ELSEVIER

Ultramicroscopy 87 (2001) 105–121

ultramicroscopy

www.elsevier.nl/locate/ultramic

A simple model of holography and some enhanced resolution methods in electron microscopy

J.M. Rodenburg*

Materials Research Institute, Sheffield Hallam University, City Campus, Sheffield S1 1WB, UK

Received 3 May 2000; received in revised form 11 September 2000

Abstract

A simple pictorial model of electron interference effects based on an extended representation of the autocorrelation function is described and developed. Unlike Abbe's theory of transmission imaging, the model incorporates fully the effect of the loss of phase that occurs in the detector plane. The aperture transfer function and information limit (envelope function) are also incorporated with reference to the simplest scattering geometry of Young's slits. The model is then applied to holography, the diffraction phase problem, ptychography, Wigner distribution deconvolution, conventional bright-field imaging, single side-band imaging and tilt-series reconstruction. Some of these methods require an understanding of four-dimensional integral functions, but the model reduces the problem into a projection of a two-dimensional space. It is hoped that the model will help material scientists who are not specialists in imaging and diffraction theory to understand some recent developments in advanced super-resolution imaging methods. © 2001 Elsevier Science B.V. All rights reserved.

1. Introduction

There are a number of advanced transmission electron imaging techniques which may obtain better resolution, or more easily interpretable information, by using unusual scattering geometries combined with inverse computation methods. As computing power becomes cheaper and detector technology is improved, these techniques have increasing potential to deliver real gains in microscope performance. Indeed, it may even be appropriate to change the whole rationale of electron microscopy and steer the instrumentation development in favour of exploiting indirect methods. However, such techniques, especially those that rely on exploiting illumination tilt series or the microdiffraction plane of the scanning transmission electron microscope (STEM), are hard to understand for non-specialists. Furthermore, much scepticism is met by any method which claims to surpass the resolution limit defined in terms of the maximum spatial frequency that can pass through a limited aperture. In this sense, Abbe's theory has become an impediment to our understanding of some indirect super-resolution methods.

In this paper I propose a different way of thinking about transmission imaging theory. The model is simply an extension of Abbe's theory, but it automatically builds in the phase problem and

*Tel.: +44-114-225-4037; fax: +44-114-225-3501.
E-mail address: j.m.rodenburg@shu.ac.uk
(J.M. Rodenburg).

0304-3991/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0304-3991(00)00097-8

the fact that other variables, such as the angle of illumination (or, in the case of STEM, the extent of the microdiffraction plane) can provide information that greatly surpasses the conventional resolution limits. The model is primarily a way of picturing the consequences of complicated interference effects that are generally expressed as rather complicated integral equations. For the purposes of clarity, I keep to single set of co-ordinates and use a simple pictorial allegory. In practice, the co-ordinates of the data are sometimes in reciprocal space and at other times in real space, but I do not express these differences in the mathematics. Note also that I do not strictly differentiate between convolution and correlation: depending on definitions of physical co-ordinates, the aperture function should in places be reflected along the x-axis, but this is not crucial to the main picture. What matters here is to give the broad scope of how some rather seemingly unrelated techniques–holography, ptychography, single side-band imaging, bright-field imaging, the classic diffraction phase problem, and Wigner distribution deconvolution, tilt series reconstruction–can all be represented in a single diagram.

In the next section, the pictorial representation is introduced without any justification or background but is used to explain the simplest reconstruction method: holography. We write down a simple version of the mathematics in Section 3. After a brief explanation of how the model relates to Young's slits, holography and the classic phase problem in Section 4, we discuss how the limits of interference and the problems of increased resolution impact upon the model in Section 5. Section 6 then describes some of the conventional and super-resolution techniques in the context of the model. Conclusions are presented in Section 7.

2. The qualitative model

Think of a paintbrush, loaded with paint, drawn diagonally across the surface of a wall, as shown in Fig. 1a. Assume the paintbrush has an uneven distribution of paint loaded onto its bristles, and that this distribution can be represented graphically by a one-dimensional plot, also shown in Fig. 1a. If the value of this one-dimensional function has a certain numerical value corresponding to a particular bristle on the paintbrush, then imagine that this value has now been painted along a line across the wall. In other words, any point in the wall that was touched by that bristle now has the numerical value associated with that bristle. Now take the same brush and perform a second diagonal brushstroke, in an inclined direction, across the first, as shown in Fig. 1b. If this was a real paintbrush, we would expect the quantity of paint at any one point on the wall to be, roughly, the *addition* of the quantities of paint left by each of the individual brushstrokes. In what follows, however, we have to imagine forming the *product* of the two numerical values painted on the wall. Lets call this the 'paint product function'. If the paintbrushes were both evenly loaded with paint, the paint product function would form a diagonal shape and would have zero value elsewhere. For more complicated functions, like a square wave distribution of paint, the product could have quite complicated structure, as shown in Fig. 1c, where dark regions represent areas of the wall which have a large numeric value. We assume the wall has a default value of zero: in other words, parts of the wall that are only painted by one of the brushstrokes form a zero product.

To proceed, we next form a horizontal integration of this strangely painted wall into another one-dimensional function. One way to imagine this is to take a dry paintbrush and to sweep it horizontally across the freshly painted wall. This paintbrush picks up paint from the wall. A bristle that passes over dense areas of paint (that is, paint formed from the product on the first two brushstrokes) builds up a large quantity of paint: each bristle integrates the paint-product function in the horizontal direction. The process is shown in Fig. 1d. In other words, we project (i.e. integrate) the paint on the wall onto a one-dimensional vertical line, which has a corresponding one-dimensional distribution of paint along it. This is a function of the vertical co-ordinate, y, and Fig. 1d, as in later figures, we can view a conventional plot of this projection by viewing at right-angles to the page the graph to the left of the main diagram.

