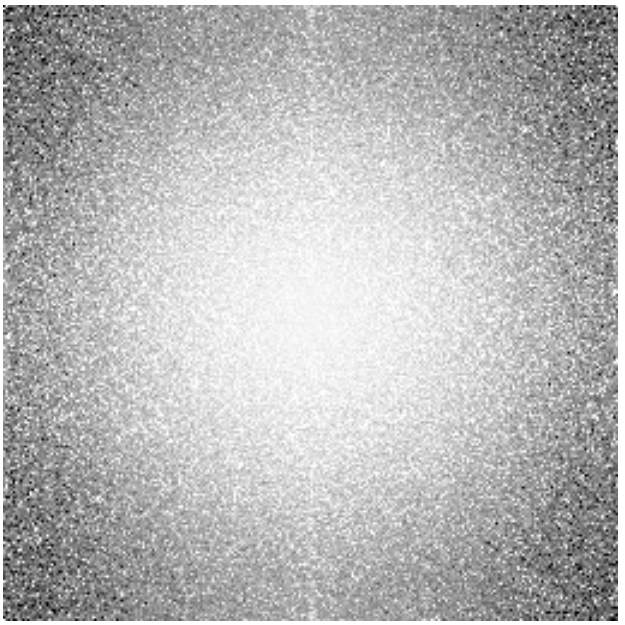


Figure 4.2: MTF extracted for out-of-focus distance $0.2mm$ corresponding to image **far3**.

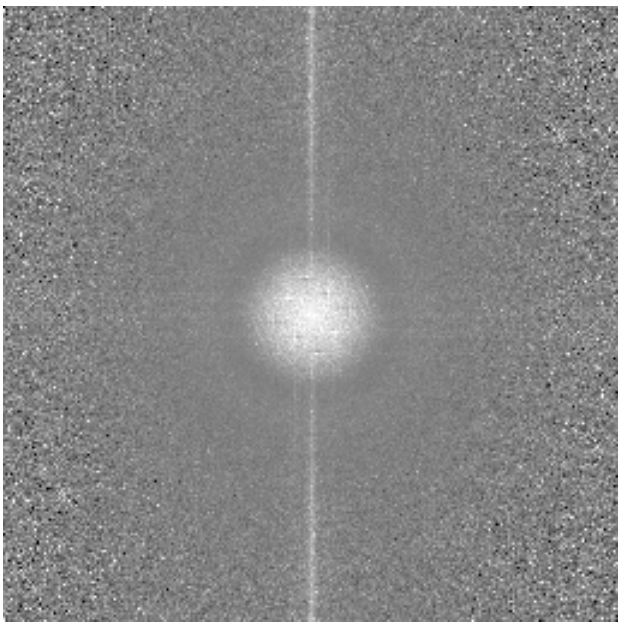
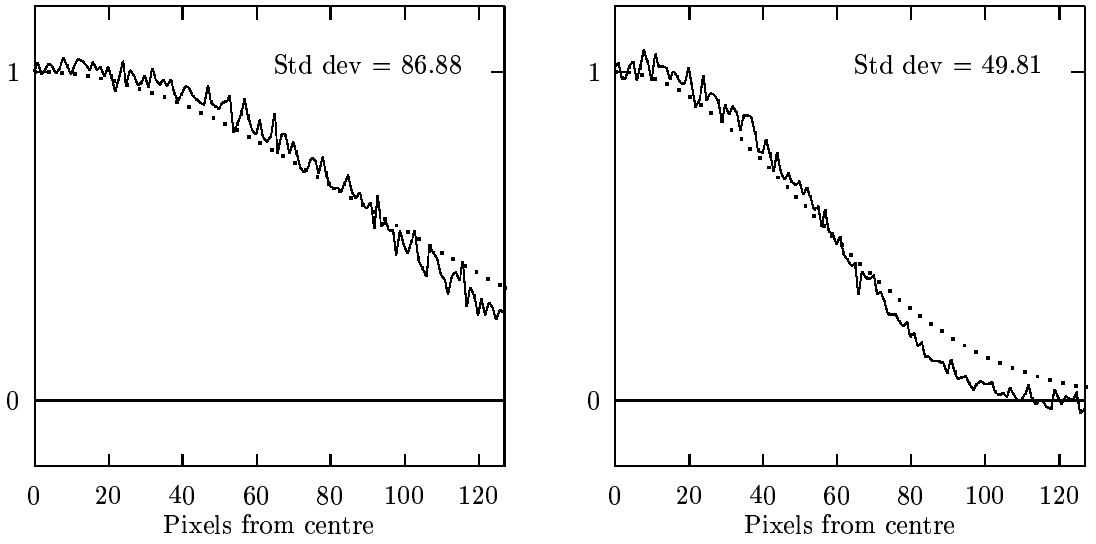
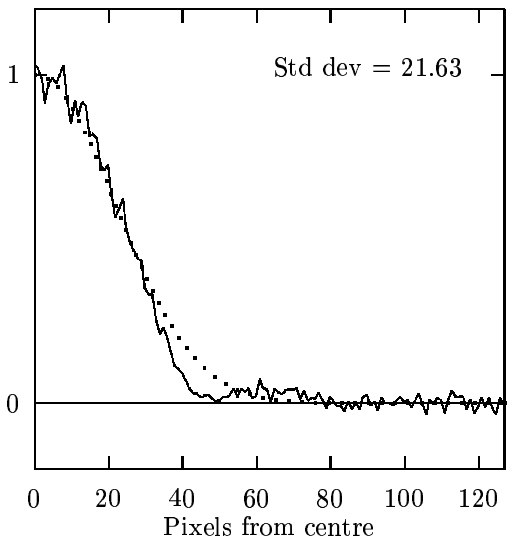


Figure 4.3: MTF extracted for out-of-focus distance $0.7mm$ corresponding to image **far6**.

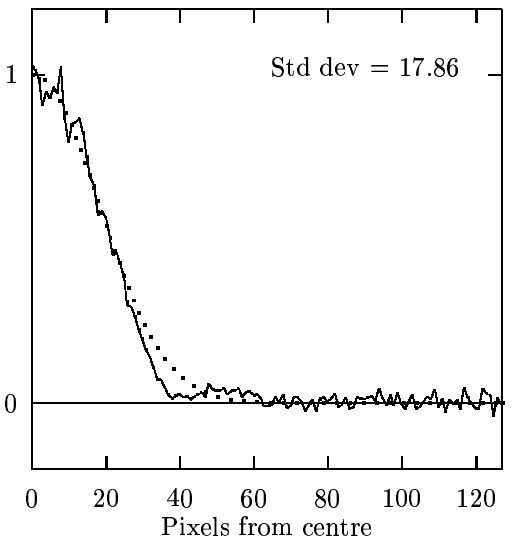


(a) far2 (0.1mm)

(b) far3 (0.2mm)



(c) far4 (0.5mm)



(d) far5 (0.6mm)

Figure 4.4: Noise-reduced profiles of MTFs corresponding to images **far2**, **far3**, **far4**, and **far5** (solid lines). Also shown (in dotted lines) are the best-fit Gaussians to these functions

between the two functions at every point for the range displayed on the graphs. This is to demonstrate the degree to which the widely made assumption of a Gaussian beam profile is accurate.

Of note is the apparent similarity between the MTFs for the different focus levels. It seems reasonable to say that each MTF is just a stretching or a compression in the horizontal direction of any other MTF. This property will be referred to as the self-similarity in the MTF at different positions along the beam. The characteristics of this self-similarity will be discussed in the following section.

4.3 Analysis of results

The results presented in the previous section will now be discussed in terms of their significance to the development of an autofocus system. The data needs to be analysed to detect any trends that exist that can be used to provide information regarding the degree of defocus for any given image.

4.3.1 Relating the spatial to the frequency domain

Before the results are presented, the nature of the relation between the spatial and the frequency domain needs to be clarified, particularly with respect to self-similarity.

Consider the Fourier transform property

$$f(ax, by) \Longleftrightarrow F(\frac{\omega_x}{a}, \frac{\omega_y}{b}) \tag{4.2}$$

where $F(\omega_x, \omega_y)$ is the Fourier transform of $f(x, y)$. $f(ax, by)$ is a function that represents $f(x, y)$ compressed by factors a and b in the x and y directions respectively. This relation now states that the Fourier transform of the function $f(ax, by)$ will be related to the function $F(\omega_x, \omega_y)$ by an overall scaling and a stretching in these two directions, by the same factors.

The effect of this is that if the MTFs are assumed to be self-similar, then the corresponding system PSFs will also be self-similar. Furthermore, if MTF_1 is twice as wide as MTF_2, then the corresponding PSF_1 will be half as wide as PSF_2.

4.3.2 Gaussian approximation

Figure 4.4 showed Gaussian curves fitted to the extracted MTFs for each out-of-focus level. It can be seen that the Gaussian only roughly captures the shape of the MTF. Apparently the

slope of the MTF is steeper than can be approximated by a Gaussian, and it has a sidelobe which the Gaussian cannot match. We will however proceed with the analysis ignoring these differences, since on the whole the Gaussian approximations seem to be reasonable indications of the width.

The standard deviations of the best-fit Gaussian curves are given in Table 4.1 for each distance from focus. The final column shows the standard deviation of these Gaussian curves as referred back to the space domain. Therefore, inasmuch as the MTFs approximate Gaussian distributions with the standard deviations given in the third column, the PSF in the space domain will approximate a Gaussian with standard deviation as given in this last column.

Image Name	Dist from focus (mm)	std dev freq (pixels)	std dev space (m)
far2	0.1	86.225	1.157e-7
far3	0.2	48.171	2.070e-7
far4	0.5	20.820	4.790e-7
far5	0.6	17.351	5.747e-7
far6	0.7	15.072	6.616e-7
far7	1.0	10.117	9.857e-7
far8	1.1	9.027	1.105e-6
far9	1.2	8.234	1.211e-6

Table 4.1: Standard deviations in spatial and frequency domains of best-fit Gaussians to the MTF for each defocus level

The conversion factor from the frequency to the spatial domain had to be found empirically from a known image. It is a scaled reciprocal relation.

Now if the standard deviation in the frequency domain is plotted as a function of out-of-focus distance (see Figure 4.5), it can be seen that the curve closely approximates a hyperbola. This becomes more evident if the standard deviation of the Gaussians in the spatial domain is plotted (see Figure 4.6); a hyperbola in the frequency domain will manifest itself as a linear relationship in this domain.

From these plots it seems reasonable to suggest that there is a linear relation between the standard deviation of the assumed Gaussian PSF and the distance from focus. This can certainly be said to be the case for the distances from focus that are being considered here.

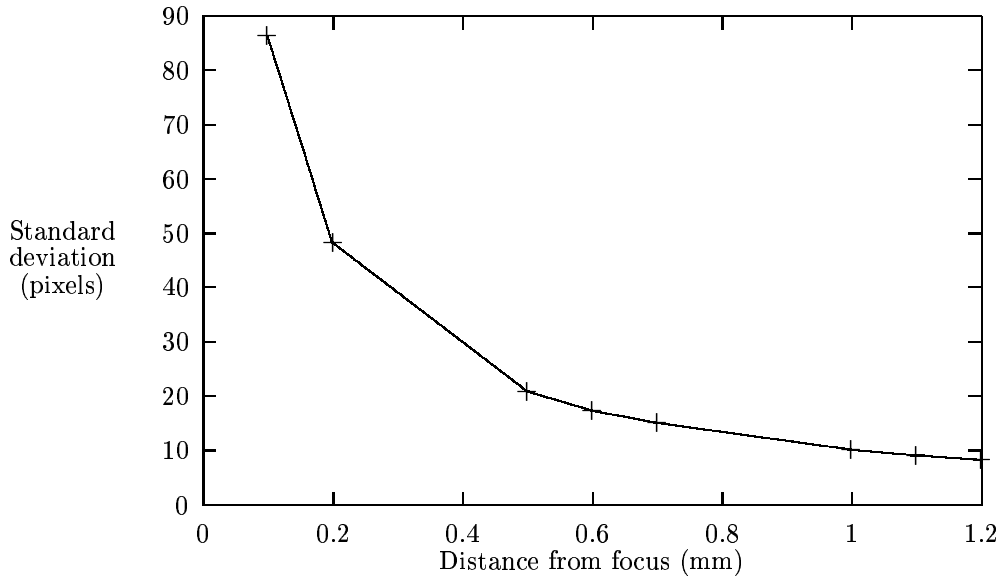


Figure 4.5: Standard deviation of best-fit Gaussians in the frequency domain plotted as a function of the distance from focus

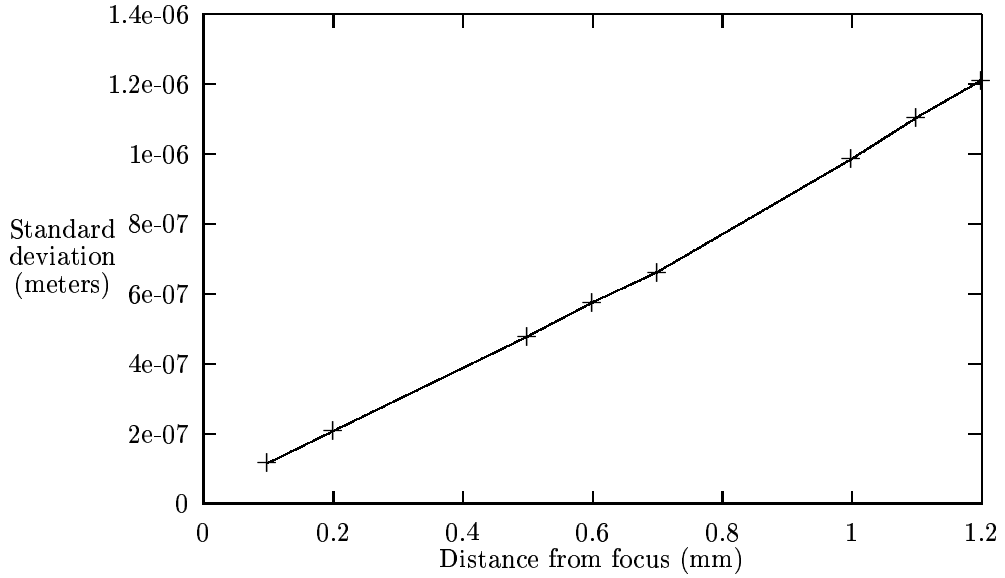


Figure 4.6: Standard deviation of best-fit Gaussians referred to the spatial domain plotted as a function of the distance from focus

4.3.3 Self-similarity approximation

Although the results from the Gaussian approximation are promising and suggest that there is a large degree of determinism in the operation of the system, it is difficult to develop further using this model. Instead, we would prefer to use the more general assumption of self-similarity of the MTFs, for which the Gaussian assumption can be said to be a specific case.

An analysis was made using the MTF corresponding to image **far3** as a template for comparing the other MTFs. For each out-of-focus level, the MTF was stretched or compressed by an amount that minimised the mean square difference between this modified function and the template MTF. For example, Figure 4.7 shows the MTF used in the formation of image **far6** as well as the MTF for image **far3** that is to be used as the template. Figure 4.8 now

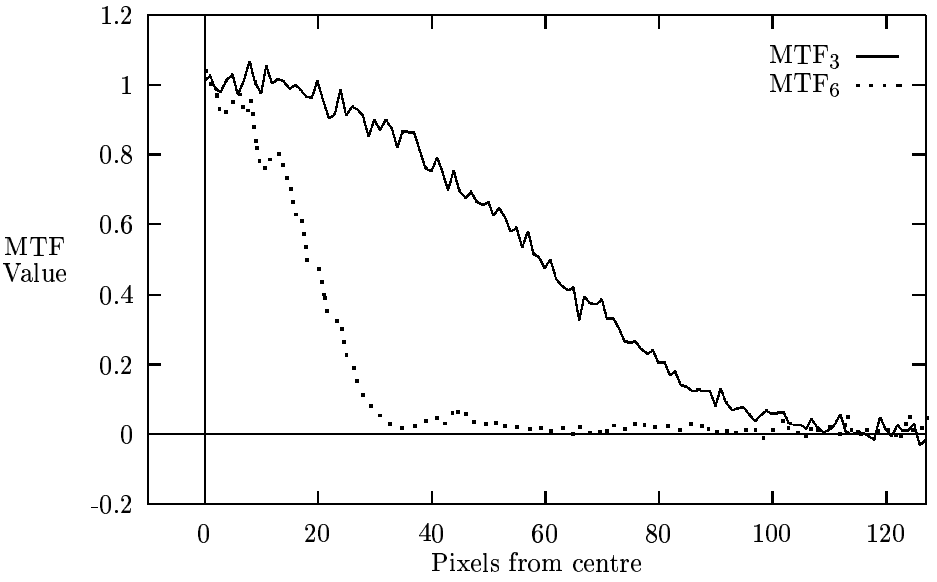


Figure 4.7: MTFs corresponding to images **far3** and **far6**

shows the same template MTF, but the MTF for image far6 has been stretched horizontally by a factor of 3.2041. This was found to be the optimal stretch for minimising the difference between these two functions.

This process of finding the optimal stretch for matching an MTF to the template was performed for the eight out-of-focus levels, and the results presented in Table 4.2. If a plot is made of this stretch value against the distance from focus, it is seen that here too a linear relationship exists (Figure 4.9).