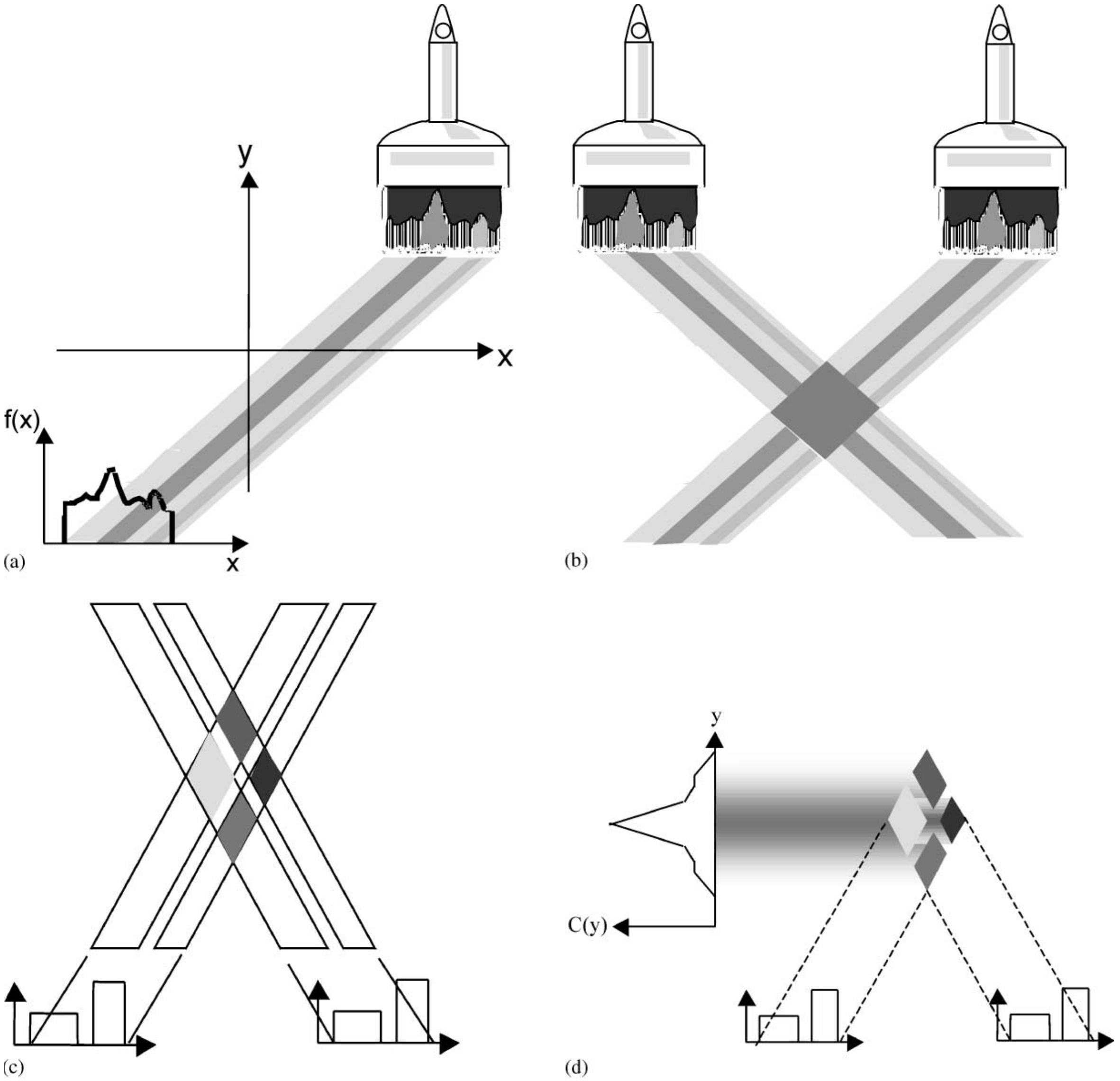

Fig. 1. (a) A single paintbrush function. A two-dimensional ‘wall’ function $w(x, y)$ has been painted across by a one-dimensional function, $f(x)$. Every point in $w(x, y)$ lying on a diagonal line has the same numerical value. Those points touched by a particular bristle of the paintbrush have the value corresponding to $f(x)$ on that bristle. (b) Two strokes from the same paintbrush cross one another. In the model, the diagonal area where the strokes overlap have a value corresponding to the product of each individual paintbrush function. (c) For a simple function consisting of two top hats of different sizes, the ‘paint product function’ has diamond-shaped areas of positive value. Darker areas represent areas where the product is large. Areas covered by just one paintbrush function have zero as product value, because we assume unpainted areas have a default value of zero. (d) The horizontal projection of the same product function in (c). To view the projection (the autocorrelation function) look at the page at right-angles to see $C(y)$ plotted as a function of y.

To summarise: we start with a one-dimensional function; we paint this in two diagonals across a wall; we form the product of the two brushstrokes; and finally, we project (integrate) this function in the horizontal direction, thus ending up with a second one-dimensional function.

All this relates to electron microscopy in the following way. The initial function we load onto the paintbrush is a one-dimensional quantum mechanical electron wave function. In practice, this would be something like the exit wave function below the specimen in a transmission electron microscope or perhaps the disturbance over the back focal plane of the objective lens. These functions would generally be two-dimensional, which means that our 'wall' would have to be four-dimensional: for the purposes of clarity, we will only consider one-dimensional functions, although all the mathematical analysis that follows can be trivially extended for two-dimensional functions. By painting the wave function across itself, we form the set of all possible interference conditions with different parts of itself. The final one-dimensional function, obtained by integrating horizontally across the two brushstrokes, is called the autocorrelation function.

Consider the case of holography, where we have a reference wave which is independent of our function of interest (image or diffraction pattern). We can regard this as our paintbrush having an errant bristle, a long way left or right of the main part of the paintbrush, as shown in Fig. 2. This errant bristle creates a narrow line of paint across the wall. Where it crosses the second brushstroke it is as if it picks out a profile of the original function. When this is subsequently projected horizontally, what we see is three regions: a central broad region and two the so-called side-bands (representations of the original function) either side of this central region. Such side-bands occur, for example, in the Fourier transform of a real-space hologram, in which case the projected function $C(y)$ (see Fig. 2) would be an estimate of the wave amplitude in the back focal plane of objective, which is itself related to the image by a forward Fourier transform propagator. The two bushstrokes, one of which must formally be the complex conjugate of the other, arises from the fact that we only measure intensity in the image plane. The main benefits of the paintbrush construction is that this loss of phase, and all the Fourier transforms, are dealt with simultaneously by the pictorial representation. For this reason, we can easily extend it into much more advanced interference methods such as ptychography and Wigner distribution deconvolution, which we explain in Section 6.

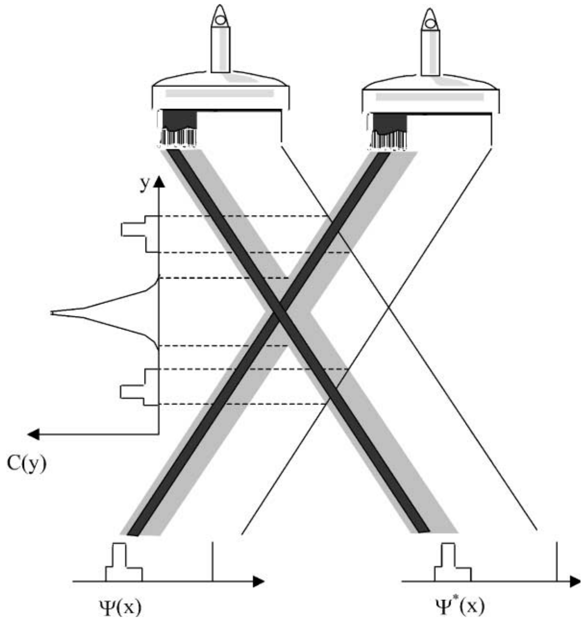

Fig. 2. The representation of holography in the model. The paintbrush has an errant bristle (the reference wave), so that $C(y)$ now clearly separates two reconstructions of the wave field in the back focal plane, $\Psi(x)$.

3. The mathematical model

The electron wave function is a complex variable. In other words, our paintbrush functions, the surface of our wall, and the final projected function must all be complex variables. Let the two-dimensional surface of the wall be given by the two-dimensional function w, wherein

$$w(x, y) = u(x, y) + \mathrm{i}v(x, y), \tag{1}$$

where i is the imaginary number and $u(x, y)$ and $v(x, y)$ are real-valued functions of real coordinates x and y which describe positions over the wall. Alternatively, we could write

$$w(x, y) = m(x, y)\exp \mathrm{i}\phi(x, y), \tag{2}$$

where

$$m(x, y) = \sqrt{(u^2(x, y) + v^2(x, y))} \tag{3}$$

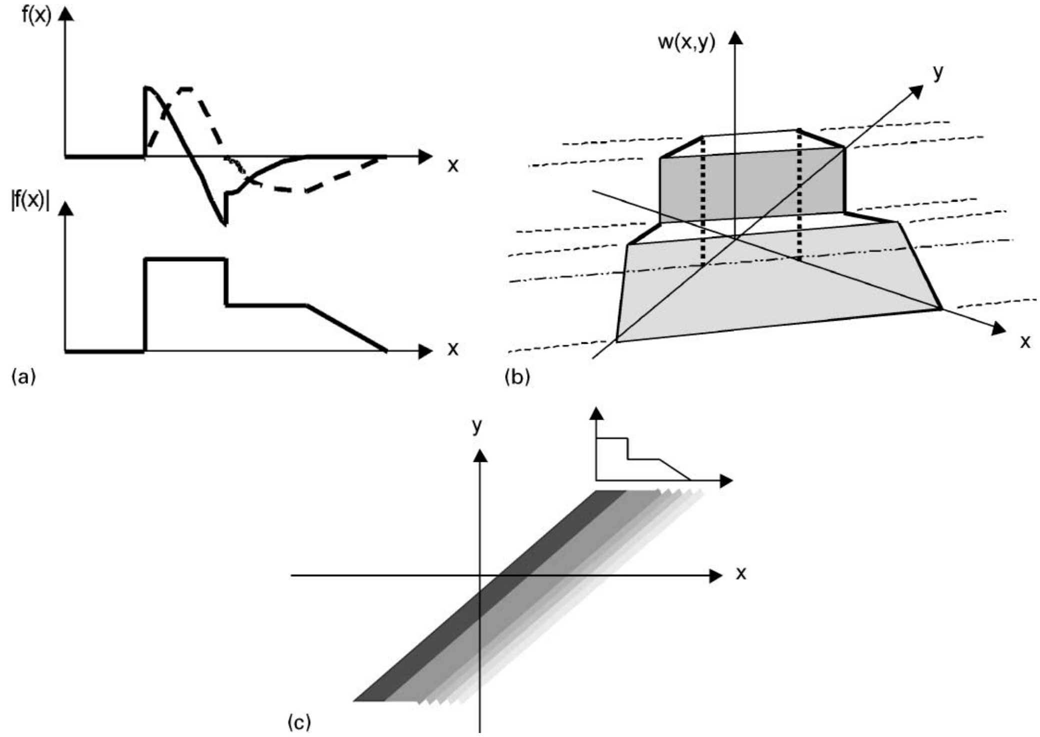

Fig. 3. (a) Conventional representation of the one-dimensional function $f(x)$, the solid line being the real part, the dotted line being the imaginary part. The modulus, $m(x)$, is shown in the lower graph. (b) The corresponding paintbrush function shown as surface, where height represents the modulus of the paintbrush function. (c) The same paintbrush function plotted (as in the rest of this paper) as a grey-scale plot, with dark tones representing high numerical values of the modulus.

is the modulus of $w(x, y)$ and

$$\phi(x, y) = \tan^{-1}(v(x, y)/u(x, y)) \tag{4}$$

is the phase or argument of $w(x, y)$.

There is no easy way to represent graphically such a function, short of resorting to two contour maps or two grey-scale images. In all the figures we show in this paper, we will adopt the convention that we will plot only the modulus of $w(x, y)$, $m(x, y)$, as a grey-scale image: we just have to remember that in this type of function also has associated with it a phase, $\phi(x, y)$.

Let $f(x)$ be the one-dimensional complex function which we apply to our paintbrush. Remember that we will use this function to represent a one-dimensional time-independent electron image, exit wave or diffraction pattern: exactly which plane of the microscope it represents will depend upon the various contexts described in the next section. We can plot this conventionally as in Fig. 3a, using a solid line to represent the real part, $u(x)$, and a dotted line to represent the imaginary part, $v(x)$. If we make our first brushstroke at 45° to vertical (Fig. 1), then we have

$$w(x, y) = f(x - y). \tag{5}$$

Fig. 3b shows $w(x, y)$ plotted as a surface embedded in a three-dimensional space (for clarity, only the modulus, $m(x, y)$ is shown). Along the x-axis (i.e. along $y = 0$), we see simply $f(x)$ plotted as a function of x. Meanwhile, along the y-axis (i.e. along $x = 0$) we see the same function again, this time plotted as a function of y, but reversed with respect to the sense of the axis. In other words, $w(0, y) = f(-y)$. In Fig. 3c we see a grey-scale plot of $w(x, y)$, where the darkness of the image corresponds to the magnitude of $w(x, y)$. All points in this image which lie on a diagonal parallel to $x = y$ have the same value. This is a single paintbrush function. More generally, we could write

$$w(x,\ y) = f(x - \alpha y - \beta), \tag{6}$$

where α and β are constants which define the gradient $(1/\alpha)$ and position of the paintbrush stroke.

In order to form the product paintbrush function described in Section 2, we need to form the product of two brushstrokes that cross over one another. We can write this as

$$w(x,\ y) = f(x - y)f^*(x + y), \tag{7}$$

where f^* represents the complex conjugate of the function f. In Section 2, we did not worry about the complex conjugate, although it is crucial when f is complex. It arises from the fact that we wish $w(x,\ y)$ to represent a set of possible interference conditions available to us experimentally. Since these are all measured in intensity (real numbers relating to quantum mechanical probabilities) then we will only encounter terms involving the original wave function times its complex conjugate.

The final step of the construction is to reduce $w(x,\ y)$ into the one-dimensional function by integrating horizontally. We will end up with a function of y only, because we have integrated over x, which we could write as

$$C(y) = \int w(x, y)\, \mathrm{d}x \tag{8}$$

If we want to retain the scaling between our original function $f(x)$ and our final function $C(y)$, we have to use a modified version of Eq. (7): what we have to do is alter slightly the angle of each brushstroke. Think of the example of holography (Fig. 2). For most combinations of brushstroke angle, the side bands of the projected function $C(y)$ will be stretched or squeezed relative to the original $f(x)$. In all real experimental situations this scaling factor will indeed occur: it is the magnification or camera length of the electron microscope, but for the purposes of symmetry and elegance, let us define $C(y)$ as

$$\begin{aligned} C(y) &= \int w(x, y)\, \mathrm{d}x \\ &= \int f(x - y/2) f^*(x + y/2)\, \mathrm{d}x, \end{aligned} \tag{9}$$

where we have chosen gradients of 2 and -2 for the paintbrush functions. $C(y)$ so defined is called the autocorrelation function. We can shear our paintbrush functions left or right without affecting the result $C(y)$ so, for example, the autocorrelation is often written

$$C(y) = \int f(x) f^*(x + y)\, \mathrm{d}x \tag{10}$$

via a simple substitution for x, equivalent in our picture of forming one vertical brushstroke, followed by one at 45° (Fig. 4). The autocorrelation is sometimes defined as the complex conjugate of Eq. (10) – see the appendix.

4. Some simple applications

4.1. Young's slits

Consider Young's slits experiment (Fig. 5). Instead of having two identical empty slits, let each slit have a modulus and a phase: the modulus of a slit corresponds to the fraction of electrons it transmits and the phase could be introduced by having a potential well or phase plate within the slit. When illuminated by a coherent plane wave, we have a wave function in the exit plane of the slits, say $\Psi(x)$, which consists of two spikes with

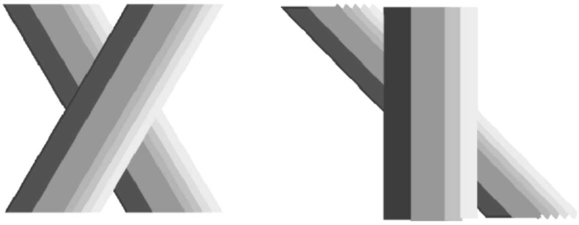

Fig. 4. We can shear the brushstrokes left and right, corresponding to the substitution of the variable of integration in Eqs (9) and (10), but this causes no difference in the horizontal projection, $C(y)$.