Thus it can be seen that the assumption of self-similarity allows for a very simple model

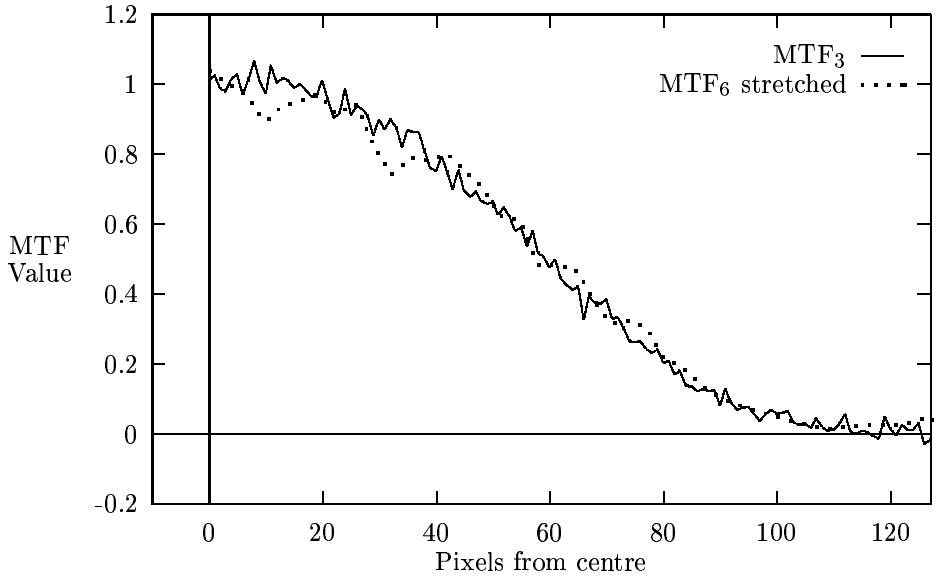


Figure 4.8: MTFs corresponding to images **far3** and **far6**, except that here the latter has been stretched horizontally by a factor of 3.2041

Image Name	Dist from focus (mm)	Best stretch
far2	0.1	0.597561
far3	0.2	1
far4	0.5	2.309
far5	0.6	2.768
far6	0.7	3.179
far7	1.0	4.707
far8	1.1	5.263
far9	1.2	5.783

Table 4.2: Optimum stretch values for matching MTFs corresponding with each out-of-focus level to that of **far3**

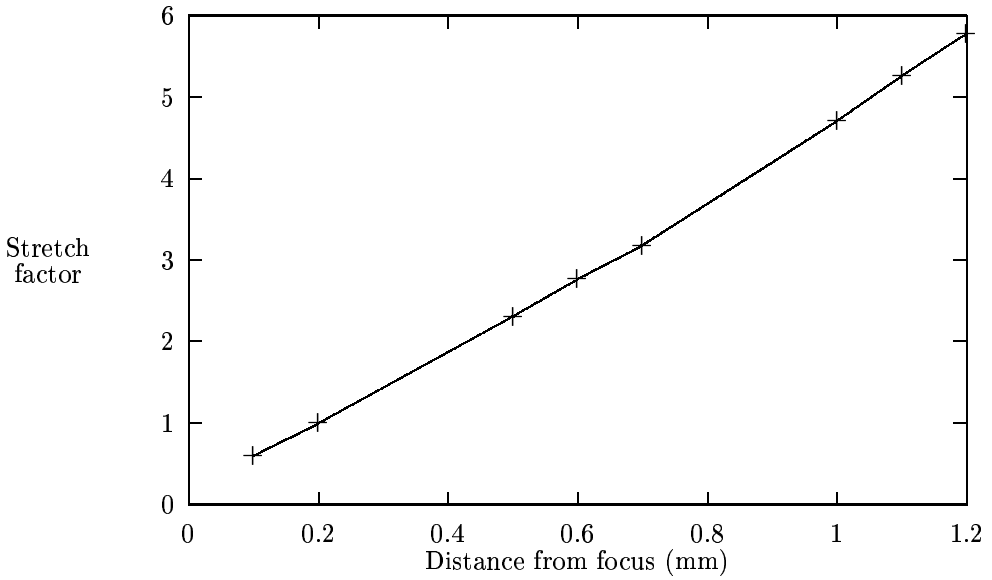


Figure 4.9: Factors by which MTFs need to be stretched, plotted as a function of the distance from focus

of image formation to be constructed. There is a linear relation between the width of the MTF and the distance from focus. This model is powerful because it eliminates the need to know anything about the shape of the MTF. It is thus more general than the Gaussian characterisation.

4.4 Construction of Beam model

The previous section demonstrated how, if the relative width of the MTFs are examined, the size of the point-spread function of the system is found to be proportional to the distance from the beam crossover. This conforms to the suggestion that the electron trajectories are straight lines, with an approximate zero crossover at the point of focus.

In order to find out more about the specifics of image formation in the SEM, it becomes important to have some idea about how the beam configuration affects the resulting images. To do this it helps to have a simple parameterised representation of the beam current density profile. This can then be used to discover trends in the image formation process. The findings of the previous section provide this simple representation. If it is found to be necessary, a more representative model can be developed.

4.4.1 Simple linear model

For purposes here an aberration-free electron beam will be defined to be ideal if there are no diffraction effects and the beam has a zero crossover at the point of focus. The trajectories of the electrons will be assumed to be straight lines. Such a configuration with a focal length f_x is shown in Figure 4.10.

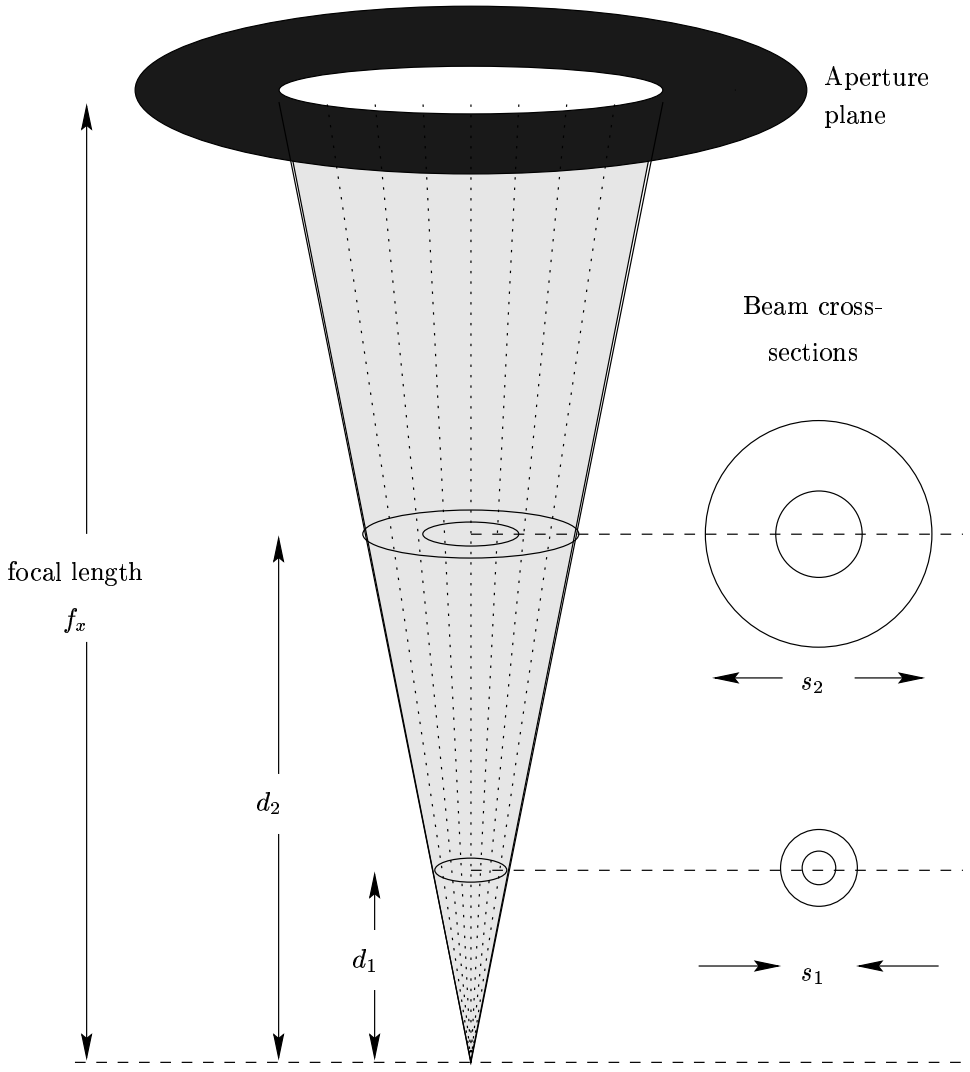


Figure 4.10: Ideal Electron Beam

Since the electron paths are straight lines, the total current through a tube with sides coinciding with electron trajectories will be conserved. If the beam profiles at distances d_1 and d_2 from focus are considered, with corresponding beam widths s_1 and s_2, then the following

relation holds

$$J_1(x,y) = (\frac{s_2}{s_1})^2 J_2(\frac{s_2}{s_1}x, \frac{s_2}{s_1}y) \tag{4.3}$$

where J_1 and J_2 are the current density distributions of the beam at those positions. The scaling factor $(s_2/s_1)^2$ is required to ensure that the total beam current is conserved. Additionally it is assumed that the beam width s is a linear function of the distance from focus d, and that at focus the width is zero. If a hypothetical current density distribution $J_0(x,y)$ is now defined for a distance f_x from focus (i.e. at the aperture), then for any distance d from focus the beam profile is

$$J(x,y) = (\frac{f_x}{d})^2 J_0(\frac{f_x}{d}x, \frac{f_x}{d}y) \tag{4.4}$$

Note that this relation will certainly not hold in practice for $d \approx f_x$ since aperture effects will dominate, nor for $d \approx 0$, where the prediction of a point beam will be an idealisation.

Linear model in the frequency domain

Under the assumptions made above, the beam in the frequency domain can be represented in terms of the distance from focus. Using Equation 4.4, in the Fourier transform domain it can be said that

$$\mathcal{F}[J(x,y)] = \mathcal{F}[k^2 J_0(kx, ky)] \tag{4.5}$$

where $k = f_x/d$ and $\mathcal{F}$ represents the Fourier transform operation. Letting $\mathcal{F}[J(x,y)] = \mathcal{J}(\omega_x, \omega_y)$ and $\mathcal{F}[J_0(x,y)] = \mathcal{J}_0(\omega_x, \omega_y)$, this becomes

$$\mathcal{J}(\omega_x, \omega_y) = k^2[\frac{1}{|k|^2}\mathcal{J}_0(\omega_x/k, \omega_y/k)] \tag{4.6}$$

and since k is real,

$$\mathcal{J}(\omega_x, \omega_y) = \mathcal{J}_0(\omega_x/k, \omega_y/k) \tag{4.7}$$

Thus it can be seen that for this model the frequency domain beam representation for different distances from focus differs only in a linear scaling in the ω_x and ω_y directions. This is consistent with the results presented in the previous section, where the relationship between MTF width was seen to be linear.

4.4.2 General representation of Gaussian

A widely made assumption in SEM literature is the assumption of a Gaussian PSF [7, 26]. For the situation here this has been shown to be a reasonable assumption. At least for low frequencies the mainlobe of the MTF approximates a Gaussian.

The general form of a 2-D Gaussian normalised to unit volume is

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{\frac{-x^2}{2\sigma_x^2}} e^{\frac{-y^2}{2\sigma_y^2}} \tag{4.8}$$

Here σ_x and σ_y are the standard deviations in the x and y directions respectively. Note that the 2-dimensional Gaussian is separable, and can be written in the form $p(x,y) = p_1(x)p_2(y)$. The frequency domain representation of this Gaussian will be

$$\mathcal{P}(\omega_x,\omega_y) = e^{\frac{-\sigma_x^2\omega_x^2}{2}} e^{\frac{-\sigma_y^2\omega_y^2}{2}} \tag{4.9}$$

This can be seen to be a 2-D Gaussian of the form

$$\mathcal{P}(\omega_x,\omega_y) = e^{\frac{-\omega_x^2}{2(\sigma_{\omega_x})^2}} e^{\frac{-\omega_y^2}{2(\sigma_{\omega_y})^2}} \tag{4.10}$$

with $\sigma_{\omega_x} = 1/\sigma_x$ and $\sigma_{\omega_y} = 1/\sigma_y$. Thus the transform of a Gaussian is simply a rescaled Gaussian with standard deviation the reciprocal of that of the original. The simple closed-form of this solution is very convenient, and it is significant to note that the transform of a Gaussian has no imaginary part.

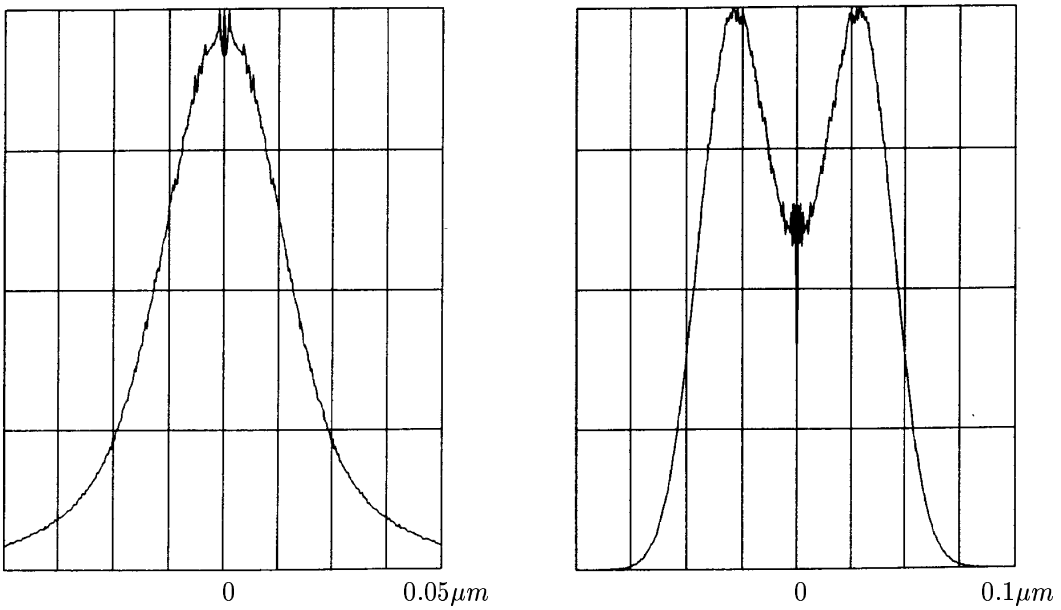
Assuming the beam current density to be Gaussian implies a nonzero current density infinitely far from the centre of the beam. Clearly this can never be the case.

4.5 Computed beam profiles

The previous sections provide for a model of the beam which exhibits some predictable characteristics. The primary assumption is that of self-similarity of the MTF profiles for varying degrees of out-of-focus.

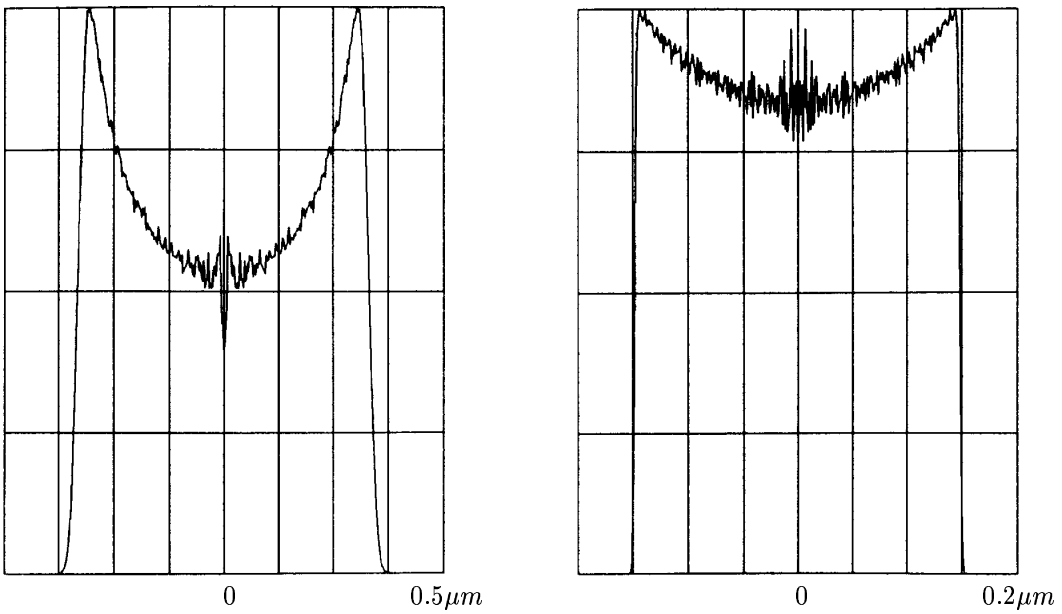
It is possible to calculate the actual electron density of the beam at different positions along its length. Appendix E provides an outline of some of the relevant theory needed to achieve this. The difficulty then becomes that of finding values of the parameters to adequately model the beam. In our case this was not necessary - Leica Cambridge Ltd contracted out to a specialist company to compute a number of beam profiles for the S440 at different positions along the beam.

Altogether 37 beam profiles were computed, starting in the Gaussian image plane and extending to a total defocus of $10mm$. Some examples of the resulting profiles are shown in Figure 4.11. The operating voltage for the calculations was $10kV$. It is interesting to note that although the profile is Gaussian near best focus, this becomes untrue as the defocus is



(a) Defocus: $0mm$

(b) Defocus: $0.01mm$



(c) Defocus: $0.05mm$

(d) Defocus: $0.2mm$

Figure 4.11: Profiles of the electron density of the beam for the S440 for varying distances from the Gaussian image plane

changed. The distribution quickly becomes bimodal, and eventually tends to uniformly flat distribution as the defocus becomes large enough.

Although these data may seem to refute the self-similarity assumption, this is not the case for two reasons:

- Despite the fact that the PSFs vary widely with changing defocus, the transformation to the frequency domain reduces this variation. Thus the MTFs may still exhibit significant self-similarity.
- The secondary electron yield has been shown to be a linear convolution of the image field with the electron density of the beam. This does not necessarily apply to the final detected signal, however, since the transfer function of the rest of the system has still to be included. The effective PSF is thus the electron density modified by this transfer function.

Chapter 5

Autofocus Method

In section 2.5 the approach to developing an autofocus algorithm was presented. It was shown that if two images are taken of the same area of the specimen but using different focal lengths, then the ratio of the transforms

$$\frac{F_1(\omega_x, \omega_y)}{F_2(\omega_x, \omega_y)} = \frac{H_1(\omega_x, \omega_y)}{H_2(\omega_x, \omega_y)} \tag{5.1}$$

might be useful in finding the position of the crossover of the beam, which is the position of best focus.

Methods were then discussed of how a reduction of noise can be effected in the formation of a 1-dimensional representation of this ratio. This 1-D signal is only representative of a radial cut through this ratio at a specified angle, due to possible astigmatism in the beam.

Finally in the previous chapter it was shown that the MTF (and hence the PSF) of the system is roughly the same for varying out-of-focus levels, except for a change in the overall width. Thus each MTF can be considered to be a stretching or compressing of any other MTF in the horizontal direction. Furthermore, it was shown that the width of the PSF is approximately proportional to the distance from the beam crossover.

In the sections which follow, all these findings are brought together in outlining ways in which knowledge of such ratios and the associated positions of their formation can be used to find the position of best focus. The methods presented in this chapter assume that the electron beam forming the images is free from astigmatism. The case for the introduction of astigmatism is deferred until later.

The chapter begins by formalising the linearity criterion. It is shown that this linearity assumption can be used even if it is only differentially valid. The self-similarity condition is added, effectively resulting in a situation that can be modelled by the equation $f(kx)/f(x) =$

$g(x)$, where f represents the MTF profile and g the ratio signal. Equations of this form are then discussed in detail. It is shown that under continuity restrictions this equation has a unique solution in f, and a numerical method is proposed for calculating this solution. It is shown how, in theory at least, a general autofocus method can be developed using just these findings. A practical implementation is described which uses the notion of an implicit template. This implementation failed because it was sensitive to noise. A description of the precise cause of this failure in the procedure is then given.

The assumption of a Gaussian MTF is used to derive a closed-form solution to the autofocus problem. Two ratios are formed using three images taken at different focal lengths. The development proceeds incrementally, introducing restrictions as they are required. It is shown that under the assumptions made, three images is the minimum required to eliminate the system-specific parameters. The solution is restructured to incorporate a search method for finding the position of best focus. The solutions obtained for this search are known to be unique under certain conditions because of the closed-form solution. The restructuring to a search allows for the assumption of a Gaussian MTF to be dropped, which results in a general solution which uses a representative template MTF.