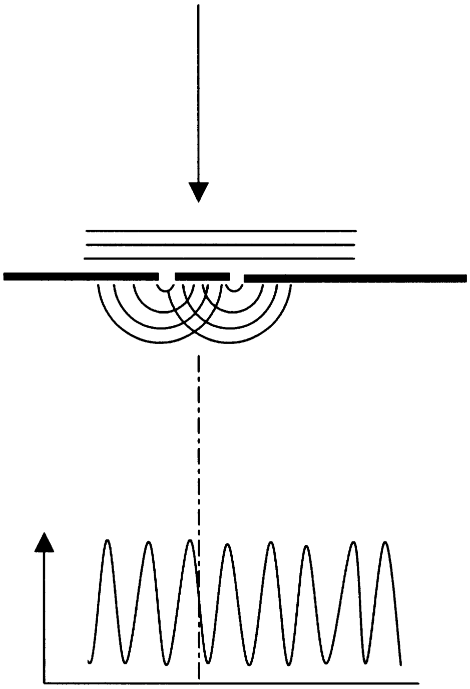

Fig. 5. Young's slits experiment. The exit wave propagates to the far-field where it forms an intensity pattern corresponding to the usual fringes. The fringes may be shifted laterally if there is a phase difference between the wave disturbances at the exit of the slits.

complex amplitudes which we will call B and C, represented graphically within Fig. 6. (We ignore time-dependent factors of e^{iwt} in the wave function, because when we come to detect an interference phenomenon these factors cancel via their complex conjugates.) This is the most elementary spatially resolved wave function we can imagine. Our aim is to measure $\Psi(x)$: i.e. to measure the position and complex amplitudes of the two spikes.

Suppose that all we are able to do is to measure the intensity of the far-field Fraunhofer diffraction fringes. In other words, we are allowed to measure the intensity of the Fourier transform of our exit wave. Not surprisingly, some information is lost in the process. We can measure four things from the interference fringes: their total intensity summed up over the entire Fraunhofer plane, which is proportional to $B^2 + C^2$; the periodicity of the fringes, which tells us the separation of the slits (but not their absolute position); the displacement of the fringe pattern relative to the optic axis, which tells us if there is a phase difference between the complex values of the two slits; and the depth of the interference fringes, which tells about the magnitude of the product BC relative to $B^2 + C^2$.

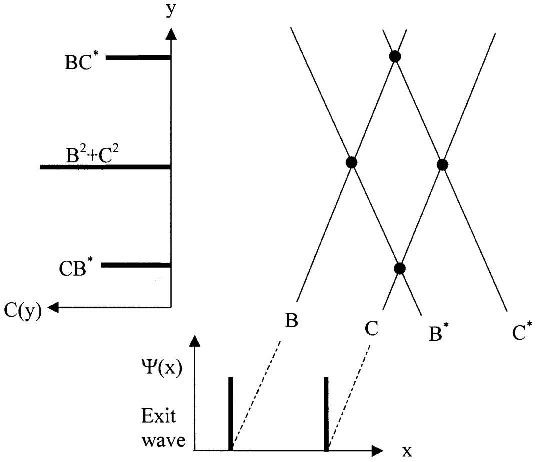

Fig. 6. The paintbrush function for Young's slits. The exit wave function consists of two spikes with unknown complex values B and C. There are four points of non-zero amplitude in $w(x, y)$. The Fourier transform of the intensity of the fringes lying in the Fraunhofer diffraction plane yields the autocorrelation function, $C(y)$, consisting of three peaks.

The same information can be obtained by taking the Fourier transform of the fringes. The process of recording the intensity of the fringes has destroyed the phase information in the wave field. Obviously, if we were able to record the full amplitude and phase of the diffraction pattern, then a single back-Fourier transform would yield the exit wave from the slits, because the Fourier transform, whether performed physically as a result of the Fraunhofer propagator or as a computation, preserves information. Taking the Fourier transform of the *intensity* of the diffraction pattern, which we can think of as the next best thing, yields the autocorrelation function (see the appendix: note that in some contexts the autocorrelation is defined as the back-Fourier transform of the intensity function), as expressed by the model described in Section 2 and by Eqs. (9) or (10).

So, for our two slits, this entire process–wave propagation into the Fraunhofer plane; the measurement of intensity in that plane; and the Fourier transformation of that intensity–is all represented by the single paintbrush stroke diagram shown in Fig. 6. It is as if our paintbrush only has two bristles, of complex value B and C. The height of the central peak in the autocorrelation function is B^2+C^2; the position of both side peaks occurs at the same distance as the separation of the slits; and the heights of these latter peaks give terms like $|BC|\exp \mathrm{i}(\phi_B-\phi_C)$, or the complex product BC^*. The symmetry of the autocorrelation function means that only one-half of it gives relevant data, because the Fourier transform of any real function (the measured intensity) is complex-conjugate symmetric.

But notice that if we have try to solve for the complex values of B and C from just the autocorrelation function (which we leave as an exercise for the reader), we find we have lost crucial information in this experiment: given only the autocorrelation function, we are unable to tell which slit has which modulus. (Of course, we have also lost the absolute phase of the wave function, but since we are not interested in time dependence, this is irrelevant.)

Before proceeding to more complicated scattering geometries, the reader should note that in this very simple case, we could measure the whole of $w(x, y)$ for the two slit problem by being allowed to perform a second experiment. If we were able to place a photographic film in the plane of the slits, or indeed in a magnified image of the slits, we would be able to measure A^2 and B^2 separately. We would therefore have been able to separate the lateral components of $w(x, y)$, at least along $y=0$, instead of simply projecting them into a one-dimensional function. Given the entirety of $w(x, y)$, it always possible to solve for the underlying paintbrush function, except for the usual loss of absolute phase. Indeed, this is exactly why the paintbrush function is a useful model because it allows us to see where extra information can be extracted from a scattering experiment.

4.2. The general phase problem

In diffraction theory, the autocorrelation function defined in Eq. (10) is called the Patterson function [1]: the Fourier transform of the diffracted intensity. In the case of a crystalline specimen where every peak can be recorded, it renders the autocorrelation of the unit cell in real space. When we replace the Young's slits with a complicated wave scattered from a large specimen, then it becomes apparent that the autocorrelation function cannot let us solve easily for the value of some general $\Psi(x)$. Cross-terms superpose upon one another as we do the projection in the $w(x, y)$ plane. $C(y)$ does not contain enough independent measurements to solve for $\Psi(x)$. The intractability of this general phase problem is mitigated when posed in two or three dimensions because of geometric constraints: other a priori information (such that the scattering medium is composed of discrete atoms, or the unit cell has a single heavy atom at a known location) can also render the problem soluble for many crystalline materials. Note that for large amorphous specimens, the diffraction pattern cannot generally be recorded at high enough angular resolution in order to capture all the necessary intensity information, and hence only average atomic spacings (the pair correlation function) can be inferred.

4.3. Holography

The holographic solution to phase problem has already been discussed informally in Section 2. A delta function can be added to the plane of $\Psi(x)$ displaced some distance away from the amplitude we are interested in. Our autocorrelation function now looks like Fig. 2, and consists of a central peak, where many cross-terms overlap, as in the classic phase problem, and two side-lobes which are complex-conjugate symmetric. In the region of the side-lobes, we find amplitudes corresponding to terms like $R\Psi^*$ and $R^*\Psi$, where R is the amplitude of the delta-function component of the wave. Within these side-lobes, values of x in $\Psi(x)$ project directly to values in y in $C(y)$ except for a constant offset determined by the separation of the reference wave. Ignoring this offset, and assigning

R to be unity and of zero phase, we can write very loosely that

$$C(y) = \Psi(-x) = \Psi(-y)$$

for values of y within the upper lobe and

$$C(y) = \Psi^*(x) = \Psi^*(y)$$

for values of y within the bottom lobe.

The reference wave has therefore cleanly separated out an estimate of $\Psi(y)$ which is equivalent to $\Psi(x)$, and which is what we want to measure.

There are many different ways of setting up an actual holographic experiment. The most commonly employed is to have a beam splitter near an image plane of the microscope, say at the selected area diffraction aperture. This serves to deflect in angle an unscattered region of the wave field with respect to that part which has passed through the specimen. In the final image plane (at the phosphor screen, film or CCD camera) the deflected beam crosses over those beams which have passed through the specimen and are destined to form the conventional image: their interference creates the hologram. If one could look up the column from the image plane, one would see the back focal plane of the objective split into two parts: the amplitude distribution diffracted from the specimen (the intensity of which corresponds to the conventional selected area diffraction pattern), and the reference beam, far off to one side (Fig. 7). This plane, the back focal plane of the microscope (albeit modified by the effects of the beam splitter) corresponds to the plane of the slits in Section 4.1. The hologram, corresponding to the interference fringes, is what we record in intensity in the image plane. Our real and reciprocal space co-ordinates have therefore swapped roles relative to the previous examples, but the effect is the same because two planes are still separated by one Fourier propagator. The Fourier transform of this image (often called the diffractogram) is therefore the autocorrelation function of the back focal plane. In practice, in order to obtain the complex value of the exit wave field from the specimen, one must slice out one side-lobe of the diffractogram; if necessary, multiply it by a phase alteration that accounts for any lens aberrations present in the objective lens; and then Fourier transform it back. However, for purposes of our discussion, everything of importance is expressed within our simple paintbrush model: the strength of holography is that it separates the intensity cross-terms so that in some other plane (in this case, lying in the back focal plane) a clean estimate of the complex wave field is available. Where this wave field actually resides (whether in real space, reciprocal or somewhere in between like the region of Fresnel diffraction that occurs as a function of defocus below the specimen) is irrelevant, provided we know what the propagator is involved to get us to the actual plane

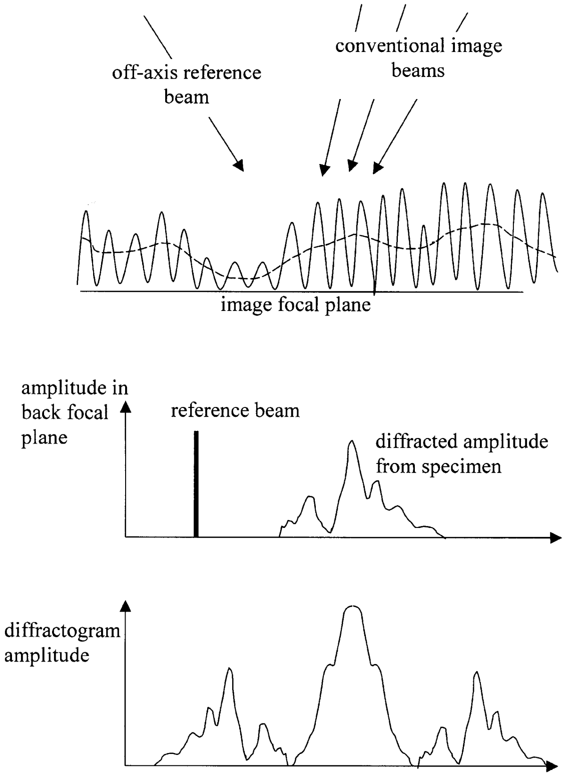

Fig. 7. Electron holography: The paintbrush function depicted in Fig. 2 represents how the Fourier transform of a hologram corresponds to the amplitude in the back focal. beams from the conventional image (top diagram) would meet at a focus in the image plane (dotted line). The reference beam interferes with these at an angle, causing high-frequency interference fringes. The effective amplitude in the back focal (middle diagram), as it would be seen looking up the column from the image plane. The diffractogram amplitude, $C(y)$, is shown in the lower diagram. Compare with Fig. 2.

of interest, say the exit wave field. Indeed, all sorts of other holographic geometries are available [2], but the principle of separation is always the same.

5. Apertures and probe functions

So far, we have not made much use of $w(x, y)$, the two-dimensional function formed by the paintbrush functions. As far as the conventional phase problem is concerned, Eq. (9) expresses everything we need to know about the consequences of measuring intensity after a Fraunhofer propagation. However, in the context of microscopy, much more than simply the projection of $w(x, y)$ is accessible in a way which allows indirect solution of the phase problem. Furthermore, if our paintbrush function, $\Psi(x)$, resides in reciprocal space, then it becomes possible to extend the domain or width of $\Psi(x)$ which is measurable, and hence increase achievable resolution, even beyond the so-called 'information limit' which arises from the difficulty of interfering electron beams of very different path length.

Let us return to the Young's slit experiment. Suppose we now have the opportunity of placing a moveable aperture in the same plane as the slits. In our one-dimensional analysis, the aperture is like a top-hat function. Its paint product function (Section 2) is just a diamond shape in $w(x, y)$ (Fig. 8a), which moves left and right as we move the aperture. If the aperture was not a real function, but also contained phase changes (which for a real electron microscope lens aperture or focussed electron probe this will always be the case), then the effect would be to retard or advance the phase of the complex numbers B and C which lie on the exit wave from the slits: in other words, we have to form the product of the slit function and the aperture before allowing the waves to propagate to the far field. So, for any one position of the aperture, the exit wave field coming from the slits and the aperture is simply the product of the slit function and the aperture function. Let the position of the aperture relative to the slits be denoted by X. We now have a correlation function, $C(y)$, for every position of the aperture X. In other words, we can measure a two-dimensional function $C(y, X)$ where

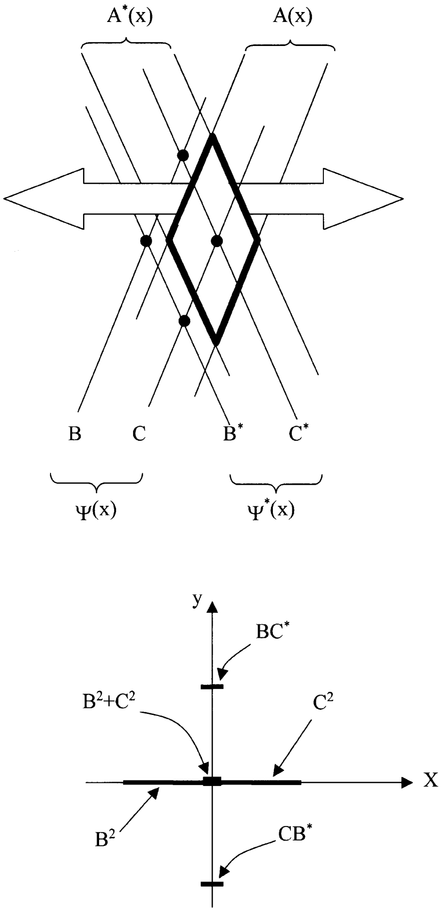

Fig. 8. (a) An aperture function, $A(x)$, crosses over $A^*(x)$, to form a diamond window within the $w(x, y)$ space. This can be moved left or right, thus isolating different sections of the underlying Young's slits paintbrush functions. (b) The complete set of all $C(y)$ functions, as a function of the aperture position, X, a two-dimensional function we call $C(y, X)$. The diamond window in (a) has been convoluted with the four points in $w(x, y)$, but there are regions where we can now measure separately all these four values, thus resolving the phase problem.

$$C(y,X) = \int \Psi(x - 1/2y)\, \Psi^*(x + \tfrac{1}{2}y)\, A(x - X - \tfrac{1}{2}y)\, A^*(x - X + \tfrac{1}{2}y)\, \mathrm{d}x \tag{11}$$

At this stage, understanding the consequences of this equation is not easy without resorting to the paintbrush analogy. We now have four paintbrush strokes multiplied by each other, as shown in Fig. 8. The $\Psi(x)$ and $A(x)$ strokes are parallel to each other, as are the $\Psi^*(x)$ and $A^*(x)$ strokes, which cross over the first two. We can think of the $A(x)A^*(x)$ strokes on their own forming a diamond shaped window in the space $w(x, y)$. Shifting the aperture by (positive) X, shifts the diamond shape to the right in $w(x, y)$. $C(y, X)$ represents the set of all lateral projections of $w(x, y)$ for all aperture shifts X.