5.1 Development assuming linear model

Before beginning the discussion it is necessary to demonstrate the general configuration of a beam for a particular out-of-focus condition. Figure 5.1 shows the situation of a beam with a focal length f impinging on a specimen a distance d from the aperture. The lightly- and darkly-shaded regions combined represents the actual electron beam. For purposes of the figure this beam is shown to be slightly curved. However, given that the beam is approximately linear in the vicinity of the specimen, it is possible to extrapolate towards the aperture and define an effective ideal beam for the case of it being perfectly linear. This is depicted by the darkly-shaded region alone. In this way it is possible to use the linear assumption even in the case where it is only differentially valid. In Figure 5.1 a is then defined to be the effective width of the beam as referred to the aperture.

By similar triangles the expression for the width s of the beam at the specimen is

$$s = a\frac{|f-d|}{f} \tag{5.2}$$

The precise definition of this width has intentionally been left unspecified. It can therefore represent either the standard deviation of the beam profile or the factor by which some reference signal must be stretched or compressed to match this profile. Both these cases were shown to conform to the linear model in the previous chapter.

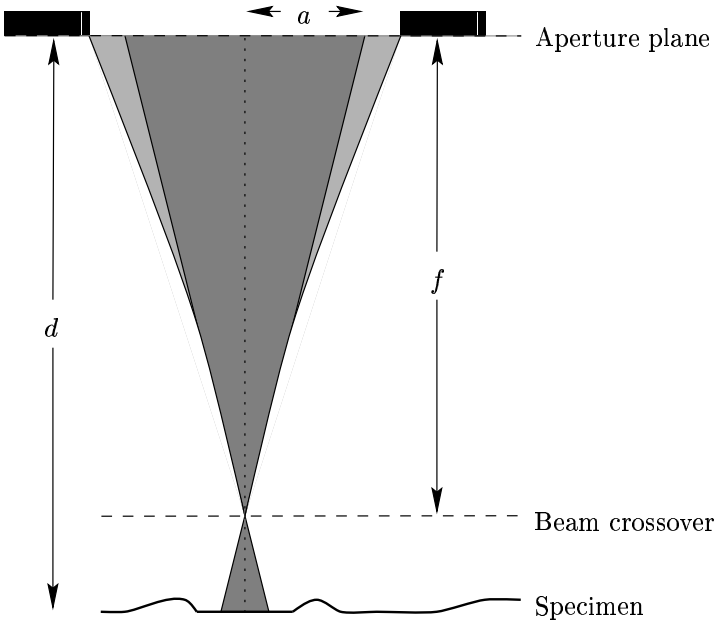


Figure 5.1: Standard deviation of Gaussian PSF with changing distance (under linear assumption)

5.1.1 Equations of the form $\mathbf{f(kx)/f(x) = g(x)}$

In order to get more information from the ratio F_1/F_2 (or equivalently H_1/H_2), it is helpful to discuss briefly the nature of the equation $f(kx)/f(x) = g(x)$, with $g(x)$ a known function and $k \in \Re, k > 1$ a specified constant. What will now be shown is that the equation has a unique solution in f under the restriction that f be continuous at the origin.

Suppose that the equation can be satisfied for two solutions $f_1(x)$ and $f_2(x)$. It is assured that

$$\begin{aligned} f_1(kx) &= f_1(x)g(x) \\ f_2(kx) &= f_2(x)g(x) \end{aligned} \tag{5.3}$$

Dividing, this then becomes

$$\frac{f_1(kx)}{f_2(kx)} = \frac{f_1(x)}{f_2(x)} \tag{5.4}$$

Now, letting

$$p(x) = \frac{f_1(x)}{f_2(x)} \tag{5.5}$$

it must be true that

$$p(kx) = p(x), \forall x \in \Re \tag{5.6}$$

Section 5.1: Development assuming linear model

Assuming continuity of $f(x)$ at 0, the ratio $p(x) = f(kx)/f(x)$ will be continuous at 0 provided that $f(0) \neq 0$ [9, p.252]. By one of the definitions of continuity it can be said that

$$\forall \epsilon > 0, \exists \delta > 0 \text{ s.t. } x \in (-\delta, \delta) \Rightarrow |p(x) - p(0)| < \epsilon \tag{5.7}$$

However, $p(x) = p(kx)$, so for the same ϵ, δ it must be the case that

$$x \in (-\delta, \delta) \Rightarrow |p(kx) - p(0)| < \epsilon \tag{5.8}$$

which is the same as

$$x \in (-k\delta, k\delta) \Rightarrow |p(x) - p(0)| < \epsilon \tag{5.9}$$

Repeating this process n times, this becomes

$$x \in (-k^n\delta, k^n\delta) \Rightarrow |p(x) - p(0)| < \epsilon \tag{5.10}$$

Now since δ is a finite constant and $k > 1$, the condition $x \in (-k^n\delta, k^n\delta)$ can be extended to include the entire real line by taking n large enough. Thus equation 5.7 can be written

$$\forall \epsilon > 0, x \in \Re, |p(x) - p(0)| < \epsilon \tag{5.11}$$

In the metric space of functions under the supremum metric

$$\sigma(f(x), g(x)) = \sup_{x \in \Re} |f(x) - g(x)| \tag{5.12}$$

it can therefore be said that the distance $\sigma(p(x), p(0)) = 0$. If this is not the case then there must be some $\nu > 0$ such that ν is a least upper bound of $|p(x) - p(0)|$. This presents a contradiction since $\nu/2$ can be shown to be a least upper bound by choosing $\epsilon = \nu/2$ in equation 5.11. By the definition of a metric space, $p(x) = p(0)$ for all $x \in \Re$. Thus p is a constant function. Going back to equation 5.5, this shows that $f_1(x) = p(0)f_2(x)$. Since for a family of MTFs $f_1(0) = f_2(0)$, this scale factor $p(0)$ equals unity, and it must be the case that $f_1(x) = f_2(x)$. Therefore under the conditions of f continuous at 0, the equation $f(kx)/f(x) = g(x)$ has a unique solution in f.

For the condition $0 < k < 1$, the entire argument can be repeated using the inverse ratio instead. Of course for $k = 1$ there will be no solution unless $g(x) = 1$, in which case there will be infinitely many.

5.1.2 Finding f(x) for discrete data

Only the case of the numerator MTF being the one with the smaller extent ($k > 1$) will be discussed. For the reverse situation, if the inverse of the ratio is considered, corresponding results can be obtained.

Let f_n be the points of the numerator MTF with the smaller extent. The denominator MTF is $f_n^\star$. Using linear interpolation, $f_n^\star$ can be expressed in terms of f_n

$$f_n^\star = f_{\lfloor n/k \rfloor} + (n/k - \lfloor n/k \rfloor)(f_{\lceil n/k \rceil} - f_{\lfloor n/k \rfloor}) \tag{5.13}$$

where $\lfloor x \rfloor$ represents the largest integer smaller than or equal to x, and $\lceil x \rceil$ the smallest integer larger than or equal to x.

The ratio of the numerator to denominator MTF is

$$\frac{f_n}{f_n^\star} = g_n \tag{5.14}$$

So

$$f_n = g_n[f_{\lfloor n/k \rfloor} + (n/k - \lfloor n/k \rfloor)(f_{\lceil n/k \rceil} - f_{\lfloor n/k \rfloor})] \tag{5.15}$$

Since $k > 1$, knowing f_n up to n points will specify $f_n^\star$ to $\lfloor nk \rfloor$ points (by linear interpolation). This interdependency can be used to recursively find point f_n given the prior points $f_0 \ldots f_{n-1}$. Thus, with an assumed starting point f_0 and given g_n and k, f_n can be uniquely deduced. There are two cases:
$\mathbf{n = \lceil n/k \rceil}$: Under this condition, equation 5.15 becomes

$$f_n = g_n[f_{\lfloor n/k \rfloor} + (n/k - \lfloor n/k \rfloor)(f_n - f_{\lfloor n/k \rfloor})] \tag{5.16}$$

which solving for f_n yields

$$f_n = \frac{f_{\lfloor n/k \rfloor}[1 + \lfloor n/k \rfloor - n/k]}{1/g_n + \lfloor n/k \rfloor - n/k} \tag{5.17}$$

$\mathbf{n > \lceil n/k \rceil}$: For this situation equation 5.15 can be used as it stands.

These equations can be used because $f_{\lfloor n/k \rfloor} \in [f_0, \ldots f_{n-1}]$. This defines a method of extracting the assumed self-similar MTF given the difference k in stretch between them.

5.1.3 General autofocus method

In the previous section a method was presented of extracting either of the assumed self-similar MTF profiles from a ratio signal given only the stretch factor between the profiles. In theory this can form the basis for a method by which the position of the specimen can be determined.

Section 5.1: Development assuming linear model

Consider the configuration depicted in Figure 5.2. This now represents a specimen being

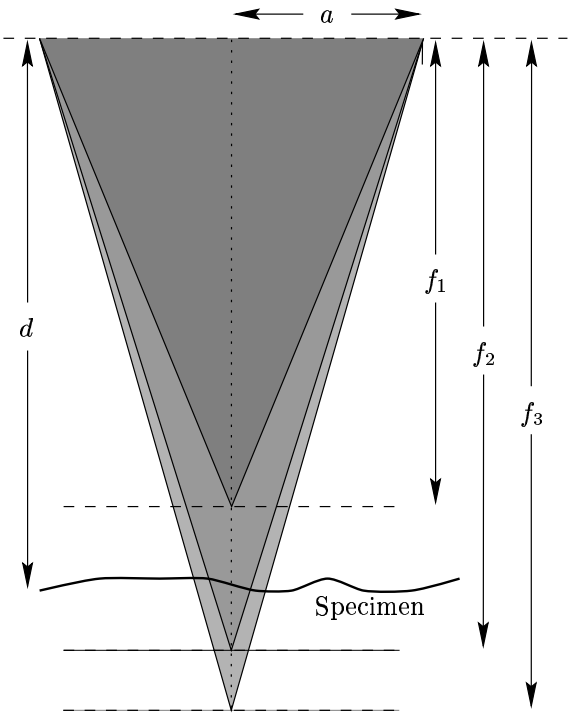


Figure 5.2: Configuration for prediction under linear assumption

imaged under three different focal lengths for the case of an ideal beam. Each dashed line indicates the position of the beam crossover during the formation of the images.

The width of the beam at the specimen for each case of the three focal lengths is

$$\begin{aligned} s_1(d) &= a\frac{|f_1 - d|}{f_1} \\ s_2(d) &= a\frac{|f_2 - d|}{f_2} \\ s_3(d) &= a\frac{|f_3 - d|}{f_3} \end{aligned} \tag{5.18}$$

where s_n is here chosen to be the width in the sense of stretch factors from a template profile as mentioned in section 5.1. This means that if $h_{ref}(x)$ is a template representing the assumed ideal PSF profile, then PSF_n is given by $h_n(\frac{x}{s_n})$.

Since PSF_1 and PSF_2 can be represented by this template stretched by factors s_1 and s_2, to a good approximation PSF_1 is just PSF_2 stretched by a factor s_1/s_2. It follows then that MTF_1 is MTF_2 stretched by the reciprocal of the ratio, s_2/s_1. Notice that these ratios are independent of the value a. The reason that this is said to be an approximation is that no

distance measure between the functions has been defined yet, and in general it will probably not be exactly true unless the profiles are precisely self-similar.

Defining $k_1(d) = s_1/s_2$ and $k_2(d) = s_3/s_2$, the ratios $\mathrm{MTF}_1/\mathrm{MTF}_2$ and $\mathrm{MTF}_2/\mathrm{MTF}_3$ can be formed. These quantities are now assumed to conform closely to the model

$$\begin{aligned} \frac{\mathrm{MTF}_1}{\mathrm{MTF}_2} &= \frac{H_2(k_1 x)}{H_2(x)} \\ \frac{\mathrm{MTF}_2}{\mathrm{MTF}_3} &= \frac{H_2(x)}{H_2(k_2 x)} \end{aligned} \tag{5.19}$$

where k_1 and k_2 depend on the unknown position of the specimen d.

If k_1 and k_2 can be found from these ratios, then equations 5.18 can be used to form two equations in d. However, unless something is known about the signs of the quantities $f_1 - d$, $f_2 - d$, and $f_3 - d$, there will be multiple solutions to these equations, some of which may be invalid.

A simpler way is to do a search through the possible specimen positions, knowing the configuration of the beam for the three focal lengths. For any assumed value d_{ass} of the specimen position, corresponding values of k_1 and k_2 can be calculated. If these values coincide with the values obtained from the ratios, then this value of d_{ass} is a possible value for the specimen position.

A complication arises in that no means has been found to obtain values of k_1 and k_2 directly from the ratios without specifying $H_2(x)$. It is however possible to use a search approach to check if a given set of ratios are consistent with an assumed specimen position d_{ass}. In order to do this the results from the previous section are required: values of k_1 and k_2 are calculated for this assumed position d_{ass}. The value obtained for k_1 is then used to find the function $H_2(x)$ from the ratio $H_1(k_1 x)/H_2(x)$. Similarly the value k_2 is used to find a second evaluation of this function $H_2(x)$, this time from the other ratio. Insofar as these two evaluations are identical, d_{ass} is a possible value for the specimen position d.

Notice that in no case is the actual MTF profile required for this procedure; it is completely general and only makes use of the assumptions that the width of the beam is differentially linear with respect to distance from crossover, and that the profiles are self-similar.

The proposed autofocus method is to find the two MTF ratios from the data, and then to search through the possible values of d_{ass}. The values for which the two estimates of $H_2(x)$ are most similar are then candidates for d. It is unclear how many such candidates there will be. Furthermore, since the function $H_2(x)$ has remained unspecified, it is impossible to assess how sensitive the estimates will be to noise in the ratios.

5.1.4 Results of general autofocus method

The search method described in the previous section was implemented and tested on a through-focus image series. It was found that it worked poorly, generally returning estimates for best focus which were very far from the actual distances.

An in-depth analysis was made of the process and the cause of this failure isolated. It was found that the process of extracting the estimate of the MTF from the ratio (given the stretch factor) was susceptible to the large degree of noise which is to be found in the MTF ratio profiles.

Figure 5.3 shows the actual MTF associated with image **far 4** ($0.5mm$ from crossover). The

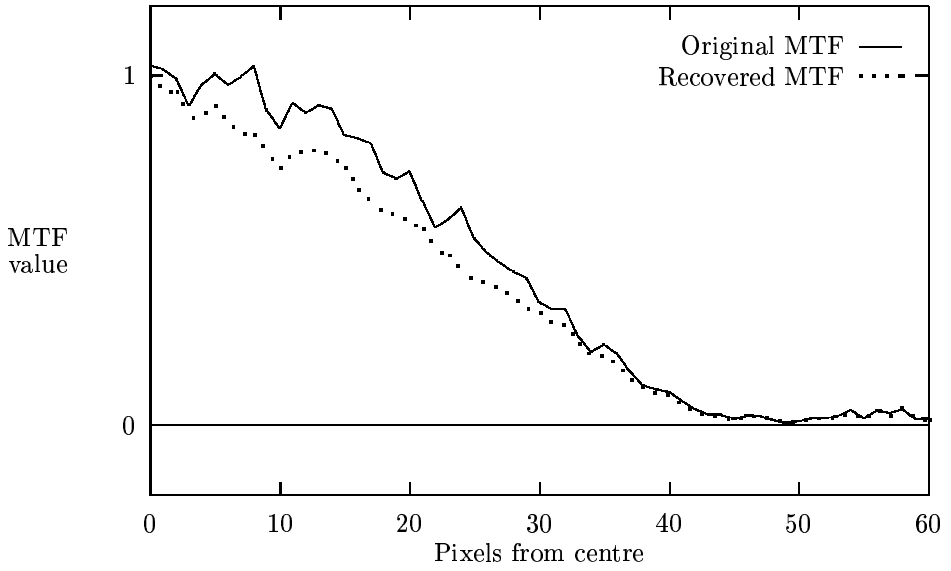


Figure 5.3: MTFs before and after extraction from ratio

ratio of this MTF to the MTF corresponding to **far 3** ($0.2mm$ from focus) was then formed, and the theory presented in section 5.1.2 used to extract what should again be the original MTF, using the knowledge that k should here be 2.5. This recovered MTF is also shown in the figure. The two MTFs should be identical. It can, however, be seen that there is a significant difference between these two profiles, particularly at low frequencies. No analysis was made as to the degree to which these profiles are different.

The difference is coming about because the self-similarity assumption is in practice not perfectly accurate. It is felt that the major cause for this is noise in the MTF ratio profile. It could also partly come about due to the rectangular windowing, which results in the image dependence not being completely removed from the ratio. The procedure might however work

for an imaging system where there is less noise present.

5.2 Development assuming Gaussian beam profile

The use of an implicit template, although theoretically sound, was shown in the previous section to be ineffective in the presence of noise. It therefore becomes necessary to explicitly specify the template before continuing further.

A widely made assumption is that the MTF has a Gaussian profile. For this assumption, the problem of autofocus takes on a closed-form solution if the relation between the standard deviation of the PSF and the distance from crossover is specified. Even if the assumption is not accurate enough to be used in a final implementation, this case is of theoretical interest in finding trends which may be of use in the development of other methods.

5.2.1 Gaussian development with no linear assumption

For now consider only the case of a rotationally symmetric beam profile (no astigmatism). In this case (see section 4.4.2) the beam current density profile can be written as

$$h(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{5.20}$$

It is assumed that some criterion has been used to sensibly and accurately relate the standard deviation $\sigma(d,f)$ of the beam to the distance d of the specimen from the lens for a beam of focal length f. In the frequency domain the MTF is then

$$H(\omega_x,\omega_y) = e^{-\frac{\sigma(d,f)^2}{2}(\omega_x^2+\omega_y^2)} \tag{5.21}$$

For the ratio of two MTFs $|H_1(\omega_x,\omega_y)|$ and $|H_2(\omega_x,\omega_y)|$ with corresponding focal lengths f_1 and f_2 and a specimen at distance d,

$$\frac{|H_1|}{|H_2|}(\omega_x,\omega_y) = e^{-\frac{(\sigma^2(d,f_1)^2-\sigma^2(d,f_2))}{2}(\omega_x^2+\omega_y^2)} \tag{5.22}$$

which has a resulting variance of

$$\sigma_{res}^2(d,f_1,f_2) = \frac{1}{\sigma^2(d,f_1)-\sigma^2(d,f_2)} \tag{5.23}$$

It is useful at this point to permit this variance to take on negative values. This will occur when the numerator MTF has a standard deviation greater than that of the denominator. In this case a Gaussian with a negative variance will just be a curve which grows exponentially

with increasing $(\omega_x^2 + \omega_y^2)$ instead of going to zero.

If the ratio of two MTFs is now considered, $\sigma_{res}(d, f_1, f_2)$ can be measured directly as the variance of this quantity. There is then an implicit relation between d, f_1 and f_2 by means of equation 5.23. This can be used directly to calculate the distance d of the specimen from the lens, and the focal length set to this.