In the case of our Youngs' slits, this new data set looks like Fig. 8b. We have not quite extracted a perfect representation of $w(x, y)$ because the aperture has finite width, especially at small values of y, so that $w(x, y)$ is convoluted by (or more precisely, correlated with) the aperture function. We noted above that in the case of the two slits, all the ambiguities of the phase problem would vanish given an independent measurement of modulus of B and C, and indeed we now have provided this information by use of an aperture function. The aperture function can also be the impulse response function of a lens system, as in the case of the Wigner distribution deconvolution method when applied to STEM microdiffraction data, as described in the next section.

6. Resolution and the limits to interference

Interference phenomena are attenuated if the source of the illumination is extended and/or when normal experimental conditions apply: mechanical vibrations, electric power supply instability, magnetic interference, earthing loops, etc. In the Young's slits experiment, the coherence of the wave immediately behind the slits is most easily sabotaged by having an extended incoherent source at a finite distance from the slits, as opposed to the usual illumination by a 'coherent plane wave'. Magnetic or power supply interference could be modelled by time-varying phase plates within the slits. Over a period of time, both effects cause the superposition in intensity of laterally shifted interference patterns; fine fringes, corresponding to well-separated slits, are quickly attenuated in these circumstances.

The exact degree of attenuation, or partial coherence, as a function of fringe periodicity is a complex issue [3]: in a microscope the propagation and/or reduction of the coherence volume is a function of the exact lens geometry employed and its aberrations. However, the overall effect of finite coherence simply reduces the width of the auto-correlation function, so that $w(x, y)$ or $C(y, X)$ has a limited extent in the y direction (this is sometimes called the ρ' cut-off [4]). Put simply, parts of the original wave which are well-separated in space are hard to interfere coherently. When we record a diffraction pattern, fine details are lost because of a finite coherence width at the specimen due to (usually) a finite source size. Similarly, when we record an image intensity, fine atomic-scale interference fringes are lost because of a finite coherence width in the back focal plane of a microscope due to (usually) lens instabilities and magnetic interference.

We can understand this easily in the case of holography. Very high frequency fringes in the image plane are difficult to record. This means that the side-bands often impinge upon the limits of the wave coherence, as in Fig. 9. Because the wave function we are attempting to solve for is in the back focal plane of the lens, this cut-off limits the extent of the Fourier transform of the real-space image. In other words, it sabotages resolution in the final image reconstruction. In fact using holography as a means of improving resolution faces more challenging problems associated with the sampling necessary in reciprocal space in order to account for the lens aberration function [5]. However, for a long time this limit to interference, sometimes known as the information limit, was regarded as the most intractable limit to improved resolution, assuming that a deconvolution could

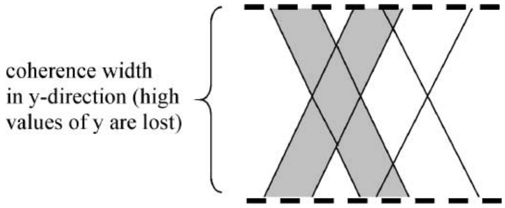

Fig. 9. The limits of coherence in the case of holography. The heavy dashed lines represent a cut-off in the Fourier transform of the image intensity, caused by partial coherence in the illuminating beam and the effects of other instabilities.

accommodate all the usual errors introduced by the electron lens.

6.1. Ptychography beyond the information limit

Consider a wave function consisting of a series of delta functions–a typical example being a conventional selected area diffraction pattern obtained in the back focal plane of the electron microscope from a crystalline specimen. If we introduce a small moveable (objective) aperture in the back focal plane, just larger in diameter than the width of the separation of the diffraction peaks, then in principle we could form a data set $C(y, X)$ that looks like Fig. 10d. The way we would do this physically would be to measure a series of images in a conventional TEM, each taken with the objective aperture in a different position. The Fourier transform of each image (the autocorrelation function, or projection of the paintbrush product function plotted as a function of y) would then have to be arranged in $C(y, X)$ along the X co-ordinate corresponding to the position of the objective aperture.

This arrangement is far removed from Hoppe's original definition of 'ptychography' [6], but the nomenclature is useful in that it is the shifting of the aperture which resolves the conventional ambiguities of the phase problem. Practically doing the experiment in this way would be exceedingly difficult and would not optimise the best use of the transfer function of the lens. A much more practical implementation is to use STEM mode [7–10], where microdiffraction patterns are collected as a function of probe position.

The important point about crystalline ptychography is that, unlike holography or any other reference wave methods, it demonstrates that it is possible in principle to access the underlying form of our function of interest $\Psi(x)$, even if $w(x, y)$ is not very wide in the y direction due to the limitations of the information limit. A little thought will show we have enough information to solve for the entirety of a paintbrush function, $\Psi(x)$, consisting of discrete bristles if we can measure $w(x, y)$ along $y = 0$ and at least one further layer of crossing brushstrokes at one further value of y (Fig. 10). Along $w(x, 0)$ we have all the intensity measurements $|\Psi(x)|^2$. The next row of points has terms like BC^*, CD^*, etc., where B, C, D, etc. are the complex values of adjacent reflections, and hence we can infer the relative phase of all the peaks, even though doing this by holography in the same instrument would be impossible (the interference term BF* shown in Fig. 10a is beyond the coherence cut-off of the microscope).

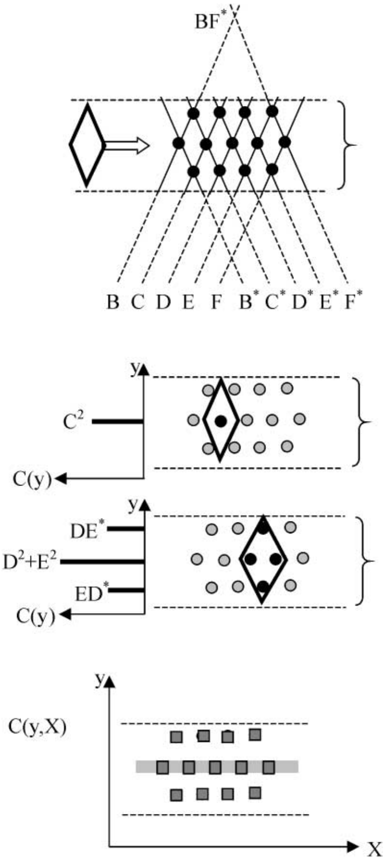

Fig. 10. The ptychography data set: (a) We assume a discrete set of points (in real or reciprocal space). Their amplitudes are labelled B,C,D, etc. Terms like BF^* are outside the information limit, but finite coherence does not limit the width of $w(x, y)$ in the x direction [4]. An aperture diamond can be shifted across the data set. (b) The aperture can select a single beam to measure the intensity (and hence modulus) of each beam. (c) The aperture can also select four points in $w(x, y)$, thus establishing the phase difference between each pair of beams. (d) The compete set of data for all aperture positions, X. Compare with Fig. 8b. The principle of Wigner distribution deconvolution is to deconvolve the blurring in the X direction, even if the wave function is not composed of delta functions.

6.2. Wigner distribution deconvolution

Ptychography is a special example in the case of a crystalline specimen of a much more general hypothesis that given $C(y, X)$ we can obtain an exact estimate of $w(x, y)$ even if the specimen exit wave (or diffraction pattern) is a completely general function and even though the information limit is finite. All that we have to do is deconvolve the diamond-shaped aperture function from $C(y, X)$. For every value of y, we can perform a separate deconvolution of the correlation integral in Eq. (9). We could do this by taking a Fourier transform of the $C(y, X)$ in the y direction and multiply it by a filter function, corresponding to the Fourier transform of the width of diamond function at that value of y, and then Fourier transform back. The filtering therefore takes place in a mixed real and reciprocal space co-ordinate system wherein the filter function is of the form of a Wigner distribution function. It can be shown that since the data set has many redundant values, it is possible to solve for both $\Psi(x)$ and $A(x)$ independently [10].

This method is moderately easy to implement with light optics [11,12], but the electron case has proved difficult to perform. The author's own experience suggests that in the case of STEM it is exceedingly difficult to sample an ordered array of probe positions across a two-dimensional specimen. The time taken to record each microdiffraction pattern is of the order of 25 ms. Scanning a reasonable field of view therefore takes tens of minutes. Notwithstanding all the usual problems of drift, contamination and damage, obtaining repeatable measurements over such a grid is, simply, very difficult. We developed various tests for ensuring that our collected data were self-consistent [13], but never managed to achieve this consistency over a wide field of view.

Konnert and D'Antonio [14] were more successful with a real space version of a similar technique, although rather than performing an inverse deconvolution, they optimised forward calculations of STEM microdiffraction data, until the experimental data set was consistent with the forward calculation. For the purposes of the present discussion, what matters is to realise that both these methods implicitly attempt to solve for $w(x, y)$, despite the blurring effects of either an aperture function or a focussed probe function, and/or the restrictions of the information limit. As detector technology progresses, it may soon become tenable to read out and store a whole diffraction pattern at much greater speed without incurring high read-out noise.

6.3. The bright-field image

We now consider some important subsets of the total scattering data set $C(y, X)$. If $X = 0$, we have an aperture function which is lying centrally on $\Psi(x)$. Furthermore, if $\Psi(x)$ is a diffraction pattern lying in reciprocal space, then the strip of data corresponding to $C(y, 0)$ is the Fourier transform of the conventional bright-field image, often called the diffractogram. Since the advent of CCD cameras, this is now routinely available in real time. (Historically, the diffractogram was formed by Fraunhofer diffraction through the micrograph plate, which yields the intensity of the Fourier transform of the image, whereas here we are referring to the complex-valued Fourier transform). Pictorially, the situation is as in Fig. 11a. Remember, in an ideal world $C(y, 0)$ should correspond to $\Psi(x)$, which is the complex wave field lying in the back focal plane. If that were the case, then the diffractogram would be identical to the complex wave field in the back focal plane of the objective lens and the bright-field image would show the complex specimen exit wave field. We know that holography can achieve this trick by having a remote reference beam which separates out the two side-bands. How does it work in bright-field imaging? The answer is 'not very well,' which explains why the bright-field image is a poor representation of the exit wave function. However, our model can serve to explain where the difficulties arise.

The usefulness of the bright-field method depends crucially upon the specimen being largely transparent–a weakly scattering object–so that the undiffracted beam at $\Psi(0)$ is strong. We now observe two strong lines in $w(x, y)$, Fig. 11. Comparing this with the case of holography above, we see that once again it is as if each

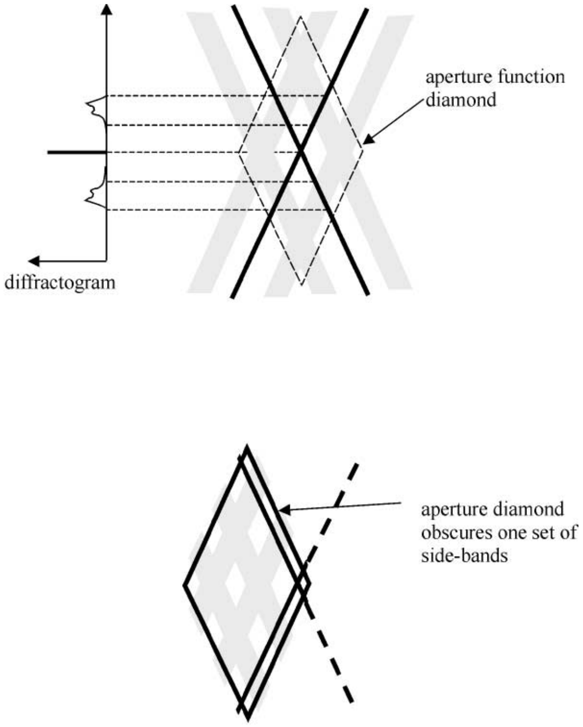
aperture function diamond
diffractogram
aperture diamond obscures one set of side-bands

Fig. 11. The paintbrush construction for the conventional bright-field image. The diffractogram is a poor representation of the amplitude in the back focal plane for two reasons. Firstly, each side-band consists of two holographic overlaps (compare with Fig. 2), and under the first born approximation these would cancel each other out, but for a phase plate introduced in the aperture. Secondly, cross-terms from regions other than that of the strong reference beam can be expressed. (b) Single side-band imaging and tilt-series reconstruction methods shift the either wave function of interest or the aperture window so as to cut out one of the overlaps in the side-band. Resolution (extent of expressed information in the y direction) is also doubled.

paintbrush function is projected conformally onto the $C(y)$, via a multiplication by the reference wave, but that now the two components overlap. If $\Psi(x)$ consists of a delta function at $x = 0$ plus a scattered amplitude $\Psi_s(x)$, which is very small relative to unity, then

$$\Psi(x) = \delta(0) + \Psi_s(x), \tag{12}$$

where $\delta(0)$ is a Dirac delta function, and we have approximately

$$C(y) = |\delta(0)|^2 + \Psi_s^*(-y) + \Psi_s(y). \tag{13}$$

It is not immediately obvious that the overlap of the two side-lobes yields anything particularly useful; this is despite the fact that the bright-field image is often cited as a sort of holographic reference-wave method. However, the First Born approximation dictates that the scattered waves in the back focal plane are proportional to the Fourier transform of the atomic scattering potential in the specimen, and that they are $\pi/2$ out of phase with the unscattered wave at $\psi(0)$. Furthermore, since the atomic potential is a real function, its Fourier transform is complex conjugate symmetric, which leads us to conclude that for a thin specimen

$$\Psi_s(x) = -\Psi_s^*(-x). \tag{14}$$

Substituting into Eq. (8) then leads to rather embarrassing conclusion that $C(y)$ consists of nothing but a delta function at $y = 0$: the two scattered terms cancel perfectly in complex amplitude. The reason is obvious; if the specimen is weak phase, then even though a rich and complicated scattering function may exist in the back focal plane, the image has zero contrast when the aperture function $A(x)$ is real, because only the phase of the image has been altered and this does not register at all on its intensity. Herein lies the image phase problem. The conventional solution is to introduce phase changes across $A(x)$, in the form of aberrations or defocus, which may suitably alter the phases of the two scattered components so that they are strongly expressed.

Even if the $A(x)$ is of the form of a perfect phase plate, so that the modulus of the two side-bands in Fig. 11a add constructively, we can see that amplitude in $w(x, y)$ which lies off the two strong lines can still add into the autocorrelation function, and hence its Fourier transform, the image. This is effect is quite independent of dynamical and three-dimensional scattering in the specimen itself–it represents a failure of the lens function to transfer a contrast which is proportional to the (weak) phase of the exit wave field–sometimes referred to as the breakdown of linear imaging approximation.

6.4. Single-side-band imaging

The principle of single-side-band imaging is to shift the aperture by exactly the right amount, as shown in Fig. 11b, so as to more closely emulate the clean separation of side-bands in holography (Fig. 2). Now the autocorrelation avoids the superposition of the two strong lines of interference in Fig. 11a, and gives an explicit measure of $\Psi(x)$, provided the cross-terms (i.e. those areas which do not lie on the strong lines) are weak. Because the two side bands do not overlap, one pair being obscured by the aperture function, we no longer have to induce artificially a defocus and/or aberration term in order to get any contrast in the image. If aberration is present, then in theory it can now be deconvolved out by dividing the autocorrelation function by the lens transfer function. But, like holography, this strategy can suffer from sampling difficulties in regions where the complex transfer function has a steep phase gradient.