5.2.2 Continued development making linear assumption

In the previous section the relation $\sigma(d, f)$ was left unspecified. In order to make the method described more concrete, the analysis will be continued with the assumption that the standard deviation of the Gaussian beam profile in the spatial domain is a linear function of the distance from focus. This was shown in section 4.3.2 to be a good approximation. Figure 5.1 is again used to demonstrate the beam width at the specimen with changing focal length. In this case the width is chosen to relate to the standard deviation of the PSF.

If σ is the standard deviation of the point-spread function, then the same relation as before holds

$$\sigma(d, f) = a\frac{|f - d|}{f} \tag{5.24}$$

Using equation 5.23, it can then be said that

$$\sigma_{res}^2(d, f_1, f_2) = \frac{1}{a^2(\frac{f_1 - d}{f_1})^2 - a^2(\frac{f_2 - d}{f_2})^2} \tag{5.25}$$

In a practical situation a could be specified beforehand. In this case the above expression is a quadratic equation in d, and there are two valid solutions. The ambiguity arises because the beam is symmetric about the plane of the crossover. The effect of this is that there are two possible specimen positions which will result in the same MTF ratio. This ambiguity would have to be resolved by making use of additional information.

The suggestion of fixing a represents a loss of generality in two ways; firstly, a has to be specified for any viewing condition for any given microscope. Furthermore, it assumes that the beam is precisely linear from the aperture plane to the specimen and the two focal planes. This makes it impossible to use the approximation of localised linearity near the focal and specimen planes without specifying a different value of a for different region along the length of the beam. Fortunately it is possible to remove these restrictions by making use of a third image and forming two separate ratios, as was done for the general autofocus case of the previous section.

Designating the three MTFs corresponding to focal lengths f_1, f_2 and f_3 by $|H_1|$, $|H_2|$ and $|H_3|$, the ratios $|H_1|/|H_2|$ and $|H_2|/|H_3|$ can then be formed. The quantities $\sigma_{res_1}^2(d, f_1, f_2)$

and $\sigma^2_{res_2}(d, f_2, f_3)$ can be found directly from these ratios. Equation 5.25 can be used to give two independent equations in d and a, from which the unknown a can be eliminated. The resulting equation yields a unique solution to the unknown distance d.

The fact that this solution is unique will be demonstrated: solving each of these equations for $1/a^2$, the following is obtained

$$\begin{aligned} \frac{1}{a^2} &= \sigma^2_{res_1}\left[\left(\frac{1}{f_1^2}-\frac{1}{f_2^2}\right)d^2-\left(\frac{2}{f_1}-\frac{2}{f_2}\right)d\right] \\ \frac{1}{a^2} &= \sigma^2_{res_2}\left[\left(\frac{1}{f_2^2}-\frac{1}{f_3^2}\right)d^2-\left(\frac{2}{f_2}-\frac{2}{f_3}\right)d\right] \end{aligned} \tag{5.26}$$

Equating the terms in $1/a^2$ and dividing by $d \neq 0$, a linear equation in d results. This can be solved, yielding

$$d = \frac{\sigma^2_{res_1}(\frac{2}{f_1}-\frac{2}{f_2})-\sigma^2_{res_2}(\frac{2}{f_2}-\frac{2}{f_3})}{\sigma^2_{res_1}(\frac{1}{f_1^2}-\frac{1}{f_2^2})-\sigma^2_{res_2}(\frac{1}{f_2^2}-\frac{1}{f_3^2})} \tag{5.27}$$

Thus for the case of a Gaussian MTF the position of best focus has a closed-form solution in terms of the focal lengths used and the standard deviations of the resulting ratios.

5.2.3 Use of search method

In the previous section an expression was obtained for the predicted specimen position in terms of the standard deviations of the MTF ratios. It is possible to reformulate this finding into a search method much like the one presented for the general autofocus method earlier in this chapter. The reason for wanting to do this is that when the restriction of a Gaussian MTF is dropped, the closed-form solutions that have been given cannot be retained.

In the same way as before the MTF ratios are assumed to conform to a model which is consistent with a linear beam. The expressions are now

$$\begin{aligned} \frac{\mathrm{MTF}_1}{\mathrm{MTF}_2} &= \frac{H_t(kk_1x)}{H_t(kx)} \\ \frac{\mathrm{MTF}_2}{\mathrm{MTF}_3} &= \frac{H_t(kx)}{H_t(kk_2x)} \end{aligned} \tag{5.28}$$

where $H_t(x)$ is assumed in this case to be a Gaussian function of arbitrary width. MTF_2 is then given by $H_t(kx)$ for some value of k which is not yet specified.

Assume however that this k is known. The same approach that was used for the general autofocus method can then be used to iterate through the possible specimen positions d_{ass}, and corresponding values calculated for k_1 and k_2. Representations of MTF_1, MTF_2 and MTF_3 can then be created by stretching the template MTF by factors $1/(kk_1)$, $1/k$, and

$1/(kk_2)$ respectively. The accuracy of each possible position d_{ass} can be assessed by forming the two ratios from these representations and comparing them to the actual ratios. The value of d_{ass} for which these two sets of ratios are identical should be the required value d, since this position has been shown to be unique.

Furthermore, it can be shown that this best distance implies a unique value for k (or at least for ka where a is a constant). Assuming that the template MTF has standard deviation σ_t, the standard deviation of MTF_2 will then be $k\sigma_t$. In terms of a model including this additional parameter k, Equation 5.24 for the standard deviation of the PSF can be written

$$\sigma(d, f) = ka\frac{|f - d|}{f} \tag{5.29}$$

Thus all the results from section 5.2.2 carry forward by simply substituting ka for a in all the equations. In particular, it can now be said that

$$\begin{aligned} \frac{1}{k^2a^2} &= \sigma_{res_1}^2 \left[\left(\frac{1}{f_1^2} - \frac{1}{f_2^2}\right) d^2 + \left(\frac{2}{f_1} - \frac{2}{f_2}\right) d\right] \\ \frac{1}{k^2a^2} &= \sigma_{res_2}^2 \left[\left(\frac{1}{f_2^2} - \frac{1}{f_3^2}\right) d^2 + \left(\frac{2}{f_2} - \frac{2}{f_3}\right) d\right] \end{aligned} \tag{5.30}$$

and as before the ratios will be identical only for a single assumed distance. The right hand side of these equations are therefore unique. The quantity k^2a^2 must then also be unique for the correct position. That is to say, in the two-dimensional space of d versus k^2a^2 there is only one point where the consistency criterion of equations 5.28 can hold. Since negative k has not been defined, the resulting negative solution to k is invalid. Therefore this unique point also maps uniquely into the space of d versus ka where k is restricted to be positive.

It has thus been demonstrated that insofar as the beam is linear and the profile Gaussian, there is only one combination of d and ka which is consistent with the actual data. In every other case the generated and actual ratios will differ from one another. Thus an exhaustive search through all possible values of d and ka will ultimately yield the required specimen position. As long as the distance measure which is applied to the ratios is exactly zero if and only if they are identical, the global minimum of the distance measure will necessarily yield this point.

Finally, as long as the distance measure is well-behaved it might be expected that it would be continuous in (d, ka)-space. This means that at worst the required point will be a local minimum in this region. If this is the case then it will facilitate 2-D searches being made in the space to find the minimum. The proofs of continuity and local and global extrema are dependent on the distance measures used, and will not be presented. The reason for this omission is because no distance measures have yet been defined. In Chapter 6 where some distance measures are introduced, these factors are discussed again.

5.3 Development using linear assumption and specified template

The method proposed for finding the specimen position in the previous section was independent of the assumption that the template be Gaussian. The search method proposed did at no stage rely on the Gaussian limitation. The assumption of a Gaussian MTF was only made use of in the proof that the method results in a unique and correct solution for the unknown aperture-to-object distance.

It is therefore possible to propose a more general method not making the assumption of a Gaussian beam profile but retaining the condition of linearity. This can be achieved by simply allowing the template MTF profile $H_t(x)$ to take on any shape. The cost of this generalisation is that it is no longer possible to prove that the solution obtained will be unique.

Section 5.3: Development using linear assumption and specified template

Chapter 6

Algorithm Implementation and Analysis

There are a number of factors that need to be discussed concerning the details of an implementation. The most important of these is undoubtably the definition of a distance measure for assessing the degree to which a generated set of MTF ratios match the actual pair. This chapter discusses the distance measures, and once suitable ones are defined, they are assessed in terms of their effect on the resulting search space.

The chapter begins by defining two distance measures: the absolute and the squared difference metrics. The Gaussian approximation is used to derive closed-form solutions to the resulting distances over the search space considered. These solutions are then used in a sensitivity analysis for a specific imaging configuration. It is seen that in this case the point of best focus corresponds to a well-defined and unique minimum in the search space. The analysis is then repeated for the general case where the Gaussian assumption is dropped, and experimental results obtained.

The metric distance measures presented do not work as required. The reason is that the ratio signal is not equally valid across its length, as the metric measures assume. Corresponding modified measures are then proposed which weight the ratio functions accordingly before finding the distances. It is shown that these measures appear to overcome the difficulty. The same analysis of the search space is then made as before, but this time using the modified distance measures.

Results are given of the effectiveness of the distance measures in predicting the point of best focus. It is seen that over a large range the accuracy is good. The conditions under which the prediction fails is discussed, and a method proposed for overcoming this.

Finally, it is explained how the search can be done using an intelligent gradient method to

reduce the number of function evaluations required.

6.1 Preliminary distance measures

For the developments of the previous chapter it was shown that a preference for the distance measure is that it is zero if and only if the generated MTF ratios are identical to the actual ratios. In this situation the position of best focus in (d, ka)-space is guaranteed to be a unique global minimum.

6.1.1 Definition of distance measures

There are definite advantages to using a metric distance measure for this purpose, since theoretical analysis is simplified by the general conclusions that may be used, particularly in demonstrating properties such as uniqueness. Perhaps the most natural and analytically tractable of these distance measures is the

$$d_a(x, y) = \int_a^b |x(t) - y(t)| dt \tag{6.1}$$

metric in function space. This corresponds to the city block metric in Euclidean space, and here $d_a(x, y)$ is just the absolute area between the curves $x(t)$ and $y(t)$ within the limits $a < t < b$. This will be referred to as the absolute difference metric. Alternately, the metric

$$d_s(x, y) = \int_a^b (x(t) - y(t))^2 dt \tag{6.2}$$

is similar but weights large differences more strongly. This will be called the squared difference metric.

It can be seen that both of these measures conform to the requirement that

$$d(x, y) = 0 \Longleftrightarrow x(t) = y(t) \text{ for } a < t < b \tag{6.3}$$

which ensures a unique solution.

6.1.2 Analysis for Gaussian case

Under the approximation that the MTF profiles be Gaussian, the MTF ratio is also Gaussian (or the inverse of a Gaussian), which is normalised to being unity at the origin. In order to apply the distance measures to the required problem, let the two MTF ratios which are derived from the actual image data have variances ρ_1^2 and ρ_2^2, while the respective ratios

dictated by the current position in (d, ka)-space have variances σ_1^2 and σ_2^2. The function $g_{\sigma^2}(x)$ will be defined to be the Gaussian in x with variance σ^2, such that

$$g_{\sigma^2}(x) = e^{-\frac{x^2}{2\sigma^2}} \tag{6.4}$$

In all cases the variances are permitted to take on negative values, as discussed in section 5.2.1. This corresponds to a curve which is at each point the reciprocal of a Gaussian with a positive variance of the same magnitude.

Using the absolute difference metric

Consider initially just the first ratio, where the actual variance is ρ_1^2 and the estimated variance σ_1^2. These ratios are strictly positive over all space. The distance between the ratios in the $d_a(x, y)$ metric in equation 6.1 is just the absolute area between the curves. If both of the variances are positive, this is equivalent to saying that it is the absolute difference between the areas under each ratio. This is the case because a Gaussian with a larger standard deviation is always greater than or equal to one with a smaller deviation for any given value of x. In this case the area under each Gaussian is finite and the limits a and b can be chosen to be negative and positive infinity respectively.

A complication arises if the variances are both negative. In this case the ratios grow without bound with increasing distance from the origin, and the area is not finite unless $\rho_1^2 = \sigma_1^2$ precisely. This problem can be circumvented by considering the distance between the inverse of the ratios, for which the integral will converge. Finally, if the variances are of opposite sign then there is no possible match between them and the distance can effectively be considered to be infinite.

By allowing a match to either the ratio or the inverse ratio, an area for the Gaussian can always be defined. Thus the distance between these two ratios can be defined to be

$$d_{a_{(ratio\ 1)}}(g_{\rho_1^2}, g_{\sigma_1^2}) = \begin{cases} |\int_{-\infty}^{\infty} g_{\rho_1^2}(x)dx - \int_{-\infty}^{\infty} g_{\sigma_1^2}(x)dx| & \rho_1^2 > 0, \sigma_1^2 > 0 \\ |\int_{-\infty}^{\infty} (g_{\rho_1^2}(x))^{-1}dx - \int_{-\infty}^{\infty} (g_{\sigma_1^2}(x))^{-1}dx| & \rho_1^2 < 0, \sigma_1^2 < 0 \\ \infty & \rho_1^2\sigma_1^2 \leq 0 \end{cases} \tag{6.5}$$

Since

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx = \sqrt{2\pi}\sigma \tag{6.6}$$

this can be simplified to

$$d_{a_{(ratio\ 1)}}(g_{\rho_1^2}, g_{\sigma_1^2}) = \begin{cases} \sqrt{2\pi}||\rho_1| - |\sigma_1|| & \rho_1^2\sigma_1^2 > 0 \\ \infty & \text{otherwise} \end{cases} \tag{6.7}$$

Similarly, for the second ratio a corresponding relation is defined for the quantity $d_{a_{(ratio2)}}$, which is the distance between the curves $g_{\rho_2^2}(x)$ and $g_{\sigma_2^2}(x)$.

Since there is no reason to favour the ratio corresponding to σ_1^2 over that of σ_2^2, an overall distance between both pairs of ratios can be defined as the sum of the distances between each pair. This leads to the expression

$$d_a = d_{a_{(ratio\ 1)}} + d_{a_{(ratio\ 2)}} \tag{6.8}$$

Using the squared difference metric

In a similar manner to the previous case the squared difference metric $d_s(x, y)$ of equation 6.2 can be applied to the two pairs of MTF ratios. Again negative values for the variance are permitted, and the use of ratios and inverse ratios is similar to that of the previous section.

In this case,

$$d_{s_{(ratio\ 1)}}(g_{\rho_1^2}, g_{\sigma_1^2}) = \begin{cases} \int_{-\infty}^{\infty}\{g_{\rho_1^2}(x) - g_{\sigma_1^2}(x)\}^2 dx & \rho_1^2 > 0, \sigma_1^2 > 0 \\ \int_{-\infty}^{\infty}\{(g_{\rho_1^2}(x))^{-1} - (g_{\sigma_1^2}(x))^{-1}\}^2 dx & \rho_1^2 < 0, \sigma_1^2 < 0 \\ \infty & \rho_1^2\sigma_1^2 \leq 0 \end{cases} \tag{6.9}$$

which after some manipulation and simplification evaluates to

$$d_{s_{(ratio\ 1)}}(g_{\rho_1^2}, g_{\sigma_1^2}) = \begin{cases} \sqrt{2\pi}\{(|\rho_1|/\sqrt{2}) - 2(1/(\frac{1}{\rho_1^2} + \frac{1}{\sigma_1^2})) + (|\sigma_1|/\sqrt{2})\} & \rho_1^2\sigma_1^2 > 0 \\ \infty & \text{otherwise} \end{cases} \tag{6.10}$$

after using equation 6.6 to calculate the area under a given Gaussian. Again the overall distance between the two pairs of ratios is defined to be the sum of the distances between the respective ratios

$$d_s = d_{s_{(ratio\ 1)}} + d_{s_{(ratio\ 2)}} \tag{6.11}$$

6.1.3 Results for Gaussian case

Using the two distance metrics that have been defined, it is possible to undertake an analysis of how sensitive the search for the position of best focus in (d, ka)-space will be. The expressions for obtaining σ_1^2 and σ_2^2 for the positions in this space are as presented in section 5.2.2, where the assumption of a linear beam is also made. Unfortunately due to the complexity of the situation there appears to be no means of assessing this sensitivity for the general case. An analysis can however be done for any specific case of focal lengths for the three images and specimen position.

Consider a theoretical case where the specimen is at a distance of $14.96mm$ from the aperture, and the three focal lengths used for image acquisition are $14.76mm$, $14.46mm$ and $14.26mm$. This configuration is roughly consistent with the configuration for images **far3**, **far4**, and **far6**. Figures 6.1 and 6.2 show contour plots of the resulting distance measures under the absolute difference metric and the squared difference metric respectively, for a range of positions in (d, ka)-space. Note that the contour levels are not evenly spaced in these plots.

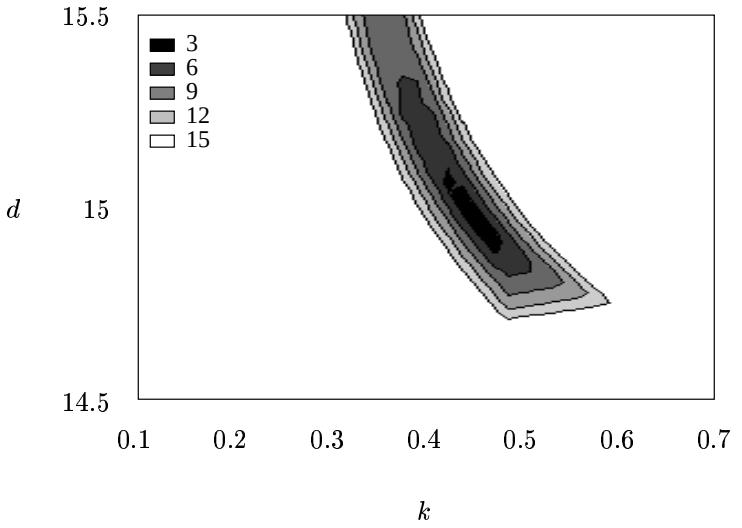


Figure 6.1: Contour plot to demonstrate sensitivity for specific case under absolute difference metric

These results represent an ideal in the sense that all MTFs used in this analysis are Gaussian, and that there is assumed to be a linear relation between the width of the beam and the distance from crossover. It can be seen that, in theory, the distance measures both exhibit a minimum at the d-position of the crossover of the beam, as has been proven to be the case in previous sections.

6.1.4 Generalisation of distance measures

The generalisation to an arbitrary template rather than a Gaussian one requires a slight reformulation of the distance measures. In particular, the ratio signal can no longer be completely described by a single width parameter; it must now be specified in relation to the template signal. Thus the conditions in equations 6.5 and 6.9 have to be adapted.