The attraction of the technique is that it solves the image phase problem without a beam-splitter and doubles the attainable resolution. Remember that large distances in $C(y)$ correspond to high-resolution information. By shifting the aperture sideways relative to the bright-field case, twice as much of the remaining side-bands become visible. In practice, in order to avoid extreme lens aberrations, $\Psi(x)$ (and not the aperture function $A(x)$) is shifted by tilting the illumination of the beam which shifts the back focal plane amplitude relative to the optic axis. However, resolution is still limited by the coherence width. Requiring an aperture to be so near the transmitted beam-inducing aperture charging that can be a serious experimental problem. Using the lens so asymmetrically also means that scattered pairs of beams that lie on achromatic rings in the objective lens–i.e. at equal radii from the optic axis where fluctuations in the objective lens supply have equal influence–tend to interfere more strongly.

6.5. Tilt-series reconstruction

A modification of the single-side-band image is to take several such images with different angles of tilt in the illuminating beam, in other words to explore different strips of $C(y, X)$ at a finite number of X. This is advantageous for a number of reasons. When the image and back focal plane are in fact two-dimensional, then the occluding aperture at any one illumination tilt can cut out significant sections of two-dimensional reciprocal space. By tilting over a two-dimensional grid of points, it becomes possible to fill in these unmeasured areas. The same data is measured several times, leading to the possibility of refinement by least squares [15,16]. However, the principle of the technique is the same as for single-side-band imaging: it has the possibility of doubling resolution provided the specimen is reasonably weak.

The equations used in the tilt-series reconstruction literature are now complicated by the introduction of the variable that we have called X–the angle of tilt. To a first approximation (ignoring three-dimensional propagation effects in the specimen) tilt simply shifts the amplitude of the scattered waves in the back focal plane, whereas the objective aperture transfer function remains stationary. This is equivalent to shifting the $\Psi(x)\Psi^*(x)$ paintbrush strokes while leaving the aperture brushstrokes $A(x)A^*(x)$ stationary.

7. Conclusions

The conventional Abbe theory of imaging considers three planes related to one another by Fraunhofer Fourier transform propagators: the back focal plane is the Fourier transform of the specimen exit wave; the image is the Fourier transform of the back focal plane. The fact that the detector lying in the image plane is only sensitive to intensity is incorporated as a last step in the analysis. In the present model, we build this fundamental loss of phase as a first step. It is then apparent that there exists an extensive data set, which we call $w(x, y)$, consisting of all possible pairs of interference combinations between any part of the image (or diffraction pattern) and itself. A holographic reference is an elegant way of extracting a true representation of the original wave function. However, in view of the model, and via the addition of moveable aperture or focussed electron probe, all sorts of other possibilities become available. In general, we can measure a function I have called here $C(y, X)$, which is $w(x, y)$ convoluted with a window function constructed similarly out of the intensity interference terms of the probe or aperture involved. In the general case, there is thus a solution to all the limitations of interference and resolution in electron microscopy. The two most favourable routes to achieve higher resolution irrespective of the limitations of the coherence width (information limit) is to process many microdiffraction patterns recorded in STEM or many tilted-illumination images recorded in TEM. However, having conceded that electron microscopy may be more opportunely improved by performing such multiple experiments, there may be other geometries which could be used more favourably, but which may not yield the convenience of the conventional image. This should be the subject of further work. In the meantime, it is hoped that the model presented in this paper may facilitate an easier understanding of these complicated issues.

Appendix

The projection of the paintbrush functions represents the autocorrelation function, which is the Fourier transform of the intensity of a Fourier transform of the original function. This result is standard, but we include a proof for completeness. Assume that all integrals are performed over infinity, and the forward Fourier transform is defined as

$$F(u) = \int f(x) \exp(\mathrm{i}2\pi xu)\,\mathrm{d}x. \tag{A.1}$$

The intensity of $F(u)$ is given by

$$|F(u)|^2 = F(u)F^*(u) \tag{A.2}$$

which substituting from Eq. (A.2) gives

$$|F(u)|^2 = \int f(x) \exp(\mathrm{i}2\pi xu)\mathrm{d}x \times \int f^*(z) \exp(-\mathrm{i}2\pi zu)\,\mathrm{d}z, \tag{A.3}$$

where we have changed the variable of integration in $F^*(u)$. Compounding the integrals we have

$$|F(u)|^2 = \int\int f(x)f^*(z) \exp(\mathrm{i}2\pi u(\mathrm{x} - z))\,\mathrm{d}x\,\mathrm{d}z. \tag{A.4}$$

Let the forward Fourier transform of $|F(u)|^2$ be

$$C(y) = \int |F(u)|^2 \exp(\mathrm{i}2\pi uy)\,\mathrm{d}u \tag{A.5}$$

so that substituting from Eq. (A.4)

$$C(y) = \int f(x)f^*(z) \exp(\mathrm{i}2\pi u(x - z + y))\,\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}u \tag{A.6}$$

The functions $f(x)$ and $f^*(y)$ do not depend on u, so we can integrate over u noting that $\int \exp(\mathrm{i}2\pi uw)\mathrm{d}u = 0$ unless $u = 0$, at which point it has infinite value and so therefore acts as a Dirac delta function $\delta(u)$, so that

$$C(y) = \int\int f(x)f^*(z)\delta(x - z + y)\,\mathrm{d}x\,\mathrm{d}z, \tag{A.7}$$

which has zero value unless $z = x + y$, and so integrating over z gives

$$C(y) = \int f(x)f^*(x + y)\,\mathrm{d}x \tag{A.8}$$

as in Eq. (10) in the main text.

Note that in many contexts in diffraction physics the autocorrelation is defined as the back Fourier transform of the intensity, in which case we have (x–z–y) in the exponential of Eq. (A.6), leading to an alternative definition

$$C(y) = \int f^*(x)f(x + y)\,\mathrm{d}x \tag{A.9}$$

which is equivalent to a reversal of the of the y coordinate.

References

[1] A.L. Patterson, Phys. Rev. 46 (1934) 372.

[2] J.M. Cowley, Ultramicroscopy 41 (1992) 335.

[3] P.D. Nellist, J.M. Rodenburg, Ultramicroscopy 54 (1994) 61.

[4] J.M. Rodenburg, R.H.T. Bates, Philos. Trans. Roy. Soc. A 339 (1992) 521.

[5] H. Lichte, Ultramicroscopy 51 (1993) 15.

[6] W. Hoppe, Acta Crystallogr. A 25 (1969) 495, 502, 508.
[7] P.D. Nellist, B.C. McCallum, J.M. Rodenburg, Nature 374 (1995) 630.
[8] P.D. Nellist, J.M. Rodenburg, Acta Crystallogr. A 54 (1998) 49.
[9] T. Plamann, J.M. Rodenburg, Acta Crystallogr. A 54 (1998) 61.
[10] B.C. McCallum, J.M. Rodenburg, J. Opt. Soc. Am. A10 (1993) 231.
[11] S.L. Friedman, J.M. Rodenburg, J. Phys D. 25 (1992) 147.
[12] B.C. McCallum, J.M. Rodenburg, Ultramicroscopy 52 (1993) 85.
[13] J.M. Rodenburg, B.C. McCallum, P.D. Nellist, Ultramicroscopy 48 (1993) 303.
[14] J. Konnert, P. D'Antonio, J.M. Cowley, A. Higgs, H.J. Ou, Ultramicroscopy 30 (1989) 384.
[15] A.I. Kirkland, R.R. Meyer, W.O. Saxton, J. Hutchison, R. Dunin-Borkowski, Inst. Phys. Conf. Ser. 161 (EMAG 99) (1999) 291.
[16] A.I. Kirkland, W.O. Saxton, K.-L. Chau, K. Tsuno, M. Kawasaki, Ultramicroscopy 57 (1995) 355.