It has been mentioned that the ratio signal may in fact not fall to zero far from the origin. This will occur if the numerator MTF has a larger spread than the denominator. For the Gaussian case, this corresponded to the ratio signal being a Gaussian with a negative vari-

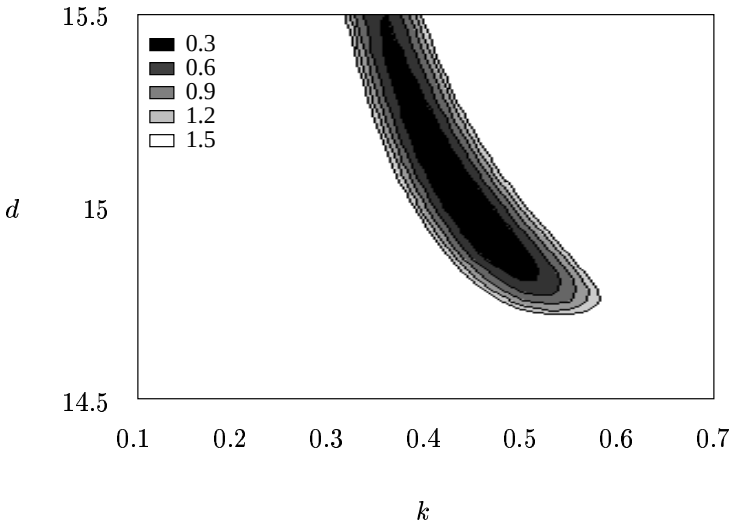


Figure 6.2: Contour plot to demonstrate sensitivity for specific case under squared difference metric

ance. The integrity of the distance measure was then retained by considering instead the distance between the inverses of both the actual ratio and the generated ratio signals. For the general case, however, it must be remembered that due to the noise reduction process in the formulation of the ratio signal, the ratio and the inverse ratio may not be simply related by a point-by-point reciprocal relation. Thus if the ratio is $R_{act\ 1}$ and the inverse ratio $R^{(inv)}_{act\ 1}$, it can in general be said that

$$R^{(inv)}_{act\ 1}(x) \neq \frac{1}{R_{act\ 1}(x)} \tag{6.12}$$

As a specific example, the action of collapsing the 2-dimensional ratio to a 1-dimensional signal that was presented in section 3.4.2 has the effect of eliminating this correspondence. If such methods are used, then representations of both the ratio and inverse ratio must be calculated separately, since the one cannot be determined from the other.

Again consider the first ratio. Let the actual ratio signal as given by the data be $R_{act\ 1}$, and the inverse ratio be $R^{(inv)}_{act\ 1}$. The ratio signal as specified by the current position in the search space is

$$R_{gen\ 1}(x) = H_t(kk_1x)/H_t(kx) \tag{6.13}$$

The distance measure in the absolute difference metric then parallels that presented in equation 6.5 if it is defined to be

$$d_{a_{(ratio\ 1)}}(R_{act\ 1}, R_{gen\ 1}) \quad = \quad \min\left\{ \left| \int_{-\infty}^{\infty} R_{act\ 1}(x) - R_{gen\ 1}(x)dx \right|, \right.$$

$$\left| \int_{-\infty}^{\infty} R_{act\ 1}^{(inv)}(x) - (R_{gen\ 1}(x))^{-1} dx \right| \Bigg\} \tag{6.14}$$

This expression can be shown to be essentially equivalent to its Gaussian counterpart if $R_{act\ 1}$ and $R_{gen\ 1}$ are restricted to be Gaussian. Then, except in a few special cases, only one of the conditions $\{(\rho_1^2 > 0, \sigma_1^2 > 0),\ (\rho_1^2 < 0, \sigma_1^2 < 0),\ (\rho_1^2 \sigma_1^2 \leq 0)\}$ will result in a finite distance, which will be the value of the function just presented.

The equation for the second ratio can be reformulated similarly, along with the two ratios in the squared difference metric. As before, the final distance measure for each metric is the sum of the values returned for each ratio.

6.2 Modified distance measures

In practice the generalised distance measures described in the previous section do not achieve exactly what is required. If a suitable template function is chosen and the search through (d, ka)-space carried out, it is found that the match is slightly compromised by the fact that the MTF ratios usually have sidelobes which cannot be ignored.

This introduces the requirement for a modification of the distance measure to eliminate the effect of these sidelobes in the comparison process. In the sections which follow the specific details of the problem will be presented, along with the changes in the distance measure that effect a solution. The way in which the search space is consequently altered will then be discussed.

6.2.1 Failings of the preliminary distance measures

The way in which the two distance measures fail to meaningfully compare the signals can be demonstrated by fitting a Gaussian to a ratio signal under one of the metrics. Consider the ratio **far4** by **far5**. Since the images are taken quite close together and far from best focus, their MTFs are fairly similar. This results in the MTF ratio having a significant secondary sidelobe beyond the range of the primary lobe. Figure 6.3 shows this ratio signal along with the Gaussian (normalised to unity at the origin) which minimises the distance between the curves under the squared difference metric presented earlier. It can be seen that after the initial minimum at approximately 40 pixels from the centre, there is a region where the signal value again grows large. This region in fact contains very little useful information: the value is large only because in this region the denominator signal is approaching zero. Thus there is significant noise in this region. The distance metric does not, however, bias this less strongly. In fitting data to the curve this region has just as much of an effect as the primary lobe which

is known to have a far lower noise contamination. This is reflected by the Gaussian fitted curve having too large a spread. Although a Gaussian is perhaps not the most suitable curve

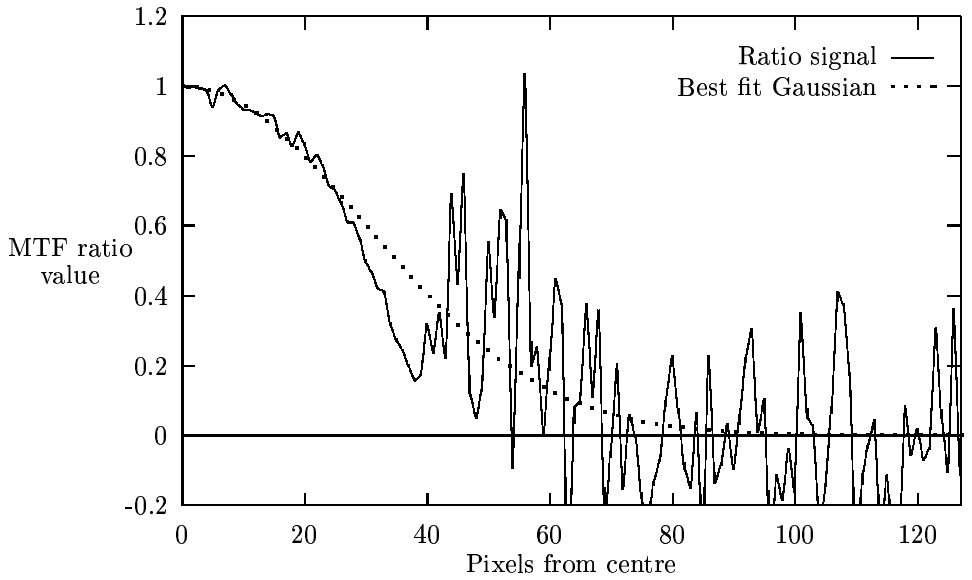


Figure 6.3: Ratio of MTFs **far5** and **far4**, along with the best-fit Gaussian to this ratio under the squared difference metric

to fit to these data, this plot demonstrates the fact that the simple metrics that have been proposed are affected by factors that should actually be insignificant. Although it will not be shown here, this also holds true for the absolute difference metric.

A solution lies in modifying the proposed distance measures to include a weighting function which ignores the signal beyond the first zero. This can be achieved without the need for detecting the position of this zero if the template itself is used as the weight.

6.2.2 Definition of the modified measure

The relation for the first generated ratio was given in equation 6.13:

$$R_{gen\ 1}(x) = H_t(kk_1x)/H_t(kx) \tag{6.15}$$

This relates the template and the position in (d, ka)-space to the generated MTF ratio. Of course, the template will always be chosen to be as accurate as possible a representation of the actual system MTF. If a restriction on the template is now made that, beyond the mainlobe, the value is forced to be zero, then the modified (weighted) absolute difference

distance measure can be defined to be

$$d_{a_{(ratio\ 1)}}(R_{act\ 1}, R_{gen\ 1}) = \min\left\{ \left| \int_{-\infty}^{\infty} (R_{act\ 1}(x) - R_{gen\ 1}(x)) H_t(kk_1 x)dx \right|, \left| \int_{-\infty}^{\infty} (R_{act\ 1}^{(inv)}(x) - (R_{gen\ 1}(x))^{-1}) H_t(kx)dx \right| \right\} \quad (6.16)$$

A similar modification can be adopted in the squared difference metric, which with the weighting function becomes

$$d_{s_{(ratio\ 1)}}(R_{act\ 1}, R_{gen\ 1}) = \min\left\{ \left| \int_{-\infty}^{\infty} (R_{act\ 1}(x) - R_{gen\ 1}(x))^2 H_t(kk_1 x)dx \right|, \left| \int_{-\infty}^{\infty} (R_{act\ 1}^{(inv)}(x) - (R_{gen\ 1}(x))^{-1})^2 H_t(kx)dx \right| \right\} \quad (6.17)$$

Some explanation is required as to the effect of including this weighting factor. It must be remembered that the signal with which the actual ratio is compared is the one where the numerator template is narrower than the denominator, and the ratio signal falls off essentially monotonically to zero (whether it is $R_{gen\ 1}(x)$ or $(R_{gen\ 1}(x))^{-1}$ depends on the position in the search space). The position of the first zero of the numerator necessarily coincides with that of the template ratio. By multiplying by this numerator the difference between the signals beyond the first zero of the generated ratio is weighted zero, and therefore makes no contribution. Thus the actual and generated ratios are only considered to be different if they differ before the first zero of the generated ratio. Furthermore, differences are weighted strongly near the origin, with the significance falling off to zero at the zero of the template ratio.

This modified squared difference distance measure was used to repeat the fit of a Gaussian to the ratio of images **far5** and **far4** that was shown in Figure 6.3. The result is shown in Figure 6.4. It can be seen that the sidelobe of the ratio data no longer has the effect of distorting the width of the best-fit Gaussian.

6.2.3 Effect of modification on search space

The effect of weighting the distance measure by the smaller MTF in the ratio will now be discussed, particularly with regard to the effect this weighting has on the search space.

It should be noted that the distance measures in function space that result from modifying the preliminary measures no longer constitute a metric. Furthermore, there is no longer a unique minimum for the distance measures over the search space. This fact can be simply demonstrated if points near the line $k = 0$ in the search space are considered. Here the numerator template tends towards a Dirac delta function. Since this numerator is also the

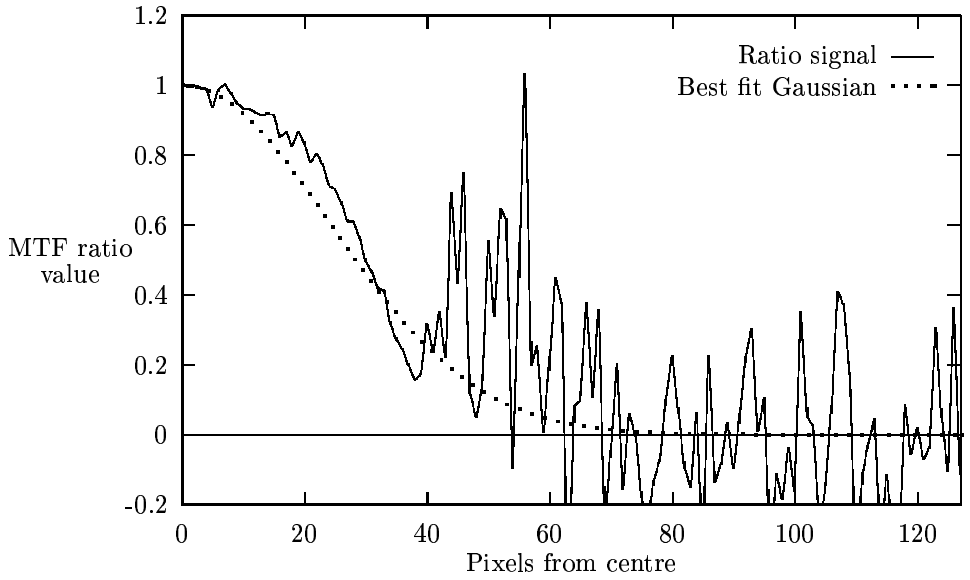


Figure 6.4: Ratio of MTFs **far5** and **far4**, along with the best-fit Gaussian to this ratio under the modified squared difference distance measure

function by which the differences are weighted, all points not at the origin receive a weighting of zero. At the origin the generated and actual ratios are necessarily approximately unity, and thus all points with $k = 0$ will return a distance of very near to zero under the modified distance measures. The search space has therefore been fundamentally modified. The point that was a global minimum in the unmodified metrics is still a global minimum in the search space under the modified distance measures, but the uniqueness has been compromised.

Figure 6.5 shows a theoretical plot of the search space for the same conditions as was presented previously, but now under the weighted squared difference distance measure. Equations similar to 6.7 and 6.10 can be derived for these cases under the Gaussian assumption, but this will not be presented. It is evident from the figure that additional local minima have arisen from this modification. The position of best focus is still however a global minimum. Also apparent is the dropping off of the distance value as the $k = 0$ line is approached.

An attempt could be made to improve the situation for those points where the width of the weighting function becomes close to zero. This can be effected by dividing the resulting distance measure by a figure which reflects the width of the weighting function. Effectively, this represents a form of normalisation. In cases where this width is small, the distance becomes increased accordingly. The problem with this solution is that now for very wide weighting functions the distance is similarly deflated, and low regions are then introduced at points far from the $k = 0$ axis.

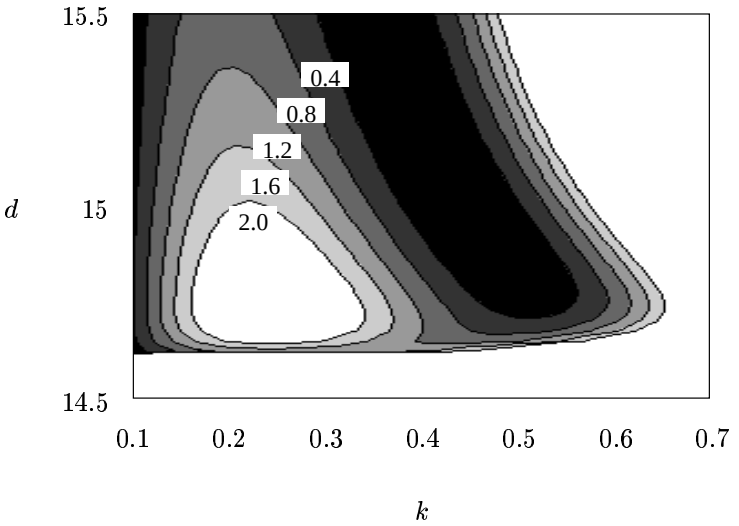


Figure 6.5: Contour plot to demonstrate sensitivity for specific case under weighted squared difference measure

In order to confirm that the theoretical search space under the Gaussian approximation does in fact resemble the actual situation where the MTF template is specified, a plot corresponding to that of Figure 6.5 was made using actual experimental data. The images used in the formation of the ratios were **far3**, **far4**, and **far6**. The distances of these images from best focus coincide with the distances that were used in the theoretical plot. The results are shown in Figure 6.6. In both cases the weighted squared difference distance measure was used. The similarity between the ideal theoretical search space and the experimental space is apparent.

6.3 Results for the autofocus algorithm

Having now defined distance measures that seem to be suitable, it is possible to generate experimental results for the proposed autofocus procedure. This is of particular interest in finalising the choice of distance measure to use, as well as in analysing the conditions under which the algorithm might be expected to work favourably.

A number of results for the test sequence **farfocus** have been generated. For every combination of three images in the set of images **far2** to **far9** the prediction procedure that has been described was carried out. This represents a total of 56 unique combinations of image distances from focus. For each case the search space in the approximate vicinity of the known point of best focus was exhaustively searched for a global minimum. The position of this point was then considered to correspond to the actual beam configuration, from which the beam crossover distance can be calculated. Figure 6.7 shows the first of these results for the

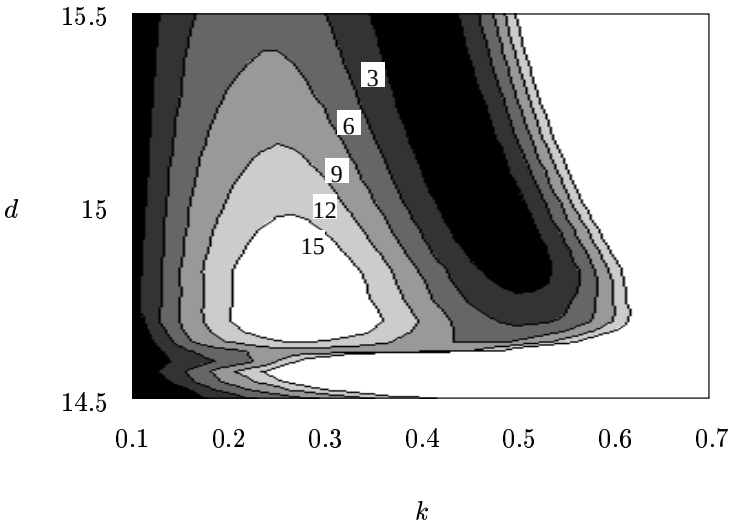


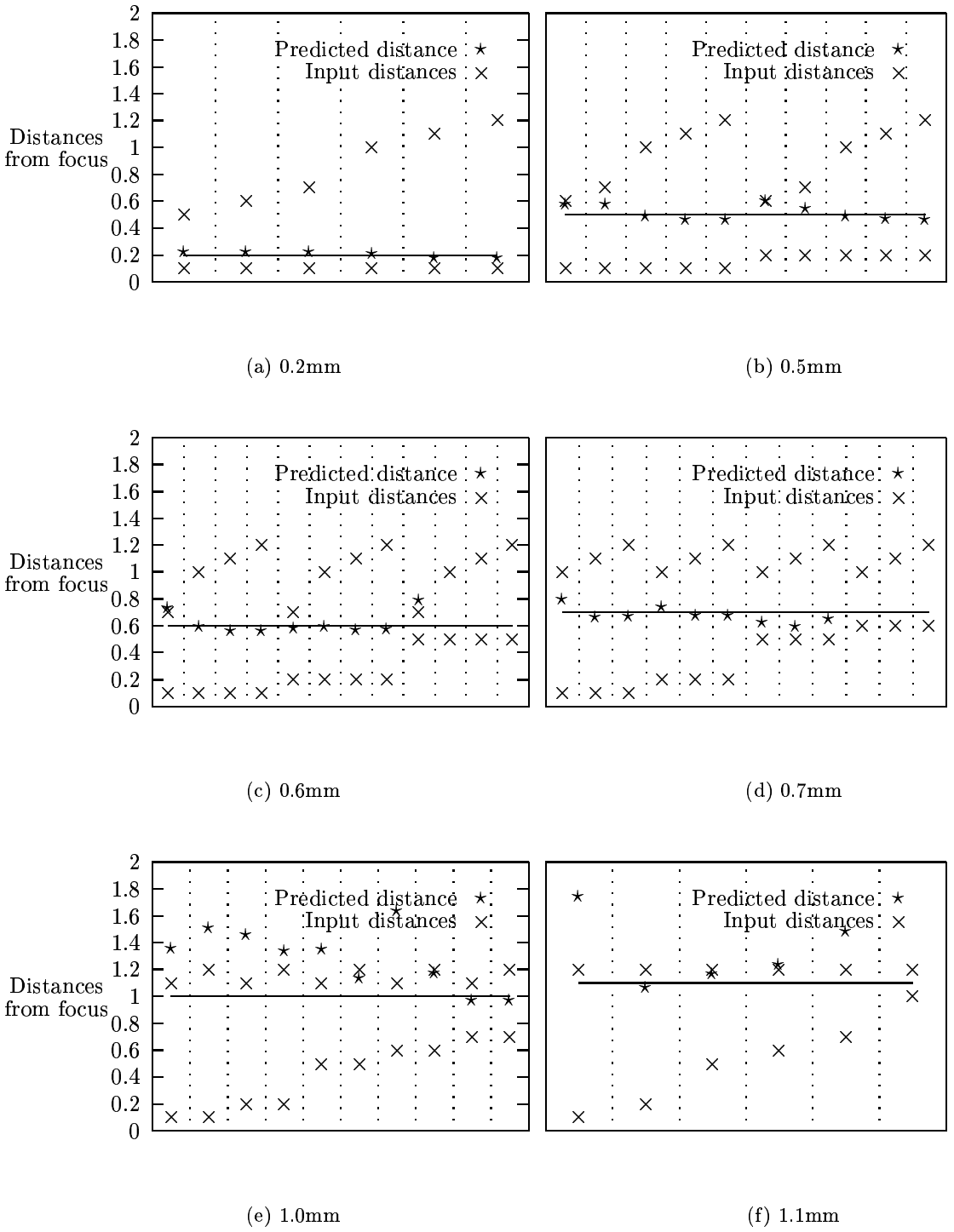
Figure 6.6: Contour plot of the actual search space under the weighted squared difference measure. The images used to generate this plot were **far3**, **far4**, and **far6**

weighted squared difference distance measure.

Each plot (a)-(f) presents the results for a fixed distance from best focus of the intermediate image used in the prediction. The solid horizontal line in each case indicates this distance. The vertical dotted lines partition the plot into separate prediction sets, each of which correspond to an experiment using different input data. Since three images are required for the predictions, the two crosses in each set indicate the distances from focus of the additional images. For example, consider plot (a) in Figure 6.7: the leftmost interval corresponds to input images at the three defocus distances $0.1mm$, $0.2mm$, and $0.5mm$, with $0.2mm$ being the distance for the intermediate input image. For each case the prediction process is applied, and the distance as calculated for this intermediate image plotted using a star. The situation for correct prediction thus corresponds to all the stars lying on the horizontal line. Note that in some cases where the prediction procedure fails, the location of the star may not appear in the plotted range, and is hence absent.

It is difficult to analyse the data in a systematic manner. However, the plot demonstrates some basic trends:

- The prediction is fairly accurate if the intermediate MTF is close to focus. For points where the actual distance is large, the reliability tends to break down.
- If the prediction fails for images which are close to focus, then generally it seems that it occurs when at least one of the pairs of images used in the formation of the ratios are close together.



(a) 0.2mm (b) 0.5mm

(c) 0.6mm (d) 0.7mm

(e) 1.0mm (f) 1.1mm

Figure 6.7: Prediction results for the **farfocus** image series under the weighted squared difference distance measure.

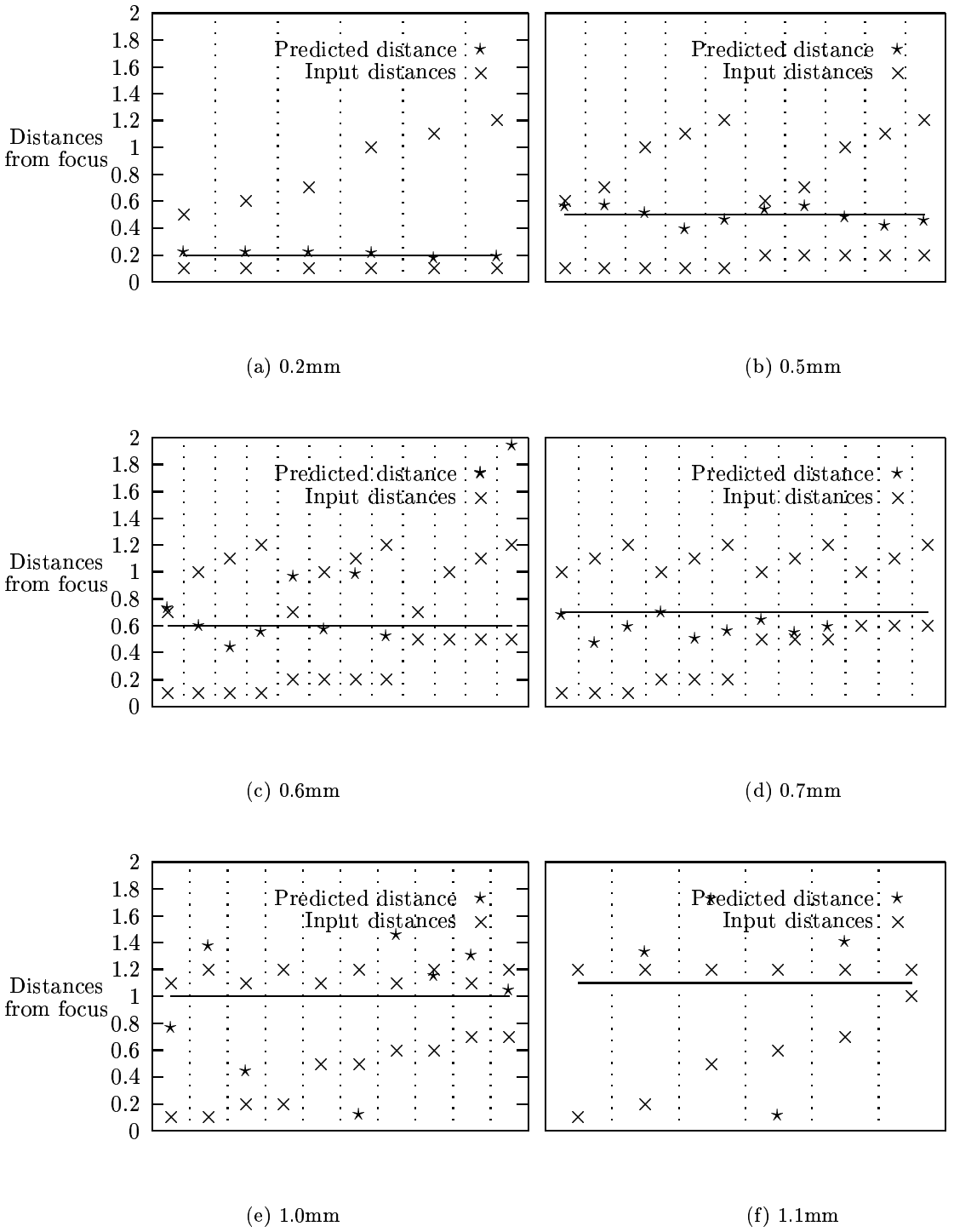
The cause of the first trend can be explained in reference to the width of the MTF as a function of the distance from the crossover. In a previous chapter it was shown that this relation is hyperbolic. Thus, near to focus the MTF is very wide, falling off rapidly as the distance is increased. The effect of the template MTF being restricted to go to zero at some point is that the first zero of the ratio will be the same as the first zero of the numerator MTF. Thus the basic shape of the ratio is dictated by the numerator. The effect of the denominator MTF is then to alter the shape of this ratio between the origin and the first zero. Clearly this is a far smaller effect to detect than the change in the zero position. If the numerator MTF now becomes narrow, the deviation caused by the denominator MTF profile becomes less significant. Huge changes in this denominator will then cause only marginal changes in the ratio signal. In the presence of noise this becomes catastrophic to the prediction process because the distance measure no longer exhibits a suitable minimum.

The observation that the prediction fails more readily if the images forming the ratio are close together can also be simply explained. In the limit as the distance between the images goes to zero, no prediction can be made because data has been irrecoverably lost. It follows then that if the images are close together then the effectiveness of the prediction will be hindered. The discussion of the first trend also makes it apparent that this closeness has to be considered in relation to the actual distance of the images from the crossover. MTFs separated by a fixed distance far from focus will differ from one another far less than MTFs separated by the same distance but nearer to the beam crossover.

In the case of the microscope, this fact that prediction fails more readily for image sets far from the crossover is not as catastrophic as it might be. This is because of the ability to change the magnification of the instrument, and hence affect the overall MTF width. The use and effect of magnification on the MTF and the autofocus algorithm will be discussed in section 7.1.

Figures 6.8 and 6.9 contain additional results based on the use of other distance measures. In particular, the results for Figure 6.8 are for the weighted absolute difference distance measure, and Figure 6.9 for the normalised weighted squared difference measure that was mentioned towards the middle of section 6.2.3. This latter plot is shown mostly for interest, since the distance measure has neither been properly defined nor discussed.

These plots serve to show that the precise details of the distance measure do not play a critical role in the prediction process. All that is required is that it takes into account the features of the signals that are being matched. No results could be generated for the unweighted metrics because of their failure to successfully match the generated ratios to the actual ratios.

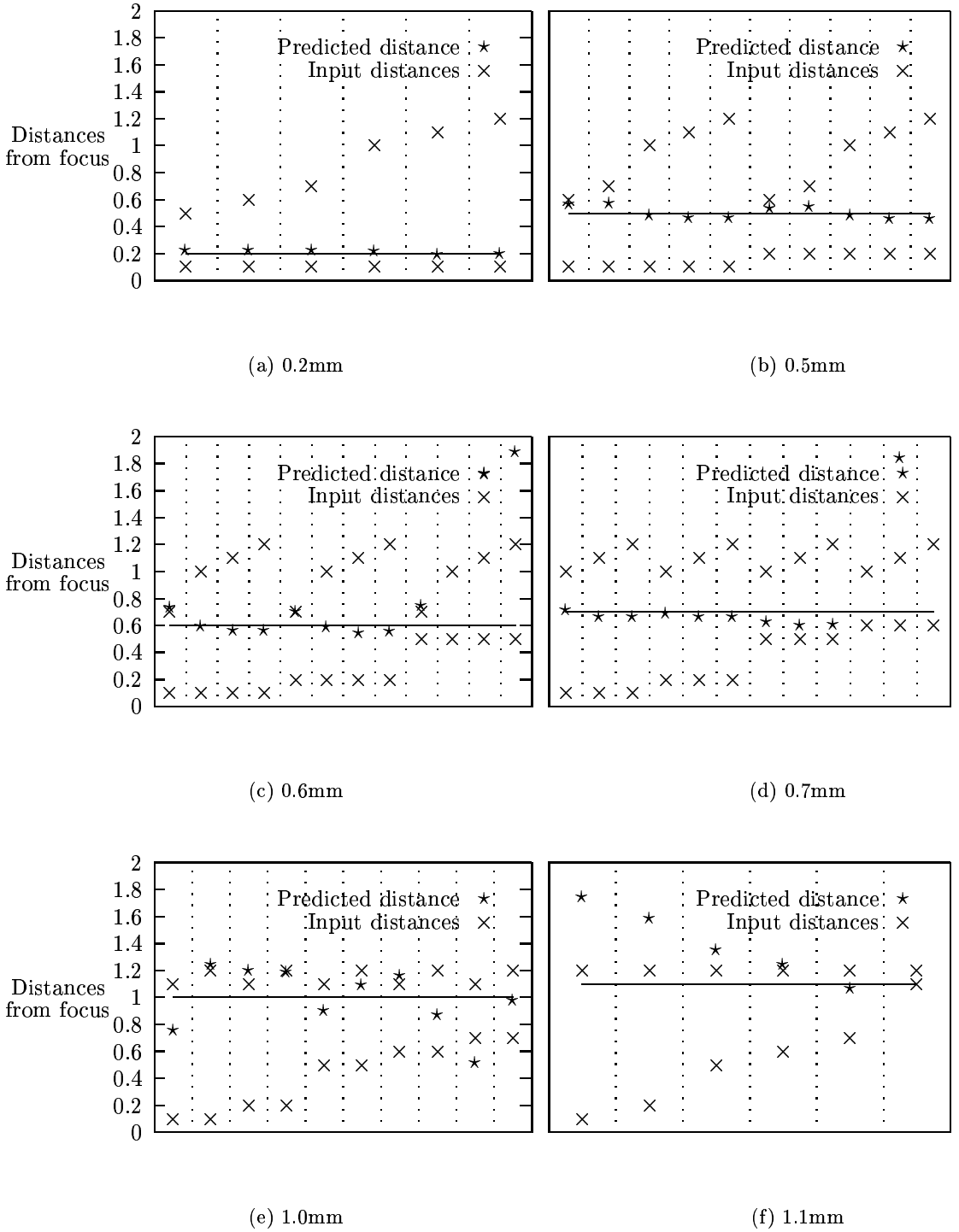


(a) 0.2mm (b) 0.5mm

(c) 0.6mm (d) 0.7mm

(e) 1.0mm (f) 1.1mm

Figure 6.8: Prediction results for the **farfocus** image series under the weighted absolute difference distance measure.



(a) 0.2mm (b) 0.5mm

(c) 0.6mm (d) 0.7mm

(e) 1.0mm (f) 1.1mm

Figure 6.9: Prediction results for the **farfocus** image series under the normalised weighted squared difference distance measure.

6.4 Extension to optimised search

The results that were generated in the previous section were obtained by an exhaustive search through the (d, ka)-space in the vicinity of the known point corresponding to focus. The process of exhaustively searching is computationally expensive, and it is possible to improve the speed by means of an intelligent 2-dimensional search.

The success of the search relies on finding a suitable starting point. This is necessitated by the fact that the search space may contain local minima, as was seen to be the case for the situation of Figure 6.6. Having verified that the Gaussian approximation results in a search space which closely resembles that of the experimental cases in the vicinity of the point of best focus, a possible solution to this problem is presented: if a best-fit Gaussian is fitted to each ratio, then a closed-form solution to the minimisation problem exists. Equation 5.27 relates the standard deviations of the ratio signals to the distance of the required point in the search space. Thus knowing d, the k-value of this point can then be calculated by means of equation 5.30.

Such a 2-dimensional search algorithm was implemented and applied to the problem. The algorithm used was a modification of the method of steepest-descent, called Praxis[4]. The results are in most cases very similar to those presented previously in this chapter, with the search failing in only a few cases where the minimum is not adequately bounded by steep enough sides in all directions. It was found that approximately 50 to 100 distance function evaluations were required in each case to find the local minimum given the calculated starting conditions. The function evaluations are sufficiently fast that the entire search typically takes considerably less than a second.

Chapter 7

Further Developments

There are a few additional sections required to clarify the use of the algorithm in a real situation. The first of these is the dependence of magnification on the image formation. The extension to astigmatism correction is also discussed.

The chapter begins by again using the Gaussian MTF assumption, this time to analyse the effects of magnification on the MTF. It is shown that in theory if the magnification is increased by a factor of two, then the MTF width (in pixels) will increase by the same factor. Experimental results are then given which demonstrate that even though this prediction is not entirely accurate in practice, the trend is at least valid.

The extension of the autofocus algorithm to the problem of astigmatism correction is then presented. The algorithm essentially requires no modification, but the procedure is now performed in three different planes through the optical axis.

The chapter concludes with a complete outline of the proposed algorithm.

7.1 Effects of magnification

It was mentioned in the previous chapter that changes in magnification can be used to improve the prediction of the point of focus. This will now be discussed in greater detail.

In the electron microscope the magnification is altered by simply changing the distance between successive samples of the specimen. Presumably reducing the distance between samples by a factor of $x > 1$ will therefore increase the magnification by this same factor. Effectively this results in the width of the PSF (in pixels) also increasing by the factor x. In the frequency domain this corresponds to the relative width of the MTF decreasing by this same factor.

7.1.1 Analysis under Gaussian assumption

In order to assess the effects of magnification on the image formation process, the assumption of a Gaussian beam profile will once again be used.

Suppose that the PSF at a distance d from the aperture has a standard deviation of $\sigma_{\mathrm{PSF}}(d)$. If this standard deviation is directly proportional to the distance from the beam crossover, then

$$\sigma_{\mathrm{PSF}}(d) = \frac{a}{d_f}|d_f - d| \tag{7.1}$$

where a is the standard deviation of the PSF as referred to the aperture plane, and d_f is the beam focal length. Defining $\delta d = d_f - d$ to be the distance of that beam position from crossover, it can be written that

$$\sigma_{\mathrm{PSF}}(\delta d) = k|\delta d| \tag{7.2}$$

with k a proportionality constant suitably defined. Although the linear dimensions of this PSF are fixed, the width in pixels depends on the magnification in use. If the magnification is M pixels/metre and the quantity $\sigma_{\mathrm{PSF}}(\delta d)$ is in metres, then the standard deviation in pixels will be

$$\sigma_{\mathrm{PSF}(pixels)}(\delta d, M) = kM|\delta d| \tag{7.3}$$

Here the dependence on magnification has been included explicitly in the relation.

It should be apparent from equation 7.3 that a change in δd can be counteracted by a particular change in M. The quantity $\sigma_{\mathrm{PSF}(pixels)}$ can be kept constant through changing δd as long as the product $M|\delta d|$ is not changed. Thus for distances far from focus where $|\delta d|$ is large, the width of the PSF can be kept within a reasonable range by reducing the magnification accordingly.

This observation can be extended to the MTF in the frequency domain as well. Increasing the magnification by a factor M will here decrease the width (in pixels) of the MTF by the same factor.

7.1.2 Experimental results

The accuracy of this previous observation can be analysed by looking at two images taken at different magnifications, and comparing the corresponding MTFs. The MTFs can be extracted from the images by the same methods as were used in section 4.3.

Figure 7.1 shows MTFs obtained from images **far3** and **far 12**. Both of these images relate to a distance of $0.2mm$ from focus, with the first at a magnification of $500\times$ and the second at $1000\times$. According to the theoretical discussion the first MTF should have twice the width

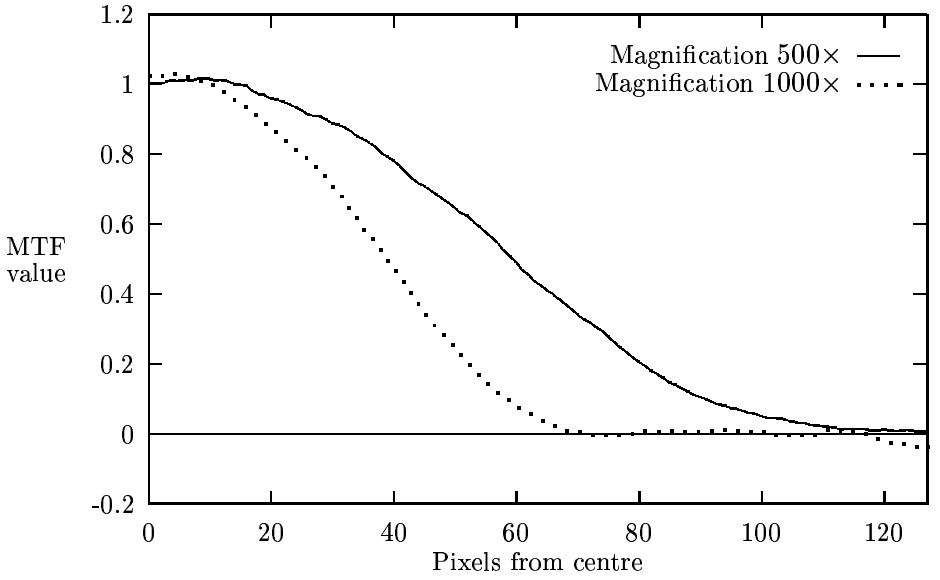


Figure 7.1: MTFs corresponding to two different magnifications at the same distance from focus ($0.2mm$)

of the second. In practice it is found that for this case the actual difference is only $1.551\times$.

The reason for this discrepancy has not been investigated. It could, however, possibly be a result of the beam not remaining stationary at the point which is being sampled. This movement may be introduced by the scanning motion. The contribution to any given pixel is then a combination of the actual desired value as well as the specimen characteristics in the close neighbourhood. This factor could cause a deviation from the theoretical case presented in the previous section. The overall principle does still however apply, namely that in terms of the MTF width the effect of a change in distance from crossover can be counteracted by a change in magnification.

In terms of the autofocus procedure, the effect of this deviation from ideal is that images that were captured at one magnification cannot be modified and reused along with images taken at a different magnification. In practice this should not present a problem.

Being able to control the width of the MTF without changing the distance can allow the accuracy of the prediction process to be enhanced. This is done by adjusting the magnification to result in MTFs that have widths most suited to the focusing procedure.

7.2 Extension to astigmatism

The autofocus procedure that has been presented thus far takes three images, and explicit information is returned about the distance of the intermediate image from crossover, as well

as the width of this MTF with respect to some template MTF profile. Also implied by this data is equivalent information regarding the MTFs corresponding to the other two images involved in the prediction.

It must be noted that this entire procedure operates in a single arbitrarily oriented plane which passes through the optical axis of the microscope. Thus the data obtained relates only to the beam as considered within this plane. In the case of the beam being circularly symmetric it is completely specified by this information, however in general this assumption cannot be made.

More generally, by considering a number of such planes at different angles through the optical axis a complete representation of the MTFs can be developed. For every extra orientation considered, more data can be built up about the actual shape of the MTFs at the position of the specimen for the three focal lengths. However, if some prior knowledge of the possible beam shapes is known, then only a small number of such samples may be required to adequately specify the beam.

In particular, in the presence of astigmatism it is no longer necessary that the PSF be circularly symmetric. Instead the beam current density profile becomes elliptical, with more blurring resulting in the direction of the major axis than in that of the minor. For the Gaussian approximation, this density profile has sometimes been assumed to be that of a two-dimensional Gaussian, with a specific eccentricity and orientation [7]. The corresponding MTF in the frequency domain will also be such a Gaussian. There are three independent parameters to describe this profile, namely an overall width parameter, an eccentricity, and an orientation. Therefore, under this model the MTF can be completely specified by analysing the beam in three independent planes, and solving for the unknowns.

It should be possible to avoid the Gaussian restriction by simply forming a more general expression for the beam width as a function of rotation angle. Good starting points for this analysis are presented in [31, 1].

Thus the autofocus method that has been presented in previous chapters can be extended to the astigmatic case by simply repeating the procedure for three different planes. The information that is thereby obtained is firstly three independent estimates of the distance of the images from crossover, and secondly the complete MTF representations of the beam at the corresponding locations. All this can be obtained from just the three images. The MTF representations can then be used to analyse the astigmatism in the beam, and to correct it.

Most autofocus algorithms which rely on searching for extrema in a criterion function can at best only optimise the focus or the astigmatism at any one time. For this reason they are forced to iterate repeatedly through cycles of alternately correcting the astigmatism and the focus [7]. For the method proposed here the beam is considered in its entirety, and no such

co-dependence exists. It is therefore possible to change both of these factors in the same pass.

7.3 Proposed autofocus algorithm

An implementation of the overall technique developed in this thesis has not been attempted. It was found that the available software for the microscope did not lend itself to easy interfacing to such an algorithm. The discussions do, however, lead towards a fairly natural implementation, the rough details of which will be discussed here.

It must be emphasised that without physically trying such an implementation it is extremely difficult to anticipate the areas where problems might occur. Also, numerically quantifying some of the factors is impossible. When putting the algorithm into practice, the particulars of the implementation would have to be arrived at experimentally. The following stages comprise the outline:

- Capture a single image and transform a portion of it. If there is enough high frequency content in this transformed sample, then the chances of success of the prediction is good, and the image can be used. If there is not sufficient high-frequency detail, then reduce the magnification until there is.
- Capture the remaining two images needed for the prediction. The amount to change the focal length from that of the first position would have to be found by some means. A possible solution to this problem would be to grab an image at some arbitrary distance and to form the MTF ratio. Once this is available, it can be decided if the ratio falls within a width range that is deemed suitable. If not, the distance can be adjusted and the process repeated. If so, the final image can just be taken the same distance from the first, but in the opposite direction.
- Once the images are captured, they can be tiled, transformed and averaged, as described earlier. The two ratios can then be formed. In order to take into account possible astigmatism, three 1-dimensional representations of these ratios can be formed at varying angles by means of the radial averaging procedure outlined.
- The autofocus procedure can then be applied to each of these cases independently. This will give three independent estimates of the position of the beam crossover in the case of a non-astigmatic beam. For an astigmatic beam the average of three should be a suitable estimate until such a time as it is corrected for. Additionally, a reasonable model of the beam can be created from the now known MTFs at the locations of the images. This can be used to correct for the astigmatism.

- After making corrections to the focal length and astigmatism, it is assumed that the instrument is now closer to best focus and zero astigmatism. The magnification can therefore be increased and the process repeated until the desired degree of accuracy is achieved. Furthermore, as successive iterations are undertaken, a progressively more in-depth model of the beam can be built-up, which can be used to intelligently refine the subsequent searches.

It is anticipated that the bulk of the time spent in this loop will be during the image capture. At least three images have to be captured for every iteration. A huge speed difference could be gained if, instead of capturing three complete images, line samples are made at the focal lengths and orientations desired. Effectively, to reproduce the situation that was obtained above, 9 line samples are required, three for each angle around the optical axis. The noise distributions will however differ for this case from the one presented in the algorithm.

Chapter 8

Conclusions

The findings of this thesis will now be summarised and the important conclusions highlighted.

The most significant result that came out of the entire investigation is the fact that it was possible in this case to generalise a commonly used model. The usual assumption of a Gaussian MTF was shown to be a special case of a model based on a more fundamental property of imaging systems, namely that of a self-similar PSF which varies linearly in size with the distance from focus. Methods were then developed for working with this new model, which although perhaps not as neat as for the Gaussian counterpart, had virtually the same flexibility. Furthermore, it was shown that this model had relevance to practical applications: An autofocus algorithm was developed on top of it, and the results obtained were fairly acceptable. It should be noted that having the closed-form Gaussian model solutions to the problems in question was particularly useful in performing this generalisation.

Also of particular interest was the noise analysis that was made on the ratio data. It was demonstrated that in this situation the commonly used methods of noise reduction would fail because of the unusual probability distribution. Various estimators for the location of the distribution had to be considered, and their characteristics analysed with regard to how successful they would be as estimators of the desired noise-free ratio. Here again a number of alternative methods had to be built up and optimised to suit the particular situation.

With regard to the autofocus algorithm that was proposed, this could be improved upon. The idea of forming an MTF ratio is a very powerful one, given that the problem of noise can be circumvented. Also, the use of a general imaging model in analysing this ratio is very effective; the usefulness of a model that can be used unaltered across all operating conditions cannot be disputed. The place where the proposed procedure falls short is possibly in the definition of the distance measure used in comparing two given ratios. This is a stage which in the development drew heavily on heuristics to make the match appear visually optimal,

but no justification was given that this is indeed the case. The problem can be seen in the search spaces under the modified distance measures: whereas in the metric measures the global minimum is steeply bounded and therefore accurately and easily found, for the modified cases this accuracy is compromised. The distortion of the search space to contain multiple global minima is also a compromise that should be addressed. There is naturally a heavy bias towards the use of commonly encountered metrics or distance measures (such as squared difference), where less known but more applicable measures might be more suitable.

Bibliography

[1] Baba, N., Oho, E., and Kanaya, K. An algorithm for on-line digital image processing for assisting automatic focusing and astigmatism correction in electron microscopy. *Scanning Microscopy*, 1(4):1507–1514, December 1987.

[2] Boddeke, F., van Vliet, L., Netten, H., and Young, I. Autofocusing in microscopy based on the OTF and sampling. *Bioimaging*, 2:193–203, 1994.

[3] Born, M. and Wolf, E. *Principles of Optics.* Pergamon Press, 1959.

[4] Brent, R. *Algorithms for Minimisation without Derivatives.* Prentice-Hall Series in Automatic Computation. Prentice-Hall, 1973.

[5] Derman, C., Gleser, L., and Olkin, I. *A Guide to Probability Theory and Application.* International Series in Decision Processes. Holt, Rinehart and Winston, Inc., 1973.

[6] Ens, J. and Lawrence, P. An investigation of methods for determining depth from focus. *IEEE Transactions on pattern analysis and machine intelligence*, 15(2):97–108, February 1993.

[7] Erasmus, S. and Smith, K. An autofocusing and astigmatism correction system for the SEM and STEM. *Journal of Microscopy*, 127:185–200, 1982.

[8] Fellgett, P. and Linfoot, E. On the assessment of optical images. *Philosophical transactions of the Royal Society of London: A*, 247:369 – 407, 1955.

[9] Fischer, E. *Intermediate Real Analysis.* Undergraduate Texts in Mathematics. Springer-Verlag, 1983.

[10] Gillespie, J. and King, R. The use of self-entropy as a focus measure in digital holography. *Pattern recognition letters*, 9(1):19–25, January 1989.

[11] Goldstein, J., Newbury, D., Echlin, P., Joy, D., Fiori, C., and Lifshin, E. *Scanning Electron Microscopy and X-Ray Microanalysis.* Plenum Press, 233 Spring Street, New York, 1981. A text for Biologists, Materials scientists and Geologists.

BIBLIOGRAPHY

[12] Gradshteyn, I. and Ryzhik, I. *Table of Integrals, Series and Products.* Academic Press, 1965.

[13] Harris, F. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, January 1978.

[14] Hopkins, H. The frequency response of a defocused optical system. *Proceedings of the Royal Society of London: A*, 231:91 – 103, 1955.

[15] Kanaya, K. and Baba, N. A method of correcting the distorted spot shape of a deflected electron probe by means of dynamic focusing and stigmator. *Journal of physics E: Scientific instruments*, 13:415 – 426, 1980.

[16] Koster, A., van den Bos, A., and van der Mast, K. Signal processing for autofocusing by beam tilt induced image displacement. In P.W.Hawkes, W.O.Saxton, F.P.Ottensmeyer, and A.Rosenfeld, editors, *Image and signal processing in electron microscopy*, pages 83–92, P.O.Box 66507, AMF O'Hare (Chicago), IL 660666, U.S.A., April 1988. Scanning Microscopy International, Scanning Microscopy International. 6th Pfefferkorn Conference, Niagara Falls, Canada.

[17] Krakow, W. Raster methods for displaying defocus and astigmatism of scanning electron microscope images. *Ultramicroscopy*, 6:387–404, 1981.

[18] Lai, S., Fu, C., and Chang, S. A generalised depth estimation algorithm with a single image. *IEEE transactions on pattern analysis and machine intelligence*, 14(4):405–411, April 1992.

[19] Lim, J. *Two-dimensional signal and image processing.* Prentice-Hall Signal Processing Series. Prentice-Hall, 1990.

[20] Linfoot, E. Information theory and optical images. *Journal of the Optical Society of America*, 45(1-12):808–819, October 1955.

[21] Nayar, S. K. and Nakagawa, Y. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, August 1994.

[22] Papoulis, A. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill series in System Science. McGraw-Hill, 1965.

[23] Pentland, A. A new sense for depth of field. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-9(4):523–531, July 1987.

[24] Reimer, L. *Scanning Electron Microscopy.* Springer Series in Optical Sciences. Springer-Verlag, Berlin, Heidelberg, 1985. Physics of image formation and microanalysis.

[25] Rempfer, G. F. and Mauck, M. S. A closer look at the effect of lens aberrations and object size on the intensity distribution and resolution in electron optics. *Journal of Applied Physics*, 63(7):2187–2199, April 1988.

[26] Smith, K. Astigmatism correction and determination of resolving power in the SEM. *Journal of Microscopy*, 139:177–186, 1985.

[27] Springer, M. *The Algebra of Random Variables.* Wiley series in probability and mathematical statistics. John Wiley and Sons, 1979.

[28] Sturrock, P. The aberrations of magnetic electron lenses due to asymmetries. *Philosophical transactions of the Royal Society of London: A*, 243:387 – 429, 1951.

[29] Subbarao, M. Parallel depth recovery by changing camera parameters. In *Second international conference on computer vision*, pages 149–155, Innisbrook Resort, Tampa, Florida, USA, December 1988. Computer society press.

[30] Subbarao, M., Choi, T., and Nikzad, A. Focusing techniques. *Optical Engineering*, 32(11):2824–2836, November 1993.

[31] Suganuma, T. A novel method for automatic measurement and correction of astigmatism in the SEM. *Journal of Physics E: Scientific Instruments*, 20:67–73, 1987.

[32] Taylor, P. and Gopinath, A. Signal processing for noise reduction in scanning electron microscopy. *Scanning electron microscopy: systems and applications*, 1973.

[33] Tee, W., Smith, K., and Holburn, D. An automatic focusing and stigmating system for the SEM. *Journal of Physics E: Scientific instrumentation*, 12:35–38, 1979.

[34] Toraldo di Francia, G. Resolving power and information. *Journal of the Optical Society of America*, 45:497 – 501, 1955.

[35] Weaver, H. J. *Applications of Discrete and Continuous Fourier Analysis.* John Wiley and Sons, Inc., 1983.

[36] Wei, T. and Subbarao, M. Continuous focusing of moving objects using DFD1F. In *Machine Vision Applications in Industrial Inspection II*, volume 2183 of *SPIE Proceedings Series*, pages 290–300, San Jose, California, February 1994. SPIE.

Appendix A

Test Images

A.1 Image series: farfocus

Two sets of nine non-astigmatic images were taken of the silver crystal specimen at magnifications of 500× and 1000×. The first image in each set was taken at best focus, using a working distance of approximately $15mm$. Subsequent images were then taken progressively further from best focus, corresponding to a greater degree of out-of-focus. The most out-of-focus image taken was at a distance of $1.2mm$ from best focus.

The microscope operating conditions were as follows:

- Operating voltage: $15kV$
- Beam current: $100pA$
- Aperture size: $30\mu m$

Noise reduction in the form of frame averaging was employed, with a complete image acquisition time of about $4.8s$. The size of each image was 1024×768 pixels.

The astigmatism was manually corrected, and the first image was taken at best focus as determined by eye. The focal length as reported by the microscope was then recorded. Eight images were then taken at progressively greater distances from this position of best focus, and the corresponding focal lengths recorded. At no point was the direction in which the focus was changed reversed. Tables A.1 and A.2 give the reported focal lengths used for the images in each set.

Figure A.1 shows the centre 384×384 pixels of the in-focus image.

Image Name	Focal length (mm)	Dist from focus (mm)
far1	14.95685428	0
far2	14.85710125	0.09975303
far3	14.75734729	0.19950699
far4	14.46030289	0.49655139
far5	14.36056383	0.59629045
far6	14.26082011	0.69603417
far7	13.96385673	0.99299755
far8	13.86412513	1.09272915
far9	13.76440191	1.19245238

Table A.1: Reported focal lengths for **farfocus** image series (500× magnification)

Image Name	Focal length (mm)	Dist from focus (mm)
far10	15.04981611	0
far11	14.95005935	0.09975676
far12	14.85030819	0.19950792
far13	14.55324236	0.49657375
far14	14.45349958	0.59631653
far15	14.35375866	0.69605745
far16	14.05677479	0.99304132
far17	13.95704225	1.09277386
far18	13.8573125	1.19250361

Table A.2: Reported focal lengths for **farfocus** image series (1000× magnification)

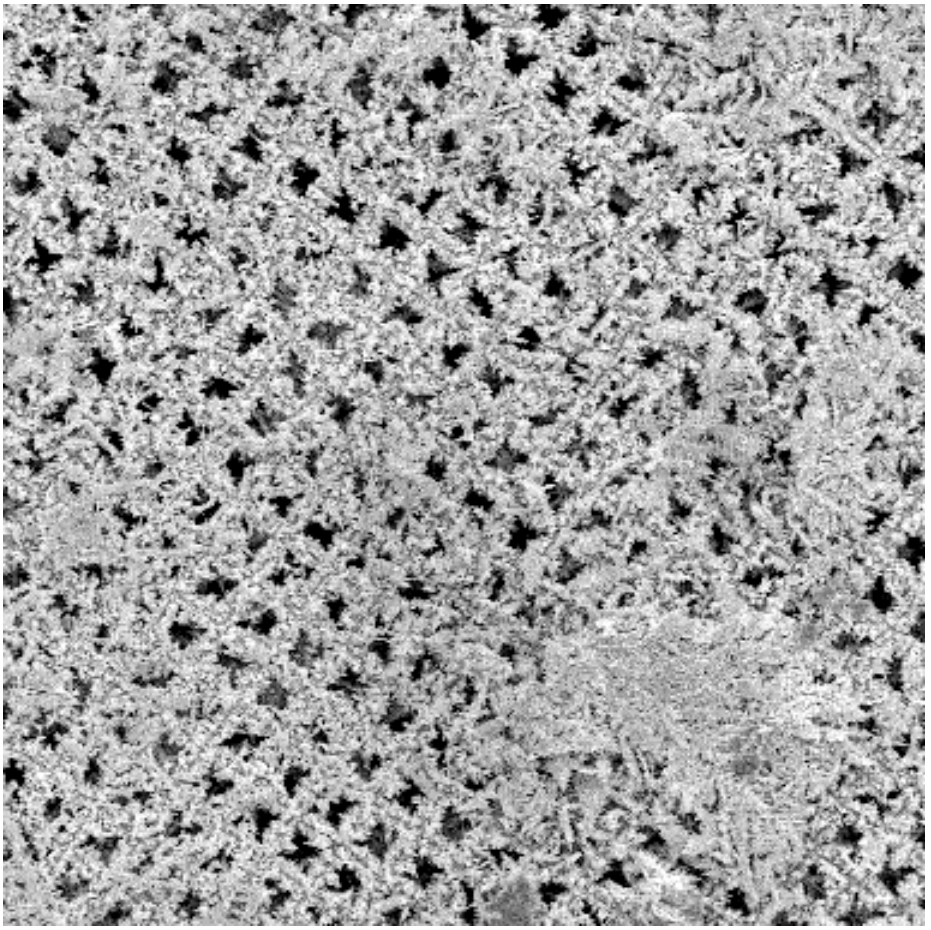


Figure A.1: Centre 384 × 384 pixels of image **far1** from image series **farfocus**

Appendix B

Additional Results

B.1 Scale in the Fourier transform domain

Given an image in the continuous spatial domain, the scale in the Fourier transform domain is unambiguously defined. However, for a sampled image, the sampling frequency needs to be known for this to be the case.

If the continuous space image is sampled such that one pixel represents m metres, then the entire image of $N \times N$ pixels represents a square of side Nm metres. For $k = 0$ to $(N-1)$, the transform coefficient k then determines the number of cycles in Nm metres. This identifies frequency f as

$$f = \frac{k}{Nm} \ [m^{-1}] \tag{B.1}$$

Thus in the frequency domain a single frequency interval (one pixel) corresponds to

$$\frac{1}{Nm} \ [m^{-1}] \tag{B.2}$$

Appendix C

Random Variables

C.1 Ratio of zero-mean normal random variables has no mean

The mean of a distribution is given by

$$E\{z\} = \int_{-\infty}^{\infty} z p_z(z) dz \tag{C.1}$$

which can be split into the two integrals

$$\begin{aligned} E\{z\} &= E_1\{z\} + E_2\{z\} \\ &= \int_0^{\infty} z p_z(z) dz + \int_{-\infty}^{0} z p_z(z) dz \end{aligned} \tag{C.2}$$

Considering just the first of these integrals, for the Cauchy distribution we have that

$$\begin{aligned} E_1\{z\} &= \int_0^{\infty} z p_z(z) dz \\ &= \frac{1}{\pi} \int_0^{\infty} \frac{z}{z^2+1} dz \end{aligned} \tag{C.3}$$

which can easily be shown to evaluate to

$$E_1\{z\} = \frac{1}{2\pi} [\ln(z^2+1)]_0^{\infty} \tag{C.4}$$

$$= \frac{1}{2\pi} \lim_{z \to \infty} \ln(z^2+1) \tag{C.5}$$

This limit does not exist, and hence the integral does not converge. This is sufficient to prove that the mean $E\{z\}$ does not exist. In a similar manner it can be shown that the second central moment (or the standard deviation) of the distribution is infinite.

Because the integrand $z/(z^2+1)$ is an odd function of z, the integral

$$\int_{-\infty}^{\infty} \frac{z}{z^2+1} dz \tag{C.6}$$

does however have a principal value of zero, since

$$\lim_{\eta \to \infty} \int_{-\eta}^{\eta} \frac{z}{z^2+1} dz = 0 \tag{C.7}$$

C.2 Ratio of nonzero-mean normal random variables has no mean

In section C.1 it was shown that

$$\int_0^{\infty} z \left\{ \frac{1}{\pi} \frac{1}{z^2+1} \right\} dz \to \infty \tag{C.8}$$

Thus, using equation 3.11, and equation 3.6 with $x_0' = 0$ and $y_0' = 0$, it must be the case that

$$\int_0^{\infty} z \left\{ k \int_{-\infty}^{\infty} |p| e^{-\frac{1}{2}p^2 z^2} e^{-\frac{1}{2}p^2} dp \right\} dz \to \infty \tag{C.9}$$

or rewriting in a more immediate form

$$k \int_0^{\infty} \int_{-\infty}^{\infty} z|p| e^{-\frac{1}{2}p^2 z^2} e^{-\frac{1}{2}p^2} dp dz \to \infty \tag{C.10}$$

Consider now finding the mean M of the general distribution in equation 3.6 by forming the integral

$$M = k \int_0^{\infty} \int_{-\infty}^{\infty} z|p| e^{-\frac{1}{2}p^2 z^2} e^{-\frac{1}{2}p^2} e^{pzx_0'} e^{py_0'} dp dz \tag{C.11}$$

This can be split into the two components

$$\begin{aligned} M &= M_1 + M_2 \\ &= k \int_0^{\infty} \int_0^{\infty} z|p| e^{-\frac{1}{2}p^2 z^2} e^{-\frac{1}{2}p^2} e^{pzx_0'} e^{py_0'} dp dz \\ &\quad + k \int_0^{\infty} \int_{-\infty}^{0} z|p| e^{-\frac{1}{2}p^2 z^2} e^{-\frac{1}{2}p^2} e^{pzx_0'} e^{py_0'} dp dz \end{aligned} \tag{C.12}$$

Taking just the term M_1, within the integration limits the conditions exist that $p \geq 0$ and $z \geq 0$. If x_0' and y_0' are now restricted to be positive numbers, then

$$e^{pzx_0'} e^{py_0'} \geq 1 \tag{C.13}$$

and

$$z|p|e^{-\frac{1}{2}p^2z^2}e^{-\frac{1}{2}p^2}e^{pzx'_0}e^{py'_0} \geq z|p|e^{-\frac{1}{2}p^2z^2}e^{-\frac{1}{2}p^2} \tag{C.14}$$

for all regions over which the integral is performed. Since by equation C.10 the integral of the right hand side tends to infinity, the integral of the left hand side over the same limits will also diverge.

It has therefore been shown that M_1 does not converge, which is sufficient to prove that the mean M does not exist. Thus the mean of the ratio of two nonzero-mean normal random variables does not exist.

Appendix D

Integration Results

This appendix describes the derivation for the closed-form solution of the probability distribution $p_Z(z)$ used in section 3.3.1.

The required integral was given by

$$p_Z(z) = K \int_{-\infty}^{\infty} |p| e^{-(\frac{1}{2}z^2+\frac{1}{2})p^2 - 2(-\frac{1}{2}zx_0' - \frac{1}{2}y_0')p} dp \tag{D.1}$$

Defining $\mu(z)$ and $\nu(z)$ as per equation 3.9, this can be written

$$\begin{aligned} p_Z(z) &= K \int_{-\infty}^{\infty} |p| e^{-\mu(z)p^2 - 2\nu(z)p} dp \\ &= K \int_{0}^{\infty} p e^{-\mu(z)p^2 - 2\nu(z)p} dp + K \int_{-\infty}^{0} -p e^{-\mu(z)p^2 - 2\nu(z)p} dp \\ &= K \int_{0}^{\infty} p e^{-\mu(z)p^2 - 2\nu(z)p} dp + K \int_{0}^{\infty} p e^{-\mu(z)p^2 + 2\nu(z)p} dp \end{aligned} \tag{D.2}$$

where this second term is obtained by a change in the sign of the variable to be integrated over. This allows for the expression to be simplified as

$$\begin{aligned} p_Z(z) &= K \int_{-\infty}^{\infty} p e^{-\mu(z)p^2} (e^{-2\nu(z)p} + e^{2\nu(z)p}) dp \\ &= 2K \int_{-\infty}^{\infty} p e^{-\mu(z)p^2} \cosh(2\nu(z)p) dp \end{aligned} \tag{D.3}$$

The result then follows from tables of integrals [12, p.365] that

$$p_Z(z) = 2K \left\{ \frac{\nu(z)}{\mu(z)} \sqrt{\frac{\pi}{\mu(z)}} e^{\frac{\nu(z)^2}{\mu(z)}} \Phi\left(\frac{\nu(z)}{\sqrt{\mu(z)}} \right) + \frac{1}{2\mu(z)} \right\} \tag{D.4}$$

Appendix E

Diffraction as a Fourier Transform

It is possible, under some restrictions, to view diffraction as a Fourier transform operation.

Suppose the optical disturbance in a given plane 1 is known, and is given by $\psi_1(\varepsilon, \eta, z)$. If the origin of the coordinate system is chosen to lie in this plane, then this disturbance and be written $\psi_1(\varepsilon, \eta, 0)$, or just $\psi_1(\varepsilon, \eta)$. It is assumed that $\psi_1(\varepsilon, \eta)$ is defined over some region of support R in the plane, which usually represents the physical aperture.

The resulting effect at a general point (x, y, z) is then given by [35]

$$\psi(x, y, z) = \int\int_R \psi_1(\varepsilon, \eta)\chi(\varepsilon, \eta, \mathbf{x})\frac{e^{ikr(\varepsilon,\eta,\mathbf{x})}}{r(\varepsilon, \eta, \mathbf{x})}d\varepsilon d\eta \tag{E.1}$$

In this equation, $r(\varepsilon, \eta, \mathbf{x})$ represents the length of the line segment from $(\varepsilon, \eta, 0)$ to (x, y, z). $\chi(\varepsilon, \eta, \mathbf{x})$ is an inclination factor which depends on the angle between this line and the normal to plane 1. k is the usually defined wave number.

If the observation points are restricted to lie in a second plane 2 a distance z from plane 1, $\psi(x, y, z)$ can be written as $\psi_2(x, y)$ and equation E.1 becomes

$$\psi_2(x, y) = \int\int_R \psi_1(\varepsilon, \eta)\chi(\varepsilon, \eta, \mathbf{x})\frac{e^{ikr(\varepsilon,\eta,\mathbf{x})}}{r(\varepsilon, \eta, \mathbf{x})}d\varepsilon d\eta \tag{E.2}$$

The distance r is given by the expression

$$r(\varepsilon, \eta, \mathbf{x}) = \sqrt{(x-\varepsilon)^2 + (y-\eta)^2 + z^2} \tag{E.3}$$

which in a different form is

$$r(\varepsilon, \eta, \mathbf{x}) = z\sqrt{\left(\frac{x-\varepsilon}{z}\right)^2 + \left(\frac{y-\eta}{z}\right)^2 + 1} \tag{E.4}$$

Under conditions of large z, the binomial approximation $\sqrt{1+\alpha} \approx 1 + \alpha/2$ can be used on equation E.4 to give

$$r(\varepsilon, \eta, \mathbf{x}) \approx z \left[1 + \frac{(x-\varepsilon)^2}{2z^2} + \frac{(y-\eta)^2}{2z^2} \right] \tag{E.5}$$

This previous approximation is used in the exponential term in equation E.2, which is a dominant term since k is a large number. For the denominator term the cruder approximation $r \approx z$ is made. Furthermore, the term $\chi(\varepsilon, \eta, \mathbf{x})$ is assumed to be constant, and approximately equal to χ. The validity of these assumptions is confirmed in [35]. The resulting expression after using these approximations is

$$\psi_2(x,y) = \frac{\chi e^{ikz}}{z} \int\int_R \psi_1(\varepsilon,\eta) y e^{\frac{ik}{2z}[(x-\varepsilon)^2+(y-\eta)^2]} d\varepsilon d\eta \tag{E.6}$$

With some manipulation this can be written in the form

$$\psi_2(x,y) = \frac{\chi e^{ikz}}{z} e^{\frac{ik}{2z}(x^2+y^2)} \int\int_R \psi_1(\varepsilon,\eta) e^{\frac{ik}{2z}(\varepsilon^2+\eta^2)} e^{-\frac{ik}{z}(x\varepsilon+y\eta)} d\varepsilon d\eta \tag{E.7}$$

Defining an aperture function

$$A(\varepsilon,\eta) = \begin{cases} 1 & (\varepsilon,\eta) \in R \\ 0 & \text{otherwise} \end{cases} \tag{E.8}$$

and from this a modified aperture function

$$\psi_{mod}(\varepsilon,\eta) = \psi_1(\varepsilon,\eta) A(\varepsilon,\eta) e^{\frac{ik}{2z}(\varepsilon^2+\eta^2)} \tag{E.9}$$

the equation E.7 can be written

$$\psi_2(x,y) = C(\mathbf{x}) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \psi_{mod}(\varepsilon,\eta) e^{-i(\frac{kx}{z}\varepsilon + \frac{ky}{z}\eta)} d\varepsilon d\eta \tag{E.10}$$

where $C(\mathbf{x})$ is

$$C(\mathbf{x}) = \frac{\chi}{z} e^{ikz} e^{\frac{ik}{2z}(x^2+y^2)} \tag{E.11}$$

The form of this last equation for $\psi_2(x,y)$ can be seen to resemble that of a Fourier transform, with respect to frequencies $(\frac{kx}{z}, \frac{ky}{z})$. If the symbol $\mathcal{F}_s$ represents the Fourier transform with the frequencies scaled by a factor $\frac{k}{z}$, then

$$\psi_2(x,y) = C(\mathbf{x}) \mathcal{F}_s \left[\psi_{mod}(\varepsilon,\eta) \right] \tag{E.12}$$

Using this form of the equation, diffraction under the conditions presented here can be represented and calculated as a modified Fourier transform operation.

E.1 PSF with aberrations

The theory of aberrations is usually discussed in terms of wavefronts. Equations for the various aberrations then represent the position of an actual wavefront with respect to a reference sphere centred on the Gaussian image point. Consider Figure E.1: W is a wavefront

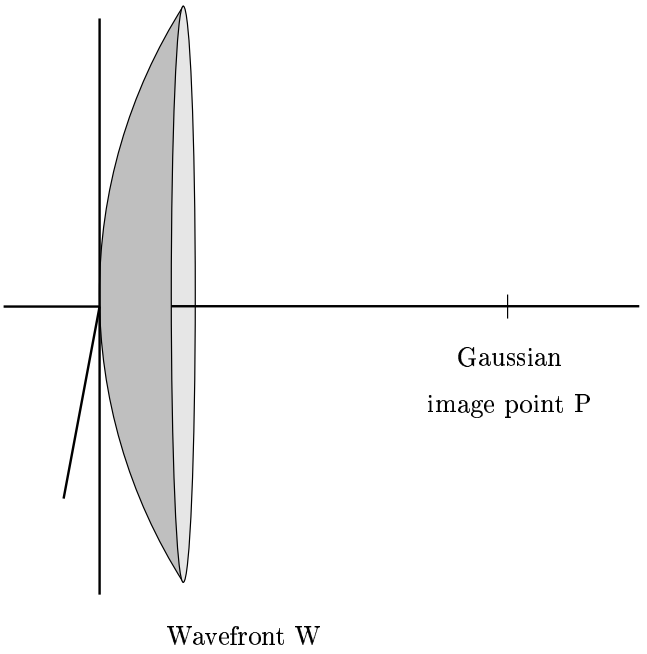


Figure E.1: Actual wavefront corresponding to ideal image point

in space which is roughly convergent on a Gaussian image point P, which is the image of a perfect point source transmitted through the imaging system. In the absence of aberrations the wavefront would be exactly circular and centred on P, and the image would also be a perfect point. However, if aberrations are introduced, the situation deviates from the ideal and this wavefront is distorted. An aberration function Φ will be defined to be the distance in the direction of wavefront propagation by which the actual position of the wavefront differs from this circular ideal.

For this purpose, a reference sphere G is defined which is centred on this image point P and passes through the centre of the aperture. P is assumed to be a distance z from the aperture. The situation is depicted in Figure E.2. $\Phi(\varepsilon, \eta)$ will be defined to be positive if the actual wavefront lies behind the reference sphere.

The primary aberrations can now be completely specified using this aberration function. These aberrations are [3]:

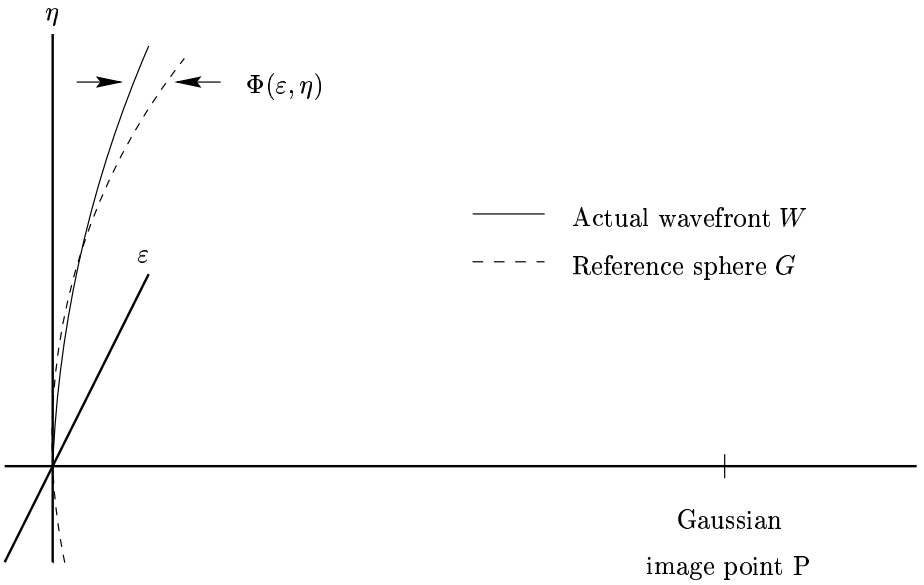


Figure E.2: Image formation in terms of wavefronts

- Spherical aberration $\Phi = -K_s\rho^4$
- Coma $\Phi = K_c\rho^3\cos(\theta)$
- Astigmatism $\Phi = -K_a\rho^2\cos^2(\theta)$
- Curvature of field $\Phi = -K_f\rho^2$
- Distortion $\Phi = K_d\rho\cos(\theta)$

where $\rho = \sqrt{\varepsilon^2+\eta^2}$ and θ is the angle of the point of interest in the (ε,η)-plane. Since these equations represent the aberrations on a point source being imaged by the system, the image resulting from the wavefront will be the PSF of the system in the presence of these aberrations. It will now be shown how the functions here can be used to calculate the PSF of the system.

In the previous section it was shown how the Fourier transform can be applied to a modified aperture function to find the resulting diffraction pattern at any point far from the aperture. What then had to be known was the disturbance function $\psi_1(\varepsilon,\eta)$ in the plane of the aperture. To present aberrations in terms of this previous discussion, it is necessary to find this disturbance function given the location of a wavefront. This can be done if the following is noted; moving from the wavefront to the reference sphere represents a phase shift of $k\Phi$, and then back to the (ε,η)-plane a shift in the opposite direction of kq, where q is the distance from the reference sphere to the aperture plane. The combined phase shift can thus be represented

by multiplication by a term

$$e^{ik(\Phi-q)} \tag{E.13}$$

where q can easily be verified to be

$$q = \frac{1}{2z}(\varepsilon^2 + \eta^2) \tag{E.14}$$

Thus if the magnitude of the disturbance on the wavefront is M, the aperture disturbance function can be written as

$$\psi_1(\varepsilon, \eta) = M e^{ik\Phi(\varepsilon,\eta)} e^{-\frac{ik}{2z}(\varepsilon^2+\eta^2)} \tag{E.15}$$

Since incoherent imaging is being considered, the system PSF is given by $|\psi_2(x, y)|^2$.